UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO - CTC
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA E GESTÃO DO CONHECIMENTO

Tatiana Tozzi

Um método para o apoio e análise de documentos textuais empregando Mineração de Texto e Taxonomia

Florianópolis
2024

Tatiana Tozzi

Um método para o apoio e análise de documentos textuais empregando Mineração de Texto e Taxonomia

Dissertação submetida ao Programa de Pós-graduação em Engenharia e Gestão do Conhecimento da Universidade Federal de Santa Catarina para a obtenção do título de Mestra em Engenharia e Gestão do Conhecimento.
Orientador: Prof. José Leomar Todesco, Dr.
Coorientador: Prof. Alexandre Leopoldo Gonçalves, Dr.

Florianópolis
2024

Ficha de identificação da obra

Tozzi, Tatiana
Um método para o apoio e análise de documentos textuais empregando Mineração de Texto e Taxonomia / Tatiana Tozzi; orientador, José Leomar Todesco, coorientador, Alexandre Leopoldo Gonçalves, 2024.
222 p.

Dissertação (mestrado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, Florianópolis, 2024.

Inclui referências.

1. Engenharia e Gestão do Conhecimento. 2. Corpus de documentos de texto. 3. Taxonomia. 4. Mineração de Texto. 5. Visualização de dados. I. Todesco, José Leomar. II. Gonçalves, Alexandre Leopoldo. III. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento. IV. Título.

Tatiana Tozzi

Um método para o apoio e análise de documentos textuais empregando Mineração de Texto e Taxonomia

O presente trabalho em nível de mestrado foi avaliado e aprovado por banca examinadora composta pelos seguintes membros:

Prof. Paulo Mauricio Selig, Dr.
Universidade Federal de Santa Catarina

Prof. Denilson Sell, Dr.
Universidade Federal de Santa Catarina

Prof. Antônio Pereira Cândido, Dr.
Instituto Federal de Santa Catarina

Certificamos que esta é a **versão original e final** do trabalho de conclusão que foi julgado adequado para obtenção do título de Mestra em Engenharia e Gestão do Conhecimento.

Documento assinado digitalmente
Roberto Carlos dos Santos Pacheco
Data: 24/06/2024 12:00:28-0300
CPF: ***.543.830-**
Verifique as assinaturas em https://v.ufsc.br

Coordenação do Programa de Pós-Graduação

Documento assinado digitalmente
Jose Leomar Todesco
Data: 05/06/2024 13:42:09-0300
CPF: ***.110.599-**
Verifique as assinaturas em https://v.ufsc.br

Prof. José Leomar Todesco, Dr.
Orientador

Florianópolis, 2024

Este trabalho é dedicado *in memoriam* a Giulio, Maximus, Leonelo, Alice, Lucca, Oliver, Noah, Théo Miguel, Lee Gon, Emília e Lupita.

AGRADECIMENTOS

Agradeço ao meu orientador, Professor Dr. José Leomar Todesco (Tite), por ter me acolhido no curso de Pós-Graduação em Engenharia e Gestão do Conhecimento e por sua dedicação, carinho e parceria ao longo do curso. Ao meu coorientador, Professor Dr. Alexandre Leopoldo Gonçalves, expresso minha gratidão pela confiança, parceria e paciência, além de seu inestimável conhecimento, essencial para a realização deste trabalho.

Agradeço também à Universidade Federal de Santa Catarina e ao Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento pela oportunidade de avançar em minha jornada acadêmica. Meus agradecimentos à CAPES pelo apoio financeiro durante parte deste mestrado, bem como aos professores e técnicos administrativos do Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento.

Aos membros da banca avaliadora, Prof. Dr. Paulo Mauricio Selig, Prof. Dr. Denilson Sell e Prof. Dr. Antônio Pereira Cândido, agradeço por aceitarem avaliar este trabalho. Suas valiosas sugestões e comentários enriqueceram significativamente minha pesquisa.

Agradeço aos professores do curso de Bacharelado em Sistemas de Informação do Instituto Federal Catarinense - Campus Camboriú (IFC), especialmente aos Professores MSc. Rodrigo Ramos Nogueira, Dr. Daniel Fernando Anderle, Dr. Alexandre de Aguiar Amaral e Dr. Aujor Tadeu Cavalca Andrade, por acreditarem em meu potencial, mesmo em momentos de dúvida pessoal.

Aos meus amigos e colegas, em especial a Patrícia Todesco, que em 2015 me incentivou a participar do processo seletivo do IFC, e àqueles que me apoiaram desde o início desta jornada e agora me acompanham em espírito (*in memoriam*).

Aos meus gatos que partiram durante esta trajetória e agora estão com São Francisco, saudáveis e livres de suas doenças e dores. A Théo Miguel por ter tido a privilegio de conviver por 16 anos e ser meu primeiro gato, e assim me tornado gateira, por ser meu norte. A Giulio e Maximus por alegria e felicidade mesmo neste curto tempo juntos. A Alice minha estrela linda tricolor. A Lucca meu formoso e encantador. A Oliver meu docinho pequeno e manhoso. A Noah, um fubázinho mimoso ronronador e maravilhoso. A meu foguete/trator Emília que foi e sempre será linda e perfeita. A Lupita que mesmo neste curto tempo juntas espero que tenha se sentido amada e aceita. E a todos meus sobrinhos gatos que partiram.

Por fim, sou profundamente grata à minha família, tanto humana quanto felina, pelo apoio incondicional e energia que me ajudaram a superar todos os desafios enfrentados durante este período.

"Num mundo saturado de informações e afogado pelo vasto volume de dados gerados, a verdadeira sabedoria reside na habilidade de discernir, filtrar e interpretar o volume abissal de dados que nos cerca, atribuindo valor à informação que realmente importa."

(Autor desconhecido, s.d)

RESUMO

A complexidade envolvida na leitura e interpretação de documentos é destacada neste trabalho, evidenciando como a compreensão dos textos pode variar significativamente em função da presença de termos técnicos e do conhecimento prévio do leitor. A introdução do estudo ressalta os desafios enfrentados pelos leitores ao se depararem com documentos que contêm linguagem especializada. Propõe-se que esses desafios possam ser mitigados por meio do uso integrado de técnicas de Mineração de Texto (*Text Mining* - TM) e Visualização de Dados. Mais especificamente, a adoção de uma taxonomia estruturada e hierárquica é sugerida como uma ferramenta essencial para melhorar a organização e a categorização de informações, tornando a análise de documentos mais acessível e intuitiva. O principal objetivo deste trabalho é desenvolver um método que emprega técnicas de Mineração de Texto e uma taxonomia definida para auxiliar na análise de documentos textuais em contextos especializados. Este método busca facilitar a interpretação e compreensão de textos com terminologias técnicas, por meio da identificação de características textuais relevantes, da especificação de uma taxonomia adequada, e da aplicação de técnicas de TM adaptadas ao domínio do conhecimento. Tal abordagem multidisciplinar, integra conhecimentos da Engenharia do Conhecimento, Representação de Conhecimento por meio de Taxonomia, Desenvolvimento de Sistemas e Engenharia de Software, demonstrando como a interação entre essas áreas pode resolver problemas complexos de análise documental. Os resultados alcançados com a implementação do método proposto indicam uma melhoria significativa na análise de documentos textuais. A aplicação prática deste método, revelou que a organização hierárquica de termos e a visualização intuitiva de dados podem proporcionar uma compreensão mais adequada e estruturada do conteúdo dos documentos. Como conclusão tem-se que a inclusão de uma taxonomia bem definida e a aplicação de técnicas de TM não apenas facilitam a identificação e compreensão de termos técnicos, mas também aprimoram significativamente a interação entre o usuário e as informações, facilitando o processo de análise documental em diversos contextos profissionais e acadêmicos.

Palavras-chave: *Corpus* de documentos de texto. Mineração de Texto. Taxonomia. Análise textual. Visualização de dados.

ABSTRACT

This work highlights the complexity involved in reading and interpreting documents, demonstrating how the comprehension of texts can vary significantly due to the presence of technical terms and the reader's prior knowledge. The study's introduction emphasizes the challenges faced by readers when encountering documents containing specialized language. It proposes that these challenges can be mitigated through the integrated use of Text Mining (TM) techniques and data visualization. More specifically, the adoption of a structured and hierarchical taxonomy is suggested as an essential tool to improve the organization and categorization of information, making document analysis more accessible and intuitive. The main goal of this work is to develop a method that employs advanced Text Mining techniques and a carefully defined taxonomy to assist in the analysis of textual documents in specialized contexts. This method aims to facilitate the interpretation and understanding of texts filled with technical terminologies, through the identification of relevant textual characteristics, the specification of an appropriate taxonomy, and the application of Machine Learning and TM techniques adapted to the knowledge domain. Such a multidisciplinary approach integrates knowledge from Knowledge Engineering, Ontologies, Systems Development, and Software Engineering, demonstrating how the interaction between these areas can solve complex problems in document analysis. The results achieved with the implementation of the proposed method indicate a significant improvement in the analysis of textual documents. The practical application of this method revealed that the hierarchical organization of terms and intuitive data visualization can provide a deeper and more structured understanding of the document content. The study concludes that the inclusion of a well-defined taxonomy and TM techniques not only facilitates the identification and understanding of technical terms but also significantly enhances the interaction between the user and the information, making the document analysis process more efficient and effective in various professional and academic contexts.

Keywords: Document Corpus. Text Mining. Taxonomy. Textual Analysis. Data Visualization.

LISTA DE FIGURAS

Figura 1 - Tipos de documentos administrativos 26
Figura 2 - Delimitações da dissertação 29
Figura 3 – Disciplinas cursadas no PPGEGC 40
Figura 4 - Exemplo de Taxonomia 52
Figura 5 - Etapas do processo de KDT 59
Figura 6 - Aplicações que utilizam *Text Mining* 61
Figura 7 - Etapas de Análise de dados 63
Figura 8 – Etapas de desenvolvimento da DSRM 76
Figura 9 - Abordagem Metodológica 76
Figura 10 - Atividades do método de investigação DSRM 77
Figura 11 - Fases da Revisão Integrativa 87
Figura 12 - Nuvem de palavras-chave 97
Figura 13 - Etapas do Método Proposto 106
Figura 14 - Trecho de textos com as informações em destaque 109
Figura 15 - Modelo de dados do módulo de análise 116
Figura 16 - Modelo de dados do módulo de análise da aplicação web 117
Figura 17 - Taxonomia de fatores condicionantes de performance 118
Figura 18 - Fatores e termos que compõem o vetor de termos 120
Figura 19 - Exemplo de coocorrência de palavras 122
Figura 20 - Conjunto de funcionalidades 130
Figura 21 - DOC *Analysis*: Página inicial 131
Figura 22 - DOC *Analysis*: Criar estudo 131
Figura 23 - DOC *Analysis*: Salvar estudo 132
Figura 24 - DOC *Analysis*: Processar estudo 132
Figura 25 - DOC *Analysis*: Resultado do processamento do texto 133
Figura 26 - DOC *Analysis*: Meus estudos 133
Figura 27 - Gráfico: Árvore de Palavras - Geral 136
Figura 28 - Gráfico: Árvore de Palavras – Fatores e Frequência 137
Figura 29 - Gráfico: Árvore de Palavras - Frequência – Termos e Frequência 137
Figura 30 - Gráfico: Barras horizontais 138

Figura 31 - Gráfico: Barras horizontais interativas 138
Figura 32 - Gráfico: Barras horizontais - Termos 139
Figura 33 - Gráfico: *Treemap* 140
Figura 34 - Gráfico: *Treemap* interativo 141
Figura 35 - Gráfico: *Zoomable Circle Packing* 141
Figura 36 - Gráfico: Nuvens de Palavras 142
Figura 37 - Gráfico - Bolhas 142
Figura 38 - Fluxograma de amostragem da revisão integrativa 187
Figura 39 - Distribuição de Publicações por ano 191
Figura 40 - Etapas de desenvolvimento de sistemas 205
Figura 41 - Caso de uso: criar novo estudo 208
Figura 42 - Caso de uso: Criar estudo 209
Figura 43 - Caso de uso: Resultados 209
Figura 44 - Caso de uso: Geral 210
Figura 45 - Paleta de cores 211
Figura 46 - Protótipo baixa fidelidade: Abrindo sistema 213
Figura 47 - Protótipo baixa fidelidade: Página inicial 214
Figura 48 - Protótipo baixa fidelidade: Abrir estudo 214
Figura 49 - Protótipo baixa fidelidade: Criar estudo 215
Figura 50 - Protótipo baixa fidelidade: Estudos recentes 215
Figura 51 - Protótipo baixa fidelidade: Resultado do estudo 216
Figura 52 – Etapas de seleção de arquivos e execução do sistema desktop 217
Figura 53 - Extração e análise do texto selecionado 218
Figura 54 - Protótipo média fidelidade: Resultado do estudo - árvore 218
Figura 55 - Protótipo média fidelidade: Resultado do estudo - barras horizontais 219
Figura 56 - Protótipo média fidelidade: Resultado do estudo - treemap 219
Figura 57 - Protótipo alta fidelidade: Página inicial 220
Figura 58 - Protótipo alta fidelidade: Criar estudo 221
Figura 59 - Protótipo alta fidelidade: Meus estudos 221
Figura 60 - Protótipo alta fidelidade: Abrir estudo 222
Figura 61 - Protótipo alta fidelidade: Resultado 222

LISTA DE QUADROS

Quadro 1 - Técnicas de *Machine Learning* 22
Quadro 2 - Técnicas de Mineração de texto 23
Quadro 3 – Trabalhos resultantes da busca com a palavra-chave: Documentos textuais 31
Quadro 4 -– Trabalhos resultantes da busca com a palavra-chave: Mineração de texto ou *text mining* 34
Quadro 5 – Trabalhos resultantes da busca com a palavra-chave: Taxonomia 35
Quadro 6 - Resultado da pesquisa conforme *string* de busca 37
Quadro 7 – Principais algoritmos de KDT 58
Quadro 8 - Trabalhos relacionados 68
Quadro 9 – Grau de intersseção entre os termos 98
Quadro 10 – Detalhes das etapas do Método Proposto 104
Quadro 11 – Descrição do Nome do Fator 128
Quadro 12 - Sistemas similares 144
Quadro 13 - *Strings* de busca utilizadas 183

LISTA DE TABELAS

Tabela 1 - Contribuição de cada Base de dados 185
Tabela 2 - Periódicos que contém os artigos selecionados 187
Tabela 3 - Autores dos artigos 191

LISTA DE ABREVIATURAS E SIGLAS

AI	*Artificial Intelligence*
BERT	*Bidirectional Encoder Representations from Transformers*
BI	*Business Intelligence*
BIM	*Building Information Modeling*
BNC	*British National Corpus*
C,T&I	Ciência, Tecnologia e Inovação
CAFe	Comunidade Acadêmica Federada
CBAPE	*Corpus* Brasileiro de Aprendizes de Português como Língua Estrangeira
CNNSs	*Convolutional Neural Networks*
COVID-19	*Coronavirus Disease* 2019
CRPC	*Corpus* de Referência do Português Contemporâneo
CSS	*Cascading Style Sheets*
CSV	*Comma-Separated Values*
DBMS	*Database Management System*
DeCS	Descritores em Ciências da Saúde
DS	*Design Science*
DSR	*Design Science Research*
DSRM	*Design Science Research Methodology*
DTMs	Desordens Temporomandibular
ECFP7FUPOL	*European Commission Seventh Framework Programme Future Policy Modeling*
EC	Engenharia do Conhecimento
EXOD	*EXploration of Open Datasets*
GAN	*Generative Adversarial Network*
GC	Gestão do Conhecimento
GPT	*Generative Pre-trained Transformer*
HTML	*HyperText Markup Language*
IoT	*Internet of Things*
JSON	*JavaScript Object Notation*
JVM	*Java Virtual Machine*

KDT	*Knowledge Discovery in Text*
KE	*Knowledge Engineering*
KNN	*k-Nearest Neighbors*
LGPD	Lei Geral de Proteção de Dados
LLM	*Large Language Model*
ML	*Machine Learning*
NER	*Named Entity Recognition*
NLG	*Natural Language Generation*
NLP	*Natural Language Processing*
NLU	*Natural Language Understanding*
OE	Objetivo Específico
PCA	*Principal Component Analysis*
PHP	*Hypertext Preprocessor*
PPGCC	Programa de Pós-Graduação em Ciências da Computação
PPGEGC	Programa de Pós-Graduação em Engenharia, Gestão e Mídia do Conhecimento
PRISMA	*Preferred Reporting Items for Systematic Reviews and Meta-Analyses*
RNNs	*Recurrent Neural Networks*
SAD	Sistemas de Apoio à Decisão
SGBDs	Sistema de Gerenciamento de Banco de Dados
SGC	Sistema de Gestão do Conhecimento
SNA	*Social Network Analysis*
SPAs	*Single Page Applications*
SQL	*Structured Query Language*
SSD	Sistemas de Suporte à Decisão
SVM	*Support Vector Machine*
TF-IDF	*Term Frequency-Inverse Document Frequency*
TI	Tecnologia da Informação
TM	*Text Mining*
UFSC	Universidade Federal de Santa Catarina
UI	*User Interface*
UML	*Unified Modeling Language*
UNL	*Universal Networking Language*

LISTA DE SÍMBOLOS

✗	Não presente
✓	Presente

SUMÁRIO

1 INTRODUÇÃO 21

1.1 CONTEXTUALIZAÇÃO DO PROBLEMA 25

1.2 PERGUNTA DE PESQUISA 27

1.2.1 Objetivo Geral 27

1.2.2 Objetivos Específicos 28

1.3 JUSTIFICATIVA E RELEVÂNCIA 28

1.4 DELIMITAÇÕES DA PESQUISA 29

1.5 ADERÊNCIA AO PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA E GESTÃO DO CONHECIMENTO 30

1.6 ORGANIZAÇÃO DOS CAPÍTULOS 40

2 FUNDAMENTAÇÃO TEÓRICA 42

2.1 *CORPUS* DE DOCUMENTOS 42

2.1.1 ANÁLISE DE DOCUMENTOS TEXTUAIS 45

2.1.1.1 Desafios na Análise Qualitativa de Documentos 45

2.2 ENGENHARIA DO CONHECIMENTO 48

2.2.1 Organização e Representação do Conhecimento 50

2.2.2 Taxonomias 51

2.2.2.1 Abordagens para a Construção de Taxonomias 52

2.2.2.2 Classificação de Taxonomias 53

2.2.2.3 Desenvolvimento de Taxonomias 54

2.2.2.4 Justificativa para o uso de Taxonomias 55

2.2.3 DESCOBERTA DE CONHECIMENTO EM TEXTOS 57

2.2.3.1 MINERAÇÃO DE TEXTO 60

2.2.3.2 ANÁLISE DE DADOS 62

2.2.3.3 VISUALIZAÇÃO DE DADOS..... 64
2.2.3.4 Tecnologia de suporte a Análise de Dados 65
2.3 TRABALHOS RELACIONADOS 66
3 METODOLOGIA DE PESQUISA 75
3.1 DESIGN SCIENCE RESEARCH METHODOLOGY 76
3.2 REVISÃO INTEGRATIVA DE LITERATURA 78
3.3 EXECUÇÃO DA *DESIGN SCIENCE RESEARCH METHODOLOGY*..... 79
3.3.1 Etapa 1 – Identificação do Problema e Motivação 80
3.3.2 Etapa 2 – Definição dos Objetivos para uma Solução..... 81
3.3.3 Etapa 3 – Projeto e Desenvolvimento..... 82
3.3.4 Etapa 4 - Demonstração 84
3.3.5 Etapa 5 - Avaliação..... 85
3.3.6 Etapa 6 – Comunicação dos Resultados 86
3.4 EXECUÇÃO DA REVISÃO INTEGRATIVA DA LITERATURA 87
3.4.1 Análise dos dados e Apresentação..... 87
4 MÉTODO PROPOSTO 101
4.1 EXPLORANDO O MÉTODO 106
4.2 INSTANCIAÇÃO DO MÉTODO PROPOSTO 112
4.3 DESENVOLVIMENTO DO BACK-END 114
4.3.1 Banco de dados..... 115
4.3.2 Leitura e Pré-processamento do Relatório de Incidente..... 118
4.3.3 Geração da Frequência de Termos dos Fatores..... 121
4.3.4 Geração de Coocorrências entre os Termos dos Fatores 121
4.3.5 Geração de Coocorrências entre Fatores e Termos..... 123
4.4 DESENVOLVIMENTO DO *FRONT-END* 124
5 AVALIAÇÃO E DISCUSSÃO DOS RESULTADOS..... 126
5.1 APRESENTAÇÃO DO CENÁRIO DE ESTUDO 126

5.2 UTILIZAÇÃO DO SISTEMA 130
5.3 FUNCIONALIDADES DE GRÁFICOS UTILIZADOS 134
5.4 SISTEMAS SIMILARES 143
5.5 ANÁLISE E RECOMENDAÇÕES FUTURAS 144
5.5.1 Considerações para Futuras Implementações 145
6 CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS 147
6.1 CONSIDERAÇÕES FINAIS 147
6.2 TRABALHOS FUTUROS 149
6.3 CONTRIBUIÇÕES DA DISSERTAÇÃO 150
REFERÊNCIAS 152
APÊNDICE A – Protocolo de Revisão Integrativa 178
APÊNDICE B – Desenvolvimento da Revisão Integrativa 183
APÊNDICE C – Tecnologias adotadas 195
APÊNDICE D – SQL – *Back-end* 200
APÊNDICE E – SQL – *Front-end* 203
APÊNDICE F – Documentação do sistema 204

1 INTRODUÇÃO

A leitura de documentos é uma atividade diária que permeia diversas áreas, onde os leitores buscam constantemente explorar novos conteúdos para aprimorar seus conhecimentos e familiarizar-se com novas áreas e práticas. A interpretação desses documentos pode variar significativamente entre os leitores, muitas vezes devido a termos técnicos desconhecidos ou lacunas de compreensão. De acordo com Pacheco e Ataide (2013), um texto pode conter várias interpretações, mas não está aberto a qualquer interpretação, apresentando pistas e estruturas de apelo que conduzem o leitor a uma leitura coerente.

A presença de termos técnicos desconhecidos pode ser um obstáculo para a compreensão eficaz do conteúdo do documento. Além disso, a interpretação pode ser influenciada pelo conhecimento de mundo, experiências, ideologias, crenças e valores individuais de cada leitor.

A interpretação de documentos textuais é um desafio que se estende além da simples leitura. A presença de termos técnicos desconhecidos e a necessidade de uma compreensão profunda do contexto muitas vezes dificultam a análise eficaz dos documentos. Nesse sentido, a integração de técnicas de Aprendizado de Máquina (do inglês M*achine Learning - ML)* e Visualização de Dados surge como uma solução promissora para facilitar a análise de documentos, especialmente com o suporte de uma taxonomia (BORGMAN, 2015).

A taxonomia, neste contexto, refere-se a um sistema de classificação que permite a organização e categorização de informações de uma maneira estruturada e hierárquica. Isso pode ser particularmente útil na análise de documentos, onde a identificação e categorização de termos técnicos podem ser desafiadoras (BORGMAN, 2015).

A integração de ML e visualização de dados pode proporcionar uma compreensão mais profunda e intuitiva dos documentos. Por exemplo, técnicas de ML podem ser usadas para identificar e classificar termos técnicos em documentos, enquanto a visualização de dados pode ajudar a apresentar essas informações de uma maneira mais acessível e compreensível (CHEN, 2017).

Existem vários exemplos de trabalhos que utilizam técnicas de ML e visualização de dados para a análise de documentos. Por exemplo, Loper e Bird (2002) desenvolveram uma biblioteca de processamento de linguagem natural que utiliza ML para a análise de texto.

O Quadro 1 apresenta uma variedade de paradigmas e técnicas de ML e suas respectivas aplicações em diversos campos, destacando a versatilidade e a eficácia em resolver problemas complexos e otimizar processos. Os paradigmas inclue o Aprendizado Supervisionado, utilizado em tarefas como previsão de vendas e diagnóstico médico; o Aprendizado Não Supervisionado, que encontra uso em segmentação de clientes e detecção de anomalias; e o Aprendizado por Reforço, aplicado em robótica e jogos; e várias formas de redes neurais, tais como Redes Neurais Convulsionais (do inglês *Convolutional Neural Networks* - CNNs) e Redes Neurais Recorrentes (do inglês *Recurrent Neural Networks* - RNNs), que são empregadas em visão computacional e modelagem de linguagem, respectivamente. Outras técnicas, como Máquinas de Vetores de Suporte (do inglês *Support Vector Machines* - SVM) e Árvores de Decisão, são utilizadas para classificação de imagens e avaliação de risco de crédito. Essa diversidade de paradigmas técnicas demonstra a abrangência do ML e sua capacidade de adaptar-se a uma ampla gama de aplicações práticas.

Quadro 1 - Técnicas de *Machine Learning*

	Técnica de *Machine Learning*	**Aplicação/Utilização**
Paradigmas de aprendizado	**Aprendizado Supervisionado**	Previsão de vendas, diagnóstico médico, reconhecimento de fala
	Aprendizado Não Supervisionado	Segmentação de clientes, análise de agrupamento, detecção de anomalias
	Aprendizado por Reforço	Jogos, robótica, otimização de rotas de entrega
Arquiteturas	**Redes Neurais (CNNs, RNNs)**	Processamento de linguagem natural, visão computacional; Reconhecimento de objetos em imagens, análise de vídeo; Previsão de séries temporais, modelagem de linguagem
	CNNs	Reconhecimento de objetos em imagens, análise de vídeo
	RNNs	Previsão de séries temporais, modelagem de linguagem
	SVM	Classificação de imagens, categorização de textos
	Árvores de Decisão	Avaliação de risco de crédito, decisões de investimento

Fonte: Elaborado pela autora com base em GÉRON (2019); BISHOP (2006); GOODFELLOW *et al*. (2016).

Essas técnicas de ML são fundamentais para muitas áreas, incluindo a Mineração de Texto (do inglês *Text Mining* – TM) é uma área de pesquisa interdisciplinar que envolve a análise de grandes volumes de texto para extrair informações relevantes e úteis. A preparação dos dados, a análise de sentimentos e a identificação de entidades nomeadas são técnicas importantes de TM. Apesar dos desafios associados ao manuseio de dados textuais, o TM tem o potencial de transformar muitas indústrias e campos de pesquisa. Por exemplo, no setor de saúde, o TM pode ser usado para analisar registros médicos para identificar padrões e tendências. No campo do *marketing*, pode ajudar a entender as opiniões dos clientes a partir de avaliações de produtos. Na pesquisa acadêmica, pode auxiliar na análise de grandes volumes de publicações para identificar novas áreas de estudo

O Quadro 2 abaixo ilustra diversas técnicas de TM e suas aplicações práticas em variados setores. Estas técnicas, que abrangem desde a Análise de Sentimento até a Análise de Tópicos, são fundamentais para extrair *insights* valiosos a partir de grandes volumes de dados textuais. Por exemplo, a Análise de Sentimento é amplamente usada para entender as emoções expressas em *feedbacks* de clientes e redes sociais, enquanto a Extração de Informação permite identificar dados estruturados em textos não estruturados. Técnicas como Sumarização Automática e Classificação de Texto são essenciais para a gestão eficiente de informações, ajudando a resumir conteúdos extensos e a categorizar automaticamente documentos. Além disso, o Reconhecimento de Entidades Nomeadas e o Agrupamento de Texto facilitam a organização e a análise de grandes conjuntos de documentos, identificando entidades específicas e agrupando textos similares, respectivamente. Estas ferramentas de mineração de texto são cruciais para desvendar padrões e facilitar a tomada de decisão baseada em dados textuais extensos.

Quadro 2 - Técnicas de Mineração de texto

Técnica de Mineração de Texto	Aplicação/Utilização
Análise de Sentimento	Monitoramento de redes sociais, análise de *feedback* de clientes
Extração de Informação	Extração de dados estruturados de textos não estruturados, como datas ou locais
Sumarização Automática	Criação de resumos de textos longos, como artigos ou relatórios
Classificação de Texto	Filtragem de spam em e-mails, categorização de documentos

Agrupamento de Texto	Descoberta de tópicos em conjuntos de dados, agrupamento de notícias similares
Reconhecimento de Entidades Nomeadas	Identificação de nomes de pessoas, organizações ou locais em textos

Fonte: Elaborado pela autora com base em MANNING, *et al.* (2008); AGGARWAL, *et al*. (2012).

Além das técnicas de mineração de texto mencionadas acima, é importante notar que a visualização de dados também desempenha um papel crucial na análise de documentos. Nesse contexto, Heer e Shneiderman (2012) discutiram várias técnicas de visualização de dados que podem ser aplicadas à análise de documentos. Entre essas técnicas, os gráficos de barras e linhas são comumente usados para representar tendências de dados ao longo do tempo. Os mapas de calor, são úteis para visualizar dados complexos em duas dimensões. Além disso, as árvores de decisão podem ser usadas para representar hierarquias e sequências de decisões.

Além do ML e da TM, existem outras técnicas que podem ser utilizadas para a análise de documentos textuais. Uma delas é a Análise Textual Discursiva, que se situa entre a Análise de Conteúdo e a Análise de Discurso. Esta técnica permite a produção de novos conhecimentos e, também, de novas compreensões acerca de fenômenos investigados em processos de aproximação, mas de diferenciação em relação aos outros dois dispositivos de pesquisa (MORAES, 2016).

Outra técnica é a Análise de Sentimento, que avalia o tom da mensagem baseado no sentimento, se o texto expressa um sentimento positivo, negativo ou neutro (LIU, 2012). A Análise de Tópicos identifica os principais assuntos abordados no texto (BLEI, 2012). A Análise de Frequência conta a frequência de repetições, palavras ou frases específicas no texto (MANNING, 2018). Além disso, a Análise Documental é um procedimento que se utiliza de métodos e técnicas para a apreensão, compreensão e análise de documentos dos mais variados tipos (BARDIN, 2016). Nesse contexto, a utilização de método e técnicas de ML e TM emergem como uma ferramenta valiosa para apoiar a análise de documentos textuais.

A visualização de dados não é apenas uma ferramenta para a análise de dados, mas também uma forma eficaz de comunicar os resultados da análise. Ao apresentar os dados de maneira visual, é possível destacar tendências, padrões e *outliers* que podem não ser imediatamente aparentes em uma tabela de dados brutos. É importante notar que a visualização de dados é uma ferramenta e não um fim em si mesma.

Os gráficos de dispersão são úteis para visualizar relações entre duas variáveis, enquanto os histogramas permitem a visualização da distribuição de frequências de um conjunto de dados. As técnicas de visualização de dados não se limitam a essas, e a escolha da técnica apropriada depende muito do tipo de dados e do objetivo da análise.

A análise de documentos textuais é um campo complexo que requer a aplicação de várias técnicas e métodos. A escolha da técnica apropriada depende do objetivo da análise e do tipo de dados disponíveis. Independentemente da técnica escolhida, é importante lembrar que a finalidade da análise é gerar insights úteis e acionáveis a partir dos dados. Portanto, a análise deve ser conduzida de maneira sistemática e rigorosa, e os resultados devem ser apresentados de maneira clara e compreensível. Nesse contexto, a visualização de dados surge como uma ferramenta poderosa para comunicar os resultados da análise de maneira eficaz. Com a rápida evolução das tecnologias de análise de dados, é provável que surjam novas técnicas e ferramentas no futuro, ampliando ainda mais as possibilidades de análise de documentos textuais.

1.1 CONTEXTUALIZAÇÃO DO PROBLEMA

A leitura de documentos é uma atividade ubíqua em nossas vidas, permeando diversos contextos e áreas de atuação. Seja para adquirir conhecimento, atualizar-se em determinado campo ou simplesmente para cumprir obrigações profissionais, a interpretação de documentos é uma habilidade essencial. A onipresença das tecnologias digitais e das redes sociais destaca como essas ferramentas transformaram a comunicação e o aprendizado, facilitando o acesso à informação em qualquer momento e lugar (ADJIN-TETTEY; SELORMEY; NKANSAH, 2022). Essa tarefa pode ser desafiadora, uma vez que a compreensão de um texto pode variar significativamente de pessoa para pessoa. Conforme destacado por Koch (2018), um texto não é um receptáculo vazio, pronto para ser preenchido com qualquer interpretação, mas sim uma construção repleta de pistas e estruturas que guiam o leitor para uma compreensão coerente. Dessa forma, a interpretação de um documento é influenciada não apenas pelo seu conteúdo explícito, mas também pelas experiências, ideologias e conhecimentos prévios do leitor. Isso é evidenciado pela teoria da recepção, que enfatiza o papel ativo dos leitores ao interpretarem textos. Hans Robert Jauss introduziu o conceito de “horizonte de expectativas”, que são os conjuntos de suposições, crenças e normas culturais que os leitores trazem ao texto, moldando suas interpretações (RECEPTION THEORY, 2023).

Um dos principais desafios na interpretação de documentos é a presença de termos técnicos e específicos de determinadas áreas do conhecimento. Esses termos podem representar barreiras significativas para o entendimento do leitor, levando a interpretações equivocadas ou incompletas do texto. Isso se aplica especialmente em áreas como a tecnologia, onde os documentos de requisitos técnicos precisam ser elaborados de maneira clara para evitar mal-entendidos que possam comprometer o projeto (SMITH, 2023). Como ilustrado na Figura 1, que apresenta uma variedade de documentos administrativos, muitos desses documentos estão repletos de terminologias especializadas, tornando a sua interpretação uma tarefa árdua para aqueles que não estão familiarizados com os conceitos. Essa dificuldade é especialmente relevante em ambientes profissionais, onde a compreensão precisa e completa dos documentos é crucial para a tomada de decisões assertivas e para a realização eficaz das atividades laborais.

Figura 1 - Tipos de documentos administrativos

ACORDÃO	AVISO DE INTIMAÇÃO	ALVARÁ	AVISO DE RECEBIMENTO	ASSENTADA	CARTA PRECATÓRIA	CARTA DE ORDEM
CERTIDÃO	CERTIDÃO DO OFICIAL	CITAÇÃO	CONTESTAÇÃO	CÓPIA PROCESSO	CÓPIA PROCESSO	DECISÃO
DEFESA	DENÚNCIA	DESPACHO	DOCUMENTO INICIAL	DOCUMENTOS SETOR ~	EDITAL	EMENTA
INFORMAÇÕES	INTIMAÇÃO	MANDADO	MEMORANDO	NOTIFICAÇÃO	OFÍCIO (GABINETE)	OFÍCIO (CORREIÇÃO)
OFÍCIO	OFÍCIO (CARTÓRIO)	OUTROS	PARECER	PETIÇÃO	PORTARIA	PROCURAÇÃO
PROVIMENTO	RELATÓRIO	REQUERIMENTO	REQUERIMENTO AVULSO	SENTENÇA	SUBSTABELECIMENTO	TERMO
		VOTO	VOTO DO RELATOR	VOTO DE VISTA		

Fonte: Elaborado pela autora com base em Tribunal de Justiça do Estado do Piauí (TJPI) [s. d.]

Além da complexidade linguística, a estrutura e a organização dos documentos também desempenham um papel fundamental na sua interpretação. Conforme ressaltado por Marcuschi (2008), a disposição dos elementos textuais, como títulos, subtítulos e parágrafos, pode influenciar significativamente a compreensão do leitor. Uma estrutura bem delineada e intuitiva facilita a assimilação do conteúdo, enquanto uma organização confusa ou caótica pode dificultar o processo de interpretação.

Outro aspecto relevante é a natureza multifacetada dos documentos, que podem conter informações de diferentes naturezas e complexidades. Por exemplo, um relatório técnico pode apresentar dados quantitativos detalhados, análises qualitativas e recomendações estratégicas,

exigindo do leitor habilidades diversas de interpretação e síntese (BAZERMAN; PRIOR, 2020).

Nesse sentido, a capacidade de discernir as informações relevantes, identificar conexões entre os diversos elementos do texto e extrair conclusões significativas é essencial para uma interpretação eficaz (FIORIN, 2016).

Além dos desafios intrínsecos aos documentos, fatores externos também podem influenciar a sua interpretação. Por exemplo, o contexto cultural, social e histórico do leitor pode moldar a sua percepção do texto, levando a interpretações divergentes entre diferentes grupos de pessoas. Da mesma forma, o propósito subjacente ao documento e a sua autoria podem influenciar a forma como ele é lido e compreendido. Como observado por Santos (2019), o conhecimento sobre o autor e o contexto de produção do texto pode fornecer *insights* valiosos para a sua interpretação.

A interpretação de documentos requer habilidades avançadas de leitura e compreensão, bem como a capacidade de aplicar conhecimentos prévios e contextuais. Isso é crucial não apenas para a compreensão dos textos em si, mas também para a tomada de decisões embasadas e assertivas no ambiente profissional. Conforme observam Silva e Melo (2020), uma interpretação precisa e contextualizada dos documentos administrativos é fundamental para a eficácia das atividades laborais e para a realização de análises que apoiem a tomada de decisões estratégicas.

1.2 PERGUNTA DE PESQUISA

Diante deste cenário, este trabalho possui a seguinte pergunta de pesquisa, a qual orienta esta dissertação: Como apoiar a análise de documentos textuais em domínios específicos por meio de técnicas de Engenharia do Conhecimento?

OBJETIVOS

1.2.1 Objetivo Geral

O objetivo deste trabalho reside em propor e desenvolver um método para apoiar a análise de documentos textuais por meio de técnicas de Mineração de Texto e Taxonomia.

1.2.2 Objetivos Específicos

A partir do objetivo geral, definiram-se os seguintes objetivos específicos (OE):

OE1. Identificar as características textuais que possibilitem o apoio a análise de documentos;

OE2. Especificar uma taxonomia com o intuito de guiar a análise de documentos textuais;

OE3. Aplicar técnicas de visualização de dados na apresentação do conhecimento de determinado domínio.

1.3 JUSTIFICATIVA E RELEVÂNCIA

O estudo apresentado nesta dissertação contribui para a análise de documentos textuais, utilizando técnicas avançadas como TM e taxonomia. Destaca-se pela combinação de tecnologias para interpretar documentos complexos, preenchendo uma lacuna na literatura. A aplicação desenvolvida identifica e compreende termos em corpora textuais, apoiada por TM, oferecendo uma abordagem para a análise de documentos.

A inclusão de uma taxonomia específica, confere características importantes ao trabalho, organizando e relacionando os termos extraídos de forma hierárquica. A abordagem multidisciplinar, combinando Engenharia do Conhecimento, Engenharia de Ontologias, Desenvolvimento de Sistemas e Engenharia de *Software*, destaca-se pela aplicabilidade, oferecendo uma solução integrada para a interpretação de documentos.

A pesquisa é útil para a academia e para aplicações práticas. Ao aplicar técnicas de mineração de texto torna-se viável a superação de desafios de interpretação e compreensão de textos. A inclusão de uma taxonomia organiza a análise dos termos extraídos, proporcionando uma compreensão mais profunda e estruturada do conteúdo.

A aplicação prática desenvolvida a partir do método proposto, demonstra sua viabilidade e relevância em diversos contextos, permitindo uma análise mais precisa e eficiente de documentos. Oferecendo estrutura organizada e visualizações intuitivas dos dados, o método proposto destaca-se como meio para auxiliar a interpretação e extração de conhecimento a partir de documentos textuais.

1.4 DELIMITAÇÕES DA PESQUISA

Esta pesquisa propõe um método que é instanciado por meio de aplicação que suporte o entendimento de termos presentes em documentos textuais (*corpus*). Para isso, o trabalho será desenvolvido em etapas, visando alcançar os objetivos propostos.

O projeto seguirá os passos da *Design Science Research Methodology* (DSRM), detalhados na Seção 3.3. Será realizado o levantamento de requisitos funcionais e não funcionais, e os principais casos de uso serão elaborados utilizando Linguagem de Modelagem Unificada (UML). A documentação completa do sistema, incluindo o levantamento de requisitos, casos de uso e demais informações, é apresentada no Apêndice F.

Não serão abordadas a implementação de ferramentas completas, técnicas avançadas além das descritas, ou a utilização de tecnologias fora do escopo estabelecido.

O desenvolvimento do protótipo utilizará linguagens de programação como Java, PHP, HTML e CSS, além de bibliotecas que auxiliem no método proposto. A Figura 2 apresenta as delimitações da dissertação, considerando o conceito, o nível de análise e a temporalidade.

Figura 2 - Delimitações da dissertação

Fonte: Elaborado pela autora.

1.5 ADERÊNCIA AO PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA E GESTÃO DO CONHECIMENTO

O presente estudo está contextualizado na área de Engenharia do Conhecimento, do Programa de Pós-Graduação em Engenharia, Gestão e Mídia do Conhecimento (PPGEGC), especificamente na linha de pesquisa Engenharia do Conhecimento aplicada às organizações, um dos objetivos fundamentais da Engenharia do Conhecimento (EC) é a conversão de dados em conhecimento, sendo caracterizada por metodologias e ferramentas que podem facilitar a modelagem e a criação desse conhecimento (ABEL, FIORINI, 2013). Segundo Pacheco (2014), a EC pode ser definida como "disciplina que se dedica à modelagem de conhecimento e a criação e inserção de sistemas de conhecimento nas organizações", em relação a linha de pesquisa tem como objetivo a "concepção, desenvolvimento e implantação de soluções da engenharia do conhecimento em organizações públicas e privadas".

Segundo o PPGEGC, o conhecimento, objeto central do programa, é definido como "conteúdo ou processo efetivado por agentes humanos ou artificiais em atividades de geração de valor" (PACHECO; SELIG, KERN, 2021). Ao considerar os documentos como fonte de conhecimento, todos os dados e informações presentes nos mesmos, a EC pode contribuir através de suas metodologias e técnicas para que os conteúdos presentes em tais documentos, sejam melhor entendidos e interpretados. Esta dissertação propõe a implementação de uma aplicação voltada para a identificação de termos em um *corpus* de documento de texto. Sendo uma pesquisa de caráter interdisciplinas uma vez que envolve as áreas de Engenharia do Conhecimento, Engenharia de Ontologias, Desenvolvimento de Sistemas e Engenharia de *Software*.

No Repositório Institucional da UFSC considerando as teses e dissertações do PPGEGC[1], foram identificados alguns trabalhos que possuem relação com esta pesquisa. Utilizando como palavra-chave o termo "Documentos textuais"[2], temos como resultado 88 trabalhos, os trabalhos listados no Quadro 3, sendo identificado 8 teses e 10 dissertações.

[1] https://repositorio.ufsc.br/handle/123456789/76395
[2] https://repositorio.ufsc.br/handle/123456789/76395/discover?query=%22documentos+textuais%22

Quadro 3 – Trabalhos resultantes da busca com a palavra-chave: Documentos textuais

Tipo	Título	Autor	Ano
Tese	Modelo de recuperação e comunicação de conhecimento em emergência médica com utilização de dispositivos portáteis	Heloise Manica	2012
Tese	Um Modelo de engenharia do conhecimento para sistemas de apoio a decisão com recursos para raciocínio abdutivo	Roberto Heinzle	2012
Tese	Criação e compartilhamento do conhecimento da área de moda em um sistema virtual integrado de informações	José Alfredo Beirão Filho	2011
Tese	Um Modelo de descoberta de conhecimento inerente à evolução temporal dos relacionamentos entre elementos textuais	Alessandro Botelho Bovo	2011
Tese	O Raciocínio abdutivo no jogo de xadrez: a contribuição do conhecimento, intuição e consciência da situação para o processo criativo	Kariston Pereira	2010
Tese	Planejamento estratégico de tecnologia da informação com ênfase em conhecimento	Paulo Henrique de Souza Bermejo	2009
Tese	Um modelo para recuperação e busca de informação baseado em ontologia e no círculo hermenêutico	Fabiano Duarte Beppler	2008
Tese	Inter-relação das técnicas *Term Extraction* e *Query Expansion* aplicadas na recuperação de documentos textuais	Raphael Winckler de Bettio	2007
Dissertação	Aplicação de ontologias para apoiar operações analíticas sobre fontes estruturadas e não estruturadas	Marcio Napoli	2011
Dissertação	Mapeamento da disposição individual de compartilhar conhecimento a partir dos níveis de consciência informados pela teoria e instrumento de Loevinger	Mário Roberto Miranda Lacerda	2011
Dissertação	Um Modelo para a visualização de conhecimento baseado em imagens semânticas	Héctor Andrés Melgar Sasieta	2011
Dissertação	Uma Arquitetura de *business intelligence* para processamento analítico baseado em tecnologias semânticas e em linguagem natural	Dhiogo Cardoso da Silva	2011
Dissertação	Sistema de conhecimento para gestão documental no setor judiciário: uma	Samuel Fernandes Ribeiro	2010

Tipo	Título	Autor	Ano
	aplicação no Tribunal Regional Eleitoral de Santa Catarina		
Dissertação	Um Modelo semi-automático para a construção e manutenção de ontologias a partir de bases de documentos não estruturados	Flávio Ceci	2010
Dissertação	Desenvolvimento de um modelo conceitual da classificação internacional da funcionalidade, incapacidade e saúde baseado na *web*	Cristiano Sena da Conceição	2007
Dissertação	A contribuição da análise do contexto organizacional na concepção de sistemas baseados em conhecimento: tecnologia KMAI®	Aline Torres Nicolini	2006
Dissertação	O acesso ao conhecimento em sistemas inteligentes de gestão e análise estratégicas: uma aplicação na segurança pública	Cristina Souza Santos	2006
Dissertação	Portal corporativo como canal para gestão do conhecimento	Flavia Maia da Nova Uriarte	2006

Fonte: Elaborado pela autora, com base no Repositório Institucional da UFSC/PPGEGC.

Nicolini (2006) propõe uma análise do contexto organizacional para a concepção de sistemas baseados em conhecimento, utilizando a tecnologia KMAI®[3]. Napoli (2011) desenvolve uma aplicação de ontologias para apoiar operações analíticas sobre fontes estruturadas e não estruturadas. Beirão Filho (2011) aborda a criação e compartilhamento do conhecimento na área de moda por meio de um sistema virtual integrado de informações. Conceição (2007) desenvolve um modelo conceitual da classificação internacional da funcionalidade, incapacidade e saúde baseado na *web*. Bettio (2007) investiga a inter-relação das técnicas *Term Extraction* e *Query Expansion* aplicadas na recuperação de documentos textuais. Lacerda (2011) mapeia a disposição individual de compartilhar conhecimento utilizando a teoria e instrumento de Loevinger. Manica (2012) propõe um modelo de recuperação e comunicação de conhecimento em emergência médica com utilização de dispositivos portáteis. A dissertação de Santos (2006) destaca a importância dos Sistemas Baseados em Conhecimento (SBC) na gestão estratégica, abordando os desafios na aquisição e representação eficiente do conhecimento. Ela enfatiza a necessidade de equipes capacitadas

[3] Representa a convergência da Gestão do Conhecimento e da Inteligência Artificial.

para explorar as funcionalidades desses sistemas, integrando tecnologia e expertise humana para alcançar resultados significativos na gestão do conhecimento em organizações complexas como a segurança pública.

Pereira (2010) investiga o raciocínio abdutivo no jogo de xadrez e sua contribuição para o processo criativo. Bermejo (2009) elabora um planejamento estratégico de tecnologia da informação com ênfase em conhecimento. Uriarte (2006) analisa o uso de um portal corporativo como canal para gestão do conhecimento. Ribeiro (2010) aborda a importância da inclusão tecnológica na gestão documental do setor judiciário, com foco nas atividades cognitivas do processo. O estudo propõe um sistema de conhecimento para auxiliar os profissionais humanos nessa tarefa, utilizando o sistema e-Docs para unificar o repositório de documentos, automatizar processos de indexação e classificação temática, e fornecer busca semântica de documentos. Os resultados obtidos com a aplicação desse modelo no Tribunal Regional Eleitoral de Santa Catarina são descritos, juntamente com propostas de pesquisa para trabalhos futuros.

Bovo (2011) propõe um modelo para a descoberta de conhecimento a partir de informações não estruturadas, levando em consideração a evolução temporal dos relacionamentos entre os elementos textuais. O modelo, dividido em fases, visa configurar temas de análise, identificar ocorrências de conceitos, realizar correlações e associações temporais, e criar um repositório de temas de análise. Um protótipo é desenvolvido para demonstrar a viabilidade do modelo, aplicando-o em um estudo de caso e realizando uma análise comparativa com outros modelos de descoberta de conhecimento em textos. Heinzle (2012) cria um modelo de engenharia do conhecimento para sistemas de apoio a decisão com recursos para raciocínio abdutivo. Melgar Sasieta (2011) apresenta um modelo para a visualização de conhecimento baseado em imagens semânticas. Beppler (2008), discute-se um modelo inovador de recuperação e busca de informação que integra ontologias e o conceito do círculo hermenêutico. Beppler propõe que os sistemas de busca podem ser aprimorados ao incorporar um processo iterativo de aprendizado e contextualização, o que permite um refinamento contínuo das buscas baseado na interação entre o usuário e a informação. Este modelo visa superar as limitações dos sistemas tradicionais que se baseiam apenas na correspondência direta de termos A dissertação de Ceci (2010) propõe um modelo semi-automático para a construção e manutenção de ontologias a partir de documentos não estruturados. O trabalho utiliza técnicas de extração de informação e agrupamento de documentos, combinadas com validação e classificação das instâncias por meio de bases de conhecimento colaborativas. Um protótipo foi desenvolvido para demonstrar a viabilidade do

modelo, aplicando-o em estudos de caso e análises comparativas, evidenciando sua eficácia tanto na construção inicial quanto na manutenção de ontologias de domínio. Silva (2011) elabora uma arquitetura de BI para processamento analítico baseado em tecnologias semânticas e em linguagem natural.

Quando pesquisado com a *string* de busca "mineração de texto" OR "*text mining*"[4], foram encontrados 38 trabalhos, no Quadro 4, são listados alguns dos trabalhos selecionados, sendo eles 4 teses e 2 dissertações.

Quadro 4 -– Trabalhos resultantes da busca com a palavra-chave: Mineração de texto ou *text mining*

Tipo	Título	Autor	Ano
Tese	Modelo de descoberta de conhecimento em texto para detecção de sinais fracos para tecnologias emergentes	Letícia Silveira Artese	2023
Tese	*Framework* conceitual de representação do conhecimento sobre o 'modelo de graduação dual'	Cleunisse Aparecida Rauen De Luca Canto	2022
Tese	Método de identificação de padrões em discurso político a partir da descoberta de conhecimento	Márcio Welter	2021
Tese	Modelo de mineração de ideias utilizando técnicas de engenharia do conhecimento	Luiz Fernando Spillere de Souza	2021
Dissertação	Redes colaborativas como dinâmica de internacionalização da educação superior: um modelo para avaliar o potencial de compartilhamento de conhecimento	Claudio de Lima	Dr. Alexandre Leopoldo Gonçalves 2023
Dissertação	Análise de agrupamentos e mineração de opinião como suporte à gestão de ideias	Guilherme Martins Alvarez	2018

Fonte: Elaborado pela autora, com base no Repositório Institucional da UFSC/PPGEGC.

Artese (2023) propõe um modelo inovador de descoberta de conhecimento em texto, focado na detecção de sinais fracos que possam indicar o surgimento de tecnologias emergentes. Este modelo utiliza técnicas avançadas de processamento de linguagem natural e análise de grandes volumes de dados para antecipar tendências tecnológicas antes que se tornem

[4]https://repositorio.ufsc.br/handle/123456789/76395/discover?rpp=10&etal=0&query=%E2%80%9Cminera%C3%A7%C3%A3o+de+texto%E2%80%9D+OR+%E2%80%9Ctext+mining%E2%80%9D

amplamente reconhecidas no mercado. Welter (2021) desenvolveu um método de identificação de padrões em discursos políticos através da descoberta de conhecimento, concentrando-se em desvendar as estruturas subliminares do discurso político para entender melhor as estratégias de comunicação utilizadas por políticos durante campanhas e mandatos, oferecendo *insights* valiosos para analistas políticos e cientistas sociais. Lima (2023) investigou as redes colaborativas na educação superior, destacando sua importância como um mecanismo dinâmico para a internacionalização. Ele apresenta um modelo para avaliar o potencial de compartilhamento de conhecimento entre instituições acadêmicas internacionais, enfatizando como essas parcerias podem enriquecer as experiências educacionais e fomentar a inovação. Em um contexto empresarial, Alvarez (2018) analisa a eficácia de agrupamentos e a mineração de opinião para suportar a gestão de ideias dentro das organizações, mostrando como essas técnicas podem ser aplicadas para melhor entender as percepções e expectativas dos stakeholders, facilitando assim a tomada de decisão estratégica e a inovação contínua nas empresas. Souza (2021) propõe um modelo de mineração de ideias utilizando técnicas avançadas de engenharia do conhecimento, focando na otimização da geração e avaliação de ideias inovadoras para permitir que organizações identifiquem e implementem soluções criativas e eficazes para enfrentar desafios complexos. Canto (2022) elaborou um *framework* conceitual para a representação do conhecimento sobre o modelo de graduação dual, visando oferecer uma base teórica e prática para compreender as nuances e o impacto desse modelo educacional, facilitando a implementação e avaliação em diferentes contextos acadêmicos.

Durante a pesquisa no repositório, é notável que diversas abordagens se concentram na elaboração e utilização de ontologias em seus estudos. É importante ressaltar que, embora a ontologia seja um elemento essencial em muitos desses trabalhos, o escopo desta pesquisa enfoca principalmente a aplicação de taxonomia na análise de documentos textuais. Quando o termo "taxonomia"[5] foi pesquisado no repositório, obteve-se o resultado de 94 trabalhos, sendo selecionados alguns trabalhos para serem explorados nesta dissertação (Quadro 5), sendo 4 teses e 3 dissertações.

Quadro 5 – Trabalhos resultantes da busca com a palavra-chave: Taxonomia

Tipo	Título	Autor	Ano
Tese	Avaliação dos portais turísticos governamentais quanto ao	Alexandre Augusto Biz	2009

[5] https://repositorio.ufsc.br/handle/123456789/76395/discover?query=taxonomia&submit=Ir&rpp=10

	suporte à gestão do conhecimento		
Tese	Padrão de projeto de ontologias para inclusão de referências do novo serviço público em plataformas de governo aberto	José Francisco Salm Junior	2012
Tese	Exploração do espaço de *design* das interações humano-computador: uma abordagem da gestão do conhecimento ergonômico	Richard Faust	2013
Tese	Um modelo de gerenciamento da qualidade de experiência para a provisão de serviços cientes de contexto	Madalena Pereira da Silva	2017
Dissertação	Uma ontologia para representação do conhecimento jurídico-penal no contexto dos delitos informáticos	Hélio Santiago Ramos Júnior	2008
Dissertação	Processo para recuperar produtos de inteligência competitiva a partir da memória organizacional	Rodrigo Garcia Rother	2009
Dissertação	Um método para a construção de taxonomias utilizando a DBpedia	Mateus Lohn Andriani	2017

Fonte: Elaborado pela autora, com base no Repositório Institucional da UFSC/PPGEGC.

A dissertação de Andriani (2017) explora a utilização da DBpedia na construção de taxonomias, destacando uma abordagem automatizada que facilita a criação dessas estruturas essenciais para a organização do conhecimento em ambientes digitais, reduzindo os recursos necessários que muitas vezes limitam tais projetos. Rother (2009) em sua dissertação, desenvolve um método para recuperar produtos de inteligência competitiva utilizando a memória organizacional, demonstrando como a integração de processos e ferramentas tecnológicas pode potencializar o acesso e a gestão do conhecimento dentro das empresas. Faust (2013) em sua tese, aborda a gestão do conhecimento ergonômico no design de interações humano-computador, propondo um método que integra taxonomias de argumentos ergonômicos para melhorar a usabilidade e a experiência do usuário em plataformas digitais. Silva (2017) desenvolve uma tese que propõe um modelo de gerenciamento da qualidade de experiência, orientado para serviços cientes de contexto em redes de comunicação, enfatizando a necessidade de uma abordagem multidimensional para avaliar a satisfação do usuário de forma mais eficaz. Salm Junior (2012) na sua tese, explora o desenvolvimento de ontologias

para incluir referências do novo serviço público em plataformas de governo aberto, destacando como essas estruturas podem facilitar a participação cidadã e melhorar a transparência governamental. Ramos Júnior (2008) foca em sua dissertação na criação de uma ontologia para representar o conhecimento jurídico-penal, especificamente em relação aos delitos informáticos, com o objetivo de esclarecer as complexidades desses crimes para cidadãos e profissionais do direito. Biz (2009), em sua tese, avalia os portais turísticos governamentais e sua eficácia na gestão do conhecimento, revelando deficiências significativas na estrutura e uso desses portais como ferramentas estratégicas para a gestão de destinos turísticos.

Utilizando a *string* de busca: taxonomia AND documento AND ("*text mining*" OR "mineração de texto")[6] no repositório institucional da Universidade Federal de Santa Catarina (UFSC), resultou em 12 trabalhos, que são apresentados no Quadro 6.

Quadro 6 - Resultado da pesquisa conforme *string* de busca

Tipo	Título	Autor	Ano
Tese	**Um Método de aquisição de conhecimento para customização de modelos de capacidade/maturidade de processos de *software***	**Jean Carlo Rossa Hauck**	**2011**
Tese	Método de construção de ontologias multilíngues com associação de conceitos a objetos em espaço 3D	César Ramirez Kejelin Stradiotto	Dr. José Leomar Todesco 2011
Tese	Um Modelo de descoberta de conhecimento inerente à evolução temporal dos relacionamentos entre elementos textuais	Alessandro Botelho Bovo	2011
Tese	Desenvolvimento de uma base de conhecimento de casos clínicos de pacientes portadores de desordem temporomandibular	Bertholdo Werner Salles	2009
Tese	Um modelo para recuperação e busca de informação baseado em ontologia e no círculo hermenêutico	Fabiano Duarte Beppler	2008
Tese	Um Modelo para recuperação e comunicação do conhecimento em documentos médicos	Rafael Andrade	2011
Dissertação	Uma proposta de modelo de aquisição de conhecimento para identificação de	Roberto Fabiano Fernandes	2012

6 https://repositorio.ufsc.br/handle/123456789/76395/discover?query=taxonomia+AND+documento+AND+%28%22text+mining%22+OR+%22minera%C3%A7%C3%A3o+de+texto%22%29+&submit=Ir&filtertype_0=dateIssued&filter_relational_operator_0=contains&filter_0=&rpp=10

	oportunidades de negócios nas redes sociais		
Dissertação	Produção científica em área de construção interdisciplinar: educação a distância no Brasil	Fernanda Schweitzer	2010
Dissertação	Um Modelo semi-automático para a construção e manutenção de ontologias a partir de bases de documentos não estruturados	Flávio Ceci	2010
Dissertação	Sistema de conhecimento para gestão documental no setor judiciário: uma aplicação no Tribunal Regional Eleitoral de Santa Catarina	Samuel Fernandes Ribeiro	2010
Dissertação	Sistemas baseados em conhecimento e ferramentas colaborativas para a gestão pública: uma proposta ao planejamento público local	Thiago Paulo Silva de Oliveira	2009
Dissertação	O acesso ao conhecimento em sistemas inteligentes de gestão e análise estratégicas: uma aplicação na segurança pública	Cristina Souza Santos	2006

Fonte: Elaborado pela autora, com base no Repositório Institucional da UFSC/PPGEGC.

Hauck (2011) desenvolveu um método de aquisição de conhecimento que visa a customização de modelos de capacidade e maturidade de processos de software. A pesquisa de Hauck é importante porque propõe a integração de melhores práticas em *frameworks* específicos, mostrando sua aplicabilidade em dois modelos de processo. Schweitzer (2010) analisou a produção científica sobre educação a distância no Brasil, revelando que, apesar de sua relevância, esta área carece de uma maior consolidação no meio acadêmico. Schweitzer destaca que a interdisciplinaridade da educação a distância representa um desafio para a sistematização de conhecimentos e sua publicação. Stradiotto (2011) investigou a criação de ontologias multilingues associadas a objetos em 3D, contribuindo para a universalização do conhecimento e sua apresentação em múltiplos idiomas. Stradiotto enfatiza a importância de tecnologias como Linguagem Universal de Rede (do inglês *Universal Networking Language* - UNL) e do Linguagem de Modelagem Unificada (inglês *Unified Modeling Language* – UML) na representação de conceitos complexos em um formato acessível e compreensível globalmente. Salles (2009) propôs a organização de uma base de conhecimento para casos clínicos de desordem temporomandibular, oferecendo uma ferramenta de apoio significativa para profissionais e acadêmicos da área odontológica. Salles evidencia como a colaboração e a

interatividade podem potencializar o diagnóstico e o tratamento de Desordens Temporomandibular (DTMs).

Oliveira (2009) explora a integração de sistemas baseados em conhecimento e ferramentas colaborativas para aprimorar o planejamento público local, destacando como essas tecnologias podem melhorar a interação entre gestores e cidadãos. A pesquisa de Oliveira sublinha a importância de facilitar o acesso e a organização das informações para fomentar a participação popular nas decisões governamentais, proporcionando um planejamento mais eficaz que reflete as demandas reais da população. Fernandes (2012) aborda a importância das redes sociais como fonte rica em informações e conhecimentos que podem ser cruciais para organizações em busca de inovação e vantagens competitivas. Sua pesquisa propõe um modelo de aquisição de conhecimento que integra técnicas como análise de conteúdo e metodologias como CESM e CommonKADS, com o objetivo de sistematizar e aproveitar os dados disponíveis em redes sociais para identificar oportunidades de negócio.

O trabalho de Andrade (2011) apresenta um modelo para a recuperação de conhecimento em documentos médicos, visando superar os desafios encontrados na busca por informações precisas nesse contexto. A pesquisa propõe técnicas avançadas, como a utilização de detecção de ativos de conhecimento da ontologia DeCS (Descritores em Ciências da Saúde) e de dicionários linguísticos, para ampliar o universo de pesquisa dos usuários e criar uma base de conhecimento reutilizável. Os resultados mostram uma significativa melhoria na eficácia da pesquisa, com uma média de 90% de precisão nos resultados obtidos pelo modelo proposto, em comparação com os 60% do modelo booleano. Além disso, destaca-se a rapidez na obtenção de resultados, com uma redução significativa no tempo de resposta médio, de 49 minutos para 0,6 segundos, demonstrando a eficiência do novo modelo na obtenção de informações médicas relevantes em um curto espaço de tempo.

No decorrer do Mestrado no PPGEGC, foram cursadas disciplinas visando contribuir no desenvolvimento deste trabalho. Na Figura 3 constam as disciplinas cursadas, sendo 5 (cinco) disciplinas na área de concentração de Engenharia do Conhecimento (EC), 2 (duas) disciplinas na área de concentração de Gestão do Conhecimento (GC) e as 3 (três) disciplinas obrigatórias (O). Além destas, a disciplina *Design* de Interface de Usuário com *Design Thinking*, foi cursada no Programa de Pós-Graduação em Ciências da Computação (PPGCC) na modalidade isolada.

Figura 3 – Disciplinas cursadas no PPGEGC

Fonte: Elaborado pela autora.

1.6 ORGANIZAÇÃO DOS CAPÍTULOS

Neste primeiro capítulo, estabelecem-se as bases para compreender o contexto da pesquisa, os motivos da escolha do tema, e a contextualização do problema abordado. Define-se claramente a questão norteadora da pesquisa, seus objetivos e os limites do escopo investigativo. Discute-se também a originalidade, a contribuição e a aderência ao Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento.

No segundo capítulo, realiza-se uma revisão da literatura sobre as principais técnicas e conceitos relacionados à análise de documentos textuais. Abordam-se temas como Processamento de Linguagem Natural (do inglês *Natural Language Processing* - NLP), ML, taxonomias e visualização de dados. Essa revisão permite uma compreensão mais aprofundada do estado da arte nessa área e embasa o método proposto neste trabalho.

No terceiro capítulo, apresenta-se a metodologia utilizada neste trabalho. Utiliza-se a Metodologia de Pesquisa em Ciência do Design (do inglês *Design Science Research Methodology* - DSRM) como abordagem metodológica, que consiste em um processo iterativo de design, implementação e avaliação de soluções para problemas específicos.

No quarto capítulo da dissertação, apresentam-se detalhes abrangentes sobre o método proposto para análise de documentos de texto e a instanciação desse método no sistema DOC *Analysis*. Neste capítulo, exploram-se os passos específicos do método, incluindo a análise da taxonomia definida pelos usuários, a associação dos documentos de texto a essa taxonomia, a identificação de correspondências entre os tópicos abordados nos documentos e os termos da taxonomia, bem como a geração de gráficos visuais intuitivos para visualizar essas relações de forma clara e compreensível. Ao longo do capítulo, discutem-se os processos e algoritmos utilizados para cada etapa do método, fornecendo uma compreensão detalhada de como o sistema realiza a análise e a visualização dos documentos de texto.

No quinto capítulo da dissertação, conduz-se uma avaliação abrangente dos resultados obtidos pelo sistema DOC *Analysis*. Neste capítulo, discutem-se e avaliam-se os resultados da análise de documentos de texto em relação aos objetivos e requisitos estabelecidos. Apresentam-se métricas e critérios de avaliação utilizados para medir a eficácia e a precisão do método proposto. Além disso, discutem-se as limitações e desafios encontrados durante o processo de avaliação, bem como as possíveis melhorias e sugestões para trabalhos futuros. A discussão dos resultados abrange tanto aspectos quantitativos quanto qualitativos, fornecendo uma análise aprofundada do desempenho do sistema DOC *Analysis* e sua relevância para a análise de documentos de texto e visualização de informações.

Por fim, no sexto capítulo, apresentam-se as conclusões desta pesquisa, ressaltando-se as contribuições alcançadas. Além disso, apontam-se sugestões para trabalhos futuros, encerrando o trabalho de forma prospectiva. Por último, listam-se todas as referências bibliográficas utilizadas ao longo do trabalho, bem como os apêndices que fornecem informações adicionais relevantes para a compreensão do trabalho.

2 FUNDAMENTAÇÃO TEÓRICA

No contexto do atual cenário tecnológico, a convergência entre MT, Análise de Dados, Visualização de Dados e Engenharia de *Software* desempenha um papel crucial na compreensão e aproveitamento de grandes conjuntos de dados. A MT, destaca-se como uma abordagem computacional, automatizando a análise de volumes expressivos de texto para extrair informações relevantes em diversas áreas, como finanças, marketing e ciências sociais. Ao enfrentar desafios associados à natureza não estruturada do texto, a preparação dos dados torna-se essencial, envolvendo *tokenização*, *stop words* irrelevantes e normalização de termos.

Paralelamente, a Análise de Dados, processo que envolve coleta, organização e interpretação de informações, destaca-se como uma habilidade essencial em setores variados. O uso de técnicas estatísticas e de ML possibilita a descoberta de *insights* valiosos para fundamentar decisões informadas. A Visualização de Dados, através da tradução de grandes volumes de dados em gráficos e diagramas, assume um papel crucial na identificação de padrões e tendências, sendo a escolha cuidadosa de técnicas de visualização, cores e elementos interativos vital para uma comunicação clara e persuasiva.

A integração dessas práticas com o suporte tecnológico, incluindo ferramentas de *Business Intelligence*, Sistemas de Gerenciamento de Bancos de Dados e plataformas de *Big Data*, fornece a base essencial para auxiliar na tomada de decisão a partir de dados. Por fim, a Engenharia de *Software*, dedicada ao desenvolvimento de metodologias e ferramentas, garante a qualidade e efetividade no processo de construção de sistemas de software. Coletivamente, essas disciplinas moldam um panorama abrangente, capacitando indústrias e campos de pesquisa a enfrentar os desafios do futuro com base em dados sólidos e sistemas eficientes. Nas subseções a seguir serão apresentados cada um dos termos relevantes a esta pesquisa.

2.1 *CORPUS* DE DOCUMENTOS

Os *corpora* de documentos e de texto desempenham papéis cruciais no campo da linguística computacional, fornecendo a base para uma variedade de pesquisas e aplicações. O termo "*corpus*" refere-se a uma coleção sistemática de textos ou documentos utilizados como amostras representativas de uma determinada língua ou domínio linguístico. Um *corpus* de documentos é uma compilação organizada de diversos tipos de registros, como artigos,

relatórios, e-mails e outros materiais textuais. Já o *corpus* de texto é mais restrito, focando-se apenas nas palavras e estruturas linguísticas presentes nos documentos, frequentemente desconsiderando aspectos mais amplos e contextuais.

O diferencial entre um *corpus* de documentos e um *corpus* de texto é crucial para pesquisadores e desenvolvedores de tecnologia da linguagem. Enquanto o primeiro permite uma análise mais abrangente do conteúdo, o segundo se concentra especificamente nas características linguísticas. Essa distinção é fundamental para investigar questões que vão desde a análise de padrões gramaticais até o desenvolvimento de algoritmos de NLP. Como destaca Smith (2018), "o *corpus* de documentos fornece uma visão holística da linguagem em uso, enquanto o *corpus* de texto permite *insights* mais profundos sobre a estrutura e o funcionamento da linguagem".

No âmbito da pesquisa, o foco principal muitas vezes varia entre o estudo de fenômenos linguísticos específicos e a criação de modelos de ML mais eficazes. Como aponta Jones *et al.* (2020), "a escolha entre *corpus* de documentos e *corpus* de texto depende das metas da pesquisa, sendo crucial compreender a natureza das informações desejadas e a profundidade da análise linguística pretendida". Esta decisão estratégica direciona o escopo da pesquisa e determina as abordagens metodológicas adotadas. Essas abordagens complementares desempenham um papel essencial na compreensão da linguagem humana e na criação de tecnologias que dependem do processamento eficaz de dados linguísticos.

A criação e o uso de corpora de textos têm sido amplamente utilizados em diversas áreas da linguística, incluindo a lexicografia, a análise de discurso e a linguística de corpus. Seguindo a abordagem de McEnery e Wilson (1996), o uso do *corpus* como método para coletar dados em estudos linguísticos tornou-se amplamente aceito e tem ganhado maior reconhecimento nos últimos anos. As pesquisas recentes de Baker (2018) e Thompson (2020) mostram que o uso de corpora em estudos linguísticos tem crescido de forma significativa nas últimas décadas, evidenciando sua aceitação e reconhecimento pela comunidade acadêmica. Além disso, a quantidade de publicações e trabalhos que empregam corpora como fonte de dados tem aumentado constantemente, refletindo o papel fundamental desse método na pesquisa linguística contemporânea.

Os *corpora* podem ser classificados de acordo com diferentes critérios, como o tamanho, o tipo de texto e o objetivo da pesquisa. Alguns exemplos incluem o *Corpus* de Referência do Português Contemporâneo (CRPC), representado por textos da língua portuguesa de diferentes gêneros e estilos; o British National *Corpus* (BNC), um *corpus* em inglês britânico

com mais de 100 milhões de palavras; e o *Corpus* Brasileiro de Aprendizes de Português como Língua Estrangeira (CBAPE), voltado ao aprendizado de português como língua estrangeira.

Os corpora podem ser usados para uma variedade de propósitos, como a identificação de padrões linguísticos, a criação de recursos lexicais, a análise de gêneros discursivos e a avaliação de materiais didáticos. De acordo com Baker (2006), "o uso de corpora como base para a pesquisa linguística está mudando o modo como os pesquisadores olham para a língua". Além disso, o uso de corpora também tem se expandido para áreas fora da linguística, como a computação e a inteligência artificial, onde podem ser utilizados para treinar algoritmos de ML e sistemas de NLP. Segundo Bird e Simons (2003), "os corpora são vistos como uma fonte de conhecimento para a construção de modelos de linguagem computacional".

A criação de corpora também envolve desafios, como a seleção adequada dos textos e a codificação dos dados. Além disso, a disponibilidade pode ser limitada em algumas línguas e áreas de estudo. Conforme Biber (1993), a construção de corpora é um processo custoso e trabalhoso, podendo representar um desafio em algumas áreas de pesquisa devido à dificuldade de obtenção dos dados necessários. A criação e o uso de corpora também envolvem desafios e limitações que precisam ser considerados pelos pesquisadores e desenvolvedores. Um dos principais desafios é a seleção apropriada dos textos que compõem o *corpus*. Como observado por Biber (1993), a construção de corpora pode ser um processo caro e trabalhoso. Além disso, encontrar uma amostra representativa que capture a diversidade linguística e textual de interesse pode ser uma tarefa complexa e exigente.

Outro desafio enfrentado pelos pesquisadores é a codificação e anotação dos dados no *corpus*. Esse processo envolve a aplicação de metadados e marcações linguísticas, o que demanda tempo e expertise linguística. Bird e Simons (2003) ressaltam que a qualidade da anotação dos corpora é crucial para garantir a confiabilidade e utilidade dos dados para análise. Além disso, a disponibilidade de corpora pode ser limitada em algumas línguas e áreas de estudo. Isso pode ser um obstáculo para pesquisadores que trabalham em contextos linguísticos menos estudados ou em domínios específicos de pesquisa. Baker (2006) destaca a importância de expandir o acesso a corpora em diferentes línguas e áreas de estudo para promover a diversidade e inclusão na pesquisa linguística.

Esta dissertação terá como foco um *corpus* de documentos textuais, enfatizando a análise abrangente do conteúdo e a compreensão holística da linguagem em uso.

2.1.1 ANÁLISE DE DOCUMENTOS TEXTUAIS

A análise de documento é uma abordagem valiosa na pesquisa linguística aplicada, oferecendo *insights* significativos sobre padrões, variações e tendências linguísticas. De acordo com Stubbs (1996), um *corpus* pode ser definido como uma coleção sistemática de textos autênticos que serve como base para investigações linguísticas. Ao examinar essa amostragem representativa da linguagem, pesquisadores podem compreender a complexidade e a diversidade das expressões linguísticas em contextos específicos.

Um aspecto fundamental na análise de *corpus* de documento é a seleção cuidadosa de fontes e gêneros textuais. Biber (1993) destaca a importância de incluir uma variedade de registros, como textos acadêmicos, literários, jornalísticos e conversacionais, para garantir uma representação abrangente da linguagem em uso. Essa diversidade permite a identificação de características linguísticas distintas em diferentes contextos, enriquecendo a compreensão da linguagem em sua totalidade.

Ao explorar um *corpus* de documento, é possível identificar padrões lexicais e semânticos. Para Stubbs (2001), a análise de frequência de palavras e a identificação de colocações são estratégias essenciais nesse processo. Essas técnicas fornecem uma visão quantitativa e qualitativa do vocabulário, revelando nuances e sutilezas na escolha lexical que podem passar despercebidas em abordagens menos sistemáticas.

Assim, a análise de *corpus* de documento também desempenha um papel crucial na pesquisa em análise de discurso. Fairclough (2003) destaca que a identificação de padrões discursivos em contextos específicos permite uma compreensão mais profunda das práticas sociais e das relações de poder que permeiam a linguagem. Essa abordagem é particularmente relevante em estudos que exploram a linguagem em ambientes políticos, midiáticos e institucionais.

No campo da tecnologia da informação, a análise de *corpus* de documento é fundamental para o desenvolvimento de algoritmos de NLP. Manning e Schütze (2008) ressaltam que a compreensão dos padrões linguísticos presentes nos documentos é essencial para aprimorar a eficácia de sistemas de tradução automática, *chatbots* e outras aplicações de NLP.

2.1.1.1 Desafios na Análise Qualitativa de Documentos

A análise qualitativa de documentos é uma abordagem valiosa na pesquisa social, permitindo uma compreensão aprofundada de contextos, significados e construções discursivas. Entretanto, este método enfrenta desafios significativos que afetam sua aplicação efetiva, especialmente ao lidar com grandes volumes de documentos sem escalabilidade. Vincular essa abordagem à interpretação humana torna-se complexo devido à necessidade de lidar com a quantidade massiva de dados (JONES, 2015). Aqui estão algumas considerações importantes:

- **Volume de Documentos:** Lidar com grandes volumes de documentos pode ser demorado e desafiador para os pesquisadores humanos. A capacidade de processar uma quantidade significativa de dados de maneira eficiente torna-se uma preocupação, pois pode resultar em análises superficiais ou na perda de informações relevantes.
- **Padronização e Consistência:** A interpretação humana pode variar e a análise qualitativa requer consistência na aplicação de critérios e categorias. Em grandes conjuntos de documentos, manter uma abordagem padronizada pode ser difícil, levando a resultados pouco confiáveis. A falta de padronização na produção de documentos é um desafio adicional. A variação na linguagem, estilo e estrutura dos documentos pode complicar a análise, exigindo abordagens flexíveis. Conforme mencionado por Miles e Huberman (2014), os pesquisadores devem ser capazes de lidar com a heterogeneidade dos documentos para extrair significados válidos. Documentos produzidos em diferentes contextos podem usar terminologias, formatos e estruturas variados, o que pode dificultar a aplicação de uma metodologia uniforme e consistente em toda a análise.
- **Dificuldade na Identificação de Padrões Ocultos:** Com grandes volumes de documentos, a identificação de padrões subjacentes ou *insights* significativos pode ser desafiadora para os pesquisadores. A capacidade humana de processamento pode ser limitada nesse contexto, resultando na possível perda de informações cruciais.
- **Necessidade de Ferramentas de Apoio:** Para superar esses desafios, é essencial incorporar ferramentas de análise de texto e Inteligência artificial (do inglês *Artificial Intelligence* - AI) (BROWN, 2018). Essas ferramentas podem ajudar na categorização, extração de informações-chave e identificação de padrões, aliviando a carga sobre os pesquisadores humanos e permitindo uma análise mais abrangente.
- **Validação e Confiabilidade:** A interpretação humana muitas vezes requer validação inter-codificadores para garantir a confiabilidade dos resultados (JOHNSON, 2019).

Isso pode ser particularmente difícil de realizar em grandes volumes de documentos, exigindo estratégias adicionais para garantir a precisão e a validade das conclusões.

- **Desafios na Análise em Relação à Questão Legal - LGPD:** Além dos desafios metodológicos e técnicos, a análise qualitativa de documentos enfrenta questões legais significativas, como a conformidade com a Lei Geral de Proteção de Dados (LGPD) no Brasil. A LGPD impõe rigorosas obrigações sobre o tratamento de dados pessoais, incluindo a necessidade de obter consentimento explícito dos indivíduos cujos dados estão sendo analisados, garantir a segurança desses dados e respeitar os direitos dos titulares. A não conformidade pode resultar em penalidades severas. Assim, os pesquisadores devem estar cientes dessas exigências legais e incorporar práticas de conformidade rigorosas em seu processo de análise para proteger a privacidade e os direitos dos indivíduos (BRASIL, 2018).

A análise qualitativa de documentos oferece uma compreensão aprofundada (SMITH, 2010). Integrar essa abordagem com grandes volumes de documentos sem escala requer a combinação inteligente de capacidades humanas e tecnológicas para superar desafios de eficiência e confiabilidade. Um estudo realizado pela Universidade de Berkeley revelou que a avalanche de dados enfrentada pela sociedade contemporânea é ainda mais significativa do que se pensava (BROWN, 2018). Conduzido nos Estados Unidos, a pesquisa evidenciou que as pessoas dedicam quase 12 horas diárias ao consumo de informações, absorvendo mais de 100 mil delas por meio da leitura ou audição.

A crescente influência e saturação de dados na vida cotidiana têm sido amplamente discutidas na literatura acadêmica. Como destacado por Mayer-Schönberger e Cukier (2013), vivemos em uma era de "*big data*", onde enormes volumes de informações são gerados e coletados em tempo real, influenciando diversos aspectos da sociedade, desde a tomada de decisões individuais até políticas públicas e estratégias de negócios. Essa avalanche de dados é resultado do uso generalizado de tecnologias digitais, como smartphones, redes sociais, dispositivos Internet das Coisas (em inglês *Internet of Things* – IoT) e sistemas de monitoramento, que geram um fluxo incessante de informações.

A compreensão e o gerenciamento consciente desse volume massivo de dados tornaram-se imperativos na sociedade contemporânea. Conforme ressaltado por Boyd e Crawford (2012), a análise crítica e reflexiva dos dados é essencial para evitar armadilhas como vieses algorítmicos, privacidade invasiva e discriminação algorítmica. Além disso, a

capacidade de discernir informações relevantes em meio ao ruído de dados é fundamental para tomar decisões informadas e desenvolver estratégias eficazes em diversos contextos.

Um dos desafios principais na análise qualitativa de documentos é a subjetividade na interpretação dos dados. Como aponta Silverman (2013), os pesquisadores podem estar sujeitos a viés interpretativo ao analisar documentos, influenciando a forma como interpretam e atribuem significado aos textos. Isso destaca a necessidade de uma reflexão constante sobre as posições e perspectivas do pesquisador durante o processo analítico.

Outro obstáculo enfrentado na análise qualitativa de documentos é a falta de contexto. Documentos muitas vezes são produzidos sem a intenção de serem utilizados em pesquisas acadêmicas, o que pode resultar em informações incompletas ou ambíguas. Conforme destacado por Krippendorff (2018), a ausência de informações contextuais pode limitar a compreensão profunda do significado subjacente nos documentos analisados.

A diversidade de formatos documentais também apresenta desafios consideráveis. Livros, artigos, relatórios, cartas e outros tipos de documentos exigem abordagens analíticas específicas. Flick (2018) ressalta que adaptar as técnicas de análise para diferentes gêneros documentais é crucial para evitar generalizações inadequadas e garantir a validade dos resultados.

Outro ponto a ser considerado é a questão ética na análise qualitativa de documentos. Documentos frequentemente contêm informações sensíveis ou confidenciais, exigindo uma abordagem ética na coleta, análise e divulgação dos dados. Lincoln e Guba (1985) argumentam que os pesquisadores devem buscar maneiras de preservar a confidencialidade e a privacidade das informações, garantindo uma prática ética de pesquisa.

2.2 ENGENHARIA DO CONHECIMENTO

A Engenharia do Conhecimento (do inglês *Knowledge Engineering* - KE) é um campo interdisciplinar, conforme Motta (1998) que se dedica à captura, representação e utilização do conhecimento em sistemas computacionais. Segundo o autor, esse processo envolve a identificação, coleta e estruturação do conhecimento relevante para um determinado problema ou domínio, com o objetivo de modelá-lo em formatos computacionais, como ontologias, regras de inferência e redes semânticas. Motta (1998) destaca ainda que a engenharia do conhecimento

trabalha na integração e aplicação desse conhecimento em sistemas inteligentes, bem como na manutenção e evolução do conhecimento ao longo do tempo.

Segundo Fensel *et al*. (2001, p.2), "a engenharia do conhecimento é um processo sistemático e disciplinado para desenvolver sistemas baseados em conhecimento". Esses sistemas podem ser utilizados em diversas áreas, como por exemplo a medicina, finanças e educação.

A captura do conhecimento (do inglês *Knowledge Capture*) é um dos principais desafios enfrentados pela KE. Segundo Chibelushi e Khan (2002, p.22), a captura de conhecimento envolve "a extração de informações relevantes do especialista, sua validação e transformação em modelos computacionais". Esse processo pode ser realizado por meio de entrevistas com especialistas, análise de documentos e observação do ambiente, entre outras técnicas.

Uma vez capturado o conhecimento, inicia-se a etapa para representá-lo de maneira adequada, a fim de ser utilizado por sistemas computacionais. Segundo Brachman e Levesque (2004, p.7), a representação do conhecimento implica na seleção de uma linguagem simbólica, na identificação dos conceitos pertinentes e na especificação das relações entre eles. Dessa forma, é possível criar estruturas e modelos de conhecimento que possam ser utilizados em sistemas computacionais.

A utilização do conhecimento é outro desafio enfrentado pela KE. De acordo com Uschold e Gruninger (2004, p.17), a utilização do conhecimento envolve "o processamento das informações representadas, a inferência de novas informações a partir das informações existentes e a tomada de decisões baseadas nessas informações". Para isso, é necessário utilizar técnicas de raciocínio e de inferência para obter novas informações a partir das informações existentes.

A KE tem um papel importante na evolução da inteligência artificial. Brachman (2006, p.6), afirma que "[...] a engenharia do conhecimento é um meio crucial para tornar a inteligência artificial mais efetiva, eficiente e acessível". Sendo assim, a utilização de conhecimentos humanos aliados com sistemas computacionais pode levar a resultados mais precisos e confiáveis em diversas áreas.

Ademais, a KE é uma área em constante evolução em que novas técnicas e abordagens são desenvolvidas ou atualizadas continuamente. Fensel *et al*. (2001, p.2), afirmam que " a Engenharia do Conhecimento é uma área em constante mudança, e novos métodos e técnicas estão sendo desenvolvidos regularmente para superar os desafios enfrentados". É imperativo

que profissionais estejam propensos a explorar e adotar as mais recentes abordagens e técnicas. Nesse contexto em evolução, a organização e representação do conhecimento desempenham um papel crucial na estruturação e aplicação eficaz dessas inovações, garantindo que o conhecimento humano seja capturado, representado e utilizado de maneira cada vez mais eficiente e efetiva.

2.2.1 Organização e Representação do Conhecimento

A organização e representação do conhecimento desempenham papéis cruciais na disseminação de informações e na tomada de decisões. Segundo Nascimento e Garcia (2014), a organização do conhecimento permite que se tenha uma visão mais clara e estruturada das informações, o que facilita o acesso e o uso delas.

Uma das principais formas de organização e representação do conhecimento é por meio da taxonomia. De acordo com Hodge (2000), a taxonomia é uma classificação sistemática dos conceitos, baseada em suas características e relações, que permite a organização e o acesso às informações.

Além da taxonomia, outras formas de representação do conhecimento incluem ontologias, tesauros e mapas conceituais. Para Nascimento e Garcia (2014), essas formas de representação são importantes para garantir a precisão e a clareza das informações, além de permitir a interoperabilidade entre sistemas e a busca por informações relacionadas.

A estruturação e representação do conhecimento desempenham um papel crucial na gestão eficaz da informação e na tomada de decisões dentro das organizações. De acordo com Hodge (2000), a organização do conhecimento é um processo em curso que possibilita a criação de estruturas facilitadoras para a recuperação e aplicação do conhecimento, impulsionando a inovação e o progresso organizacional.

As ontologias, conforme definidas por Fonseca *et al.* (2002), são modelos formais de representação de conhecimento que descrevem conceitos, suas propriedades e relações em um domínio específico. Cada vez mais comuns em sistemas de informação, as ontologias possibilitam uma representação compartilhada e precisa do conhecimento, facilitando a tomada de decisões e a resolução de problemas em diversos domínios. Segundo Studer (1998), uma ontologia é uma especificação explícita de uma conceitualização compartilhada, sendo uma ferramenta crucial na construção de sistemas de informação eficientes.

Uschold e Gruninger (1996) destacam que a utilização de ontologias possibilita aos sistemas de informação compreender o significado dos dados manipulados, facilitando a integração de diversas fontes de informação e a realização de inferências. Assim, a construção de ontologias torna-se uma atividade fundamental na Engenharia do Conhecimento, permitindo a representação clara e precisa do conhecimento de um determinado domínio, conforme ressaltado por Fonseca *et al.* (2002) e Studer (1998).

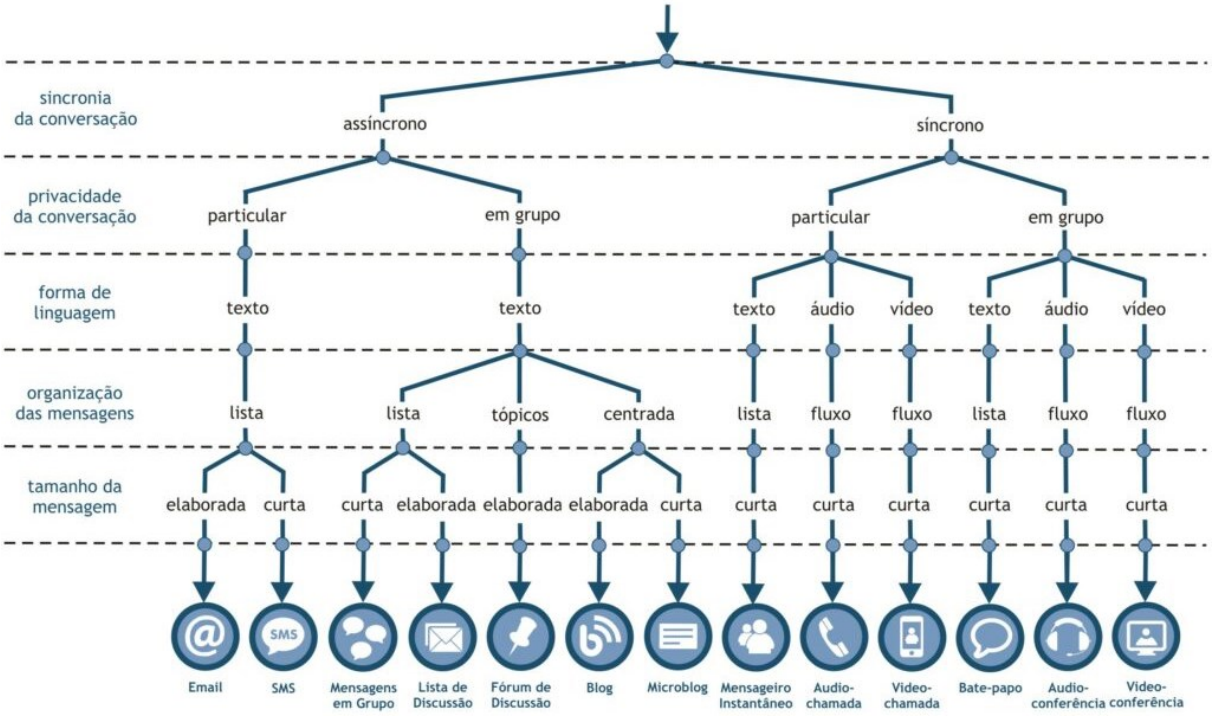
2.2.2 Taxonomias

Taxonomias são estruturas organizacionais de classificação que permitem a categorização de conceitos em diferentes níveis de abstração, agrupando itens semelhantes e permitindo sua recuperação e navegação de forma mais eficiente (SMITH *et al.*, 2019). Essas estruturas são utilizadas em diversas áreas, como na biologia, na biblioteconomia e na ciência da informação, e têm sido amplamente aplicadas em sistemas de informação para organização e recuperação de informações.

Segundo Zhang e Gao (2020), a utilização de taxonomias é uma estratégia eficiente para facilitar a navegação e a busca de informações em sistemas de informação, pois permite que os usuários possam se orientar e encontrar conteúdos relevantes de forma mais rápida e intuitiva. Além disso, as taxonomias também auxiliam no processo de indexação e recuperação de informações, permitindo que diferentes usuários possam encontrar conteúdos relevantes utilizando diferentes termos de busca.

É importante destacar que a construção de taxonomias requer um cuidadoso processo de análise e categorização dos conceitos a serem classificados, levando em consideração as características e relações entre eles (SMITH *et al.*, 2019). Além disso, as taxonomias devem ser atualizadas constantemente para refletir as mudanças no domínio em que são utilizadas, garantindo sua relevância e utilidade. A Figura 4 representa um exemplo de taxonomia, o qual foi usado uma árvore invertida para representar graficamente a taxonomia de diferenciação dos meios de conversação.

Figura 4 - Exemplo de Taxonomia



Fonte: Calvão; Pimentel; Fuks (2014).

A utilização de taxonomias em sistemas *web* é uma estratégia comum para organizar e classificar informações em ambientes digitais. Segundo Liddy e Teredesai (2005), as taxonomias podem ser aplicadas em diferentes tipos de sistemas, como portais corporativos, sistemas de comércio eletrônico e motores de busca, permitindo que os usuários possam encontrar informações relevantes de forma mais eficiente.

De acordo com Broughton (2006), as taxonomias são particularmente úteis em sistemas *web* que lidam com grande volume de informações, pois permitem a organização e categorização dos conteúdos de forma hierárquica e intuitiva, facilitando a navegação e a busca de informações pelos usuários. Além disso, as taxonomias também auxiliam no processo de indexação e recuperação de informações, permitindo que diferentes usuários possam encontrar conteúdos relevantes utilizando diferentes termos de busca.

2.2.2.1 Abordagens para a Construção de Taxonomias

A construção de taxonomias é um processo complexo que requer a utilização de abordagens específicas para garantir sua efetividade na organização de informações. Segundo Marcondes e Del Nero (2011), as abordagens mais comuns para a construção de taxonomias são a *bottom-up* e a *top-down*. Na abordagem *bottom-up*, as categorias são identificadas a partir das informações existentes, enquanto que na abordagem *top-down*, as categorias são definidas previamente e as informações são classificadas de acordo com essas categorias.

Além das abordagens *bottom-up* e *top-down*, existem outras abordagens que podem ser utilizadas na construção de taxonomias, como a abordagem de modelagem de conceitos e a abordagem de análise de agrupamentos (*clusters)*. Segundo Liu e Li (2015), a abordagem de modelagem de conceitos consiste em identificar os conceitos e as relações entre eles, enquanto a abordagem de análise de *clusters* agrupa as informações com base em suas similaridades.

Conforme Broughton (2006), a escolha da abordagem para a construção de taxonomias depende do contexto em que serão utilizadas e dos objetivos a serem alcançados. A abordagem *bottom-up*, por exemplo, é mais indicada quando há grande variedade de informações e é importante identificar novas categorias. Já a abordagem *top-down* é mais indicada quando há uma estrutura pré-definida e é importante garantir a consistência e a uniformidade na classificação das informações.

2.2.2.2 Classificação de Taxonomias

Uma abordagem comum na classificação de taxonomias é a utilização de algoritmos de ML. De acordo com Menczer e Ahn (2011), algoritmos como o KNN[7] (*k-Nearest Neighbors*) e o SVM[8] (*Support Vector Machine*) podem ser utilizados para classificar taxonomias de acordo com suas características. Esses algoritmos consideram a similaridade entre as taxonomias, analisando as relações entre os termos e suas frequências.

Outra abordagem para a classificação de taxonomias é a utilização de técnicas de análise semântica. A análise semântica tem se mostrado uma ferramenta poderosa para identificar a similaridade entre as taxonomias com base na semântica dos termos. De acordo com Salloum, Khan e Shaalan (2020), a análise semântica, incluindo a análise semântica explícita e a análise semântica latente, contribui para o aprendizado de linguagens naturais e

[7] K-ésimo vizinho mais próximo ou algoritmo de vizinho mais próximo (tradução nossa).
[8] Máquina de vetores de suporte (tradução nossa).

textos, permitindo que os computadores processem linguagens naturais. Além disso, Suzen, Gorban, Levesley e Mirkes (2021) destacam que a análise semântica pode prever o impacto futuro de artigos científicos com base na semântica das palavras usadas no texto. Portanto, a análise semântica desempenha um papel crucial na identificação da similaridade entre as taxonomias. Essa abordagem utiliza técnicas de NLP para analisar as descrições dos termos e suas relações.

Além disso, é possível utilizar abordagens híbridas na classificação de taxonomias. Segundo Deng *et al*. (2019), uma abordagem que combina análise semântica e ML pode ser mais eficaz na classificação de taxonomias do que abordagens que utilizam apenas uma dessas técnicas.

Outra abordagem para avaliar taxonomias é por meio da aplicação de medidas de avaliação. Conforme apontado por Liao *et al.* (2019), métricas como precisão, *recall* e F-*measure* são empregadas para avaliar o desempenho da classificação das taxonomias. Essas medidas possibilitam a análise da eficácia da classificação em relação às taxonomias reais e auxiliam na identificação de eventuais problemas durante o processo de classificação.

Além das medidas mencionadas, outras métricas de avaliação, como a entropia e a pureza das classes, também podem ser utilizadas para avaliar a qualidade da classificação das taxonomias. A entropia mede a incerteza associada à distribuição das classes em um conjunto de dados, enquanto a pureza avalia a homogeneidade das classes em cada categoria da taxonomia. Essas métricas adicionais fornecem uma visão mais abrangente do desempenho da classificação e ajudam a identificar áreas específicas que podem exigir ajustes ou melhorias. Combinadas, essas medidas de avaliação oferecem uma abordagem abrangente para garantir a precisão e a eficácia das taxonomias em diversos contextos de aplicação.

2.2.2.3 Desenvolvimento de Taxonomias

O desenvolvimento de taxonomias é um processo crucial para a organização e classificação de informações em diferentes áreas, incluindo biblioteconomia, ciência da informação e tecnologia da informação. Segundo Lopes e Barbosa (2019), o processo de desenvolvimento de taxonomias envolve a definição de conceitos, a criação de hierarquias e a determinação de relações entre os termos.

Uma abordagem comum para o desenvolvimento de taxonomias é a utilização de especialistas no domínio em questão para a identificação e organização dos termos (WANG *et al*., 2019). Porém, essa abordagem pode ser limitada pela subjetividade dos especialistas e pela falta de padronização no processo de classificação (BAKER *et al*., 2018).

Uma alternativa para a construção de taxonomias é a utilização de técnicas automatizadas, como a mineração de texto e a análise de redes (LI *et al*., 2018). Essas técnicas permitem a identificação de relações entre os termos de forma objetiva e eficiente, além de possibilitar a análise de grandes quantidades de dados.

É importante ressaltar que o desenvolvimento de taxonomias é um processo contínuo e que as taxonomias devem ser revisadas e atualizadas periodicamente para garantir sua relevância e precisão (LOPES; BARBOSA, 2019). Além disso, a colaboração entre diferentes especialistas e a utilização de padrões e vocabulários controlados também são importantes para a construção de taxonomias eficazes (BAKER *et al*., 2018).

Portanto, o desenvolvimento de taxonomias é um processo complexo e importante para a organização e classificação de informações em diferentes áreas. A utilização de técnicas automatizadas pode ser uma alternativa eficiente, porém é necessário garantir a revisão e atualização constante das taxonomias para que sejam precisas e relevantes, o que pode ser uma tarefa desafiadora sem os meios e ou sistemas adequados.

2.2.2.4 Justificativa para o uso de Taxonomias

O uso de taxonomias é justificado, pois essas estruturas de classificação proporcionam uma maneira organizada e sistemática de categorizar informações, facilitando a recuperação e a compreensão de dados complexos. Como mencionado por diversos estudiosos no campo da ciência da informação, a aplicação de taxonomias permite a padronização da terminologia e a criação de relações claras entre os diferentes conceitos, promovendo a interoperabilidade e a integração de sistemas de informação (SVENONIUS, 2000; MAI, 2015; LAM, 2018). Além disso, as taxonomias possibilitam a análise e a visualização de padrões e tendências, contribuindo para a tomada de decisões mais embasadas e eficazes. As taxonomias são ferramentas poderosas para mapear o conhecimento em várias disciplinas. Como Bloom (1956) argumentou, uma taxonomia bem estruturada pode facilitar a organização do conhecimento de maneira hierárquica, tornando o aprendizado mais eficiente e eficaz. Da mesma forma, Anderson e Krathwohl (2001) expandiram essa ideia, sugerindo que as taxonomias podem ser

usadas não apenas para mapear o conhecimento, mas também para promover o pensamento crítico e a compreensão profunda. Portanto, as taxonomias desempenham um papel crucial na estruturação e no mapeamento do conhecimento.

Outra justificativa para o uso de taxonomias é a sua capacidade de auxiliar na tomada de decisão e na resolução de problemas. Ao estruturar o conhecimento de uma determinada área, as taxonomias podem ser utilizadas para identificar *gaps* e lacunas no conhecimento, auxiliando na identificação de soluções para problemas específicos (BECHHOFER *et al*., 2010). As taxonomias continuam a ser uma ferramenta valiosa em várias disciplinas, ajudando na organização e classificação de informações. De acordo com Anderson e Krathwohl (2001), a taxonomia pode promover a aprendizagem ao fornecer uma estrutura clara para a compreensão dos conceitos. Além disso, Hotho *et al.* (2005) sugerem que as taxonomias podem auxiliar na tomada de decisões ao fornecer um quadro de referência para a avaliação e comparação de diferentes opções. Portanto, as taxonomias desempenham um papel crucial na estruturação do conhecimento e na facilitação do processo de tomada de decisões.

O uso de taxonomias é justificado não apenas pela sua capacidade de organizar e classificar informações, mas também pelo seu potencial de aumentar a eficiência em diversos contextos. Segundo Bloom *et al.* (2001), as taxonomias podem aumentar a eficiência do processo de aprendizagem ao fornecer uma estrutura clara para a compreensão dos conceitos. Além disso, de acordo com Hotho *et al.* (2005), as taxonomias podem aumentar a eficiência na tomada de decisões ao fornecer um quadro de referência para a avaliação e comparação de diferentes opções.

As taxonomias podem ser utilizadas para facilitar a comunicação entre diferentes áreas do conhecimento, permitindo que profissionais de diferentes áreas possam se comunicar de forma mais eficiente e precisa (HEERY, 2004). A estrutura hierárquica da taxonomia permite que conceitos e termos sejam organizados de forma clara e objetiva, possibilitando uma melhor compreensão dos diferentes tópicos e áreas do conhecimento.

Por fim, o uso de taxonomias também é justificado pela sua capacidade de permitir a integração de diferentes fontes de informação em um único sistema. As taxonomias podem ser utilizadas para integrar informações de diferentes formatos e tipos, permitindo uma busca mais abrangente e completa das informações relevantes em um determinado contexto (LUO *et al*., 2015).

2.2.3 DESCOBERTA DE CONHECIMENTO EM TEXTOS

A Descoberta do Conhecimento em Textos (*Knowledge Discovery in Text* - KDT) é um processo que envolve a extração de informações úteis e desconhecidas de grandes conjuntos de documentos de texto (SMITH; JOHNSON, 2020; CHEN; LIU, 2021; WANG; ZHANG, 2022).

Segundo Konchady (2001), "a KDT é uma subárea da KDD que lida com a extração de conhecimento útil de textos". A KDT é uma área de pesquisa em rápido desenvolvimento que tem o potencial de transformar muitas indústrias e campos de pesquisa. KDT também envolve a aplicação de técnicas de mineração de dados em grandes conjuntos de dados textuais para extrair informações úteis e desconhecidas (FELDMAN; SANGER, 2007). O objetivo é "descobrir conhecimento que possa ser útil para tomadores de decisão" (FELDMAN; SANGER, 2007). Isso pode incluir a identificação de tendências, padrões, sentimentos e relações em documentos de texto. Utilizando uma variedade de algoritmos, a KDT transforma dados não estruturados em conhecimento acionável, auxiliando na tomada de decisões informadas e na geração de *insights* estratégicos. O Quadro 7 apresenta um resumo sobre os principais algoritmos de KDT, suas utilizações e exemplos reais onde são aplicados.

Quadro 7 – Principais algoritmos de KDT

Algoritmo	Utilização	Exemplos Reais
Mineração de Texto (*Text Mining*)	Extração de padrões e tendências a partir de grandes volumes de texto.	Análise de artigos científicos para descobrir novos padrões de pesquisa.
Análise de Sentimento (*Sentiment Analysis*)	Determinação da atitude emocional expressa em um texto.	Monitoramento de opiniões de consumidores em redes sociais sobre produtos.
Classificação de Texto (*Text Classification*)	Organização e categorização automática de documentos.	Filtragem de spam em e-mails; categorização de notícias em portais de notícias.
Extração de Entidades (*Entity Extraction*)	Identificação e extração de entidades nomeadas, como pessoas, locais e organizações.	Identificação de personagens em livros; mapeamento de empresas mencionadas em artigos de negócios.
Agrupamento de Documentos (*Document Clustering*)	Agrupamento de textos semelhantes para facilitar a análise.	Agrupamento de resenhas de produtos por similaridade de opinião.
Detecção de Tópicos (*Topic Detection*)	Identificação dos principais temas abordados em um conjunto de documentos.	Descoberta de tópicos recorrentes em fóruns de discussão online.
Resumo de Texto (*Text Summarization*)	Criação de resumos automáticos de documentos longos.	Geração de resumos automáticos de artigos de notícias e relatórios de pesquisa.
Reconhecimento de Entidades Nomeadas (do inglês *Named Entity Recognition* - NER)	Identificação e classificação de entidades mencionadas em um texto.	Análise de documentos legais para identificar nomes de partes envolvidas.
Análise de Coocorrência de Termos (*Term Co-occurrence Analysis*)	Identificação de termos que frequentemente aparecem juntos.	Análise de coocorrência de palavras em literatura médica para encontrar relações entre sintomas e doenças.
Classificação de Textos com Redes Neurais (*Text Classification with Neural Networks*)	Uso de redes neurais para melhorar a precisão da classificação de textos.	Implementação de sistemas de recomendação em plataformas de *streaming*.

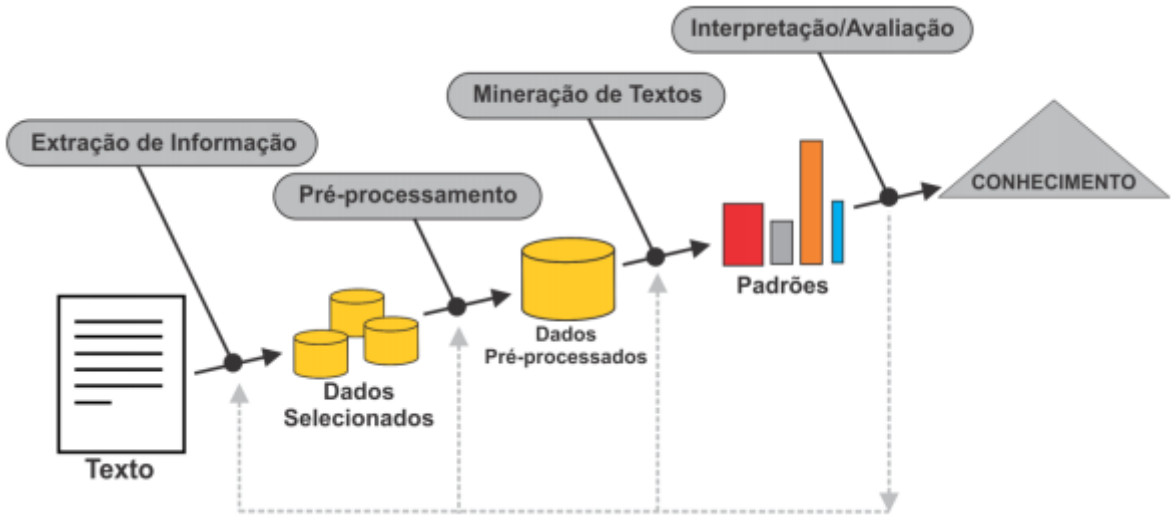
Fonte: Elaborado pela autora com base em TAN, *et al.* (2017); PANG, LEE (2008); SEBASTIAN (2002); BORTHWICK *et al.* (1998).

Um dos primeiros passos do processo de KDT é a preparação dos dados. Segundo Witten, Frank e Hall (2016), a preparação dos dados é o processo de converter os documentos de texto em um formato que possa ser processado por algoritmos de KDT. Isso envolve a *tokenização* dos documentos em palavras ou termos, a eliminação de palavras irrelevantes (*stop words*[9]) e a normalização dos termos (por exemplo, reduzindo todas as palavras ao seu radical). A preparação dos dados é importante porque ajuda a reduzir o tamanho do conjunto de dados e aumenta a eficácia dos algoritmos de KDT.

Outro passo importante do processo de KDT é a extração de recursos. Segundo Manning, Raghavan e Schütze (2008), a extração de recursos é o processo de identificar características úteis dos documentos de texto que possam ser usadas para descobrir padrões e relações nos dados. Isso pode incluir a extração de palavras-chave, a identificação de entidades nomeadas (como nomes de pessoas, locais e organizações) e a análise de sentimentos (para determinar a opinião expressa em um documento). A extração de recursos é importante porque ajuda a reduzir a dimensão do conjunto de dados e aumenta a qualidade dos resultados.

Após a preparação dos dados e a extração de recursos, é possível aplicar algoritmos de KDT para descobrir padrões e relações nos dados. Segundo Aggarwal e Zhai (2012), existem muitos algoritmos de KDT disponíveis, incluindo modelos de tópicos, análise de sentimento, agrupamento de documentos e classificação de documentos. Cada algoritmo é adequado para um conjunto diferente de problemas de KDT, e a escolha do algoritmo depende do objetivo da análise e das características dos dados. A Figura 5 representa as etapas do processo de KDT.

Figura 5 - Etapas do processo de KDT



Fonte: UFS – GISI (2015)

Uma das aplicações mais comuns da KDT é a análise de sentimentos em mídias sociais. Segundo Liu (2012), a análise de sentimentos é o processo de determinar a opinião expressa em um documento, como positiva, negativa ou neutra. A análise de sentimentos pode ser usada para monitorar a opinião do público sobre uma empresa ou produto, prever tendências de mercado e identificar influenciadores em mídias sociais.

Logo, a KDT é uma área de pesquisa importante que tem o potencial de transformar muitas indústrias e campos de pesquisa. A preparação dos dados, a extração de recursos e a aplicação de algoritmos de KDT são os principais passos do processo de KDT. Existem muitos algoritmos de KDT disponíveis, cada um adequado para diferentes problemas de análise de texto.

2.2.3.1 MINERAÇÃO DE TEXTO

A Mineração de Texto (do inglês *Text Mining* – TM) é um campo de pesquisa interdisciplinar que envolve a análise de grandes volumes de texto para extrair informações relevantes e úteis (FELDMAN; SANGER, 2007). Segundo Gómez-Rodríguez *et al.* (2017), "a mineração de texto é uma abordagem computacional que permite analisar grandes volumes de textos de forma automática e extrair informações relevantes". O TM tem aplicações em uma ampla variedade de campos, incluindo finanças, *marketing*, ciência política e ciências sociais, a Figura 6 ilustra alguns dos exemplos de aplicações que utilizam TM para explorar recursos.

Figura 6 - Aplicações que utilizam *Text Mining*

Fonte: Shutterstock (2023)

Um dos principais desafios da TM é lidar com a natureza não estruturada do texto. Segundo Feldman e Sanger (2007), "o texto é um tipo de dados não estruturado e, portanto, não é fácil de ser processado por computadores". Isso significa que a preparação dos dados é um passo crucial na Mineração de Texto, que envolve a *tokenização* dos documentos em palavras, a remoção de palavras irrelevantes (*stop words*) e a normalização dos termos.

Outro desafio da MT é lidar com a complexidade dos dados. Segundo Berry (2015), "os dados textuais são altamente complexos e podem ser difíceis de serem analisados usando técnicas tradicionais de mineração de dados". Isso significa que é necessário utilizar técnicas específicas de MT, como a análise de sentimentos, a identificação de entidades nomeadas e a análise de tópicos, para lidar com essa complexidade.

Além disso, a diversidade linguística e a ambiguidade inerentes aos dados textuais podem representar desafios adicionais. Em muitos casos, os textos podem conter gírias, regionalismos, abreviações e variações gramaticais que dificultam a análise automatizada. A interpretação correta do contexto e das nuances linguísticas é essencial para extrair *insights* precisos e relevantes dos dados textuais

O Processamento de Linguagem Natural (do inglês *Natural Language Processing* - NLP) emerge como uma disciplina interdisciplinar que visa capacitar as máquinas a compreender, interpretar e responder à linguagem humana de maneira inteligente.

Fundamentada em avanços na inteligência artificial e na linguística computacional, a NLP tem transformado a maneira como interagimos com sistemas tecnológicos.

A análise semântica, um componente crucial da NLP, permite que sistemas compreendam o significado subjacente às palavras e frases. Abordagens como análise de sentimentos (PANG; LEE, 2008) capacitam as máquinas a discernir emoções expressas em textos, tornando possível uma resposta mais personalizada e adaptada às necessidades do usuário. Além disso, com os avanços em processamento de linguagem natural multimodal, os sistemas podem interpretar e gerar informações a partir de diferentes modalidades, como texto, imagem e fala. Essa capacidade de processar múltiplas formas de entrada amplia as possibilidades de interação homem-máquina (WANG *et al.*, 2020).

Para o método proposto, o suporte do NLP foi essencial. Na execução do trabalho, foram necessárias diversas tarefas de processamento de texto, incluindo a execução de parsers, a separação e determinação de *chunks* relevantes, como os parágrafos, a localização de termos da taxonomia dentro do texto e o processamento de contagem e concorrência. Essas tarefas, que envolvem processamento léxico e sintático, são parte integrante das funcionalidades do NLP. O uso dessas técnicas permitiu a normalização dos dados, extração de parágrafos e localização de padrões dentro do texto, contribuindo significativamente para a eficácia do trabalho.

2.2.3.2 ANÁLISE DE DADOS

A análise de dados é um processo que envolve a coleta, organização, interpretação e comunicação de informações úteis a partir de conjuntos de dados (CHIANG, *et al.*, 2019). Segundo a *Harvard Business Review*, a análise de dados é uma das habilidades mais importantes para profissionais de negócios nos dias de hoje (DAVENPORT; KIM, 2013). O objetivo da análise de dados é descobrir *insights* e padrões nos dados que possam ser usados para tomar decisões informadas.

Um dos primeiros passos da análise de dados é a coleta e organização dos dados. Segundo Kelleher e Tierney (2018), a coleta de dados é o processo de obter informações de diferentes fontes, como bancos de dados, arquivos de texto e mídias sociais. A organização dos dados envolve a estruturação dos dados de forma que possam ser facilmente acessados e analisados. Isso pode incluir a criação de tabelas para representar visualmente os dados.

Após a coleta e organização dos dados, é possível aplicar técnicas estatísticas e de ML para descobrir *insights* nos dados. A análise estatística é uma ferramenta poderosa que permite a interpretação de dados de maneira significativa. Segundo Hastie *et al.* (2009), técnicas de análise estatística, como regressão linear e análise de variância, são fundamentais para entender as relações entre variáveis. Além disso, Tibshirani *et al.* (2013) destacam a importância da regularização na análise estatística para evitar o sobreajuste dos modelos. A análise estatística, portanto, desempenha um papel crucial na tomada de decisões informadas em diversas áreas, desde a ciência até a economia. Isso pode incluir a identificação de tendências e padrões nos dados. Já a aprendizagem de máquina é o processo de desenvolver algoritmos que podem aprender a partir dos dados e realizar tarefas complexas, como reconhecimento de padrões e classificação de dados. A Figura 7, ilustra as etapas da análise de dados, sendo crucial seguir uma abordagem metodológica para alcançar resultados eficazes

Figura 7 - Etapas de Análise de dados

Fonte: Terra (2023)

A primeira etapa consiste na definição clara do problema ou da questão a ser investigada. Nesta fase, é fundamental compreender os objetivos da análise e identificar as variáveis relevantes para o estudo. Uma vez que o problema está definido, passa-se para a segunda etapa: a coleta de dados. Segundo Smith (2019), a coleta de dados é o processo de

obtenção de informações de diversas fontes, como bancos de dados, arquivos de texto e registros históricos.

Após a coleta, os dados precisam ser tratados para garantir sua qualidade e consistência. Isso inclui a limpeza dos dados, a remoção de valores ausentes ou inconsistentes e a padronização das variáveis. Como ressaltado por Johnson (2020), o tratamento de dados é uma etapa crítica para garantir a confiabilidade dos resultados da análise.

Com os dados limpos e preparados, é possível avançar para a etapa de análise. Neste estágio, uma variedade de técnicas estatísticas e algoritmos de aprendizado de máquina podem ser aplicados para explorar padrões, identificar correlações e realizar inferências sobre os dados. Segundo Li *et al.* (2021), a análise de dados é uma etapa fundamental para transformar dados brutos em informações úteis e acionáveis.

Por fim, a visualização dos dados desempenha um papel crucial na comunicação dos resultados da análise. Gráficos, tabelas e outros tipos de representações visuais são utilizados para apresentar os *insights* de forma clara e compreensível. Conforme destacado por Chen (2018), a visualização dos dados facilita a interpretação dos resultados e auxilia na tomada de decisões informadas.

Essa abordagem estruturada permite que as organizações extraiam valor máximo de seus dados, promovendo uma tomada de decisão informada e orientada por evidências. A análise de dados pode ser aplicada em diversas áreas, incluindo negócios, saúde, ciências sociais e governo. Segundo Chen e Liu (2014), a análise de dados é uma ferramenta essencial para a tomada de decisões em muitas indústrias. Por exemplo, a análise de dados pode ser usada para prever tendências de mercado, identificar oportunidades de crescimento e otimizar processos de produção.

Podemos concluir que, a análise de dados é uma ferramenta essencial para a tomada de decisões informadas em muitas áreas. A coleta e organização dos dados são os primeiros passos do processo de análise de dados, seguidos pela aplicação de técnicas estatísticas e de aprendizado de máquina para descobrir *insights* nos dados. A análise de dados é uma ferramenta poderosa que pode ser aplicada em muitas indústrias para prever tendências, identificar oportunidades e otimizar processos.

2.2.3.3 VISUALIZAÇÃO DE DADOS

A Visualização de Dados é uma prática essencial para apresentar informações de forma clara e intuitiva. De acordo com Cairo (2019), trata-se de uma linguagem visual que traduz grandes volumes de dados em gráficos e diagramas de fácil interpretação pelos usuários. Essa técnica desempenha um papel fundamental na identificação de padrões, tendências e anomalias nos dados, tornando-se indispensável em diversas áreas de atuação.

A seleção adequada da técnica de visualização é crucial para extrair *insights* significativos dos dados. Conforme destacado por Few (2013), a escolha do gráfico ou diagrama deve ser guiada pelo tipo de dados em questão e pelas perguntas que se deseja responder. Por exemplo, um gráfico de barras é útil para representar a distribuição de uma variável categórica, enquanto um gráfico de dispersão é mais apropriado para ilustrar a relação entre duas variáveis contínuas.

A cor desempenha um papel crucial na visualização de dados, como ressaltado por Wong (2019). A escolha da paleta de cores deve considerar a natureza dos dados e os objetivos da visualização. Por exemplo, uma paleta de cores divergentes é útil para destacar diferenças significativas entre duas extremidades, enquanto uma paleta sequencial é mais adequada para visualizar tendências ou variações em um conjunto de dados.

A interatividade é uma característica cada vez mais valorizada na visualização de dados, como observado por Kosara e Mackinlay (2016). A interatividade permite que os usuários explorem os dados em profundidade e descubram novos *insights*. Recursos como zoom, filtro, destaque e exploração de detalhes enriquecem a experiência do usuário e contribuem para uma análise mais completa e precisa dos dados.

A visualização de dados está se tornando cada vez mais relevante em diversas áreas, como negócios, ciência, jornalismo e governo. Conforme apontado por Segel e Heer (2010), essa prática tem o potencial de transformar a maneira como tomamos decisões e compreendemos o mundo ao nosso redor. Ao explorar grandes conjuntos de dados, comunicar informações de maneira clara e persuasiva e identificar padrões e tendências ocultas, a visualização de dados desempenha um papel crucial na geração de *insights* e na promoção da compreensão e do avanço do conhecimento.

2.2.3.4 Tecnologia de suporte a Análise de Dados

A tecnologia de suporte à análise de dados vem se tornando cada vez mais essencial no contexto empresarial atual. O uso de ferramentas como Inteligência Empresarial (do inglês

Business *Intelligence* – BI), *Analytics*, ML*,* DM e AI permitem que as empresas obtenham *insights* valiosos a partir de seus dados (WITTEN; FRANK, 2016). Essas ferramentas capacitam as organizações a explorar padrões, identificar tendências, prever comportamentos futuros e tomar decisões mais informadas (HAN *et al.*, 2011).

Em conjunto, essas ferramentas proporcionam às empresas uma vantagem competitiva significativa ao transformar dados em conhecimento acionável. O objetivo final do BI é possibilitar que as organizações tomem decisões melhores e mais informadas (KIMBALL; ROSS, 2013).

Dentre as principais ferramentas de BI, destacam-se os sistemas de gerenciamento de bancos de dados (do inglês *Database Management System* - DBMS), responsáveis por armazenar, organizar e gerenciar grandes volumes de dados. Esses sistemas são fundamentais para suportar a análise de dados em larga escala, fornecendo uma infraestrutura robusta para a gestão dos dados (ELMASRI; NAVATHE, 2015).

Além dos DBMS, outra tecnologia de suporte à análise de dados é a computação em nuvem, que oferece uma infraestrutura escalável e elástica para processar grandes volumes de dados. Isso permite que as empresas se concentrem na análise e tomada de decisão, em vez de lidar com a infraestrutura (BUYYA *et al.*, 2009).

No contexto da computação em nuvem, destaca-se o uso de plataformas de *Big Data*, como o *Apache* Hadoop e o *Spark*. Essas ferramentas são poderosas para a análise de grandes volumes de dados, fornecendo um modelo de programação escalável e tolerante a falhas para processar e armazenar dados em larga escala (ZAHARIA *et al.*, 2010).

É importante mencionar também o papel dos algoritmos de ML como tecnologia de suporte à análise de dados. Esses algoritmos permitem que as empresas extraiam padrões e *insights* dos seus dados, facilitando a tomada de decisão orientada por dados (KELLEHER; TIERNEY, 2018). Dessa forma, as tecnologias de suporte à análise de dados são essenciais para aprimorar a capacidade de tomada de decisão das empresas, permitindo que elas extraiam valor de seus dados

2.3 TRABALHOS RELACIONADOS

Os resultados de uma revisão integrativa da literatura que abrange uma ampla gama de trabalhos relacionados em campos interdisciplinares, enfocando principalmente a

Visualização de Informações, Análise Visual e Mineração de Texto. Esta revisão buscou compreender e analisar criticamente as contribuições recentes e significativas na pesquisa, explorando diferentes abordagens metodológicas e temas emergentes.

Os trabalhos selecionados representam uma diversidade de perspectivas e metodologias, oferecendo *insights* valiosos sobre o estado atual dessas áreas de estudo. O protocolo da pesquisa e os resultados são abordados nos Apêndices A e B, respectivamente. Incluem desde análises sistemáticas da literatura até estudos mais específicos sobre técnicas de análise visual na educação online e desafios associados à desinformação em visualizações. Cada trabalho contribui para a compreensão aprofundada dos avanços recentes e das lacunas existentes.

O Quadro 8, apresenta uma visão geral das principais características de cada trabalho relacionado, incluindo a descrição do estudo, as metodologias utilizadas, os resultados alcançados e os recursos utilizados.

Quadro 8 - Trabalhos relacionados

Título	Autores	Descrição	O que foi usado	Recursos Usados
Best-Practice Piloting Based on an Integrated Social Media Analysis and Visualization for E-Participation Simulation in Cities	Dirk Burkhardt, Kawa Nazemi, Egils Ginters	Este artigo apresenta um método piloto de melhores práticas usando análise de mídia social e visualização para simulação de e-participação em cidades.	Análise Integrada de Mídia Social, Visualização, Simulação de E-Participação	Ferramentas de Análise de Mídia Social, Ferramentas de Visualização
Applications of Natural Language Techniques to Enhance Curricular Coherence	Adrian S. Barb, Nil Kilicay-Ergin	O artigo discute a aplicação de técnicas de linguagem natural para melhorar a coerência curricular.	Técnicas de Linguagem Natural, Coerência Curricular	Ferramentas de Processamento de Linguagem Natural
On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges	Jänicke, S., Franzini, G., Cheema, M.F., Scheuermann, G.	Esta pesquisa explora os métodos e desafios da leitura próxima e distante nas Humanidades Digitais.	Leitura Próxima, Leitura Distante, Humanidades Digitais	Ferramentas de Análise Textual, Ferramentas de Humanidades Digitais
A taxonomy of Twitter data analytics techniques	Hamzah, Muzaffar, Vu, Tuong Thuy	O artigo apresenta uma taxonomia de técnicas para análise de dados do Twitter.	Técnicas de Análise de Dados do Twitter, Taxonomia	Ferramentas de Análise de Dados, Ferramentas de Taxonomia
Visualization of Swedish News Articles: A Design Study	Kucher, Kostiantyn, Engstrom, Nellie, Axelsson, Wilma, Savas, Berkant, Kerren, Andreas	Este estudo demonstra a visualização de artigos de notícias suecas.	Visualização, Artigos de Notícias	Ferramentas de Visualização
Mining implicit 3D modeling patterns from unstructured temporal BIM log text data	Yarmohammadi, Saman, Pourabolghasem, Reza, Castro-Lacouture, Daniel	O artigo discute a mineração de padrões implícitos de modelagem 3D a partir de dados de log de BIM temporais não estruturados.	Mineração de Dados, Dados Textuais, Padrões de Modelagem 3D	Ferramentas de Mineração de Dados, Ferramentas de Mineração de Texto, Software BIM

Quadro 8 - Trabalhos relacionados

Título	Autores	Descrição	O que foi usado	Recursos Usados
Supporting Story Synthesis: Bridging the Gap between Visual Analytics and Storytelling	Chen, Siming, Li, Jie, Andrienko, Gennady, Andrienko, Natalia, Wang, Yun, Nguyen, Phong H., Turkay, Cagatay	Este artigo explora a ponte entre análises visuais e narração de histórias para suportar a síntese de histórias.	Análise Visual, Narração de Histórias, Síntese de Histórias	Ferramentas de Análise Visual, Ferramentas de Narração de Histórias
Getting over High-Dimensionality: How Multidimensional Projection Methods Can Assist Data Science	Ortigossa, Evandro S., Dias, Fábio Felix, Do Nascimento, Diego Carvalho	O artigo discute como os métodos de projeção multidimensional podem ajudar a lidar com a alta dimensionalidade na ciência de dados.	Métodos de Projeção Multidimensional, Ciência de Dados	Métodos de Redução de Dimensionalidade, Ferramentas de Ciência de Dados
Bridging Text Visualization and Mining: A Task-Driven Survey	Liu, Shixia, Wang, Xiting, Collins, Christopher, Dou, Wenwen, Ouyang, Fangxin, El-Assady, Mennatallah, Jiang, Liu, Keim, Daniel A.	Esta pesquisa explora a abordagem orientada a tarefas para unir visualização e mineração de texto.	Visualização de Texto, Mineração de Texto, Abordagem Orientada a Tarefas	Ferramentas de Visualização de Texto, Ferramentas de Mineração de Texto, Métodos Orientados a Tarefas
Rule-based Visual Mappings - With a Case Study on Poetry Visualization	Abdul-Rahman, A., Lein, J., Coles, K., Maguire, E., Meyer, M., Wynne, M., Johnson, C.R., Trefethen, A., Chen, M.	O artigo apresenta mapeamentos visuais baseados em regras com um estudo de caso sobre visualização de poesia.	Mapeamentos Visuais Baseados em Regras, Visualização de Poesia	Ferramentas de Visualização, Sistemas Baseados em Regras
Big complex biomedical data: Towards a taxonomy of data	Holzinger, Andreas, Stocker, Christof, Dehmer, Matthias	Este artigo visa estabelecer uma taxonomia para grandes dados biomédicos complexos.	Dados Biomédicos, Taxonomia	Ferramentas de Taxonomia de Dados, Ferramentas de Dados Biomédicos
Visual and Interactive Exploration of a Large Collection of Open Datasets	Liu, T., Ahmed, D. Bangash, Bouali, F., Venturini, G.	O artigo apresenta a exploração visual e interativa de uma grande coleção de conjuntos de dados abertos.	Visualização, Exploração Interativa, Conjuntos de Dados Abertos	Ferramentas de Visualização, Plataformas de Dados Abertos

Quadro 8 - Trabalhos relacionados

Título	Autores	Descrição	O que foi usado	Recursos Usados
A comprehensive review of visualization methods for association rule mining: Taxonomy, challenges, open problems	Fister, Iztok, Fister, Iztok, Fister, Dušan, Podgorelec, Vili, Salcedo-Sanz, Sancho	O artigo fornece uma revisão abrangente dos métodos de visualização para mineração de regras de associação.	Métodos de Visualização, Mineração de Regras de Associação, Revisão	Técnicas de Visualização, Ferramentas de Mineração de Regras de Associação
Visualizing patterns of appraisal in texts and corpora	Almutairi, Bandar Alhumaidi A.	O artigo discute a visualização de padrões de avaliação em textos e corpora.	Visualização de Texto, Padrões de Avaliação	Ferramentas de Visualização de Texto
Methods and visualization tools for the analysis of medical, political and scientific concepts in Genealogies of Knowledge	Luz, Saturnino, Sheehan, Shane	O artigo discute métodos e ferramentas de visualização para análise de conceitos médicos, políticos e científicos nas Genealogias do Conhecimento.	Ferramentas de Visualização, Análise de Conceitos, Genealogias do Conhecimento	Ferramentas de Visualização, Métodos de Análise de Conceitos
VisForum: A visual analysis system for exploring user groups in online forums	Fu, Siwei, Wang, Yong, Yang, Yi, Bi, Qingqing, Guo, Fangzhou, Qu, Huamin	O artigo apresenta o VisForum, um sistema de análise visual para explorar grupos de usuários em fóruns *online*.	Sistema de Análise Visual, Fóruns Online, Grupos de Usuários	Sistemas de Visualização, Plataformas de Fóruns *Online*
Visual and interactive analysis of a large collection of open data with the relative neighborhood graph	Liu, T., Bouali, F., Venturini, G.	O artigo discute a análise visual e interativa de uma grande coleção de dados abertos usando o gráfico de vizinhança relativa.	Visualização, Análise Interativa, Dados Abertos	Ferramentas de Visualização, Métodos de Análise Interativa
Cohort comparison of event sequences with balanced integration of visual analytics and statistics	Malik, Sana, Du, Fan, Monroe, Megan, Onukwugha, Eberechukwu, Plaisant, Catherine, Shneiderman, Ben	Este estudo compara coortes de sequências de eventos com uma integração equilibrada de análises visuais e estatísticas.	Sequências de Eventos, Análise Visual, Estatísticas	Ferramentas de Análise Estatística, Ferramentas de Análise Visual

Quadro 8 - Trabalhos relacionados

Título	Autores	Descrição	O que foi usado	Recursos Usados
Twitter sentiment analysis approaches: A survey	Adwan, Omar Y., Al-Tawil, Marwan, Huneiti, Ammar M., Shahin, Rawan A., Abu Zayed, Abeer A., Al-Dibsi, Razan H.	O artigo revisa várias abordagens de análise de sentimento no Twitter.	Análise de Sentimento, Twitter, Revisão	Ferramentas de Análise de Sentimento, API do Twitter
Beyond Explanation: A Case for Exploratory Text Visualizations of Non-Aggregated, Annotated Datasets	Havens, Lucy, Bach, Benjamin, Terras, Melissa, Alex, Beatrice	O artigo defende visualizações textuais exploratórias de conjuntos de dados não agregados e anotados.	Visualização Textual Exploratória, Conjuntos de Dados Não Agregados	Ferramentas de Visualização Textual
Visual data mining in software repositories: A survey	Eteläaho, Anna, Soini, Jari, Jaakkola, Hannu, Mattila, Anna-Liisa	Esta pesquisa explora a mineração visual de dados em repositórios de software.	Mineração de Dados, Repositórios de Software, Análise Visual	Ferramentas de Mineração de Dados, Técnicas de Visualização
A Survey on Event-Based News Narrative Extraction	Keith Norambuena, Brian Felipe, Mitra, Tanushree, North, Chris	O artigo apresenta uma pesquisa sobre extração de narrativa de notícias baseada em eventos.	Extração de Narrativa de Notícias, Análise Baseada em Eventos	Ferramentas de Mineração de Texto, Métodos de Análise de Eventos
Semantics Visualization as a User Interface in Business Information Searching	Dudycz, Helena	O artigo discute a visualização de semântica como uma interface de usuário na busca de informações empresariais.	Visualização de Semântica, Interface de Usuário, Informação Empresarial	Ferramentas de Visualização, Design de Interface de Usuário
The four dimensions of social network analysis: An overview of research methods, applications, and software tools	Camacho, David, Panizo-LLedot, Ángel, Bello-Orgaz, Gema, Gonzalez-Pardo, Antonio, Cambria, Erik	O artigo fornece uma visão geral dos métodos de pesquisa, aplicações e ferramentas de software na análise de redes sociais.	Análise de Redes Sociais, Métodos de Pesquisa, Ferramentas de Software	Ferramentas de Análise de Redes Sociais, Métodos de Pesquisa
EXOD: A tool for building and exploring a large graph of open datasets	Liu, Tianyang, Bouali, Fatma, Venturini, Gilles	O artigo apresenta o EXOD, uma ferramenta para construir e explorar um grande gráfico de conjuntos de dados abertos.	Construção de Gráficos, Exploração de Conjuntos de Dados, Visualização	Ferramentas de Visualização de Dados, Ferramentas de Análise de Gráficos

Quadro 8 - Trabalhos relacionados

Título	Autores	Descrição	O que foi usado	Recursos Usados
An Empirical Study on How Well Do COVID-19 Information Dashboards Service Users' Information Needs	Li, Xinyan, Wang, Han, Chen, Chunyang, Grundy, John	Este estudo empírico avalia a eficácia dos painéis de informação sobre COVID-19 em atender às necessidades de informação dos usuários.	Painéis de COVID-19, Necessidades de Informação do Usuário	Painéis de Dados, Avaliação da Experiência do Usuário
Overview Visualizations for Large Digitized Correspondence Collections: A Design Study	Swietlicki, Laura, Cubaud, Pierre	Este estudo apresenta visualizações panorâmicas para grandes coleções de correspondências digitalizadas.	Visualização, Correspondência Digitalizada, Estudo de Design	Técnicas de Visualização, Métodos de Design
Usability of business information semantic network search visualization	Dudycz, Helena	O artigo discute a usabilidade da visualização de busca em rede semântica de informações empresariais.	Busca em Rede Semântica, Avaliação de Usabilidade	Teste de Usabilidade, Ferramentas de Rede Semântica
A Tamil lyrics search and visualization system	Ranganathan, Karthika, Barani, B., Geetha, T.V.	O artigo apresenta um sistema de busca e visualização de letras em tâmil.	Busca de Letras, Sistema de Visualização, Tâmil	Ferramentas de Recuperação de Informação, Técnicas de Visualização
Data abstraction for visualizing large time series	Shurkhovetskyy, G., Andrienko, N., Andrienko, G., Fuchs, G.	O artigo discute técnicas de abstração de dados para visualização de grandes séries temporais.	Abstração de Dados, Visualização de Séries Temporais	Técnicas de Visualização, Métodos de Abstração de Dados
Of Course it's Political! A Critical Inquiry into Underemphasized Dimensions in Civic Text Visualization	Baumer, Eric P. S., Jasim, Mahmood, Sarvghad, Ali, Mahyar, Narges	O artigo examina criticamente dimensões subestimadas na visualização de textos cívicos.	Visualização de Textos Cívicos, Análise Política	Ferramentas de Análise Textual, Ferramentas de Engajamento Cívico
Communicating Uncertainty in Digital Humanities Visualization Research	Panagiotidou, Georgia, Lamqaddam, Houda, Poblome, Jeroen, Brosens, Koen	O artigo discute a comunicação da incerteza na pesquisa de visualização em humanidades digitais.	Comunicação da Incerteza, Humanidades Digitais, Pesquisa de Visualização	Técnicas de Comunicação, Métodos de Pesquisa em Visualização

Quadro 8 - Trabalhos relacionados

Título	Autores	Descrição	O que foi usado	Recursos Usados
UIWGViz: An architecture of user interest-based web graph visualization	Saleheen, Shibli; Lai, Wei	O artigo apresenta uma arquitetura para produzir e visualizar o gráfico da web baseado no interesse do usuário (UIWG).	Modelagem de Interesse do Usuário, Análise de Dados da Web, Geração de Gráficos da Web	Ferramentas de Visualização, Métodos de Coleta de Dados da Web, Análise de *Feedback* do Usuário
Ontology-Based Interactive Visualization of Patient-Generated Research Questions	Borland, David; Christopherson, Laura; Schmitt, Charles	Este trabalho apresenta uma ferramenta de visualização interativa que ajuda os pesquisadores a navegar efetivamente pelo conteúdo dos fóruns.	Desenvolvimento de Ontologia, Visualização Interativa, Análise de Conteúdo de Fórum	Ferramentas de Visualização, Ferramentas de Ontologia, Plataformas de Fóruns
Analysis and data visualization in bibliometric studies	Alhuay-Quispe, Joel; Estrada-Cuzcano, Alonso; Bautista-Ynofuente, Lourdes	O artigo explora métodos e ferramentas usados por pesquisadores de bibliometria através de uma análise descritiva e textual.	Análise de Co-ocorrência de Palavras, Extração de Dados Bibliométricos, Análise Descritiva	Ferramentas Bibliométricas, Métodos de Análise Textual
Competitive analysis of online reviews using exploratory text mining	Amadio, William J.; Procaccino, J. Drew	O artigo explora a utilidade de ferramentas de mineração de texto e análises visuais para análise competitiva usando avaliações online.	Mineração de Texto, Análise SWOT, Análise Visual	Ferramentas de Mineração de Texto, Software de Análise Visual
A survey of visual analytics techniques for machine learning	Jun Yuan, Changjian Chen, Weikai Yang, Mengchen Liu, Jiazhi Xia & Shixia Liu	Esta pesquisa revisa técnicas de análise visual para aprendizado de máquina, destacando desafios de pesquisa e oportunidades futuras.	Análise Visual, Aprendizado de Máquina, Metodologia de Pesquisa	Ferramentas de Análise Visual, *Frameworks* de Aprendizado de Máquina
New Cybercrime Taxonomy of Visualization of Data Mining Process	Babic, M.; Jerman-Blazic, B.	O artigo propõe uma taxonomia de técnicas de visualização para processos de mineração de dados de crimes cibernéticos.	Mineração de Dados, Visualização, Taxonomia de Crimes Cibernéticos	Técnicas de Visualização, Métodos de Mineração de Dados

Quadro 8 - Trabalhos relacionados

Título	Autores	Descrição	O que foi usado	Recursos Usados
BLASTGrabber: a bioinformatic tool for visualization, analysis and sequence selection of massive BLAST data	Neumann, RS; Kumar, S; Haverkamp, THA; Shalchian-Tabrizi, K	O BLASTGrabber é uma ferramenta bioinformática para visualizar e analisar dados de saída do BLAST, integrando identificação de taxonomia e mineração de texto.	Análise BLAST, Mineração de Texto, Visualização	Ferramentas de Bioinformática, Software BLAST, Técnicas de Mineração de Texto
Intelligent information extraction from government on-site inspection reports of construction projects: A graph-based text mining approach	Liu, MY; Luo, XW; Wang, GB; Lu, WZ	O artigo apresenta uma estrutura de mineração de texto para extração de informações de relatórios de inspeção governamental de projetos de construção.	Mineração de Texto, Análise de Gráficos, Extração de Informações	Ferramentas de Mineração de Texto, Software de Análise de Gráficos

Fonte: Elaborado pela autora.

A revisão integrativa é abordada detalhadamente no Capítulo 3 na Seção 3.2, onde são mostrados as informações iniciais da pesquisa e os resultados obtidos; além disso, o Apêndice A inclui os detalhes do protocolo da revisão integrativa da literatura.

3 METODOLOGIA DE PESQUISA

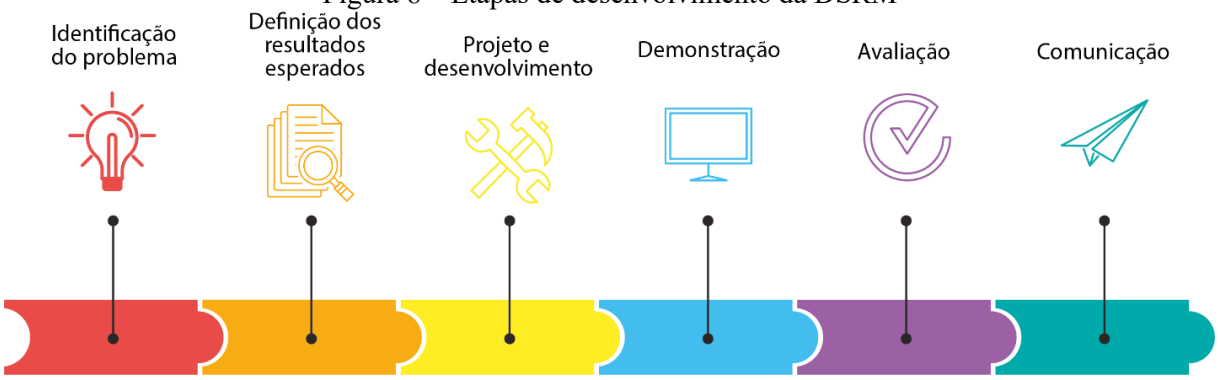
Este projeto de dissertação terá uma abordagem metodológica que adota a visão de mundo com a perspectiva Pragmática (projetada). Tem como foco a solução do problema que foi identificado ou levantado, assim essa abordagem é direcionada ao problema (SILVA; et. al., 2017). Em relação à modalidade, considera-se uma Pesquisa Tecnológica, na qual a pesquisa tecnológica é direcionada à produção de algo novo (Cupani, 2006). Segundo Cupani (2006), a pesquisa tecnológica concentra-se na produção de algo novo, sendo o campo do conhecimento que se ocupa de projetar artefatos, planejar sua construção, operação, configuração, manutenção e acompanhamento, com base no conhecimento científico.

A abordagem metodológica geral desta dissertação adota o paradigma científico da Pesquisa em Design (do inglês *Design Science* - DS), a qual busca identificar um problema e propor uma solução por meio da criação de novos artefatos. A DS representa uma abordagem inovadora na busca por soluções eficazes para problemas complexos no campo do design. Nesse contexto, uma revisão integrativa desempenha um papel crucial, pois oferece uma análise abrangente e holística das contribuições existentes no domínio. Esta revisão, a ser apresentada nesta seção, busca consolidar o conhecimento atual, identificar lacunas no entendimento e destacar tendências emergentes no âmbito da DS. Ao integrar *insights* de diversas fontes e experiências, a revisão integrativa visa proporcionar uma visão aprofundada do estado atual da pesquisa em DS, oferecendo assim um guia valioso para os envolvidos nesse domínio dinâmico.

> A *Design Science* é a ciência que procura desenvolver e projetar soluções para melhorar sistemas existentmees, resolver problemas ou, ainda, criar novos artefatos que contribuam para uma melhor atuação humana, seja na sociedade, seja nas organizações. Logo, a natureza deste tipo de pesquisa costuma ser pragmática e orientada à solução (DRESCH, LACERDA, JÚNIOR, 2015, p. 57).

O desenvolvimento desta dissertação foi realizado em etapas, seguindo a proposta de Metodologia de Pesquisa em Ciência do Design (do inglês *Design Science Research Methodology* - DSRM) de Peffers *et al.* (2007). Esta metodologia oferece um quadro estruturado para a condução de pesquisas na área de design, permitindo a criação sistemática e rigorosa de soluções inovadoras para problemas complexos. Ao seguir as diretrizes estabelecidas pela DSRM, esta dissertação buscará identificar claramente os problemas de pesquisa, propor soluções baseadas na construção de artefatos e avaliar criticamente a eficácia dessas soluções por meio de rigorosos métodos de avaliação. A DSRM é composta por seis etapas distintas, como ilustrado pela Figura 8.

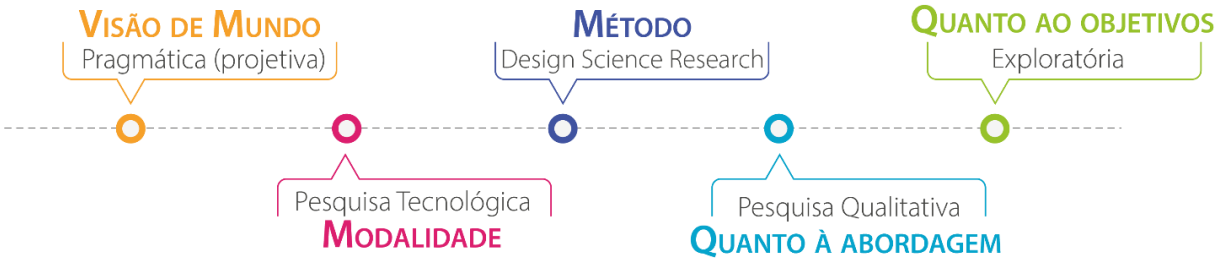
Figura 8 – Etapas de desenvolvimento da DSRM



Fonte: Elaborado pela autora com base no trabalho de Peffers, *et. al.* (2007)

Quanto à abordagem este projeto é definido como uma pesquisa qualitativa, pois permite que os dados venham a ser coletados no ambiente que ocorre a questão de pesquisa, tendo como foco a perspectiva de quem está sendo estudado (MERRIAM, 2009; CRESWELL, 2010). Já quanto aos objetivos, é classificada como exploratória. Araújo e Oliveira (1997) afirmam que estudos exploratórios visam procurar desenvolver, esclarecer e modificar ideias e conceitos. A Figura 9, resume a abordagem metodológica adotada.

Figura 9 - Abordagem Metodológica



Fonte: Elaborado pela autora.

3.1 DESIGN SCIENCE RESEARCH METHODOLOGY

A pesquisa em *design science research methodology* tem se consolidado como uma metodologia cada vez mais utilizada na área de sistemas de informação. Segundo Gregor e Jones (2007), o objetivo da pesquisa em *design science* é desenvolver e avaliar artefatos, tais como modelos, métodos, processos e sistemas, que possam ser utilizados para solucionar

problemas específicos e contribuir para a melhoria do conhecimento e prática em uma determinada área.

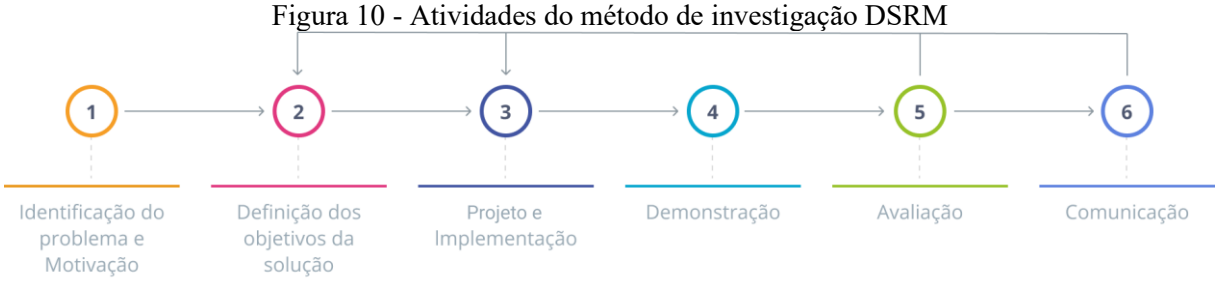
March e Smith (1995) afirmam que a pesquisa em *design science* tem uma abordagem prática e aplicada, em que a produção de conhecimento é derivada da construção de artefatos que são avaliados em termos de sua efetividade na solução de problemas. Já Peffers *et al.* (2007) enfatizam que a pesquisa em *design science* tem como base uma abordagem pragmática, em que o objetivo é a criação de artefatos que possam ser utilizados para resolver problemas práticos.

Os artefatos desenvolvidos na DSRM passam por um processo de avaliação em que são verificados sua efetividade e utilidade em solucionar problemas reais. Segundo Hevner *et al*. (2004), esse processo de avaliação é realizado por meio de critérios específicos que levam em consideração tanto os objetivos da pesquisa quanto as características do artefato desenvolvido. Alguns critérios comuns de avaliação incluem a relevância, a validade, a efetividade, a eficiência e a satisfação do usuário.

A DSRM é uma metodologia que busca produzir conhecimento aplicado por meio da criação e avaliação de artefatos que possam ser utilizados para solucionar problemas práticos. A abordagem pragmática e aplicada da DSRM, aliada ao processo de avaliação dos artefatos desenvolvidos, torna essa metodologia uma alternativa efetiva para a solução de problemas em sistemas de informação.

Segundo Peffers *et al*. (2007), a DSRM é composta por seis fases: identificação do problema e motivação, definição dos objetivos para a solução do problema, desenho e desenvolvimento da solução, demonstração, avaliação e comunicação, a Figura 10 demonstra o fluxo de interação entre as fases.

Figura 10 - Atividades do método de investigação DSRM



Fonte: Adaptado pela autora a partir de Peffers *et al*. 2007.

A fase inicial da Metodologia de Pesquisa em Ciência do Design (DSRM) consiste na identificação do problema, um passo crucial para o sucesso da pesquisa. De acordo com Gregor

et al. (2013), "a escolha de um problema de pesquisa significativo é essencial para a validade e relevância da investigação". Nesta fase, o pesquisador deve identificar um problema real que possa ser abordado e solucionado por meio da aplicação da DSRM.

A segunda fase da DSRM consiste na definição dos objetivos para a solução do problema. De acordo com Peffers *et al*. (2007), nessa fase, o pesquisador deve definir os objetivos da solução, que devem ser consistentes com as necessidades e expectativas dos usuários, bem como com as restrições e limitações do ambiente em que a solução será implementada.

A terceira fase da Metodologia de Pesquisa em Ciência do Design (DSRM) é a fase de design e desenvolvimento da solução. Nesta etapa, o pesquisador é responsável por criar um artefato que possa solucionar o problema identificado. Segundo Gregor *et al.* (2013), o artefato deve ser desenvolvido de forma rigorosa, utilizando métodos e técnicas apropriados, e deve ser validado empiricamente para garantir sua eficácia e utilidade.

A quarta fase da DSR é a demonstração da solução. De acordo com Peffers *et al*. (2007), nessa fase, o pesquisador deve demonstrar que o artefato criado é capaz de solucionar o problema identificado. A demonstração pode ser realizada por meio de protótipos, experimentos ou estudos de caso.

A quinta fase da DSRM é a fase de avaliação. Segundo March, Smith e Hinrichs (2016), nesta etapa, o pesquisador deve avaliar a efetividade e eficiência do artefato em relação aos objetivos definidos na segunda fase. A avaliação deve ser rigorosa e envolver a coleta de dados e análise estatística para garantir resultados confiáveis e significativos.

Por fim, última fase da DSRM é a comunicação. Nessa fase, o pesquisador deve comunicar os resultados da pesquisa de forma clara e precisa. De acordo com Peffers *et al*. (2007), a comunicação pode ser realizada por meio de publicações em conferências e revistas científicas, apresentações em conferências e workshops, e relatórios técnicos.

3.2 REVISÃO INTEGRATIVA DE LITERATURA

A revisão integrativa de literatura é uma técnica que tem como objetivo integrar evidências de diferentes estudos a fim de gerar uma síntese abrangente sobre um tema específico (TORRES; RIBEIRO; OLIVEIRA, 2020). De acordo com a definição de Whittemore e Knafl (2005), a revisão integrativa é um método que permite a inclusão de

diversos tipos de estudos, como revisões sistemáticas, estudos experimentais e qualitativos, possibilitando a análise de diferentes perspectivas sobre um mesmo tema.

Uma revisão integrativa de literatura é composta por várias etapas, incluindo a seleção da pergunta de pesquisa, definição dos critérios de inclusão e exclusão dos estudos, busca nas bases de dados, seleção e avaliação dos estudos, extração dos dados, análise e síntese dos resultados (PRADO; SOUZA; CAMPONOGARA, 2018). É importante ressaltar que a revisão integrativa não se limita apenas a uma síntese descritiva dos estudos, mas deve buscar identificar tendências e padrões, além de apontar lacunas na literatura (TERRA; GLINA; FERRAZ, 2012).

A revisão integrativa é uma técnica de pesquisa que busca sintetizar o conhecimento científico sobre um determinado tema, a partir da análise crítica de diferentes estudos. Ela é amplamente utilizada em diversas áreas do conhecimento, incluindo a saúde, a engenharia e a administração (TORRACO, 2010; WHITTEMORE e KNAFL, 2005). Para Rodrigues *et al*. (2021), a revisão integrativa é uma abordagem importante para a produção de conhecimento científico e para orientar tomadas de decisão em diferentes contextos. Segundo *et al*. (2019), a revisão integrativa pode ser uma alternativa viável para a síntese de evidências quando não há disponibilidade de revisões sistemáticas ou quando o objetivo da pesquisa é avaliar diferentes perspectivas sobre um tema.

Com foco na execução do DSRM, na revisão integrativa da literatura e na proposição e desenvolvimento do método está dissertação, busca-se uma abordagem robusta e sistemática. O DSRM guiou o desenvolvimento do método proposto, enquanto a revisão integrativa fundamentou o trabalho nas contribuições existentes. Neste sentido, a execução do método proposto seguirá uma metodologia clara, garantindo a coerência com o DSRM. Essas etapas visam integrar teoria e prática para contribuir com esta pesquisa.

3.3 EXECUÇÃO DA *DESIGN SCIENCE RESEARCH METHODOLOGY*

A DSRM (PEFFERS *et al*., 2007) estabelece um arcabouço estruturado composto por seis fases distintas, delineando um caminho sistemático para a concepção, desenvolvimento e avaliação de soluções inovadoras para problemas complexos. Sendo assim, proporciona um guia abrangente que vai além da tradicional pesquisa científica, buscando ativamente criar e avaliar artefatos que abordam desafios práticos em contextos específicos.

As servem como um roteiro estratégico, desde a identificação do problema até a comunicação dos resultados. Cada etapa desempenha um papel crucial no processo, desde a compreensão aprofundada do contexto até a entrega de soluções tangíveis e aplicáveis. Oferece assim, uma estrutura robusta para pesquisadores e profissionais envolvidos na criação e avaliação de soluções inovadoras, assegurando uma abordagem sistemática e fundamentada em evidências ao longo de todo o ciclo de pesquisa.

A colaboração entre academia e indústria representada por esse projeto é essencial para preencher a lacuna entre a teoria e a prática, resultando em melhorias substanciais na segurança operacional e no bem-estar dos trabalhadores na indústria de óleo e gás.

Esta dissertação se caracteriza por abordar o tipo de lacuna onde a realidade prática é adaptada para o contexto acadêmico. Ou seja, utiliza as necessidades e desafios reais enfrentados pelas empresas da indústria de óleo e gás como base para desenvolver soluções acadêmicas viáveis e aplicáveis. Ao fazer isso, não só promove uma compreensão mais profunda dos fatores humanos e práticas de resiliência no setor, mas também contribui para a criação de metodologias e práticas que podem ser implementadas efetivamente, fortalecendo a cultura de segurança organizacional.

3.3.1 Etapa 1 – Identificação do Problema e Motivação

Nesta etapa inicial da metodologia DSRM, os objetivos da solução são definidos. O principal objetivo é desenvolver um método que será instanciado. Este método visa oferecer funcionalidades avançadas para a análise de documentos de texto, utilizando taxonomia para a organização dos dados, visualizações gráficas para facilitar a interpretação dos resultados e a capacidade de salvar consultas para futura referência e reutilização.

Para fundamentar o desenvolvimento do método, um dos objetivos é realizar uma revisão de sistemas similares, conforme apresentado na Seção 5.4. Esta revisão busca identificar as funcionalidades dos sistemas existentes, comparar suas características e destacar as inovações e vantagens do sistema proposto. A revisão de sistemas similares permitirá contextualizar o desenvolvimento do DOC *Analysis* e garantir que ele atenda às necessidades e expectativas dos usuários de maneira superior.

A análise de documentos textuais representa um desafio significativo em muitos contextos, especialmente em setores altamente complexos, como a indústria petrolífera

offshore. Nesse cenário, a segurança operacional é uma preocupação primordial devido aos riscos envolvidos nas operações em plataformas marítimas. A crescente quantidade de dados textuais disponíveis, como relatórios de incidentes, procedimentos de segurança e registros de manutenção, demanda abordagens inovadoras que superem as limitações das técnicas tradicionais de análise. Especificamente, a análise de documentos é crucial para identificar padrões e tendências que possam contribuir para a prevenção de acidentes e melhorias na segurança das plataformas.

A motivação para abordar esse problema reside na importância cada vez maior da compreensão de grandes volumes de texto em áreas críticas, como a segurança operacional em plataformas *offshore*. A capacidade de extrair significado dos documentos textuais não apenas impulsiona a eficiência operacional, mas também oferece uma base sólida para inovações em setores que dependem da interpretação inteligente de informações textualmente complexas. Na indústria petrolífera *offshore*, a falta de soluções integradas, que combinem efetivamente técnicas de Processamento de Linguagem Natural (PLN), visualização de dados e taxonomia, cria obstáculos para uma análise profunda e contextualizada.

Assim, esta pesquisa busca preencher essa lacuna, propondo uma solução que integre harmoniosamente técnicas avançadas de PLN, visualização de dados e taxonomia, proporcionando uma visão holística para a análise de documentos textuais específicos desse domínio. Ao endereçar a complexidade inerente a esses dados, a pesquisa não apenas visa superar desafios práticos na interpretação de documentos, mas também contribuir para a evolução do conhecimento na interseção entre PLN, visualização de dados e taxonomia.

3.3.2 Etapa 2 – Definição dos Objetivos para uma Solução

Esta pesquisa tem como foco o desenvolvimento de um método integrado que combina técnicas de TM e visualização de dados, apoiado por uma taxonomia, com o objetivo de aprimorar a análise de documentos textuais em diversos domínios. Para alcançar este objetivo, a pesquisa segue três objetivos principais.

O primeiro objetivo é obter uma compreensão abrangente dos conceitos fundamentais relacionados a corpus de texto, taxonomia, mineração de textos, análise e visualização de dados. Essa base teórica é essencial para o desenvolvimento de uma abordagem prática e eficaz na análise de documentos textuais.

O segundo objetivo é identificar e aplicar as tecnologias relevantes para a extração de textos, identificação de termos-chave e visualização de dados. Ao explorar ferramentas e métodos emergentes, a pesquisa visa garantir que a solução proposta esteja alinhada com as últimas inovações tecnológicas, proporcionando resultados precisos e atualizados.

O terceiro objetivo é aplicar uma taxonomia ao contexto do método proposto. Essa taxonomia funcionará como uma estrutura organizacional, facilitando a categorização eficiente e a interpretação contextualizada dos dados textuais. Ao integrar esses objetivos, a pesquisa busca não apenas criar uma solução técnica, mas também estabelecer um arcabouço conceitual que contribua significativamente para a análise de documentos textuais em diversos domínios.

3.3.3 Etapa 3 – Projeto e Desenvolvimento

A taxonomia dos fatores humanos pode ser compreendida em três níveis principais: Indivíduo, Trabalho e Organização, cada um com seus conceitos específicos.

No nível do indivíduo, consideramos as competências técnicas, que são as habilidades e conhecimentos necessários para a realização de tarefas específicas. A comunicação é a capacidade de transmitir e receber informações de forma eficaz. O trabalho em equipe refere-se à habilidade de colaborar com outros para alcançar objetivos comuns, enquanto a capacidade relacional é a aptidão para interagir de maneira positiva e produtiva com outros. A tomada de decisão envolve a capacidade de avaliar situações e escolher ações apropriadas, e a consciência situacional é a percepção e compreensão do ambiente e dos eventos ao redor. A atitude diz respeito às inclinações mentais e emocionais que afetam o comportamento, e as condições psicológicas se referem ao estado mental e emocional do indivíduo. As condições fisiológicas englobam o estado físico e a saúde, e as condições sociais são as influências sociais que afetam o comportamento do indivíduo.

No nível do trabalho, o design do trabalho é a estruturação e organização das tarefas e responsabilidades. O design de interfaces refere-se ao projeto de sistemas e ferramentas de interação do usuário. As condições internas incluem o ambiente de trabalho interno, como temperatura, iluminação e ergonomia, enquanto as condições externas abrangem fatores como clima e localização. A liderança diz respeito ao estilo e à eficácia da liderança no ambiente de trabalho.

No nível da organização, a cultura de segurança abrange os valores e práticas organizacionais que promovem a segurança. A aprendizagem refere-se à capacidade da organização de aprender e melhorar continuamente. A gestão de pessoas inclui as práticas e políticas para recrutar, desenvolver e reter funcionários, e a gestão envolve os processos e estruturas organizacionais que definem como a organização é gerida.

As relações entre esses níveis são fundamentais. A interação entre indivíduos é crucial para a comunicação, o trabalho em equipe e a capacidade relacional. As competências técnicas, atitudes e condições (psicológicas, fisiológicas e sociais) de um indivíduo influenciam diretamente como ele desempenha seu trabalho e se adapta à organização. O design do trabalho e as condições de trabalho estão intimamente ligados à cultura de segurança, à aprendizagem organizacional e à gestão de pessoas e processos.

Esses conceitos formam uma estrutura interligada onde cada nível e fator influencia os outros, criando um sistema dinâmico que afeta o desempenho e o bem-estar dos indivíduos, bem como a eficácia das organizações.

O desenvolvimento do método de análise de texto baseado em uma taxonomia envolve várias etapas essenciais para extrair *insights* significativos de grandes volumes de informações textuais. Primeiro, obtém-se um documento de texto relacionado ao domínio da taxonomia. Em seguida, esse texto é processado através de técnicas de processamento de linguagem natural (NLP), analisando-o léxica e sintaticamente. A carga da taxonomia fornece uma estrutura de referência organizada para categorizar os termos e conceitos no texto.

Os dados processados são então armazenados de forma estruturada, incluindo informações sobre a frequência e coocorrência de termos e fatores. Tabelas específicas são usadas para registrar dimensões, fatores, termos e parágrafos, bem como as relações entre eles. Finalmente, a visualização dos dados permite uma compreensão mais profunda e organizada do conteúdo textual, utilizando gráficos e mapas de conceitos, adaptados ao objetivo da análise.

Cada etapa desempenha um papel crucial no desenvolvimento de *insights* valiosos a partir da análise textual, garantindo uma abordagem sistemática e eficaz.

A criação de um protótipo composto por um sistema em Java complementado por uma aplicação *web* dedicada à exploração visual dos gráficos gerados a partir dos arquivos JSON e CSV, resultantes da análise textual baseada na taxonomia de domínio aplicada.

O sistema será responsável pela manipulação dos dados extraídos e pela geração dos arquivos intermediários. A utilização de bibliotecas e *frameworks* na linguagem Java

especializados em processamento de texto e TM proporcionará a base técnica necessária para lidar com a complexidade dos documentos textuais no contexto da taxonomia.

A aplicação *web*, por sua vez, oferecerá uma interface intuitiva para os usuários explorarem visualmente os gráficos gerados. Utilizando tecnologias *web* recentes, como HTML5, CSS3 e JavaScript, a aplicação objetiva proporcionar uma experiência interativa e amigável. Os dados provenientes dos arquivos JSON e CSV serão dinamicamente incorporados aos gráficos, permitindo uma análise mais profunda e contextualizada.

Ao integrar eficientemente o sistema Java e a aplicação *web*, essa etapa visa não apenas criar uma solução técnica funcional, mas também assegurar que os usuários finais possam explorar de forma eficaz e visualmente atraente os resultados da análise textual, promovendo uma compreensão mais ampla dos documentos textuais no âmbito do domínio em que está se aplica.

3.3.4 Etapa 4 - Demonstração

Nesta fase a pesquisa se materializa em uma apresentação concreta da solução proposta. Envolve a exposição prática e funcional da solução proposta, demonstrando a operacionalização conjunto para apoiar a análise de documentos textuais com base em uma taxonomia de domínio. O projeto "Integração de Fatores Humanos e Resiliência para o Fortalecimento da Cultura de Segurança na Indústria de Óleo e Gás" foi elaborado para responder a uma necessidade explícita do consórcio LIBRA, cujo foco está no aprimoramento da segurança operacional na indústria offshore. A pesquisa enfatiza a importância dos fatores humanos como pilares para consolidar uma cultura de segurança robusta, destacando que tais fatores são cruciais para adaptar e melhorar a resiliência dos trabalhadores em ambientes de alta complexidade e risco (MIRANDA JUNIOR, *et al.* 2023).

Este estudo abordou a dinâmica dos sistemas sociotécnicos na indústria de óleo e gás, revelando como estratégias operacionais e pessoais são empregadas para lidar com as demandas cotidianas e organizacionais. A análise sociológica destas práticas mostrou-se essencial para a identificação de oportunidades de melhoria e para a implementação de estratégias eficazes (REIF, *et al.* 2023). A inclusão dos fatores humanos não apenas fortalece a segurança, como também promove um ambiente de trabalho mais adaptável e resiliente.

O caráter transdisciplinar do projeto é evidenciado pela participação de diversos profissionais, que colaboram para desenvolver soluções inovadoras e práticas. Esta abordagem colaborativa é essencial para enfrentar os desafios da indústria de óleo e gás, permitindo a criação de um ambiente mais seguro e responsivo (MIRANDA JUNIOR, *et al.* 2023).

O relatório de sustentabilidade da Petrobras de 2020 destacou a importância dos fatores humanos na construção de uma cultura de segurança madura e sustentável. Este relatório serviu como um guia e uma referência importante para o projeto da PUCRS, fornecendo *insights* valiosos sobre as melhores práticas na indústria de óleo e gás (PETROBRAS, 2020).

A avaliação e mitigação de riscos relacionados a fatores humanos são cruciais, já que muitos acidentes na indústria de óleo e gás são atribuídos a falhas humanas. Ao focar nesses aspectos, o projeto não só busca reduzir a ocorrência de acidentes, mas também melhorar o bem-estar e a eficiência dos trabalhadores. Especialistas enfatizam a importância de uma abordagem holística que inclui o bem-estar dos trabalhadores como uma prioridade.

A implementação de práticas de gestão que incentivem a participação ativa de todos os níveis da organização é outro pilar fundamental do projeto. Criar um ambiente onde os funcionários se sintam responsáveis e capacitados para identificar e relatar riscos é essencial para o sucesso da iniciativa. Este empoderamento dos trabalhadores não só fortalece a cultura de segurança, mas também promove um ambiente de trabalho mais colaborativo e proativo.

A demonstração iniciará com a apresentação do sistema Java, destacando suas capacidades de processamento de texto, extração de dados e geração de arquivos JSON e CSV. Em seguida, a atenção será direcionada à aplicação *web* do Método, proporcionando uma exploração guiada dos gráficos gerados. Os participantes poderão interagir com a interface, explorar visualmente as tendências identificadas pelo sistema Java e compreender como a taxonomia contribui para a categorização significativa dos dados textuais.

A demonstração não se limitará à funcionalidade técnica. A eficiência na navegação, a clareza na apresentação dos dados e a capacidade de personalização serão aspectos fundamentais demonstrados durante essa etapa. Espera-se que a solução apresentada não apenas cumpra os objetivos estabelecidos na pesquisa, mas também evidencie sua utilidade prática, impacto potencial e contribuição significativa para a análise avançada de documentos textuais em domínios específicos.

3.3.5 Etapa 5 - Avaliação

Na fase de Avaliação, a solução desenvolvida será submetida a um teste prático em um caso de uso específico, tendo como cenário de estudo os problemas de fatores humanos em plataformas *offshore*, como acidentes e incidentes de trabalho (conforme é apresentado na Seção 5.1), proporcionando uma avaliação contextualizada e relevante para a aplicação real. Esta abordagem direcionada permite não apenas avaliar a eficácia geral da solução, mas também verificar sua adaptabilidade e utilidade em situações específicas do mundo real.

O caso de uso escolhido representará um cenário típico ou desafiador dentro do domínio específico abordado pela pesquisa. Durante esse teste, a solução será submetida a documentos textuais representativos desse cenário, permitindo uma avaliação aprofundada da capacidade do sistema Java em processar e categorizar corretamente a informação de acordo com a taxonomia estabelecida.

O método proposto será empregado para explorar os resultados gerados pelo sistema Java no contexto do caso de uso. A avaliação em um cenário de estudo específico do projeto de Fatores Humanos contribuirá significativamente para a avaliação prática da solução proposta, destacando suas capacidades e fornecendo dados valiosos para refinamentos adicionais. Isso garantirá que a solução não apenas atenda aos requisitos teóricos, mas também se mostre efetiva e aplicável em cenários reais em domínio específico.

3.3.6 Etapa 6 – Comunicação dos Resultados

Finalizada as etapas anteriores, a Etapa 6 enfoca a comunicação dos resultados obtidos durante a pesquisa, avaliação e aplicação prática da solução proposta. Esta etapa é crucial para disseminar o conhecimento gerado, compartilhar *insights* descobertos e destacar a contribuição significativa para o campo de estudo.

A comunicação dos resultados será realizada de maneira abrangente, utilizando diferentes meios e formatos. Por meio deste documento servirá como uma referência abrangente para a comunidade acadêmica e profissional interessada no tema. Serão preparadas apresentações visuais claras e envolventes que resumem os principais pontos da pesquisa. Essas apresentações podem ser utilizadas em conferências, seminários ou workshops, proporcionando uma plataforma para a interação e o compartilhamento de ideias com outros pesquisadores e profissionais do campo.

Espera-se realizar a publicação de artigos em revistas acadêmicas especializadas garantindo que os resultados da pesquisa alcancem uma audiência mais ampla e contribuam para o avanço do conhecimento na área específica de estudo.

3.4 EXECUÇÃO DA REVISÃO INTEGRATIVA DA LITERATURA

Esta pesquisa fez o uso de uma revisão integrativa da literatura para explorar a questão de pesquisa proposta nesta dissertação. Este tipo de revisão se caracteriza por ter todos os seus passos documentados (Apêndice A), de forma a permitir a replicabilidade da pesquisa por outros pesquisadores interessados no tema. Para tal, seguiu-se a proposta de Whitemore e Knafl (2005), a qual é realizada em cinco etapas, conforme Figura 11.

Figura 11 - Fases da Revisão Integrativa

Fonte: Elaborado pela autora (2023) com base no estudo de Whittemore e Knafl (2005).

As fases de Identificação do Problema, Busca na Literatura e Avaliação dos Dados estão documentadas no Apêndice B.

3.4.1 Análise dos dados e Apresentação

Os artigos obtidos por meio da revisão integrativa da literatura foram submetidos a uma análise minuciosa. A análise envolveu a identificação de padrões, tendências e lacunas de conhecimento presentes nos estudos selecionados. Foram examinadas características como o método de pesquisa utilizado, os principais resultados encontrados, as conclusões dos autores e a relevância dos estudos para o tema em questão. Um resumo de cada um dos artigos é apresentado a seguir.

O artigo de Burkhardt *et al*. (2015) aborda a participação eletrônica (*e-participation)* e o engajamento dos cidadãos na política. O objetivo principal é desenvolver soluções inovadoras que envolvam os cidadãos nos processos de tomada de decisão política, utilizando a análise e visualização de mídias sociais. A integração das mídias sociais permite que as opiniões dos cidadãos sejam consideradas, proporcionando uma maior transparência e credibilidade nas políticas públicas. Contudo, os desafios incluem a diversidade de linguagens temáticas e a fraca visualização dos resultados. A metodologia descrita foi aplicada no projeto de pesquisa do Programa-Quadro da Comissão Europeia para a Modelagem de Políticas Futuras (do inglês *European Commission Seventh Framework Programme Future Policy Modeling* - EC FP7 FUPOL), demonstrando sua praticidade e benefícios.

Barb e Kilicay-Ergin (2020) discutem a importância de um currículo eficaz e coerente para o sucesso dos programas acadêmicos. Utilizando técnicas de processamento de linguagem natural e ontologias, o artigo avalia a coerência curricular de um programa de Ciência da Informação. A abordagem possibilística identifica lacunas e sobreposições nos cursos centrais oferecidos, permitindo que universidades otimizem suas ofertas e conteúdo para melhorar a qualidade acadêmica. O estudo qualitativo de coerência curricular realizado em diferentes níveis de abstração ontológica é central para garantir que os alunos recebam uma educação integrada e aplicável em contextos profissionais.

O artigo de Jänicke *et al*. (2015) apresenta uma visão geral das pesquisas sobre visualizações que suportam leituras próximas e distantes de dados textuais nas humanidades digitais. A pesquisa inclui uma taxonomia de métodos aplicados para essas leituras e ilustra abordagens que combinam ambas as técnicas para uma visão multifacetada dos dados. Além disso, o artigo destaca as experiências de colaboração no desenvolvimento de visualizações e aponta desafios futuros na área. Este trabalho é fundamental para entender como diferentes técnicas de visualização podem ser usadas para explorar dados textuais de maneira eficaz.

Hamzah e Vu (2018) examinam as técnicas analíticas aplicadas a dados do *Twitter*, especialmente aqueles com informações de localização. Desde a introdução do serviço de localização do *Twitter* em 2009, tornou-se possível coletar dados geolocalizados, o que é valioso para a criação de sistemas baseados em localização. No entanto, os dados ruidosos dos usuários exigem processos de análise de texto e visualização para serem compreensíveis e úteis cientificamente. O artigo propõe uma taxonomia das técnicas de análise de dados do *Twitter*,

categorizando seus usos e potenciais aplicações, especialmente para dados baseados em localização, e discute as oportunidades e desafios futuros.

Kucher *et al*. (2024) investigam a visualização de um grande volume de artigos de notícias suecas. Utilizando NLP e métodos de visualização de texto, o estudo busca resolver a dificuldade dos usuários em encontrar informações relevantes. Três representações visuais foram projetadas e avaliadas, com um protótipo interativo final que utiliza gráficos de área para representar a evolução de tópicos. O estudo revela a preferência dos participantes por representações focadas nos principais tópicos e oferece sugestões para trabalhos futuros na visualização de dados textuais suecos.

Yarmohammadi *et al*. (2017) exploram a utilização de Modelagem da Informação da Construção (do inglês *Building Information Modeling* - BIM) para monitorar e medir o desempenho dos modeladores durante o processo de design. Utilizando registros de atividades de *software*, como o *Autodesk Revit*, os autores analisaram os padrões implícitos presentes nos arquivos de *log*. A pesquisa identificou que há diferenças estatisticamente significativas no tempo necessário para executar sequências de comandos, contribuindo para uma melhor compreensão do processo de modelagem e propondo uma nova metodologia para extrair padrões significativos de dados de *log* não estruturados.

Chen *et al*. (2020) abordam a lacuna existente entre a análise visual complexa e a comunicação dos resultados para um público leigo. O artigo propõe uma estrutura geral onde a síntese de histórias conecta a análise de dados à apresentação de resultados, organizando os conteúdos de forma significativa. Este método é aplicado a duas áreas distintas: mídia social e análise de movimento, demonstrando sua versatilidade. A abordagem foca em selecionar, montar e organizar os achados de maneira compreensível, em vez de apenas rastrear o histórico de análise.

Ortigossa *et al*. (2022) discutem técnicas de projeção multidimensional para facilitar a análise de dados complexos. O artigo fornece uma visão abrangente das técnicas de projeção, como a Análise de Componentes Principais (do inglês *Principal Component Analysis* - PCA), destacando suas capacidades e limitações. Exemplos práticos em experimentos genéticos e mineração de texto ilustram a aplicação dessas técnicas. A pesquisa promove uma compreensão mais profunda dos métodos de projeção, sua eficácia e possíveis distorções, contribuindo para o desenvolvimento de sistemas de descoberta de conhecimento em ciências de dados.

Liu *et al*. (2019) conduzem uma análise abrangente de técnicas de visualização e mineração de texto, examinando 263 artigos de visualização e 4.346 artigos de mineração. O

estudo resultou na criação de uma taxonomia para cada tipo de técnica e na extração das relações de coocorrência entre os conceitos. Este trabalho serve como um ponto de partida para pesquisadores entenderem as práticas atuais e explorarem oportunidades de pesquisa, além de facilitar a integração das técnicas de visualização e mineração de texto em análises interdisciplinares.

Abdul-Rahman *et al*. (2013) apresentam um estudo de design centrado no usuário para a visualização de poesia. Desenvolveram uma solução baseada em regras para balancear a flexibilidade na visualização de várias variáveis poéticas e reduzir a carga cognitiva na interação com o painel de controle de mapeamento visual. O estudo adota o modelo de design de *Munzner* para manter interações de alto nível com os usuários finais e examina opções de design para facilitar a visualização de poemas. Exemplos de uso da visualização de poesia em pesquisas acadêmicas são apresentados.

Holzinger *et al*. (2014) discutem os desafios enfrentados pelos profissionais das ciências da vida devido ao aumento dos conjuntos de dados complexos. A maioria desses dados é não estruturada ou fracamente estruturada, dificultando o uso de métodos tradicionais de recuperação de informações. O artigo propõe métodos avançados para lidar com dados biomédicos, considerando aspectos temporais (como entropia da informação) e espaciais (como topologia computacional). Exemplos de dados biomédicos são apresentados e uma taxonomia específica para conjuntos de dados médicos é discutida, destacando a necessidade de novos métodos para descobrir conhecimento desconhecido a partir desses dados.

Liu *et al*. (2013) abordam a criação de um mapa interativo e visual para uma grande coleção de conjuntos de dados abertos. Utilizando técnicas de mineração de texto para definir um espaço de representação e o método de k-vizinhos mais próximos para construir um gráfico de proximidade, o artigo apresenta uma visualização baseada no *software Tulip*. A pesquisa inclui 293.000 conjuntos de dados do site francês de dados abertos, limitando a visualização a 151.000 conjuntos de dados. *Clusters* descobertos são analisados e demonstrados como úteis para navegação na coleção.

Fister *et al*. (2023) revisam métodos de visualização para mineração de regras de associação, que busca relações entre atributos em bancos de dados de transações. O processo inclui técnicas de pré-processamento, mineração de regras e pós-processamento com visualização. O artigo identifica as principais técnicas publicadas, examina suas características

e aplicações, e propõe futuros passos para a pesquisa na área. A revisão destaca a importância da visualização para a compreensão dos resultados da mineração de regras de associação.

Almutairi (2013) apresenta o *AppAnn*, um sistema de visualização de texto projetado para apoiar a análise de discurso, especificamente para modelar padrões de avaliação ao longo do texto. O artigo destaca a dificuldade de rastrear relações complexas e em desenvolvimento em dados de alta dimensionalidade e como o *AppAnn* pode auxiliar linguistas a realizar análises de longo alcance. O sistema é demonstrado com exemplos de visualização de padrões de avaliação em textos, mostrando sua utilidade para a análise linguística.

Luz e Sheehan (2020) descrevem a metodologia de *co-design* utilizada para desenvolver ferramentas de visualização no projeto *Genealogies of Knowledge*. A abordagem inclui revisão de metodologias publicadas, observação in situ de estudiosos, prototipagem de software, análise de produção acadêmica e entrevistas com usuários. O artigo apresenta estudos de caso que demonstram o uso do software em análises de conceitos médicos, científicos e políticos, discutindo as implicações metodológicas no contexto da abordagem baseada em corpus do projeto.

Fu *et al*. (2018) apresentam o *VisForum*, um sistema de análise visual projetado para explorar grupos de usuários em fóruns online. Utilizando uma interface multi-coordenada e glifos inovadores, como glifos de grupo e de usuário, o sistema facilita a detecção e comparação de grupos em threads de discussão. Um algoritmo específico é implementado para reduzir ruídos e identificar membros de alto impacto nos fóruns. Estudos de caso com conjuntos de dados reais demonstram a utilidade do sistema e a eficácia dos novos designs de glifos.

Liu *et al*. (2013) abordam a criação de um mapa visual e interativo para uma grande coleção de conjuntos de dados abertos. Utilizando técnicas de mineração de texto para criar características e o método dos vizinhos relativos para construir um gráfico de proximidade, o artigo apresenta uma visualização baseada no *software Tulip*. A pesquisa inclui 300.000 conjuntos de dados do *site* francês de dados abertos, limitando a visualização a 150.000 conjuntos de dados. Os *clusters* descobertos são analisados e utilizados para navegação na coleção.

Malik *et al*. (2015) apresentam o Comparação de Coortes (do inglês *Cohort Comparison - CoCo)*, uma ferramenta de análise visual que integra estatísticas automatizadas e uma interface inteligente para comparar coortes de sequências de eventos temporais. A ferramenta auxilia os usuários na identificação de características distintas e significativas entre as coortes, equilibrando abordagens visuais e estatísticas. Estudos de caso iniciais mostram a

aplicação da ferramenta em pesquisas sobre desempenho de equipes médicas em emergências e em pesquisas farmacêuticas.

Adwan *et al*. (2020) fornecem uma visão geral dos algoritmos e abordagens utilizadas na análise de sentimento no *Twitter*. Os artigos revisados são categorizados em quatro abordagens principais, e o estudo discute direções futuras para a pesquisa, incluindo a utilização de teorias e tecnologias de outras áreas como ciência cognitiva, *Web* semântica, *big data* e visualização. A análise de sentimento no *Twitter* é valiosa para entender as opiniões e interesses dos usuários em diversos contextos.

Havens *et al*. (2022) discutem técnicas de visualização de texto para facilitar a análise de conjuntos de dados anotados não agregados, relevantes para o perspectivismo de dados. O artigo aborda os desafios das práticas atuais de criação de conjuntos de dados e plataformas de anotação, propondo técnicas de visualização que permitem uma análise intuitiva e multifacetada. Essas técnicas são úteis para pesquisadores e *stakeholders* de NLP que podem não ter habilidades em ciência de dados ou computação.

Eteläaho *et al*. (2018) abordam a mineração de dados visuais em repositórios de software, destacando a quantidade massiva de dados que muitas vezes não é utilizada devido à falta de ferramentas poderosas de visualização. A pesquisa foca em técnicas de visualização que ajudam a descobrir relações nos dados de *software*, essenciais para a manutenção, evolução e compreensão dos aspectos sociotécnicos do desenvolvimento de software. A pesquisa revisa as técnicas mais comuns de mineração de dados visuais e sua aplicação na engenharia de *software*, mostrando como essas técnicas podem facilitar o reconhecimento de informações e a descoberta de conhecimento.

Norambuena *et al*. (2023) exploram a extração de narrativas de notícias baseadas em eventos, uma subárea da inteligência artificial que utiliza técnicas de recuperação de informações e processamento de linguagem natural. A pesquisa revisou mais de 900 artigos, sintetizando 54 relevantes, e organizando-os por modelo de representação, critérios de extração e abordagens de avaliação. O estudo identifica tendências recentes, desafios abertos e linhas de pesquisa potenciais, ressaltando a importância da extração de narrativas para compreender a paisagem informacional em evolução.

Dudycz (2021) investiga a visualização de redes semânticas como interface de usuário para a busca de informações de negócios. O estudo inclui uma análise da literatura e a validação de protótipos desenvolvidos no *Protégé*. Quatro áreas de pesquisa são identificadas:

desenvolvimento de novos *softwares*, aplicação de técnicas e tecnologias para operações de visualização, verificação de design gráfico e validação do usuário. A pesquisa revela problemas potenciais relacionados ao uso da visualização de redes semânticas como interface visual na busca de informações econômicas.

Camacho *et al*. (2020) apresentam uma revisão abrangente da análise de redes sociais (do inglês *Social Network Analysis* - SNA), propondo quatro dimensões essenciais: descoberta de padrões e conhecimento, fusão e integração de informações, escalabilidade e visualização. A pesquisa avalia e classifica 20 ferramentas de SNA, oferecendo uma análise quantitativa das tecnologias de redes sociais. O estudo também realiza uma análise cientométrica para identificar as áreas de pesquisa e domínios de aplicação mais ativos, fornecendo *insights* sobre os desafios e tendências futuras na área.

Liu *et al*. (2014) apresentam o Exploração de Conjuntos de Dados Abertos (do inglês *EXploration of Open Datasets* - EXOD), uma ferramenta para análise visual de uma grande coleção de conjuntos de dados abertos. Utilizando técnicas de mineração de texto para extrair características de metadados e conteúdo, o EXOD constrói um gráfico de proximidade e utiliza o *software Tulip* para visualização. A ferramenta permite explorar interativamente os dados, facilitando a descoberta de *clusters* e informações relevantes. O EXOD processou 293.000 conjuntos de dados do site francês de dados abertos, demonstrando sua eficácia na exploração de grandes coleções de dados.

Li *et al*. (2022) realizam um estudo empírico sobre a eficácia dos *dashboards* de informações sobre a Doença por Coronavírus 2019 (do inglês *Coronavirus Disease* 2019 - COVID-19) em atender às necessidades informacionais dos usuários. O estudo compara as necessidades expressas no *Twitter* com as informações fornecidas pelos *dashboards*, identificando lacunas significativas. Além das informações disponíveis, os usuários também estão interessados na relação da COVID-19 com outros vírus, a origem do vírus, o desenvolvimento de vacinas, notícias falsas, e os impactos na educação, mulheres e negócios. Os resultados podem ajudar os desenvolvedores a otimizar os *dashboards* para melhor atender às necessidades dos usuários e melhorar futuros desenvolvimentos de *dashboards* de gerenciamento de crises.

Swietlicki e Cubaud (2022) descrevem uma ferramenta de visualização de visão geral para coleções de correspondência digitalizadas, aplicada ao arquivo Godin-Moret (20.000 cartas). A interface, desenvolvida com a ajuda de um *workshop* de co-criação online, utiliza uma matriz de correspondência coordenada com visualizações padrão, como nuvens de *tags*,

mapas e gráficos de barras. A ferramenta visa oferecer uma alternativa útil aos motores de busca em bibliotecas digitais, facilitando a navegação e exploração dos dados.

Dudycz (2015) investiga a usabilidade de uma rede semântica visual baseada em ontologias de indicadores econômicos e financeiros como interface interativa para busca de informações. O estudo utiliza um questionário de satisfação do usuário e um teste de usabilidade para avaliar a eficácia de dois protótipos de visualização. Os resultados mostram que participantes com formação apenas em economia ou Tecnologia da Informação (TI) avaliaram a visualização de busca semântica de forma mais positiva do que aqueles com formação em ambas as áreas, identificando menos problemas potenciais.

Ranganathan *et al.* (2013) desenvolvem um sistema de busca e visualização de letras de músicas em Tamil, utilizando modelagem estatística para explorar características internas e externas das letras. O sistema permite buscas por palavras-chave, busca semântica e busca por estilo do letrista, utilizando Frequência de Termo-Frequência Inversa de Documento (do inglês *Term Frequency-Inverse Document Frequency* - TF-IDF) e a Linguagem de Rede Universal (do inglês *Universal Networking Language* - UNL). A visualização das características das letras é representada por uma estrutura em forma de flor, criada com uma ferramenta *Graphics2D*, para aumentar o interesse do usuário. A eficácia da visualização foi avaliada em termos de conforto e precisão na busca de gêneros e emoções.

Shurkhovetskyy *et al.* (2018) propõem uma estrutura de classificação para métodos de abstração de dados temporais, voltada para a visualização de séries temporais grandes. A pesquisa avalia métodos de mineração de dados quanto à sua utilidade para visualização, propondo critérios essenciais para a seleção de métodos de abstração. A classificação considera propriedades dos dados, formas de representação desejadas, características comportamentais a serem estudadas, precisão e nível de detalhe necessários, além da eficiência na busca e consulta. A pesquisa também sugere direções para possíveis extensões da estrutura de classificação.

Baumer *et al.* (2022) discutem as dimensões políticas subestimadas na visualização de textos cívicos. O artigo enfatiza a importância de considerar aspectos políticos, feministas, éticos e retóricos nas decisões de design de visualização de dados. A análise crítica de pesquisas sobre visualização de textos cívicos revela que o enquadramento exclusivamente analítico pode levar a problemas como a interpretação equivocada dos dados, a ausência de vozes minoritárias e a exclusão do público dos processos de tomada de decisão. Para abordar essas questões, os

autores propõem dimensões conceituais que ajudam a revelar as implicações políticas das decisões de design na visualização cívica.

Panagiotidou *et al.* (2023) exploram a comunicação da incerteza nas visualizações utilizadas em pesquisas das humanidades digitais. A revisão de 126 publicações mostra que a incerteza permeia várias etapas do processo de pesquisa, desde os artefatos de origem até a sua dataficação. As visualizações, por vezes, falham em comunicar essa incerteza, o que pode afetar a confiança dos humanistas nas representações visuais. O artigo propõe duas taxonomias empíricas: uma de incertezas e outra de estratégias para lidar com elas, contribuindo para o desenvolvimento de visualizações mais confiáveis e conscientes da incerteza.

Saleheen e Lai (2018) apresentam o UIWGViz, uma arquitetura para a visualização de grafos da *web* baseada nos interesses do usuário. A arquitetura propõe procedimentos para modelar os interesses dos usuários, analisando tanto *feedbacks* implícitos quanto explícitos. Os interesses dos usuários são incorporados na filtragem, geração de grafos e agrupamento de documentos da web, melhorando a relevância dos resultados de busca. A interface visual do UIWGViz permite uma navegação e interação eficazes, capturando *feedbacks* dos usuários para otimizar a visualização.

Borland *et al.* (2019) desenvolvem uma ferramenta de visualização interativa baseada em ontologias para ajudar pesquisadores a explorar perguntas de pesquisa geradas por pacientes com Doença de *Crohn* e colite. A ferramenta organiza os tópicos de pesquisa em visualizações vinculadas, permitindo aos pesquisadores identificar temas comuns e conexões entre eles. Um exemplo de uso da ferramenta e o *feedback* dos usuários são discutidos, destacando a potencialidade da integração de ontologias específicas da comunidade com ferramentas de visualização interativas para melhorar a priorização das agendas de pesquisa.

Alhuay-Quispe *et al*. (2022) investigam os métodos e ferramentas utilizadas em estudos bibliométricos, focando em pesquisadores que não têm formação específica em Ciências da Informação. O estudo utiliza análise textual e de coocorrência de palavras para identificar *softwares* e ferramentas mais frequentes em estudos bibliométricos. Os resultados mostram quatro níveis de tratamento de dados: recuperação, preparação, processamento e análise, e visualização. A pesquisa destaca a importância de métodos de análise variados, como redes sociais, geoespaciais, temáticos e de acoplamento bibliográfico, para a realização de estudos bibliométricos rigorosos.

Amadio e Procaccino (2016) investigam a utilidade da mineração de texto e análise visual de avaliações online para realizar análises SWOT no setor hoteleiro. Utilizando a

ferramenta *ReviewMap*, eles transformam um arquivo de avaliações em uma hierarquia de dados, identificando recursos necessários para a competição e diferenciação entre hotéis. O estudo revela que a combinação de resumos de mineração de texto com classificações numéricas quase dobrou a eficácia analítica, destacando ações competitivas promissoras para os hotéis analisados.

Yuan *et al*. (2020) revisam sistematicamente 259 artigos publicados na última década sobre técnicas de análise visual aplicadas ao aprendizado de máquina. Eles constroem uma taxonomia com três categorias principais: técnicas antes, durante e depois da construção do modelo. Cada categoria é caracterizada por tarefas representativas de análise, exemplificadas por trabalhos influentes. O artigo discute desafios e oportunidades futuras de pesquisa, fornecendo uma visão abrangente para pesquisadores de análise visual.

Babic e Jerman-Blazic (2016) propõem uma nova taxonomia para visualização do processo de mineração de dados em atividades de cibercrime. O estudo enfatiza a importância da visualização para entender grandes volumes de dados e identificar padrões que podem não ser detectados em dados baseados em texto. A pesquisa integra técnicas de teoria dos grafos e fractais para suportar a exploração de grandes conjuntos de dados, destacando a aplicação de visualização e mineração de dados em cibercrime.

Neumann *et al*. (2014) apresentam o *BLASTGrabber*, uma ferramenta bioinformática para visualização e análise de dados massivos gerados pelo algoritmo BLAST. Implementada em Java, a aplicação é independente de sistema operacional e oferece uma interface gráfica amigável. O *BLASTGrabber* permite a visualização de alinhamentos de sequência, organização de dados em uma árvore taxonômica interativa e análise por mineração de texto. A ferramenta visa facilitar o processamento de grandes saídas de BLAST para usuários não especialistas em habilidades computacionais.

Liu *et al.* (2023) apresentam uma abordagem de mineração de texto baseada em grafos para extração de informações de relatórios de inspeção de projetos de construção. O *framework* proposto integra coleta de dados, pré-processamento e três níveis de análise de texto: palavra, sentença e documento. A análise identifica tipos de questões de não conformidade e suas inter-relações, automatizando o desenvolvimento de uma taxonomia baseada em dados. A abordagem demonstrou eficácia na classificação e identificação de características críticas em relatórios de inspeção, utilizando menos intervenção manual que métodos tradicionais.

Esses trabalhos fornecem uma base sólida para futuras pesquisas e desenvolvimentos, evidenciando oportunidades e desafios em cada área abordada. A interseção entre arte e ciência na visualização de dados, a comunicação efetiva de incertezas e a aplicação de técnicas visuais na educação online são áreas que podem se beneficiar de investigações mais aprofundadas e desenvolvimentos práticos.

Os artigos selecionados tiveram em comum, a utilização de várias palavras-chaves em comum, presentes no título, como pode ser observado pela nuvem de palavras mostrada na Figura 12. Em especial as palavras: *visualization* (15), *data* (11), *analysis* (9), *visual* (8), *mining* (7), *survey* (6), *information* (5), *taxonomy* (4), no total de 192 palavras.

Figura 12 - Nuvem de palavras-chave

Fonte: Elaborado pela autora.

Com relação ao grau de intersecção, podemos classificar os artigos com:

- **Baixo grau de intersecção:** Isso ocorre quando dois dos termos são usados como complemento um do outro. Por exemplo, se um artigo discute "mineração de texto" e "taxonomia", com pouca sobreposição entre os dois conceitos.
- **Médio grau de intersecção:** Isso acontece quando os três termos são usados em momentos distintos dentro do mesmo artigo. Por exemplo, um artigo pode abordar "mineração de texto", "taxonomia" e "visualização de dados", mas sem uma forte integração entre eles.
- **Alto grau de intersecção:** Isso ocorre quando a maioria dos termos se complementam mutuamente. Por exemplo, um artigo pode discutir "mineração de texto" com uma ênfase significativa em "taxonomia" e "visualização de dados", onde os termos se relacionam estreitamente e se complementam em todo o texto.

A avaliação do grau de intersecção entre os termos "*text mining*", "*information visualization*" e "*taxonomy*" nos quarenta artigos revela uma panorâmica diversificada das abordagens adotadas pelos pesquisadores no campo da visualização de dados, mineração de texto e aprendizado de máquina. Ao classificar o grau de intersecção em cada artigo, emerge uma compreensão mais refinada da variedade de perspectivas e ênfases presentes nas pesquisas.

O Quadro 9 classifica os artigos com base no grau de intersecção dos termos "mineração de texto", "taxonomia" e "visualização de dados". Essa análise identifica a frequência e a profundidade com que esses termos são integrados em cada artigo, categorizando-os em três graus de intersecção: baixo, médio e alto.

Quadro 9 – Grau de intersseção entre os termos

Autores	Grau de Intersecção	Termos
Burkhardt *et al*. (2015)	Baixo	visualização de dados
Barb e Kilicay-Ergin (2020)	Médio	mineração de texto, taxonomia
Jänicke *et al.* (2015)	Médio	taxonomia, visualização de dados
Hamzah e Vu (2018)	Alto	mineração de texto, taxonomia, visualização de dados

Autores	Grau de Intersecção	Termos
Kucher *et al*. (2024)	Médio	mineração de texto, visualização de dados
Yarmohammadi *et al*. (2017)	Médio	mineração de texto, visualização de dados
Chen *et al*. (2020)	Médio	mineração de texto, visualização de dados
Ortigossa *et al*. (2022)	Médio	mineração de texto, visualização de dados
Liu *et al*. (2019)	Alto	mineração de texto, taxonomia, visualização de dados
Abdul-Rahman *et al*. (2013)	Baixo	visualização de dados
Holzinger *et al*. (2014)	Médio	taxonomia, visualização de dados
Liu *et al*. (2013)	Médio	mineração de texto, visualização de dados
Fister *et al*. (2023)	Médio	mineração de texto, visualização de dados
Almutairi (2013)	Médio	mineração de texto, visualização de dados
Luz e Sheehan (2020)	Médio	visualização de dados, taxonomia
Fu *et al*. (2018)	Baixo	visualização de dados
Liu *et al*. (2013)	Médio	mineração de texto, visualização de dados
Malik *et al*. (2015)	Médio	visualização de dados
Adwan *et al*. (2020)	Médio	visualização de dados
Havens *et al*. (2022)	Médio	mineração de texto, visualização de dados
Eteläaho *et al*. (2018)	Médio	mineração de texto, visualização de dados
Norambuena *et al*. (2023)	Médio	mineração de texto, taxonomia
Dudycz (2021)	Médio	visualização de dados, taxonomia
Camacho *et al*. (2020)	Médio	visualização de dados, taxonomia
Liu *et al.* (2014)	Médio	mineração de texto, visualização de dados
Li *et al.* (2022)	Baixo	visualização de dados
Swietlicki e Cubaud (2022)	Baixo	visualização de dados
Dudycz (2015)	Médio	visualização de dados, taxonomia
Ranganathan *et al*. (2013)	Médio	mineração de texto, visualização de dados
Shurkhovetskyy *et al*. (2018)	Médio	mineração de texto, visualização de dados
Baumer *et al*. (2022)	Baixo	visualização de dados

Autores	Grau de Intersecção	Termos
Panagiotidou *et al*. (2023)	Médio	visualização de dados, taxonomia
Saleheen e Lai (2018)	Baixo	visualização de dados
Borland *et al*. (2019)	Médio	visualização de dados, taxonomia
Alhuay-Quispe *et al*. (2022)	Médio	mineração de texto, visualização de dados
Amadio e Procaccino (2016)	Médio	mineração de texto, visualização de dados
Yuan *et al*. (2020)	Alto	mineração de texto, taxonomia, visualização de dados
Babic e Jerman-Blazic (2016)	Alto	mineração de texto, taxonomia, visualização de dados
Neumann *et al*. (2014)	Alto	mineração de texto, taxonomia, visualização de dados
Liu *et al*. (2023)	Alto	mineração de texto, taxonomia, visualização de dados

Fonte: Elaborado pela autora.

Essa classificação evidencia a amplitude e a interdisciplinaridade da pesquisa contemporânea no campo, onde os pesquisadores exploram diferentes facetas da visualização de dados, mineração de texto e taxonomia. Além disso, ressalta a importância de uma abordagem integrada para avançar no entendimento e na aplicação eficaz desses conceitos inter-relacionados. A diversidade de classificações reflete o dinamismo e a evolução constante desse campo multidisciplinar, incentivando futuras pesquisas a abordar desafios e explorar oportunidades emergentes nas interseções desses domínios.

4 MÉTODO PROPOSTO

A análise de texto, apoiada por uma taxonomia como referência, é uma abordagem poderosa para extrair *insights* significativos de grandes volumes de informações textuais. Este método visa não apenas compreender o conteúdo textual, mas também relacioná-lo a uma estrutura conceitual predefinida, proporcionando uma compreensão mais profunda e organizada dos dados. A seguir, serão apresentadas as etapas desse processo, destacando a importância de cada uma para o sucesso da análise textual e a geração de *insights* valiosos.

A primeira etapa (objeto de estudo) envolve a obtenção de um documento de texto como entrada para o processo, representando o objeto de interesse a ser estudado. Esse documento pode ser obtido de várias fontes, como artigos científicos, páginas da web, documentos corporativos, entre outros. É essencial que o documento seja em formato de texto para que possa ser processado adequadamente. Além disso, é importante que o conteúdo do documento esteja relacionado com o domínio da taxonomia em estudo. A escolha criteriosa do objeto de estudo garante que a análise seja relevante e focada nos temas de interesse.

Escolher o documento de texto correto é fundamental para o sucesso da análise. A relevância do documento deve ser avaliada em relação à sua contribuição potencial para o entendimento do domínio específico da taxonomia. Documentos que não estão diretamente relacionados ao domínio podem fornecer dados irrelevantes ou confundir a análise. Portanto, a seleção do documento deve ser feita com cuidado, considerando a qualidade e a relevância da informação contida nele.

Além disso, o formato do documento é um fator crucial. Documentos em formatos processáveis, como TXT, DOCX, ou PDF, facilitam a extração de texto e a aplicação de técnicas de NLP. Documentos em formatos inadequados ou protegidos por senha podem dificultar a análise e exigir etapas adicionais de conversão ou processamento, o que pode introduzir erros ou distorções nos dados analisados.

Após a obtenção do documento, inicia-se o processamento do texto (etapa 2) utilizando técnicas de Processamento de Linguagem Natural (NLP). Nesta etapa, o texto é analisado tanto lexical quanto sintaticamente. A análise lexical envolve a divisão do texto em palavras individuais, enquanto a análise sintática foca na estruturação do texto em unidades gramaticais, como frases e parágrafos. Esse processamento inicial é crucial para preparar o texto para as análises subsequentes, facilitando a identificação e extração de informações relevantes.

A análise lexical é o primeiro passo para decompor o texto em suas unidades básicas. Este processo, conhecido como tokenização, envolve a segmentação do texto em palavras e outras unidades, como pontuações e espaços em branco. Isso permite que o texto seja manipulado de maneira mais granular, facilitando a identificação de padrões e a aplicação de outras técnicas de NLP. A tokenização é uma etapa crítica, pois uma segmentação inadequada pode levar a erros nas etapas subsequentes de análise.

A análise sintática, por sua vez, foca na estrutura gramatical do texto. Nesta fase, o texto é dividido em frases e parágrafos, e as relações gramaticais entre as palavras são identificadas. Isso permite uma compreensão mais profunda da estrutura do texto, facilitando a análise das relações entre diferentes partes do texto. A análise sintática é essencial para a identificação de padrões mais complexos e para a extração de informações que dependem da estrutura gramatical do texto.

Uma etapa essencial do processo é a carga da taxonomia (etapa 3), que deve conter informações detalhadas, como Dimensão, Nome do Fator, Pref Termo, Alt Termo e Termos. A utilização de uma taxonomia bem estruturada é crucial, pois proporciona uma referência organizada que permite uma análise mais precisa e sistemática dos documentos textuais. A taxonomia facilita a categorização dos termos e conceitos presentes nos documentos, contribuindo para uma compreensão mais profunda e eficaz do conteúdo. É importante que a taxonomia seja abrangente e reflita de maneira fiel o domínio de interesse.

A taxonomia deve ser carregada de maneira que seja facilmente acessível durante a análise. Isso pode envolver a criação de um banco de dados ou a utilização de arquivos estruturados que contenham todas as informações necessárias. A estrutura da taxonomia deve ser clara e bem definida, com cada dimensão, fator e termo devidamente categorizado e descrito. Isso garante que a análise possa ser realizada de maneira consistente e que os resultados sejam comparáveis entre diferentes estudos.

O alinhamento da taxonomia com o documento é uma etapa crítica. É necessário mapear os termos e conceitos do documento em relação à taxonomia para garantir que a análise seja precisa e relevante. Isso pode envolver a utilização de técnicas de correspondência de termos e a revisão manual para garantir que os termos sejam corretamente identificados e categorizados. Esse alinhamento é fundamental para que a análise forneça *insights* significativos e acionáveis.

Nesta etapa de Armazenamento de Dados, é fundamental compreender os dados armazenados para suportar a etapa de visualização. Isso inclui a frequência dos termos presentes na taxonomia, as coocorrências de termos, as coocorrências de fatores e termos, a anotação de fatores nos parágrafos, entre outros. As tabelas de armazenamento são projetadas com objetivos específicos: a tabela "*study*" registra os dados do estudo realizado; a "*dimension*" armazena as dimensões da taxonomia; a "*factor*" armazena os fatores relacionados a cada dimensão; a "*term*" contém os termos vinculados a cada fator; a "*paragraph*" guarda os parágrafos extraídos dos relatórios de incidentes, enquanto a "*paragraph_term*" estabelece o relacionamento entre parágrafos e termos. Além disso, as tabelas "*relation*" e "*factor_term_relation*" registram as relações entre os termos e entre os termos e fatores, respectivamente. Esses dados são essenciais para uma análise detalhada e compreensiva do conteúdo dos documentos textuais.

O projeto das tabelas deve ser feito de maneira a facilitar a extração e a análise dos dados. Cada tabela deve ser projetada para armazenar informações específicas de maneira estruturada e eficiente. A tabela "*study*" deve incluir metadados sobre o estudo, como o título, a data e o autor. A tabela "*dimension*" deve listar todas as dimensões da taxonomia, enquanto a tabela "*factor*" deve incluir os fatores associados a cada dimensão. A tabela "*term*" deve conter todos os termos vinculados a cada fator, e a tabela "*paragraph*" deve armazenar os parágrafos extraídos dos documentos analisados. As tabelas "*relation*" e "*factor_term_relation*" devem registrar as relações entre os termos e entre os termos e fatores, respectivamente.

A correta armazenagem dos dados é crucial para a etapa de visualização, pois permite a criação de visualizações precisas e informativas. Dados mal armazenados ou estruturados de maneira inadequada podem dificultar a análise e levar a conclusões incorretas. Portanto, é essencial que o armazenamento dos dados seja feito de maneira rigorosa e bem planejada, garantindo que todas as informações relevantes sejam capturadas e organizadas de maneira eficiente.

Com a base de dados preenchida, é disponibilizada a visualização dos dados (etapa 5). A partir deste modelo, podem-se criar visualizações que permitem compreender melhor o conteúdo do texto e as relações entre os conceitos presentes nele. Essas visualizações podem assumir diversas formas, como gráficos de rede, mapas de conceitos, diagramas de árvore, entre outros. A escolha da visualização depende do objetivo da análise e das características do texto e da taxonomia utilizada. Visualizações eficazes ajudam a identificar padrões, tendências e *insights* que não seriam facilmente perceptíveis apenas com a leitura do texto bruto, proporcionando uma compreensão mais profunda e acionável dos dados analisados.

A criação de visualizações é uma etapa crucial para transformar dados complexos em *insights* acionáveis. Visualizações como gráficos de rede podem mostrar as conexões entre diferentes termos e fatores, revelando padrões ocultos e relações significativas. Mapas de conceitos podem fornecer uma visão geral dos principais temas presentes no texto, enquanto diagramas de árvore podem detalhar a hierarquia dos conceitos e como eles se relacionam uns com os outros. Cada tipo de visualização tem suas vantagens e pode ser escolhido com base no tipo de informação que se deseja destacar.

Além de escolher o tipo de visualização, é importante garantir que as visualizações sejam claras e compreensíveis. Isso pode envolver o uso de cores, legendas e outras ferramentas de design para destacar informações importantes e facilitar a interpretação dos dados. Visualizações mal projetadas podem confundir o usuário e dificultar a extração de *insights*, por isso é essencial investir tempo e recursos na criação de visualizações eficazes e informativas.

O Quadro 10, resume o detalhamento de cada etapa, proporcionando uma visão clara e estruturada do método proposto:

Quadro 10 – Detalhes das etapas do Método Proposto

Etapa	Detalhamento
Etapa 1: Objeto de Estudo	**Seleção do Documento:** Escolha criteriosa do documento de texto que será analisado. A relevância e a adequação do documento ao domínio da taxonomia são cruciais. **Formato do Documento:** Garantia de que o documento esteja em um formato de texto processável, como TXT, DOCX, ou PDF, facilitando a extração de texto. **Relacionamento com o Domínio:** Verificação de que o conteúdo do documento está relacionado com o domínio da taxonomia, garantindo que a análise seja focada e relevante.
Etapa 2: Processamento do Texto	**Análise Lexical:** Segmentação do texto em unidades básicas (tokens), como palavras e pontuações, utilizando técnicas de tokenização. **Análise Sintática:** Estruturação do texto em frases e parágrafos, identificando relações gramaticais entre palavras para formar sentenças coerentes. **Preparação para Análise Posterior:** Garantia de que o texto esteja devidamente processado para facilitar a identificação e extração de informações relevantes nas etapas subsequentes.
Etapa 3: Carregamento da Taxonomia	**Estrutura da Taxonomia:** Detalhamento das dimensões, fatores e termos incluídos na taxonomia.

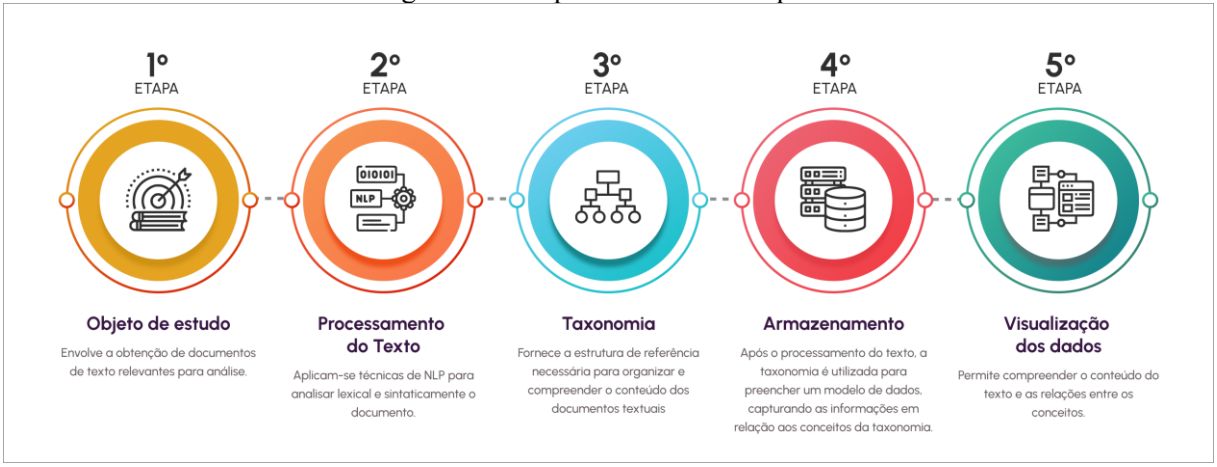
Etapa	Detalhamento
	Carregamento da Taxonomia: Implementação de um sistema para carregar e acessar a taxonomia durante a análise. **Alinhamento com o Documento:** Verificação e mapeamento dos termos e conceitos do documento em relação à taxonomia para garantir a relevância e a precisão da análise.
Etapa 4: Armazenamento de Dados	**Projeto de Tabelas:** Criação de tabelas específicas para armazenar dados do estudo, dimensões, fatores, termos, parágrafos e suas relações. **Registro de Coocorrências:** Documentação das frequências e coocorrências de termos e fatores, permitindo análises quantitativas e qualitativas. **Armazenamento Estruturado:** Garantia de que os dados sejam armazenados de maneira estruturada e eficiente, facilitando a extração e análise subsequente.
Etapa 5: Visualização dos Dados	**Criação de Visualizações:** Desenvolvimento de gráficos e diagramas que representem visualmente os dados e suas inter-relações. **Escolha da Visualização:** Seleção do tipo de visualização mais adequado com base no objetivo da análise e nas características do texto e da taxonomia. **Interpretação dos Dados:** Análise das visualizações para identificar padrões, tendências e *insights* relevantes para o objeto de estudo.

Fonte: Elaborado pela autora.

As etapas envolvidas no processo de análise de texto utilizando uma taxonomia como referência são fundamentais para a compreensão e extração de informações relevantes. Desde a obtenção do documento de texto até a visualização dos dados, cada etapa desempenha um papel crucial no desenvolvimento de *insights* valiosos a partir do texto analisado. A integração dessas etapas no processo de análise textual com uma taxonomia como referência é essencial para extrair *insights* valiosos e proporcionar uma compreensão aprofundada do texto analisado. Cada etapa desempenha um papel crucial, desde a obtenção do documento de texto até a visualização dos dados, contribuindo para o desenvolvimento de *insights* significativos a partir da análise textual.

Neste contexto, a Figura 13 ilustra visualmente o fluxo de trabalho envolvido no processo de análise de texto utilizando uma taxonomia como referência.

Figura 13 - Etapas do Método Proposto



Fonte: Elaborado pela autora.

As etapas iniciais do processo, que envolvem a definição do objeto de estudo e a carga da taxonomia, desempenham um papel crucial na preparação dos dados para análise avançada. Nesse sentido, o objeto de estudo, representado por documentos textuais relevantes, é submetido ao NLP após a carga da taxonomia. O NLP, por sua vez, analisa e estrutura o texto de acordo com padrões léxicos e sintáticos, preparando-o para a etapa subsequente. Uma vez processado pelo NLP, os dados são armazenados em um formato estruturado que captura informações relevantes em relação aos conceitos da taxonomia. Esses dados armazenados, por sua vez, servem como entrada para a etapa de visualização, onde são transformados em representações visuais compreensíveis. Assim, as etapas iniciais do processo fornecem a base essencial para a análise textual avançada, preparando os dados para serem explorados, através da visualização dos dados

4.1 EXPLORANDO O MÉTODO

No estágio inicial do processo de análise de texto com o suporte de uma taxonomia, é essencial adquirir um documento de texto relevante e adequado ao escopo do estudo. Por exemplo, se estivermos interessados na análise de textos relacionados à inteligência artificial, podemos buscar artigos científicos de revistas especializadas, relatórios técnicos de conferências ou documentos oficiais de organizações de pesquisa renomadas. A escolha das fontes deve ser criteriosa, garantindo que os documentos obtidos sejam confiáveis, atualizados e abordem temas pertinentes à área de interesse. Além disso, é fundamental que os documentos

estejam disponíveis em formato de texto legível, facilitando sua manipulação e processamento durante as etapas subsequentes da análise.

Após a seleção dos documentos de texto, a próxima etapa é garantir que eles estejam adequadamente formatados e prontos para serem processados. Isso envolve a verificação da integridade dos arquivos, a remoção de quaisquer elementos de formatação indesejados e a padronização do texto, se necessário. Por exemplo, podemos utilizar ferramentas de pré-processamento para remover caracteres especiais, pontuações excessivas e espaços em branco desnecessários. Esse processo de limpeza e preparação dos documentos de texto é crucial para garantir que os dados de entrada sejam consistentes, coesos e livres de ruídos, o que facilita análises posteriores mais precisas e significativas.

Uma vez que os documentos de texto tenham sido coletados e preparados, é possível avançar para as etapas subsequentes do processo de análise, como o processamento do texto e a construção de modelos de dados. A qualidade e a relevância dos documentos de texto selecionados desempenham um papel fundamental no sucesso global da análise, pois fornecem a base sobre a qual todas as análises posteriores serão construídas.

No estágio de processamento do texto, as técnicas de Processamento de Linguagem Natural (NLP) desempenham um papel crucial na preparação e análise dos documentos de texto. Uma das primeiras etapas é a tokenização, onde o texto é dividido em unidades linguísticas significativas, como palavras, frases ou parágrafos. Por exemplo, um algoritmo de tokenização pode dividir uma frase em uma lista de palavras individuais, removendo espaços em branco e pontuações. Esse processo permite uma análise mais granular do texto, facilitando a identificação de padrões e informações relevantes.

Além da tokenização, o processamento do texto envolve a remoção de *stopwords*, que são palavras comuns que geralmente não contribuem para o significado do texto, como "o", "a", "de" e "para". Essas palavras são frequentemente removidas para reduzir o ruído nos dados e concentrar a análise nas palavras mais importantes e significativas. Por exemplo, ao analisar um texto sobre inteligência artificial, as *stopwords* comuns podem ser removidas para destacar os termos mais relevantes, como "algoritmo", "aprendizado de máquina" e "processamento de linguagem natural".

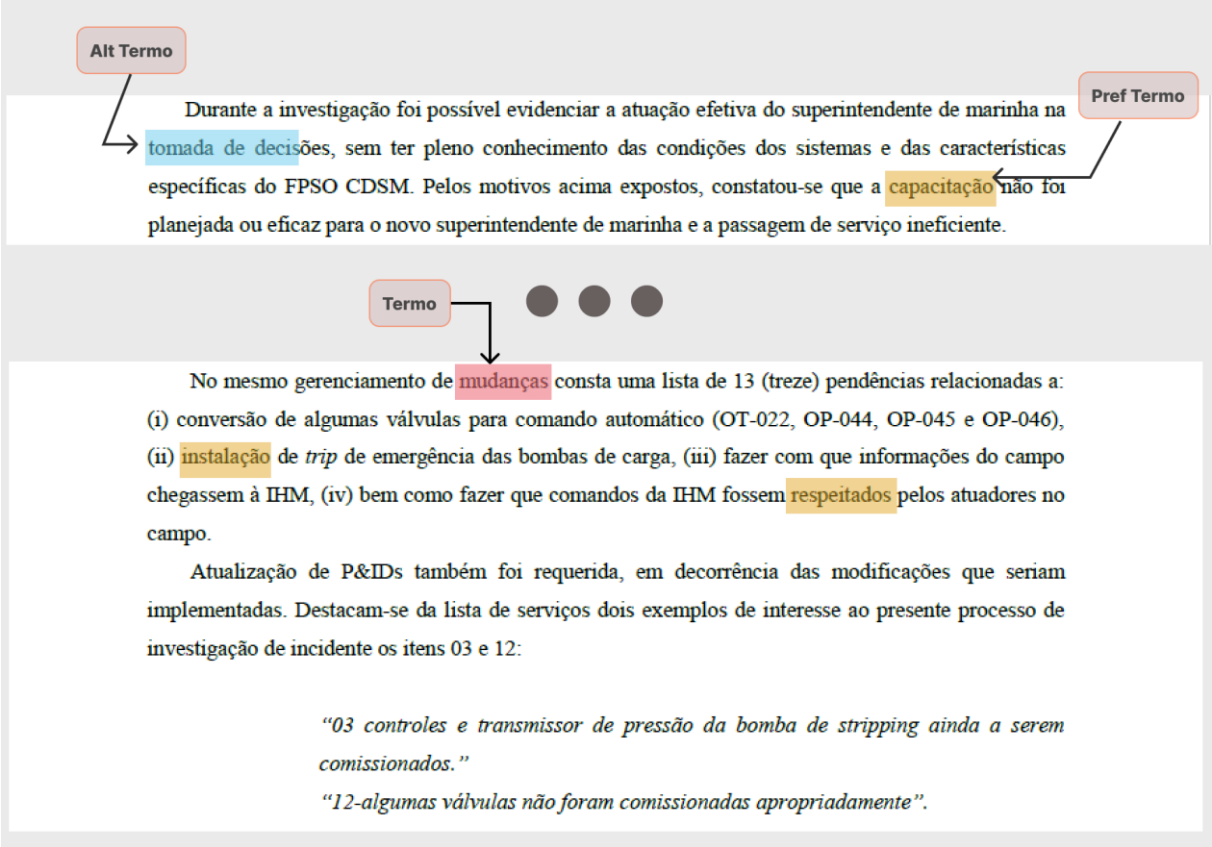
Além disso, durante o processamento do texto, são aplicadas técnicas de normalização, como a lematização e a *stemização*[10], que visam reduzir as palavras a sua forma base ou raiz. Por exemplo, as palavras "correndo", "correu" e "correrá" podem ser reduzidas ao seu lema "correr", facilitando a análise ao agrupar palavras relacionadas sob uma única forma. Essas técnicas ajudam a simplificar o texto e a capturar melhor seu significado subjacente, tornando-o mais adequado para análises subsequentes, como a modelagem de tópicos ou a extração de informações.

A etapa de Taxonomia desempenha um papel fundamental no processo, exigindo a carga de informações essenciais, como Dimensão, Nome do Fator, Pref Termo, Alt Termo e Termos. A Figura 14 demonstra trechos de texto extraídos do relatório de investigação do incidente de explosão ocorrido em 11/02/2015 no FPSO Cidade de São Mateus, desenvolvido pela Superintendência de Segurança Operacional e Meio Ambiente (SSM) em agosto de 2015. Nesta figura, são destacados no texto as informações Pref Termo, Alt Termo e Termos.

Na etapa de definição da taxonomia a ser utilizada, os itens de Fatores Humanos da camada condicionantes de performance, que conta com as três dimensões - Indivíduo, Trabalho e Organização - foram estruturados em uma taxonomia que totaliza 19 fatores.

[10] Processo de reduzir palavras flexionadas (ou às vezes derivadas) ao seu tronco (*stem*), base ou raiz, geralmente uma forma da palavra escrita.

Figura 14 - Trecho de textos com as informações em destaque



Fonte: Elaborado pela autora, com base o relatório de SSM (2015).

Cada uma destas informações são componentes para a taxonomia, o quais são melhor explorados na sequência:

- **Dimensão:** Representa as categorias ou conjuntos de categorias que organizam os fatores ou conceitos em um sistema hierárquico. Por exemplo, em um estudo sobre saúde mental, as dimensões podem abranger áreas como "Sintomas", "Tratamentos" e "Fatores de Risco".
- **Nome do Fator:** Serve como o rótulo ou identificador atribuído a um conceito específico dentro de uma dimensão. Por exemplo, dentro da dimensão "Sintomas" em um estudo de saúde mental, um fator pode ser "Depressão".
- **Pref Termo (Termo Preferido):** Refere-se à forma preferencial ou principal de um termo ou conceito, geralmente escolhido como o termo padrão para representar esse conceito. Por exemplo, para o fator "Depressão", o termo preferido pode ser simplesmente "Depressão".

- **Alt Termo (Termo Alternativo):** Consiste em uma forma alternativa ou sinônimo do termo preferido, permitindo uma maior flexibilidade na análise e busca de informações. Por exemplo, um termo alternativo para "Depressão" pode ser "Tristeza Profunda".
- **Termos:** Representam as expressões específicas que compõem os conceitos dentro de cada fator da taxonomia. Por exemplo, os termos relacionados ao fator "Depressão" podem incluir "Tristeza Persistente", "Falta de Interesse" e "Sentimentos de Desespero". Esses termos são os elementos individuais que compõem o conjunto de conceitos dentro de cada fator da taxonomia.

A presença de uma taxonomia é crucial para o método, pois oferece uma estrutura de referência organizada, permitindo uma análise mais precisa e sistemática dos documentos textuais. Ao facilitar a categorização dos termos e conceitos presentes nos documentos, a taxonomia contribui significativamente para uma compreensão mais profunda e eficaz do conteúdo, fornecendo uma base sólida para a identificação de padrões e *insights* relevantes.

Além disso, a taxonomia também ajuda na padronização e consistência da análise, garantindo que os mesmos termos sejam interpretados e utilizados de maneira consistente ao longo do processo. Isso promove uma maior confiabilidade nos resultados obtidos e facilita a comunicação e colaboração entre os membros da equipe envolvidos na análise.

Neste passo de modelo de dados, a taxonomia desempenha um papel fundamental na organização e estruturação das informações extraídas do texto processado. A taxonomia, que consiste em um conjunto hierárquico de categorias e subcategorias, fornece uma estrutura conceitual para representar os diferentes temas e conceitos abordados no documento. Por exemplo, em um contexto de análise de textos sobre saúde, a taxonomia pode incluir categorias como "doenças", "tratamentos" e "sintomas", com subcategorias específicas para cada uma delas.

Ao utilizar a taxonomia como entrada, um modelo de dados é construído para capturar e organizar as informações relevantes do texto de acordo com os conceitos definidos na taxonomia. Isso envolve mapear os termos e conceitos identificados no texto para as categorias correspondentes na taxonomia. Por exemplo, se um documento descreve diferentes tipos de doenças, o modelo de dados associaria cada doença às categorias apropriadas na taxonomia, como "cardiovascular", "respiratória" ou "neurológica".

A etapa de Armazenamento desempenha um papel crucial no processo global, uma vez que é responsável por capturar e organizar os dados necessários para suportar a etapa de visualização. Esses dados abrangem uma variedade de informações, desde a frequência dos termos presentes na taxonomia até as coocorrências entre termos e fatores. Ao armazenar essas informações de forma estruturada em tabelas específicas, o sistema é capaz de fornecer uma base sólida para análises posteriores.

Cada tabela no banco de dados desempenha um papel específico e complementar. A tabela "*study*" registra os detalhes de cada estudo realizado, fornecendo informações sobre os dados utilizados e o contexto da análise. A tabela "*dimension*" armazena as dimensões da taxonomia, proporcionando uma visão geral das categorias que organizam os conceitos. Por sua vez, a tabela "*factor*" contém os fatores associados a cada dimensão, identificando conceitos específicos dentro de cada categoria.

Além disso, as tabelas "*term*", "*paragraph*" e "*paragraph_term*" são cruciais para a análise textual detalhada. A tabela "*term*" mantém os termos vinculados a cada fator, enquanto a "*paragraph*" guarda os parágrafos extraídos dos relatórios de incidentes. Já a tabela "*paragraph_term*" estabelece o relacionamento entre parágrafos e termos, permitindo uma compreensão mais precisa da distribuição dos conceitos nos documentos textuais. Em suma, a estruturação e organização cuidadosa desses dados são fundamentais para garantir uma análise detalhada e compreensiva do conteúdo dos documentos textuais, possibilitando a geração de *insights* valiosos e informados.

Na etapa de visualização dos dados, o modelo de dados previamente preenchido com as informações relevantes do texto é o ponto de partida. Com base nesse modelo estruturado, são criadas representações visuais que permitem explorar e compreender melhor o conteúdo do texto, bem como as relações entre os conceitos presentes nele. Essas visualizações desempenham um papel crucial na interpretação e análise dos dados, fornecendo *insights* valiosos de maneira intuitiva e acessível.

As visualizações podem assumir diversas formas, dependendo do objetivo da análise e das características do texto e da taxonomia utilizada. Por exemplo, gráficos de rede podem ser empregados para representar conexões entre diferentes conceitos ou entidades, enquanto mapas de conceitos podem ser úteis para visualizar a hierarquia e as relações entre termos específicos. Além disso, diagramas de árvore podem ser empregados para ilustrar a estrutura hierárquica da taxonomia e sua relação com os conceitos extraídos do texto.

A escolha da visualização adequada é crucial para comunicar eficazmente as informações contidas no texto e facilitar a compreensão por parte dos usuários. Dessa forma, as visualizações geradas a partir do modelo de dados preenchido oferecem uma maneira poderosa e eficiente de explorar, analisar e interpretar os dados textuais, permitindo *insights* valiosos para diferentes aplicações.

A consideração dos Fatores Humanos na avaliação de desempenho é essencial para compreender e aprimorar a eficiência operacional em diversos contextos. Esta abordagem, muitas vezes, é organizada em uma taxonomia que abrange três dimensões fundamentais: Indivíduo, Trabalho e Organização. Essas dimensões são essenciais para uma compreensão abrangente dos elementos que influenciam a performance humana em ambientes de trabalho complexos.

4.2 INSTANCIAÇÃO DO MÉTODO PROPOSTO

A instanciação do método proposto é um processo crucial no desenvolvimento da solução, onde o desenvolvimento de sistemas desempenha um papel fundamental na modernização e aprimoramento das práticas de pesquisa. Esses sistemas são concebidos para manipular dados complexos, realizar análises avançadas e oferecer suporte às atividades cruciais no contexto científico. Para garantir que tais ferramentas atendam aos rigorosos padrões de confiabilidade, precisão e eficiência exigidos pela comunidade científica, é essencial adotar uma metodologia de desenvolvimento robusta e especializada.

Esse processo metodológico abrange desde a fase inicial de levantamento de requisitos até a implantação prática do sistema, considerando elementos fundamentais como validação, integração com ferramentas científicas existentes, segurança e treinamento dos usuários. A abordagem iterativa e colaborativa, associada às metodologias ágeis, se destaca como uma estratégia eficaz para garantir que o sistema evolua em consonância com as necessidades dinâmicas da pesquisa científica.

No contexto da análise de documentos de texto e visualização de informações relevantes, o sistema DOC *Analysis* emerge como uma solução abrangente. Neste capítulo, serão explorados os elementos fundamentais desse sistema, delineando seu escopo, requisitos funcionais e não funcionais, casos de uso, arquitetura, tecnologias adotadas e detalhes sobre a base de dados. Adicionalmente, são apresentadas as fases de desenvolvimento, identidade

visual, design system, protótipos e a evolução do sistema por meio de diferentes versões. Por fim, o texto aborda a fase de testes e implementação, fornecendo uma visão completa do ciclo de vida do DOC *Analysis*.

A análise do sistema DOC *Analysis* foi conduzida com precisão, seguindo as diretrizes estabelecidas para garantir sua relevância. O processo de desenvolvimento passou por diversas etapas, desde a definição do escopo até a implementação final do sistema. As fases de execução do desenvolvimento foram as seguintes:

- **Definição de Escopo e Requisitos:** O escopo do DOC *Analysis* foi delineado considerando a necessidade de proporcionar uma solução abrangente para análise de documentos de texto e visualização de tópicos relevantes. Os requisitos funcionais e não funcionais foram identificados de forma a garantir a usabilidade, segurança e desempenho desejados.
- **Desenvolvimento e Tecnologias Adotadas:** O desenvolvimento do sistema foi conduzido utilizando tecnologias como Java, PHP, HTML, CSS e JavaScript. A escolha dessas tecnologias visou garantir a portabilidade, segurança e eficiência do DOC *Analysis*. A implementação da arquitetura seguiu as melhores práticas contemporâneas, assegurando escalabilidade e responsividade.
- **Prototipagem e Design Visual**: O processo de prototipagem envolveu a criação de protótipos de baixa, média e alta fidelidade, proporcionando uma visualização clara e iterativa da interface do usuário. O design system foi desenvolvido considerando a identidade visual definida, resultando em protótipos finais que refletem efetivamente as funcionalidades e a estética desejadas.
- **Testes e Implementação:** A etapa de testes foi crucial para avaliar a funcionalidade e segurança do sistema. Protótipos de baixa fidelidade permitiram testes preliminares, enquanto protótipos de média e alta fidelidade foram utilizados para avaliações mais aprofundadas. A implementação final do sistema ocorreu após ajustes e melhorias com base nos resultados dos testes.
- **Resultados e Recomendações:** Os resultados obtidos demonstraram a eficácia do DOC *Analysis* na análise de documentos e visualização de tópicos. Recomendações para melhorias contínuas foram identificadas, destacando a

importância de atualizações regulares para atender às demandas em constante evolução.

O ciclo de vida do DOC *Analysis*, desde a concepção até a implementação, reflete um processo de desenvolvimento robusto e orientado para atender às expectativas e necessidades dos usuários finais. A abordagem iterativa adotada durante a execução permitiu adaptações contínuas, resultando em um sistema que se destaca pela sua eficiência e usabilidade. A documentação do sistema pode ser explorada através do Apêndice F.

A instanciação do método visa tornar a visão do método proposto pragmática e funcional, dividindo-se em duas etapas distintas: a parte *Back-end* e a parte *Front-end*. A implementação dessas duas etapas é fundamental para concretizar o método, sendo a parte *back-end* responsável pela lógica de processamento dos dados e análise de texto, enquanto a parte *front-end* se concentra na interface do usuário e na apresentação dos resultados. Essa abordagem permite uma implementação eficiente e coerente do método proposto, garantindo sua viabilidade e utilidade prática na análise de documentos de texto.

4.3 DESENVOLVIMENTO DO BACK-END

Esta etapa do método realiza a leitura do relatório de incidente em formato PDF e, utilizando um analisador compatível, extrai o conteúdo sem formatação, facilitando determinados processamentos necessários à população do banco de dados. De maneira geral, a etapa possui as seguintes fases:

a) Armazenamento das informações utilizadas na execução do módulo, ou seja, o nome do estudo, a localização do arquivo da taxonomia, a localização do arquivo referente ao relatório de incidentes e a data do estudo.
b) Leitura do arquivo referente ao relatório de incidente em formato PDF indicado durante a execução do módulo.
c) Extração do conteúdo do relatório para o formato textual, ou seja, do formato PDF para o formato de texto simples (*plain text*).
d) Análise do texto para a identificação de parágrafos e, considerando cada parágrafo, seu armazenamento em uma estrutura apropriada, no caso do projeto do sistema, na tabela *paragraph*.

4.3.1 Banco de dados

MySQL é um sistema de gerenciamento de banco de dados relacional de código aberto amplamente utilizado em todo o mundo. Ele é considerado um dos bancos de dados mais populares do mercado e é utilizado por grandes empresas como Facebook, Twitter e YouTube (MYSQL, 2021). O MySQL é conhecido por sua alta escalabilidade, desempenho e confiabilidade.

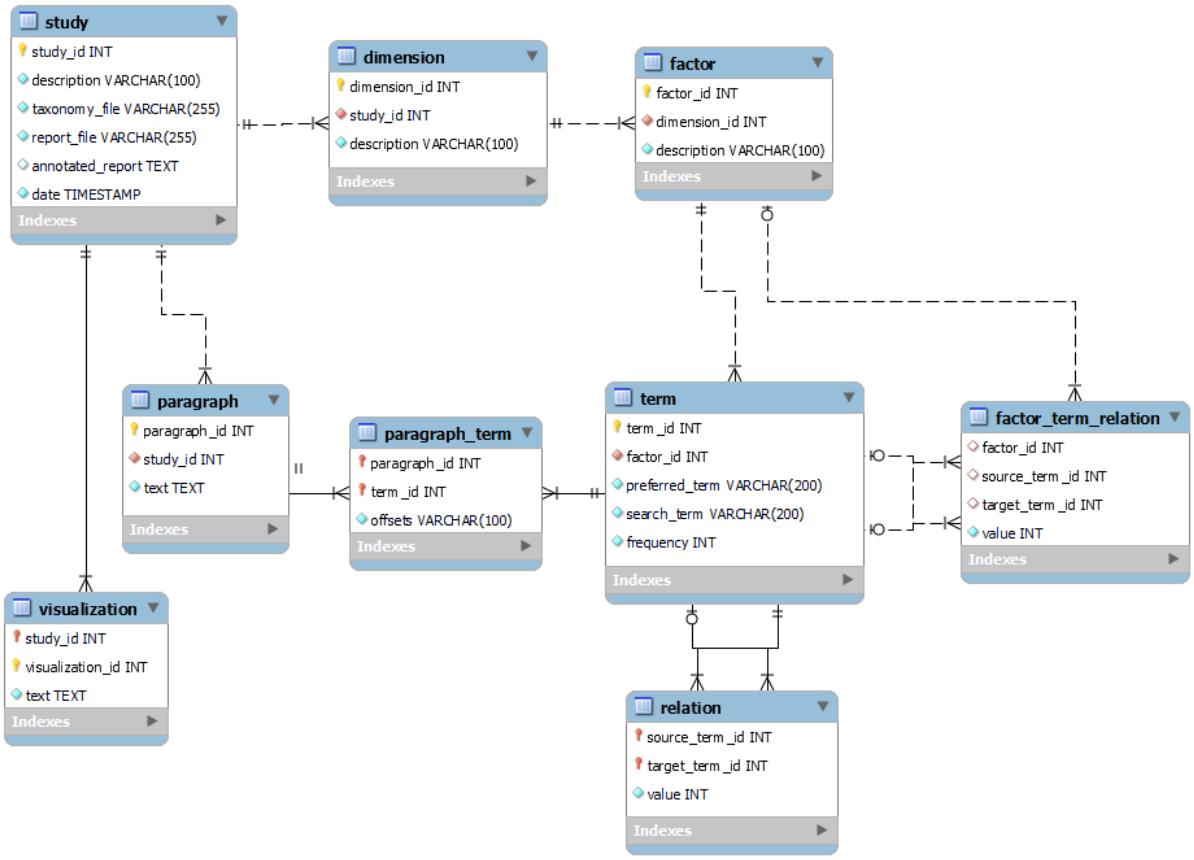
Uma das principais vantagens do MySQL é a sua capacidade de suportar grandes quantidades de dados. Ele pode lidar com milhões de registros sem afetar o desempenho do sistema (BORTOLOTTI, 2020). Além disso, o MySQL possui recursos avançados de segurança que protegem os dados armazenados no banco de dados.

Outra vantagem do MySQL é a sua capacidade de trabalhar com diferentes linguagens de programação. Ele suporta PHP, Java, Python, C++, entre outras linguagens (MYSQL, 2021). Isso permite que os desenvolvedores escolham a linguagem de programação que melhor se adapta às suas necessidades.

Por fim, o MySQL é uma ferramenta de baixo custo e fácil de usar. Ele pode ser instalado em diferentes sistemas operacionais, como Windows, Linux e Mac OS (BORTOLOTTI, 2020). Além disso, existem diversas ferramentas disponíveis para gerenciar o MySQL, como o phpMyAdmin, que permite gerenciar facilmente o banco de dados através de uma interface web.

Para suportar o módulo de *back-end* foi proposto um modelo de dados relacional (Figura 15) composto por um conjunto de tabelas (Apêndice C) que possibilitam o armazenamento dos diversos dados extraídos a partir de um relatório de incidente tendo como base uma taxonomia.

Figura 15 - Modelo de dados do módulo de análise



A seguir são apresentados os objetivos de cada uma das tabelas visando promover um entendimento do modelo como um todo:

a) *study*: permite armazenar os dados do estudo realizado, composto por uma identificação do estudo, o arquivo da taxonomia de fatores, o relatório de incidentes, o relatório anotado com os fatores e a data do estudo. De modo geral, cada vez que o módulo é executado este recebe três parâmetros, sendo, o nome do estudo, o caminho do arquivo da taxonomia e o caminho do relatório de incidentes. Neste sentido, um estudo é visto como uma execução que irá considerar determinado relatório de incidentes e os dados da taxonomia que possibilita a avaliação dos fatores com base no relatório.

b) *dimension*: armazena as dimensões que constam na taxonomia de fatores.

c) *factor*: armazena os fatores relacionados a determinada dimensão que constam na taxonomia de fatores.

d) *term*: armazena os termos vinculados a cada fator que consta na taxonomia que possui dois campos para descrever determinado termo, sendo o primeiro, o termo preferido (*preferred_term*) (será utilizado nas visualizações e análises) e o segundo, o termo de busca (search_term) (utilizado para localização do padrão no relatório de incidente). Além disso, armazena a quantidade de ocorrências de determinado termo no campo *value*.

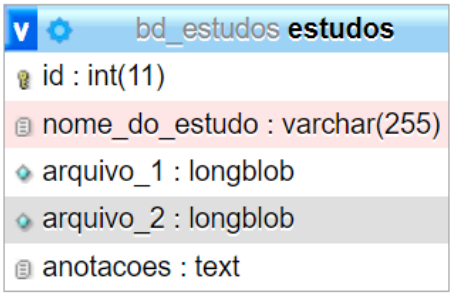
e) *paragraph*: armazena os parágrafos extraídos do relatório de incidentes e que, em processamento posterior, são unidos novamente com as anotações dos termos que ocorrem em cada parágrafo.

f) *paragraph_term*: permite armazenar o relacionamento entre determinado parágrafo e um termo guardando o(s) deslocamento(s) em que o termo ocorre no campo *offsets*, ou seja, armazena a(s) posição(ões) no parágrafo em que determinado termo foi localizado.

g) *relation*: armazena as relações entre termos independente de fator, bem como a frequência conjunta entre dois termos quaisquer nos diversos parágrafos que compõem um relatório de incidente registrando o resultado no campo *value*. Objetiva viabilizar representações em rede demonstrando como os termos se interconectam.

h) *factor_term_relation*: armazena as relações entre fatores e termos, bem como a frequência conjunta entre dois termos quaisquer nos diversos parágrafos que compõem um relatório de incidente registrando o resultado no campo *value*. Já para módulo *front-end* foi proposto um modelo de dados (Apêndice D), para armazenamento dos estudos criados, conforme representado pela Figura 16.

Figura 16 - Modelo de dados do módulo de análise da aplicação web



Fonte: Elaborado pela autora

4.3.2 Leitura e Pré-processamento do Relatório de Incidente

Os itens de Fatores Humanos da camada condicionantes de performance, que conta as três dimensões, Indivíduo, Trabalho e Organização, foi estruturado em uma taxonomia em um total de 19 fatores, conforme Figura 17.

Figura 17 - Taxonomia de fatores condicionantes de performance

Dimensão	Fator
Indivíduo	Competências técnicas
Indivíduo	Comunicação
Indivíduo	Trabalho em equipe
Indivíduo	Capacidade relacional
Indivíduo	Tomada de decisão
Indivíduo	Consciência situacional
Indivíduo	Atitude
Indivíduo	Condições psicológicas
Indivíduo	Condições fisiológicas
Indivíduo	Condições sociais
Trabalho	Design do trabalho
Trabalho	Design de interfaces
Trabalho	Condições internas
Trabalho	Condições externas
Trabalho	Liderança
Organização	Cultura de segurança
Organização	Aprendizagem
Organização	Gestão de pessoas
Organização	Gestão

Fonte: Elaborado pela autora

Para cada fator, foram definidos termos que o representam, de maneira que cada fator tenha um vetor de termos que resgatam o fator. De posse desse vetor, foram aplicadas técnicas de TM para anotar documentos. Conforme mencionado na seção 5.1, foram utilizados relatórios de acidentes, de forma a identificar os fatores latentes no documento em análise.

A Figura 19 apresenta parte da taxonomia com os fatores e respectivos termos que os representam. A inclusão dessas informações nas tabelas proporcionará uma representação mais completa e estruturada dos fatores identificados, permitindo uma interpretação mais aprofundada dos resultados obtidos. Além disso, essa abordagem facilitará a compreensão dos padrões subjacentes e contribuirá para a elaboração de estratégias preventivas mais eficazes.

Estes vetores serão utilizados posteriormente na ferramenta para análise dos documentos e para a visualização.

A apresentação desta taxonomia requer um cuidado especial, uma vez que por razões de sigilo estratégico, optamos por compartilhar apenas uma parcela representativa da taxonomia elaborada para o consórcio Libra e PUC. Essa decisão visa preservar informações sensíveis e garantir a integridade do projeto. Cabe ressaltar que a elaboração desta taxonomia foi possível graças à colaboração de diversos pesquisadores vinculados à Pontifícia Universidade Católica (PUC) e à Universidade Federal de Santa Catarina (UFSC), que desempenharam um papel fundamental no projeto, contribuindo com sua expertise e conhecimento.

A confidencialidade da taxonomia completa é essencial para preservar a propriedade intelectual e manter a competitividade do projeto. Este processo colaborativo entre instituições acadêmicas e consórcios industriais destaca a importância da sinergia entre a pesquisa acadêmica e o desenvolvimento prático, resultando em avanços significativos no entendimento e categorização de fenômenos complexos no âmbito do projeto.

Figura 18 - Fatores e termos que compõem o vetor de termos

Dimensão	Nome do Fator	Pref Termo	Alt Termo
Indivíduo	Competências técnicas	Habilidade	habilidade*
Indivíduo	Competências técnicas	Saber	saber*
Indivíduo	Competências técnicas	Experiência	experimencia*
Indivíduo	Competências técnicas	Competência	compet$nci*
Indivíduo	Competências técnicas	Experimento	experiment*
Indivíduo	Competências técnicas	Conhecimento	conheci*
Indivíduo	Competências técnicas	Prática	prátic*
Indivíduo	Competências técnicas	Percepção	percep*
Indivíduo	Competências técnicas	Vivência	vivência*
Indivíduo	Competências técnicas	Treinamento	treinamento*
Indivíduo	Competências técnicas	Capacitação	capacitaç*
Indivíduo	Competências técnicas	Capacidade	capacidad*
Indivíduo	Competências técnicas	Expertise	expertise
Indivíduo	Competências técnicas	Análise prévia	análise* prévia*
Indivíduo	Competências técnicas	Análise crítica	análise* crítica*
Indivíduo	Comunicação	Comunicação	comunic*
Indivíduo	Comunicação	Informação	inform*
Indivíduo	Comunicação	Compartilhamento	compartilh*
Indivíduo	Comunicação	Feedback	feedback
Indivíduo	Comunicação	Transmissão	transmi*
Indivíduo	Comunicação	Repertório	reperti*
Indivíduo	Comunicação	Objetividade	objetivi*
Indivíduo	Comunicação	Escuta	escut*
Indivíduo	Comunicação	Entendimento	entendim*
Indivíduo	Trabalho em equipe	Trabalho em equipe	trabalho em eq*
Indivíduo	Trabalho em equipe	Equipe	equipe*
Indivíduo	Trabalho em equipe	Time	time*
Indivíduo	Trabalho em equipe	Colega	coleg*
Indivíduo	Trabalho em equipe	Grupo	grup*
Indivíduo	Trabalho em equipe	Confiança	confianç*
Indivíduo	Trabalho em equipe	Conflito	conflit*
Indivíduo	Trabalho em equipe	Cooperação	cooperaç*
Indivíduo	Trabalho em equipe	Proativo	proativ*

Fonte: Elaborado pela autora

Como pode ser visto na Figura 18 o fator “Competências técnicas” da dimensão “Indivíduo” tem um vetor composto por 15 termos (entre eles Habilidade, Saber, Experiência e Análise crítica). Na coluna “Pref Termo” está o termo escrito em uma de suas variações e na coluna “Alt Termo” está o termo como será buscado no texto. O uso do símbolo “*” objetva permitir a busca pelo termo e suas variações. Por exemplo, “transmi*” retornarão as variações desse termo como “transmissão”, “transmissões”, “transmitir”, “transmitido”, entre outras.

Além disso, o desenvolvimento do *back-end* de TM separou o texto em parágrafos de maneira a contabilizar a frequência de ocorrências de determinado fator ou termo no parágrafo e a coocorrência dos fatores e termos. Dessa forma, é possível com a identificação dessas frequências apresentar os resultados da análise de um documento de maneira gráfica no front-end. O *front-end* é responsável em apresentar de maneira visual, gráficos, nuvens de termos, grafos e outras maneiras de modo a apoiar o usuário a analisar o acidente a partir do seu relato

ou outros documentos de seu interesse. Ou seja, identificar os fatores latentes na ocorrência de um acidente ou incidente ou mesmo documentos diversos.

4.3.3 Geração da Frequência de Termos dos Fatores

Esta etapa do módulo tem o intuito de identificar, no texto de determinado relatório de incidente vinculado a um estudo em particular, um termo da taxonomia e a frequência com que este termo é mencionado. De maneira geral, a etapa possui os seguintes passos:

a) Criação de uma lista com os termos de busca (campo *search_term*).

b) A partir de cada instância do termo é identificada a quantidade de vezes que o termo aparece em determinado parágrafo. Para tal, é aplicado um algoritmo de alinhamento de *strings*, em que, tanto o parágrafo quanto o termo de busca (pode ser composto por mais de uma palavra), são transformados em vetores. Após isso, o termo de busca é alinhado em relação ao vetor com o objetivo de identificar em quantos pontos existe uma similaridade entre ambos os vetores. Estes pontos representam a posição em que o termo ocorre em determinado parágrafo gerando um vetor de deslocamentos sendo então armazenados no campo *offsets* da tabela *paragraph_term*.

c) Utilizando o vetor de deslocamento tem-se a quantidade de vezes que determinado termo foi identificado em um parágrafo do relatório de incidente, sendo este dado armazenado na tabela *term*, coluna *frequency*.

Analisando o resultado do conteúdo armazenado na tabela *term* são vislumbradas algumas possibilidades de consumo dos dados. Individualmente, formas de visualização suportadas pelas descrições dos termos (*preferred_term*) e suas frequências podem ser utilizadas, por exemplo, através de nuvens de palavras (neste caso, os termos da taxonomia) e histogramas com os termos ou fatores que mais ocorrem em determinado estudo. Levando-se em conta dois ou mais estudos, diferentes visualizações podem ser empregadas com o intuito de proporcionar análises comparativas.

Referente à tabela *paragraph_term*, que também é populada nesta etapa, a possibilidade de uso de seus dados.

4.3.4 Geração de Coocorrências entre os Termos dos Fatores

Nesta etapa do módulo são produzidos os dados que preenchem a tabela *relation*. De modo geral, o processo ocorre por meio da obtenção da coocorrência de dois termos quaisquer de maneira distinta a partir do texto do relatório de incidente em um determinado estudo. Para tal, um parâmetro importante é a janela em que estes dois termos serão analisados, ou seja, a quantidade de palavras que são necessárias para conectar os dois termos. Se um par de termos estiver contido na janela a ocorrência é computada. Ao final todos os pares termos que possuírem coocorrência maior ou igual a 1 (um) são armazenados no banco de dados. A Figura 19, demonstra um parágrafo de texto, a coocorrência de palavras relacionadas ao tema de acidentes rodoviários é evidente. Palavras como "acidente", "rodoviário", "velocidade", "distração", "condutor", "socorristas", "assistência médica", "análise", "dados", "padrões recorrentes", "medidas de prevenção", "segurança viária" e outras estão interconectadas, formando um contexto coeso sobre o tema. A coocorrência dessas palavras reflete a inter-relação semântica entre os elementos do texto, contribuindo para a compreensão global do tópico abordado.

Figura 19 - Exemplo de coocorrência de palavras

"No trágico cenário de um acidente rodoviário, a combinação de fatores como alta velocidade, distração do condutor e más condições climáticas frequentemente contribui para colisões devastadoras. Os socorristas rapidamente respondem à cena do acidente, prestando assistência médica e tentando restabelecer a ordem. A análise dos dados do acidente revela padrões recorrentes, fornecendo insights valiosos para medidas de prevenção. A segurança viária depende da conscientização coletiva, regulação eficaz e tecnologias inovadoras para mitigar os riscos e garantir estradas mais seguras para todos."

Fonte: Elaborado pela autora

No contexto do processo descrito, são delineadas cinco etapas fundamentais que compõem a análise e identificação de termos em documentos textuais. Os seguintes passos são executados:

a) Criação da lista de termos a partir da taxonomia.

b) Definição da janela para análise da coocorrência entre os termos.

c) A partir de um termo de origem (*source*), é efetuada a localização dos parágrafos que mencionam o termo.

d) Considerando os termos seguintes na lista, é verificado em cada parágrafo se determinado termo de destino (*target*) também existe e se o limite da janela definido no módulo é respeitado, ou seja, se a distância em palavras entre o termo de origem e o termo de destino é menor ou igual ao tamanho da janela. Em caso afirmativo, a coocorrência entre os dois termos é adicionada.

e) Persistência de todos os pares de termos em que a coocorrência for maior do que 0 (zero) na tabela *relation*.

As etapas delineadas fornecem um roteiro claro e sistemático para a análise e identificação de termos em documentos textuais, abrangendo desde a criação da lista de termos até a persistência dos pares de termos em uma tabela específica. Essa abordagem estruturada permite uma análise abrangente e eficaz da coocorrência entre os termos, proporcionando *insights* valiosos para a compreensão e organização do conteúdo textual.

4.3.5 Geração de Coocorrências entre Fatores e Termos

Nesta etapa do módulo são produzidos os dados que preenchem a tabela *factor_trem_relation*. De modo geral, o processo inicia pela ligação de todos os termos de determinado fator que possuem frequência superior a 0 (zero). Após isso são analisados pares de termos que possuem coocorrência para possibilitar a geração dos dados de relacionamento. Neste sentido, os seguintes passos são executados:

a) Geração das relações entre fatores e termos. Para tal, utilizam-se os dados previamente gerados e armazenadas na tabela *term* em que a frequência de determinado termo seja superior a 1 (um). Cada entrada na tabela será composta pelo código do fator, pelo código do termo que representará o termo de origem (*source_term_id*) e o valor, obtido a partir da frequência do termo. Neste momento, caso fosse realizada a visualização existiriam vários grafos (redes) conectando o fator, que se encontraria ao centro, com seus termos nas extremidades.

b) Leitura de pares de termos de fatores distintos, ou seja, cada um dos termos necessários para formar uma relação devem estar conectados a diferentes fatores. Este dado encontra-se disponível na tabela *relation*, sendo necessária então a sua localização e obtenção da coocorrência.

c) Persistência de todos os pares de termos com suas respectivas coocorrências no campo *value*.

A etapa de produção de dados para a tabela *factor_term_relation* é essencial para o processo de análise e relacionamento entre os termos e fatores. Inicialmente, são estabelecidas as relações entre os fatores e os termos, com base na frequência dos termos em cada fator. Posteriormente, são analisados os pares de termos provenientes de fatores distintos, garantindo a diversidade e abrangência das relações. Por fim, os pares de termos são persistidos na tabela, juntamente com suas coocorrências, permitindo uma representação visual clara e detalhada das conexões entre fatores e termos. Essa etapa é crucial para fornecer *insights* valiosos sobre a inter-relação dos elementos dentro do contexto estudado, contribuindo significativamente para uma análise mais profunda e abrangente dos dados.

4.4 DESENVOLVIMENTO DO *FRONT-END*

O processo de desenvolvimento do sistema DOC *Analysis* foi conduzido de forma abrangente, seguindo uma abordagem iterativa e colaborativa. O ciclo de desenvolvimento envolveu diversas etapas, desde a concepção da ideia até a implementação final, com um foco especial no desenvolvimento do *front-end*, que desempenha um papel fundamental na experiência do usuário.

A primeira etapa do desenvolvimento do *front-end* foi dedicada à definição da identidade visual do sistema. Isso incluiu a criação de elementos gráficos, a escolha de cores, tipografia e ícones que representassem a marca DOC *Analysis* de maneira consistente e atraente. A definição da identidade visual ocorreu em estreita colaboração com os stakeholders para garantir que refletisse os valores da empresa e proporcionasse uma experiência visualmente agradável aos usuários.

Protótipo de Baixa Fidelidade Com a identidade visual estabelecida, avançou-se para a etapa de prototipagem de baixa fidelidade. Nessa etapa, foram criados esboços e *wireframes* para visualizar a estrutura geral do *front-end*. Esses protótipos permitiram avaliar a disposição dos elementos na interface e testar a usabilidade de maneira rápida e econômica.

Protótipo de Média e Alta FidelidadeA terceira etapa consistiu no desenvolvimento de protótipos de média e alta fidelidade. Isso envolveu a transformação dos *wireframes* em designs mais detalhados, considerando interações, animações e responsividade. Essa etapa foi

essencial para refinar o design, garantir uma melhor experiência do usuário e adaptar a interface para diferentes dispositivos e tamanhos de tela.

Ao longo dessas etapas, a documentação foi uma parte integrante do processo. Todo o código *front-end* foi documentado de maneira abrangente, e as decisões de *design* foram registradas para facilitar futuras manutenções e expansões. A documentação completa está disponível no Apêndice F, proporcionando uma referência detalhada para facilitar a compreensão e o desenvolvimento contínuo do sistema DOC *Analysis*.

5 AVALIAÇÃO E DISCUSSÃO DOS RESULTADOS

Neste capítulo é realizada uma análise detalhada dos resultados obtidos a partir da aplicação do sistema DOC *Analysis* projetado com base no método proposto. Este capítulo visa avaliar a eficácia e o desempenho do sistema na análise de documentos de texto e na visualização de informações relevantes. A avaliação abrange diversos aspectos, a qualidade das visualizações geradas e a comparação com métodos tradicionais de análise de documentos. Além disso, são discutidas as principais descobertas e conclusões derivadas da aplicação do sistema, oferecendo *insights* sobre sua utilidade e aplicabilidade em cenários práticos de análise de texto.

5.1 APRESENTAÇÃO DO CENÁRIO DE ESTUDO

A mitigação dos problemas de fatores humanos em plataformas *offshore*, como acidentes e incidentes de trabalho, é um desafio crucial que exige abordagens eficazes para análise e prevenção. Nesse contexto, o método proposto oferece uma estrutura sistemática para lidar com esses problemas, fornecendo uma visão abrangente e organizada do cenário. Ao adotar uma abordagem sistemática, é possível identificar os principais pontos de falha, compreender suas causas subjacentes e implementar medidas preventivas que visem reduzir o risco de ocorrência de acidentes e incidentes.

Para isso, é fundamental empregar uma taxonomia específica que possibilite a categorização e análise detalhada dos eventos ocorridos. Uma taxonomia bem desenvolvida permite a classificação dos incidentes de acordo com sua natureza, gravidade, causas imediatas e raízes subjacentes. Isso facilita a identificação de padrões recorrentes e tendências, fornecendo *insights* valiosos para a formulação de estratégias preventivas mais eficazes. Além disso, a análise de documentos textuais, como relatórios de incidentes passados, pode oferecer uma rica fonte de informações para entender os contextos específicos em que as falhas humanas ocorrem e os fatores que as influenciam.

Um exemplo de documento que pode ser utilizado para análise é o relatório de investigação de incidentes elaborado após um acidente em uma plataforma *offshore*. Esse tipo de relatório geralmente contém uma descrição detalhada do evento, incluindo fatores como condições operacionais, ações dos trabalhadores envolvidos, falhas de equipamentos e

quaisquer outras circunstâncias relevantes. Ao examinar cuidadosamente esses relatórios e aplicar uma taxonomia adequada, é possível identificar padrões comportamentais, lacunas de treinamento, deficiências de procedimentos e outras áreas que requerem atenção para evitar incidentes semelhantes no futuro. Assim, a combinação de uma abordagem sistemática, uma taxonomia bem definida e a análise de documentos textuais pode fornecer uma base sólida para a prevenção de falhas humanas em plataformas *offshore*.

O relatório de investigação do incidente de explosão ocorrido em 11/02/2015 no FPSO Cidade de São Mateus, desenvolvido pela Superintendência de Segurança Operacional e Meio Ambiente (SSM) em agosto de 2015, oferece uma análise detalhada dos eventos que levaram à tragédia (SSM, 2015). O incidente teve início durante uma tentativa de drenagem de resíduo líquido do tanque de carga central número 6 (6C), resultando em vazamento de condensado dentro da casa de bombas da plataforma. Apesar dos alarmes dos detectores de gás fixos, várias equipes foram enviadas ao local, mesmo com a confirmação da presença de atmosfera explosiva (SSM, 2015).

O relatório destaca uma série de falhas críticas, desde a gestão de segurança até a execução de procedimentos de resposta à emergência. A falta de liderança das equipes de brigada, a inadequação dos procedimentos de resposta e a ausência de previsão de cenários identificados em estudos de risco foram alguns dos principais pontos levantados (SSM, 2015). Além disso, a análise revelou uma série de decisões gerenciais ao longo do ciclo de vida da plataforma que introduziram riscos não gerenciados, culminando nas condições necessárias para o acidente (SSM, 2015).

Por fim, o relatório oferece uma série de recomendações destinadas a toda a indústria *offshore* de petróleo e gás natural, visando evitar a recorrência de acidentes semelhantes (SSM, 2015). Essas recomendações são consideradas mandatórias e destacam a importância de uma abordagem holística para a segurança operacional, desde a fase de projeto até a operação e manutenção das instalações *offshore*. O incidente serve como um lembrete dos riscos envolvidos na operação *offshore* e da necessidade contínua de vigilância e melhoria dos padrões de segurança.

Para abordar esses desafios complexos, é essencial entender a natureza das falhas humanas e desenvolver estratégias eficazes para reduzir sua ocorrência. A taxonomia, ou classificação sistemática, das falhas humanas oferece uma estrutura analítica que pode ajudar a identificar e compreender as causas subjacentes dessas falhas. O Quadro 11, apresenta a descrição de cada nome do fator conforme a sua dimensão. As dimensões individuo, trabalho e

organização representam pilares fundamentais na gestão eficaz de recursos humanos e na promoção de um ambiente de trabalho seguro e produtivo. A dimensão indivíduo abrange as competências, habilidades e condições pessoais dos colaboradores, como sua capacidade de comunicação, trabalho em equipe e saúde física e mental. Já a dimensão trabalho diz respeito à organização e estrutura das tarefas, ao design do ambiente de trabalho e às condições internas e externas que impactam diretamente à execução das atividades laborais. Por fim, a dimensão organização engloba a cultura, valores, práticas de segurança e gestão de pessoas adotadas pela empresa, refletindo na forma como os colaboradores interagem e se desenvolvem dentro do contexto organizacional. O equilíbrio e integração entre essas dimensões são essenciais para garantir a eficiência operacional e o bem-estar dos trabalhadores.

Quadro 11 – Descrição do Nome do Fator

Indivíduo	
Competências técnicas	Refere-se ao conjunto de habilidades e conhecimentos técnicos que um indivíduo possui para desempenhar suas funções. São fundamentais para garantir um desempenho eficaz e consistente nas tarefas atribuídas.
Comunicação	Envolve a capacidade de transmitir e receber informações de forma clara e eficaz. Uma comunicação eficiente é essencial para evitar mal-entendidos e promover a colaboração dentro de uma equipe.
Trabalho em equipe	Diz respeito à habilidade de colaborar e interagir harmoniosamente com colegas de trabalho para alcançar objetivos comuns. A capacidade de trabalhar em equipe é crucial para o sucesso de projetos e para a criação de um ambiente de trabalho produtivo.
Capacidade relacional	Refere-se à habilidade de estabelecer e manter relacionamentos interpessoais positivos. Relacionamentos saudáveis no ambiente de trabalho contribuem para um clima organizacional positivo e para o bem-estar dos colaboradores.
Tomada de decisão	Envolve a habilidade de analisar informações e escolher a melhor alternativa entre várias opções. Uma tomada de decisão eficaz é essencial para resolver problemas e alcançar objetivos de forma eficiente.
Consciência situacional	Refere-se à capacidade de compreender e reagir adequadamente às condições e eventos do ambiente de trabalho. Uma boa consciência situacional permite aos indivíduos anteciparem problemas e agirem proativamente para evitá-los.

Atitude	Diz respeito à postura mental e emocional de um indivíduo em relação ao trabalho e às circunstâncias. Uma atitude positiva pode influenciar significativamente o desempenho e a satisfação no trabalho.
Condições psicológicas	Envolve o estado mental e emocional de um indivíduo, incluindo fatores como estresse, motivação e bem-estar. O cuidado com as condições psicológicas dos colaboradores é importante para garantir um ambiente de trabalho saudável e produtivo.
Condições fisiológicas	Refere-se ao estado físico e saúde de um indivíduo, incluindo fatores como fadiga, sono e saúde geral. As condições fisiológicas dos colaboradores podem impactar diretamente sua capacidade de desempenhar suas funções de maneira eficaz e segura.
Condições sociais	Diz respeito ao ambiente social e relacional no qual um indivíduo está inserido, incluindo fatores como cultura organizacional e relacionamentos interpessoais. Um ambiente social positivo promove a colaboração, a motivação e o bem-estar no trabalho.
Trabalho	
Design do trabalho	Refere-se à organização e estrutura das tarefas e responsabilidades atribuídas aos indivíduos. Um bom design do trabalho pode aumentar a eficiência e a satisfação no trabalho.
Design de interfaces	Envolve o desenvolvimento e a usabilidade de sistemas e interfaces com os quais os trabalhadores interagem. Interfaces bem projetadas facilitam a realização de tarefas e contribuem para a eficácia do trabalho.
Condições internas	Diz respeito ao ambiente físico e ergonômico dentro do local de trabalho. Condições internas adequadas são essenciais para garantir o conforto e a segurança dos colaboradores durante a execução de suas atividades.
Condições externas	Refere-se ao ambiente externo ao local de trabalho que pode afetar o desempenho e a segurança dos trabalhadores. Fatores externos, como clima e transporte, podem influenciar diretamente a produtividade e o bem-estar dos colaboradores.
Liderança	Envolve a habilidade dos líderes em orientar, motivar e influenciar positivamente suas equipes. Uma liderança eficaz é fundamental para inspirar confiança, promover o engajamento e alcançar os objetivos organizacionais.
Organização	
Cultura de segurança	Refere-se às normas, valores e práticas organizacionais relacionadas à segurança no trabalho. Uma cultura de segurança forte é essencial para prevenir acidentes e promover um ambiente de trabalho saudável e seguro.

Aprendizagem	Envolve os processos e práticas organizacionais relacionadas ao desenvolvimento e aquisição de conhecimento. Investir em programas de aprendizagem contínua é fundamental para manter os colaboradores atualizados e competitivos no mercado de trabalho.
Gestão de pessoas	Diz respeito às estratégias e práticas adotadas pela organização para gerenciar e desenvolver seus recursos humanos. Uma boa gestão de pessoas promove o desenvolvimento profissional, a motivação e o engajamento dos colaboradores.
Gestão	Refere-se às atividades e processos relacionados à administração e governança da organização como um todo. Uma gestão eficaz é fundamental para garantir o sucesso e a sustentabilidade do negócio a longo prazo.

Fonte: Elaborado pela autora

5.2 UTILIZAÇÃO DO SISTEMA

A DOC *Analysis* segue um conjunto de funcionalidades conforme Figura 20. Para se criar um estudo, primeiro deve-se criar um nome para o estudo, na sequência selecionar o arquivo de texto e em seguida a taxonomia nos arquivos do computador e clicar no botão 'Executar estudo', assim o sistema irá realizar a análise do texto conforme a taxonomia, sendo apresentado o Resultado do estudo, em que o usuário pode explorar os gráficos.

Figura 20 - Conjunto de funcionalidades

Fonte: Elaborado pela autora

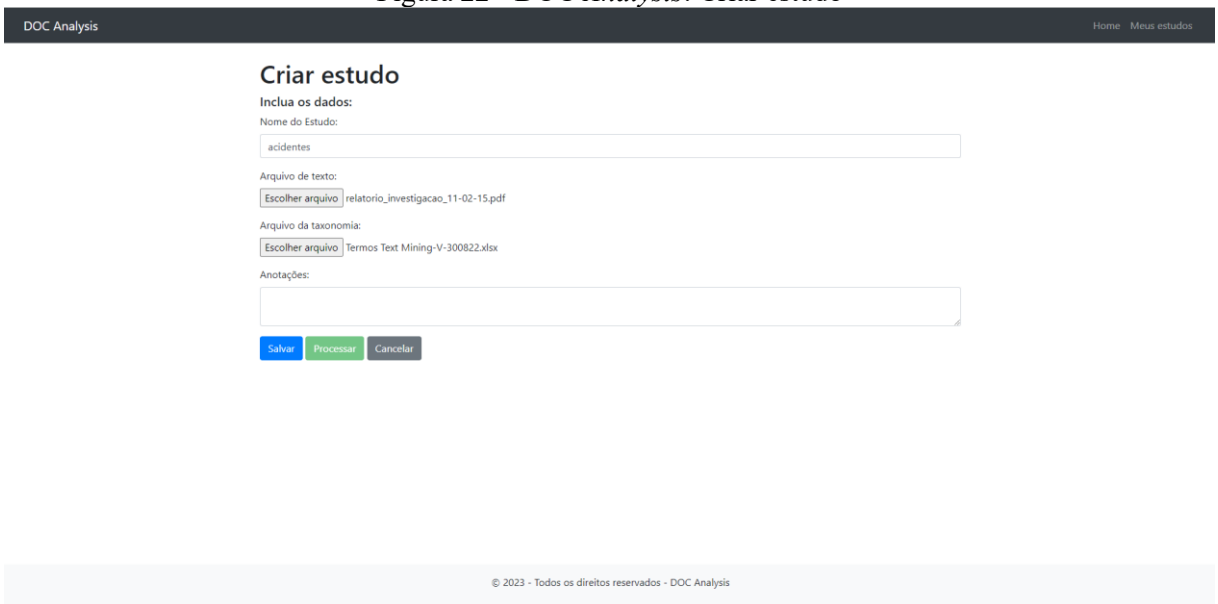
A Figura 21, demonstra a página principal do sistema. Nesta página o usuário pode criar um estudo ou acessar todos os estudos já criados (Meus estudos).

Figura 21 - DOC *Analysis*: Página inicial

Fonte: Elaborado pela autora

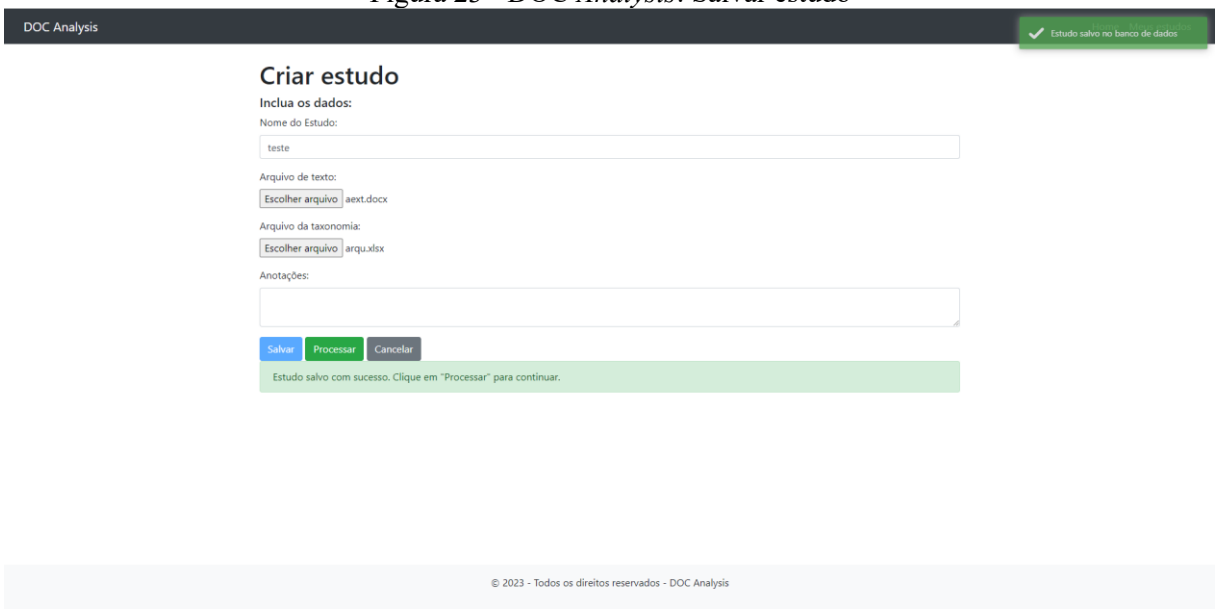
Ao acessar o botão ‘Criar estudo’, o usuário será redirecionado para a respectiva página (Figura 22), na qual deverá nomear o estudo e selecionar os arquivos para upload (arquivo de texto e taxonomia).

Figura 22 - DOC *Analysis*: Criar estudo



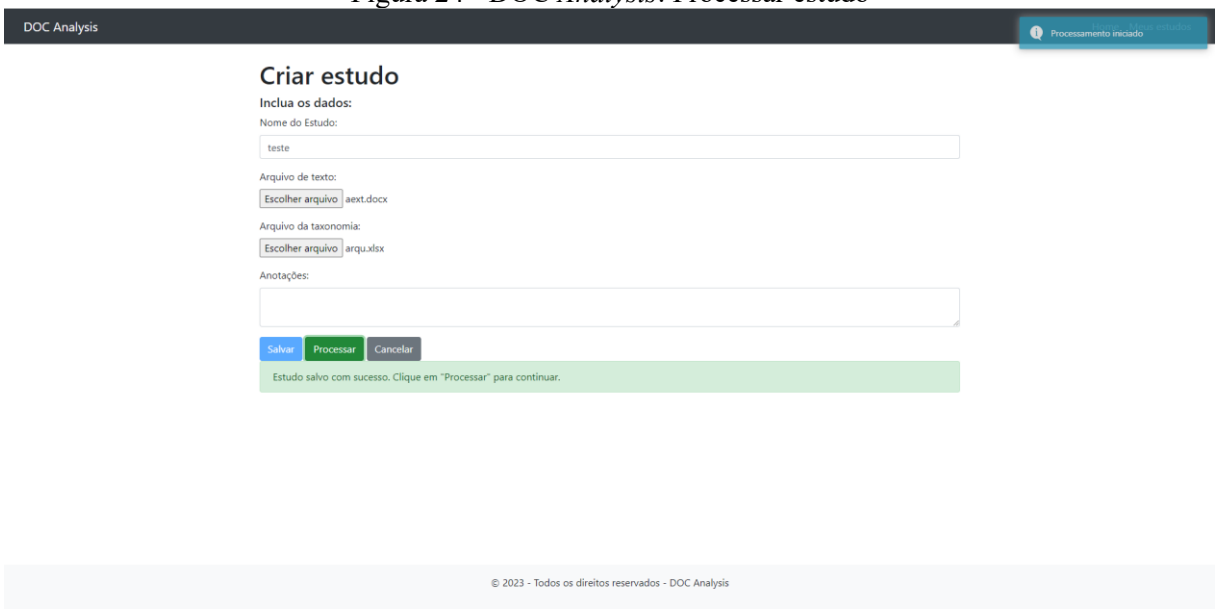
Ao clicar em ‘Salvar’ o estudo é salvo no banco de dados (Figura 23). o módulo *back-end* irá analisar o texto.

Figura 23 - DOC *Analysis*: Salvar estudo



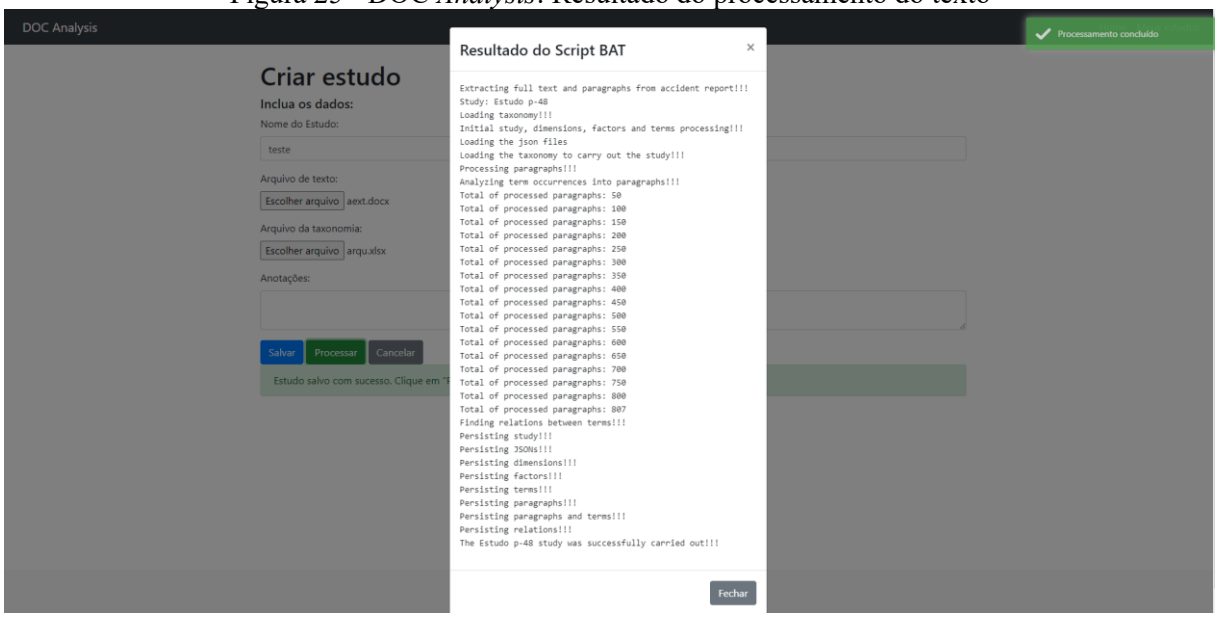
Uma vez salvo o estudo no banco de dados, o usuário de clicar no botão ‘Processar’ (Figura 24), o sistema aciona o módulo *back-end* para iniciar a análise do documento de texto.

Figura 24 - DOC *Analysis*: Processar estudo



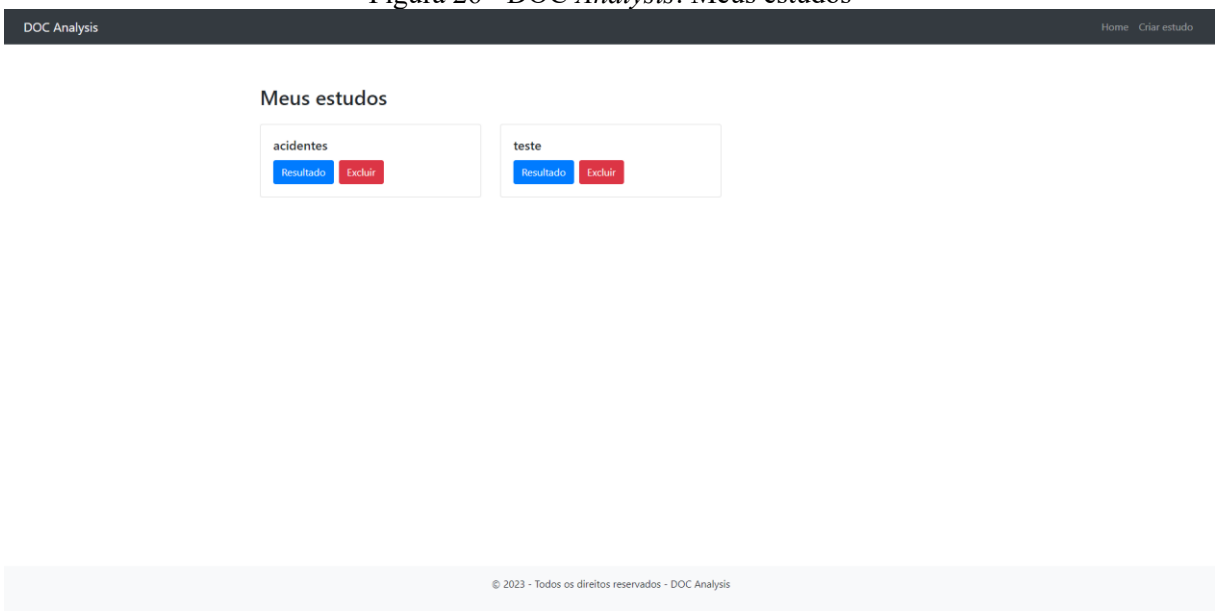
Finalizado o processamento da análise do documento salvo, o resultado do processamento é exibido (Figura 25). Quando o usuário utiliza o botão ‘Fechar’, a página é redirecionada para Meus estudos.

Figura 25 - DOC *Analysis*: Resultado do processamento do texto



Ao acessar ‘Meus estudos’, o usuário poderá visualizar a página (Figura 26) contendo todos os estudos já criados, podendo acessar o resultado ou excluindo o estudo.

Figura 26 - DOC *Analysis*: Meus estudos



A facilidade na criação de estudos, a execução eficiente da análise com base em taxonomias e a visualização clara dos resultados por meio de gráficos contribuem para a usabilidade do sistema.

O DOC *Analysis* oferece uma solução valiosa para a análise de texto, destacando-se não apenas por sua eficiência técnica, mas também por sua interface amigável, tornando a tarefa de análise textual mais acessível e eficaz para uma variedade de usuários.

O DOC *Analysis* oferece uma solução valiosa para a análise de texto, destacando-se não apenas por sua eficiência técnica, mas também por sua interface amigável, tornando a tarefa de análise textual mais acessível e eficaz para uma variedade de usuários.

5.3 FUNCIONALIDADES DE GRÁFICOS UTILIZADOS

A visualização de dados desempenha um papel crucial na compreensão e interpretação de informações complexas. Gráficos são ferramentas eficazes que transformam dados brutos em representações visuais acessíveis e significativas, facilitando a identificação de padrões, tendências e *insights* valiosos. Segundo Tufte (2001), um renomado especialista em visualização de dados, a visualização de dados é uma forma de comunicação, cuja ação principal de qualquer gráfico é revelar dados de maneira mais eficiente e clara.

Ao explorar diferentes tipos de gráficos, desde os tradicionais como barras horizontais e árvores de palavras até técnicas mais avançadas como *Zoomable Circle Packing*, é crucial reconhecer a importância da escolha apropriada da representação visual. Como expresso por Few (2009), gráficos eficazes mostram as verdades dos dados, não a verdade do gráfico. Neste contexto, examina-se a aplicação de cada tipo de gráfico em diversas situações, destacando sua utilidade e eficácia na transmissão de informações complexas de maneira clara.

Ao longo deste estudo, é essencial considerar a orientação de especialistas em visualização de dados para garantir que as representações escolhidas maximizem a clareza, promovam a compreensão e conduzam a interpretações precisas. Conforme discutido por Cairo (2013), a visualização eficaz é aquela que é clara, precisa e eficiente na comunicação da informação. Os gráficos que foram utilizados neste projeto foram:

- Árvore de Palavras (*Word Tree*): um gráfico que exibe palavras organizadas hierarquicamente, conectando-as com linhas para representar relações entre elas.
- Barras Horizontais (*Horizontal Bar Chart*): um gráfico de barras onde as barras são dispostas horizontalmente. Cada barra representa uma categoria e o comprimento da barra é proporcional ao valor que ela representa.

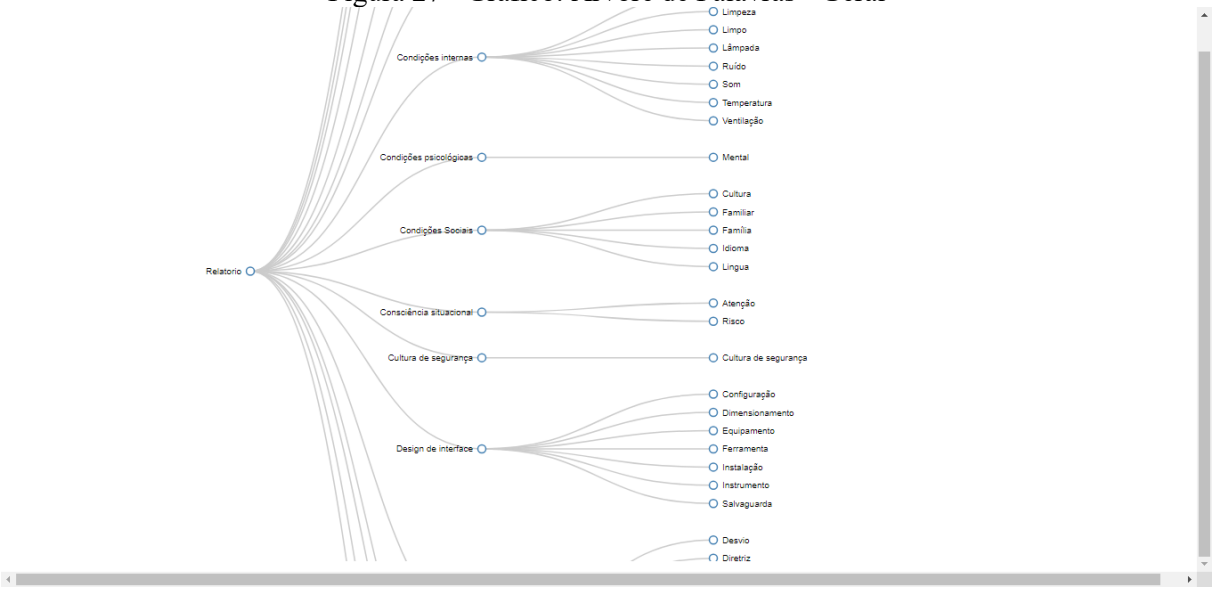
- Barras Horizontais Interativas (*Interactive Horizontal Bar Chart*): uma versão interativa do gráfico de barras horizontais, permitindo que os usuários interajam com os dados, como filtrar ou realçar categorias específicas.
- Histograma Horizontal (*Horizontal Histogram*): um gráfico que apresenta a distribuição de frequência de dados em intervalos contínuos. As barras são dispostas horizontalmente, onde cada barra representa um intervalo e o comprimento da barra indica a frequência ou a densidade de dados nesse intervalo. É uma ferramenta eficaz para visualizar a distribuição de dados e identificar padrões, tendências e outliers em conjuntos de dados contínuos.
- *Treemap*: um gráfico que exibe hierarquias de dados como retângulos aninhados. Cada retângulo representa uma categoria e seu tamanho é proporcional ao valor que ela representa.
- *Zoomable Circle Packing*: uma técnica de visualização que exibe hierarquias de dados como círculos aninhados. Os círculos são interativos, permitindo que os usuários ampliem ou reduzam para explorar diferentes níveis da hierarquia.
- *Word Clouds* (Nuvens de Palavras): uma representação visual de palavras onde o tamanho das palavras é proporcional à sua frequência. Palavras mais frequentes são exibidas em tamanho maior.
- Gráfico de Bolhas (*Bubble Chart*): um gráfico que utiliza círculos para representar dados. O tamanho do círculo indica o valor da variável, e a posição no gráfico pode representar outras dimensões.

A Figura 27, apresenta o resultado usando o gráfico de árvore de palavras, mostrando o resultado geral, ou seja, as dimensões, fatores e termos que ocorreram no documento em análise. Por meio do gráfico de árvore de palavras, oferece uma visão abrangente dos fatores e termos presentes no documento em análise, conforme a taxonomia estabelecida. A estrutura da árvore revela a organização hierárquica dos elementos identificados, começando pelos fatores gerais e se desdobrando em termos mais específicos.

No nível mais alto da árvore, encontramos os fatores que influenciam a performance humana, como competências técnicas, comunicação, trabalho em equipe, entre outros. Esses fatores representam dimensões fundamentais para entender o desempenho dos indivíduos em um ambiente de trabalho.

Ao nos aprofundarmos na árvore, observamos os pref termos, que são subcategorias dentro de cada fator. Por exemplo, dentro do fator "Indivíduo", temos subcategorias como "Competências técnicas", "Comunicação" e "Atitude". Esses pref termos representam aspectos específicos relacionados ao indivíduo que podem afetar sua performance no trabalho.

Figura 27 - Gráfico: Árvore de Palavras - Geral

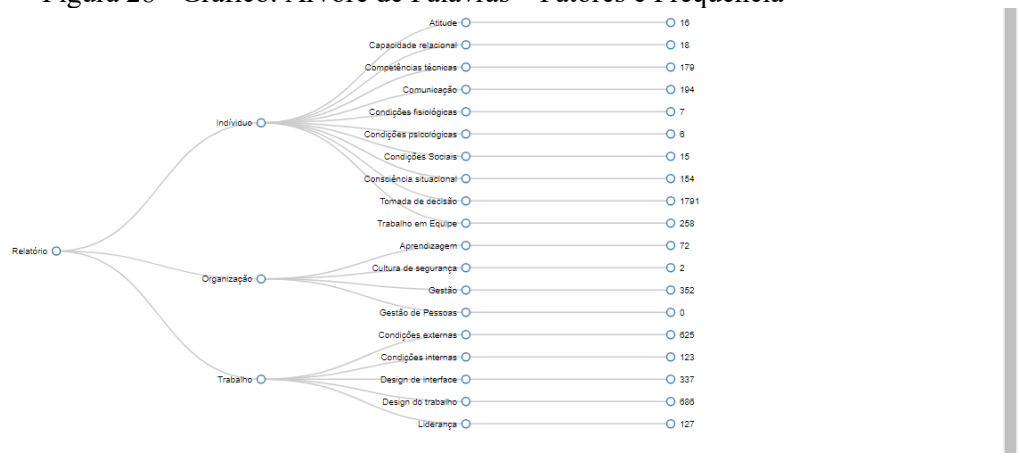


Fonte: Elaborado pela autora

A segunda árvore de palavras, representada na Figura 28, proporciona uma visão mais detalhada das dimensões, fatores e suas respectivas frequências de ocorrência no documento em análise. Nesse contexto, as dimensões referem-se aos diferentes aspectos ou categorias que estão sendo analisados, enquanto os fatores são os termos específicos que compõem cada dimensão. A frequência de ocorrência indica com que frequência cada termo ou fator aparece no texto do documento. Além disso, a frequência de ocorrência de cada fator fornece uma indicação da sua importância relativa no contexto do documento.

A análise dessas árvores de palavras pode ajudar na identificação de padrões, tendências e tópicos-chave presentes no documento, auxiliando na tomada de decisões informadas e na extração de conhecimento significativo. Essa abordagem visual é especialmente útil para resumir e sintetizar grandes volumes de texto, tornando mais fácil para os pesquisadores e profissionais identificarem informações relevantes e tirarem conclusões embasadas a partir do documento analisado.

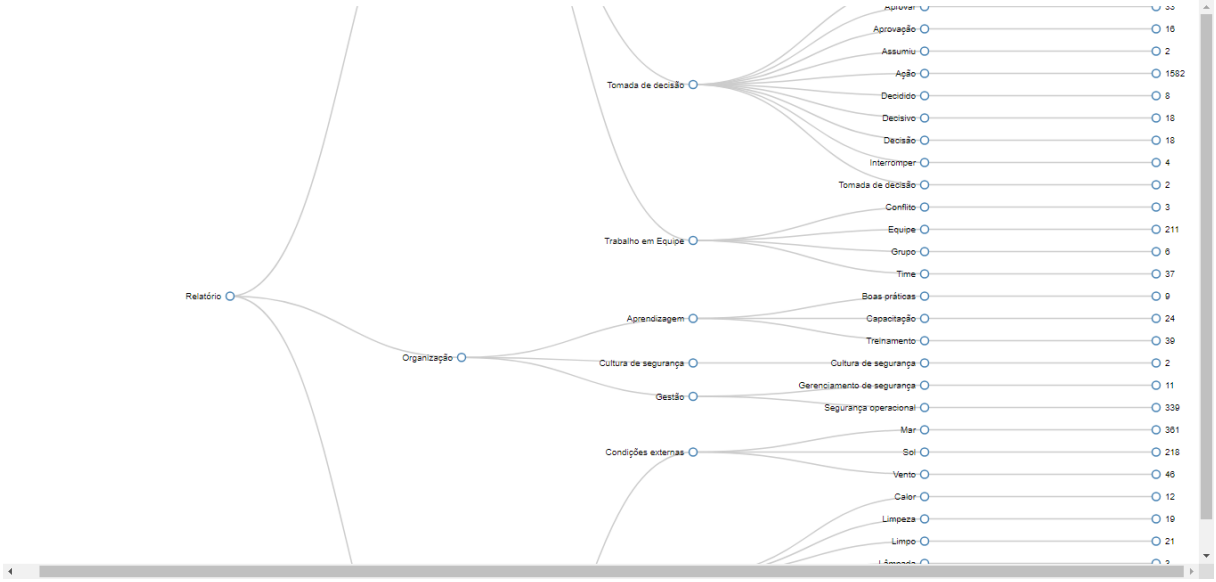
Figura 28 - Gráfico: Árvore de Palavras – Fatores e Frequência



Fonte: Elaborado pela autora

A terceira árvore de palavras, ilustrada na Figura 29, desempenha um papel crucial ao fornecer uma visão detalhada das dimensões, fatores, termos e suas respectivas frequências de ocorrência no documento em análise. Nesse contexto, as dimensões representam as categorias amplas ou aspectos abordados no documento, enquanto os fatores são os termos específicos que compõem cada dimensão, refletindo os subtemas ou conceitos dentro dessas categorias.

Figura 29 - Gráfico: Árvore de Palavras - Frequência – Termos e Frequência

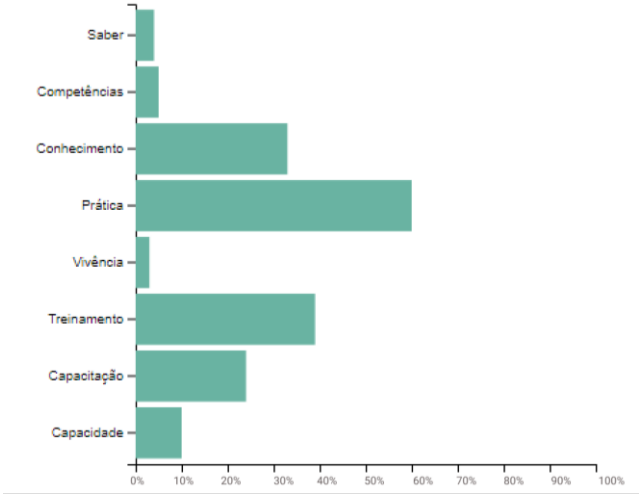


Na Figura 30, é apresentado um gráfico de histograma que demonstra todos os termos encontrados no documento, representados por barras cuja altura corresponde à frequência de

ocorrência de cada termo. Esse tipo de visualização é comumente utilizado em análises de texto para destacar os termos mais frequentes e sua distribuição no documento.

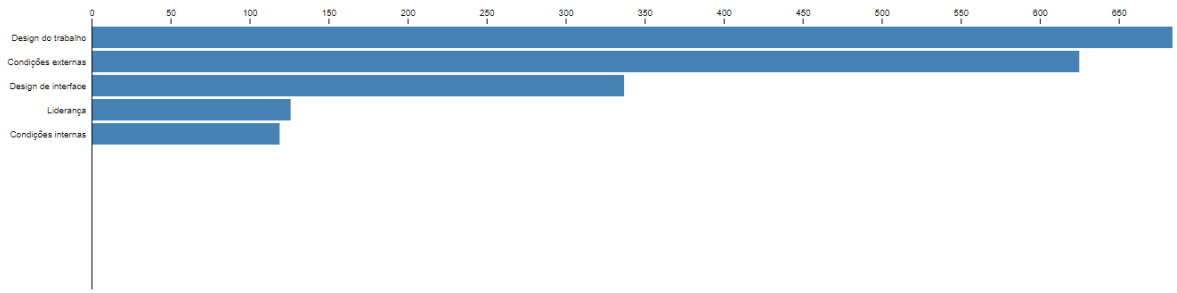
Cada barra no gráfico de histograma representa um termo específico encontrado no texto, enquanto a altura da barra indica com que frequência esse termo ocorre no documento. Termos mais frequentes são representados por barras mais altas, enquanto termos menos comuns têm barras mais baixas.

Figura 30 - Gráfico: Barras horizontais



Já a Figura 31, representa o resultado do estudo utilizado um gráfico de barras horizontais hierárquicas interativo, servindo para apresentar a frequência de ocorrência das dimensões em um gráfico de barras com a possibilidade de navegar na hierarquia clicando na barra.

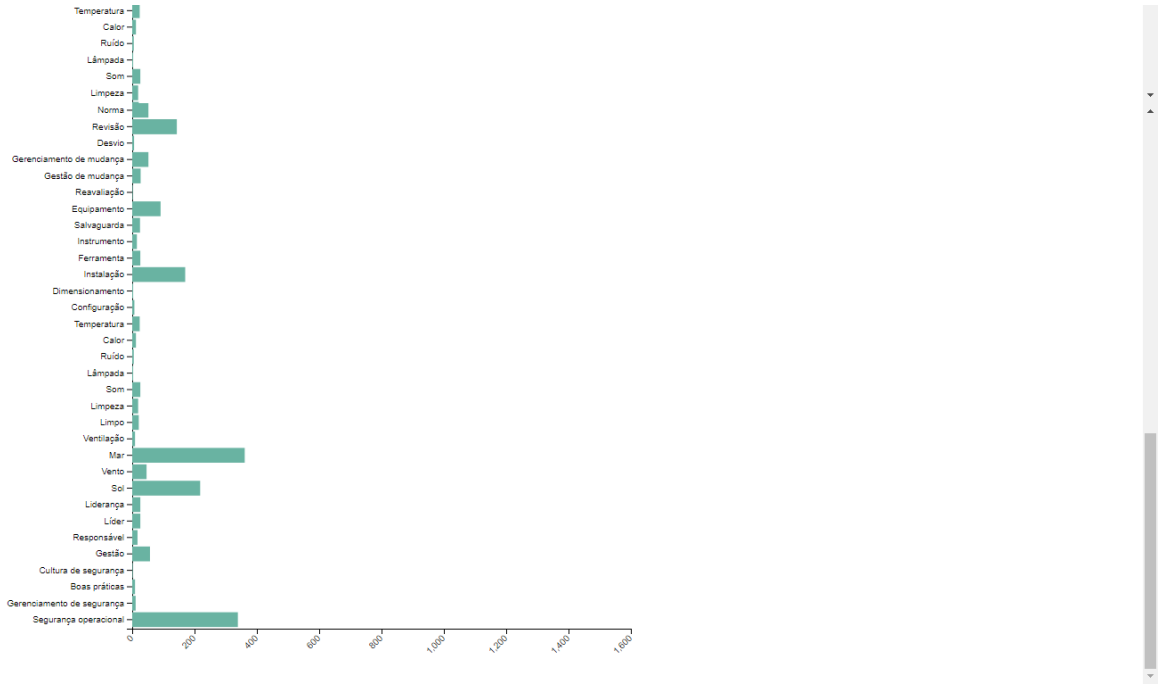
Figura 31 - Gráfico: Barras horizontais interativas



No gráfico de barras horizontais (Figura 32), a frequência de todos os termos presentes no documento é exibida ao longo do eixo horizontal. Cada barra horizontal representa um termo

específico encontrado no texto, enquanto o comprimento da barra indica com que frequência esse termo ocorre no documento.

Figura 32 - Gráfico: Barras horizontais - Termos



Na Figura 33, é apresentado um gráfico *treemap*, uma ferramenta visual que exibe as frequências de ocorrências dos termos de forma hierárquica usando retângulos aninhados. Nesse tipo de gráfico, cada retângulo representa um termo, sendo o tamanho proporcional à sua frequência de ocorrência no documento analisado.

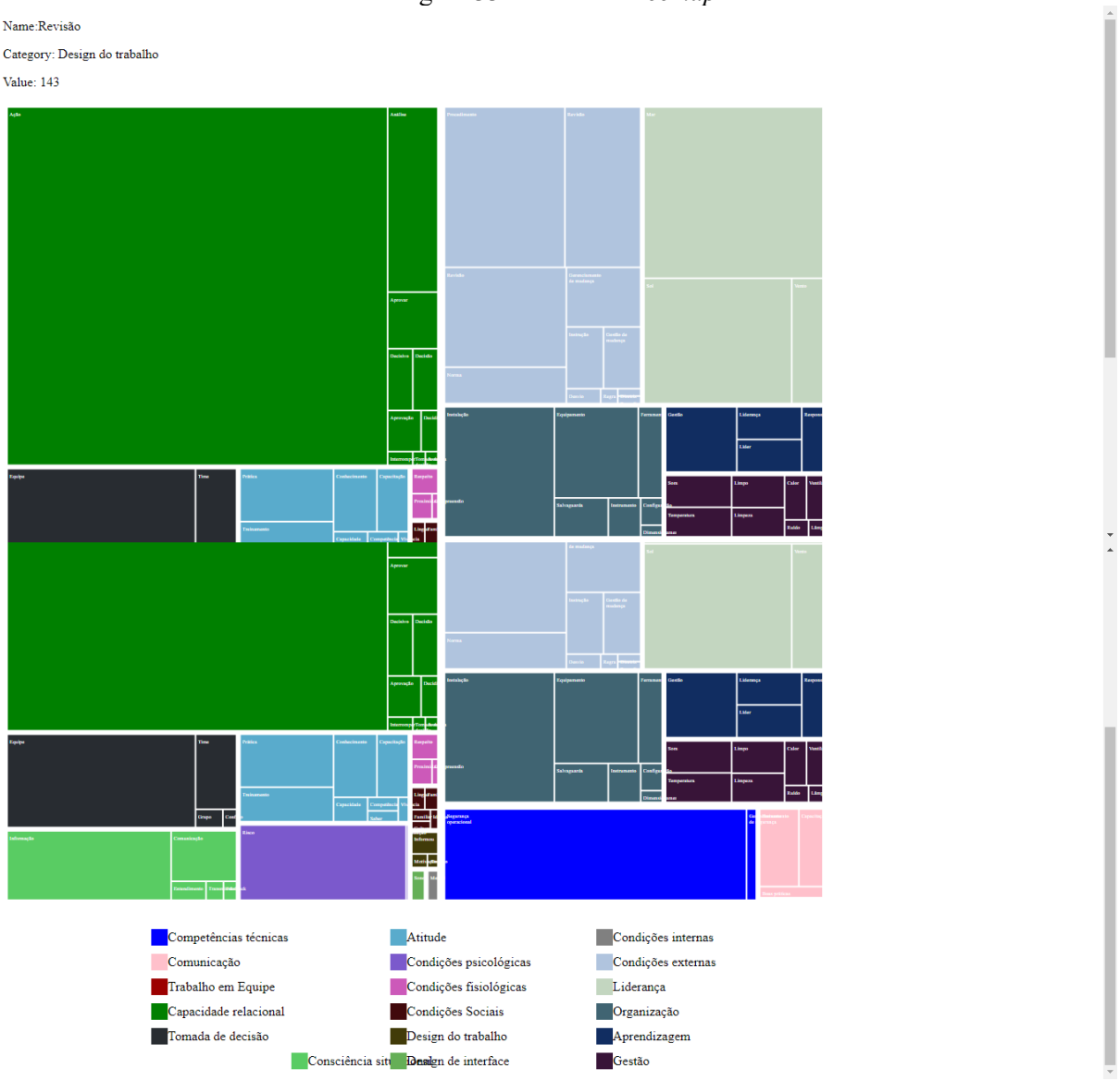
A hierarquia é representada através da organização dos retângulos maiores em áreas subdivididas por retângulos menores. Essa subdivisão permite visualizar não apenas a frequência dos termos individualmente, mas também como eles se relacionam e se agrupam dentro do documento.

Os retângulos maiores correspondem aos termos mais frequentes, enquanto os retângulos menores dentro deles representam termos menos frequentes, formando uma estrutura hierárquica. Essa representação visual permite uma análise intuitiva da distribuição dos termos ao longo do documento, destacando os termos mais relevantes e suas relações contextuais.

Por meio do gráfico *treemap*, os analistas podem identificar padrões, tendências e *insights* sobre a estrutura e conteúdo do documento de forma rápida e eficiente. A disposição hierárquica dos retângulos aninhados facilita a compreensão da distribuição da frequência dos

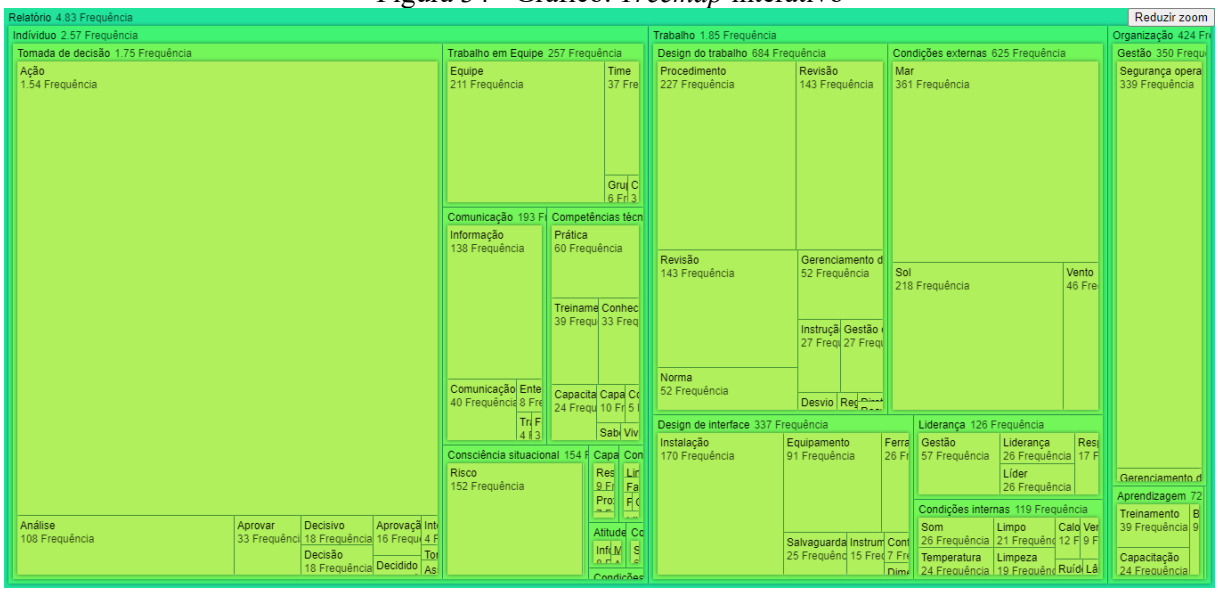
termos, proporcionando uma visão abrangente do vocabulário utilizado e de sua importância relativa no texto.

Figura 33 - Gráfico: *Treemap*



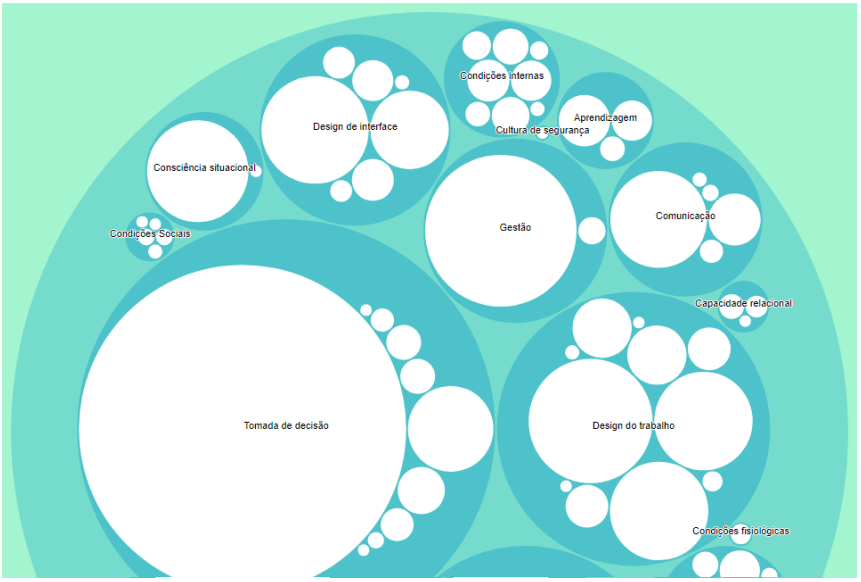
Já o gráfico *treemap*, representado pela Figura 34 também apresentar o gráfico das frequências de ocorrências dos termos de maneira hierárquica usando retângulos aninhados, permitindo que o usuário interaja entre as dimensões.

Figura 34 - Gráfico: *Treemap* interativo



O gráfico, apresentado pela Figura 35 utiliza o tipo de gráfico *zoomable circle packing*, usado para visualizar o gráfico das frequências de ocorrências dos termos de forma hierárquica, o método emprega o conceito de "*zoomable circle packing*" (empacotamento de círculos ampliável). Essa abordagem se baseia na representação visual por meio de círculos aninhados, proporcionando uma representação visual eficaz e interativa da hierarquia das informações.

Figura 35 - Gráfico: *Zoomable Circle Packing*

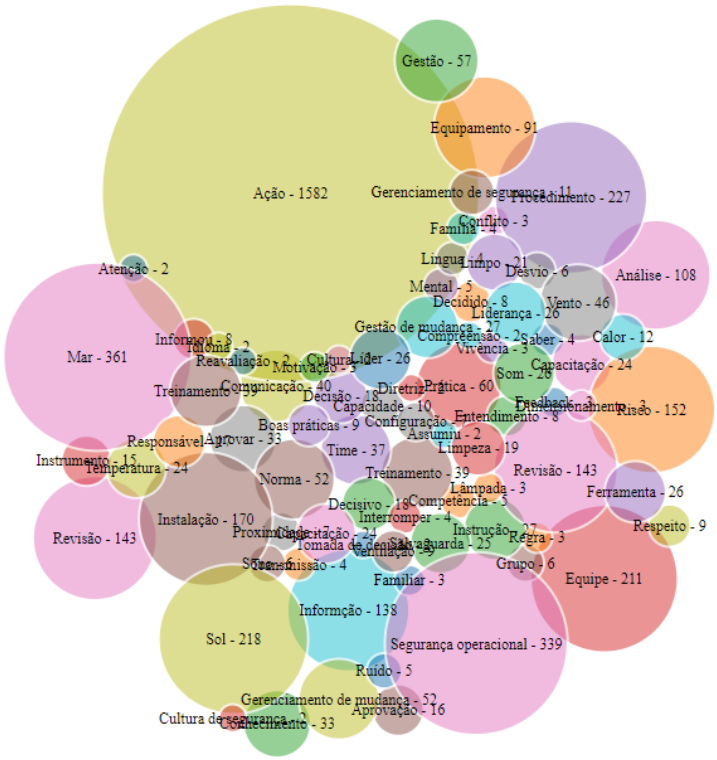


A Figura 36, representa um gráfico de nuvens de palavras, utilizado para identificar a frequência dos termos no documento analisado.

Figura 36 - Gráfico: Nuvens de Palavras

Através da Figura 37, um gráfico de bolha é utilizado para apresentar resultados. Os gráficos de bolhas consistem em círculos compactos e não hierárquicos, nos quais a área de cada círculo reflete proporcionalmente o seu valor, como o tamanho do arquivo. A natureza orgânica desses diagramas pode despertar interesse e curiosidade.

Figura 37 - Gráfico - Bolhas



A visualização de dados emerge como uma ferramenta vital na decodificação de informações complexas. A variedade de gráficos, desde os convencionais até os inovadores, oferece oportunidades para uma compreensão mais profunda e acessível dos dados. A análise

de sentimento, somada a essas representações visuais, amplia ainda mais a compreensão, fornecendo *insights* valiosos sobre emoções e opiniões expressas em dados textuais.

5.4 SISTEMAS SIMILARES

No decorrer desta pesquisa, foram encontrados sistemas similares ao implementado nesta dissertação a partir do modelo proposto. A seguir, são identificados os sistemas e suas diferenças em relação ao DOC *Analysis*.

O sistema *Voyant Tools* permite que sejam enviados documentos e *links* para serem analisados. Após o carregamento, são apresentados gráficos e visualizações dos dados do conteúdo analisado. Este sistema, no entanto, não utiliza taxonomia e não possui a funcionalidade de salvar consultas, o que pode limitar a organização e recuperação de informações específicas pelo usuário.

O *Doc Analyzer* também permite o envio de arquivos de texto. Diferentemente do *Voyant Tools*, após o envio do arquivo, o usuário pode fazer perguntas sobre o conteúdo analisado, proporcionando uma interação direta com os dados textuais. Contudo, assim como o *Voyant Tools*, não utiliza taxonomia e não possui visualizações gráficas dos dados. Uma característica distintiva do *Doc Analyzer* é a possibilidade de salvar consultas realizadas pelo usuário, permitindo maior controle e reutilização das informações pesquisadas.

O sistema *Petal* se destaca por permitir a interação com o sistema através de perguntas sobre o texto analisado, além de possibilitar que o usuário faça comentários sobre o texto. Apesar dessa funcionalidade interativa, o Petal não utiliza taxonomia e não oferece visualizações gráficas dos dados. Além disso, similar ao *Voyant Tools*, o *Petal* não permite salvar consultas, o que pode ser uma limitação para usuários que necessitam acessar frequentemente consultas anteriores.

Por fim, o DOC *Analysis*, desenvolvido nesta dissertação, permite o uso de arquivos de texto e utiliza taxonomia para análise, proporcionando uma estrutura organizada para a categorização dos dados. O DOC *Analysis* oferece visualização dos dados através de gráficos e outras representações, facilitando a interpretação dos resultados. Além disso, diferencia-se dos outros sistemas por permitir salvar consultas realizadas pelo usuário, uma funcionalidade que aprimora a organização e a recuperação das informações ao longo do tempo. O Quadro 12, identifica os sistemas e suas diferenças em relação ao DOC *Analysis*.

Quadro 12 - Sistemas similares

Nome	*Uso* de arquivo de texto	*Uso* de taxonomia	Visualização dos dados	Salvar consulta
DOC *Analysis*	✓	✓	✓	✓
Voyant *Tools*[11]	✓	✗	✓	✗
Doc *Analyzer*[12]	✓	✗	✗	✓
Petal[13]	✓	✗	✗	✓

Fonte: Elaborado pela autora.

Assim, ao comparar esses sistemas, nota-se que o DOC *Analysis* combina várias funcionalidades desejáveis, como o uso de taxonomia, visualização de dados e a capacidade de salvar consultas, superando as limitações encontradas nos outros sistemas analisados.

5.5 ANÁLISE E RECOMENDAÇÕES FUTURAS

A avaliação do DOC *Analysis, instanciação do método proposto,* revela uma solução promissora para análise de texto, caracterizada por sua abordagem intuitiva e eficiente. Para aprimorá-lo ainda mais e garantir uma evolução contínua, algumas considerações e recomendações podem ser ponderadas.

O DOC *Analysis* se destaca por sua abordagem clara e eficaz na análise de texto, a busca constante por melhorias na usabilidade, funcionalidade e personalização é essencial. Ao implementar essas recomendações, o sistema poderá oferecer uma experiência mais rica e adaptável, atendendo às crescentes demandas de um cenário dinâmico e diversificado de usuários. A flexibilidade e a relevância são cruciais para assegurar que o DOC *Analysis* permaneça uma ferramenta para análise de texto.

[11] https://voyant-tools.org/
[12] https://docanalyzer.ai/
[13] https://www.petal.org/

Para atingir o objetivo proposto para o cenário de estudo foi necessário seguir um fluxo de utilização:

- **Criação de Estudo:** os analistas iniciam um novo estudo no DOC *Analysis*, nomeando-o como "Análise de Incidentes - Projeto Resiliência Óleo e Gás".
- ***Upload* do documento de texto:** fazem o *upload* dos relatórios de incidentes acumulados ao longo do projeto.
- **Seleção da Taxonomia:** utilizam uma taxonomia específica, previamente definida para fatores humanos e resiliência na indústria de óleo e gás, para categorizar os elementos relevantes nos relatórios.
- **Execução da Análise:** clicam em "Executar estudo" para iniciar a análise com base na taxonomia selecionada. O DOC *Analysis* processa os relatórios, identificando padrões, fatores humanos e outros elementos críticos.
- **Visualização dos Resultados:** exploram os gráficos gerados pelo sistema, como árvores de palavras hierárquicas ou barras horizontais interativas, para visualizar a distribuição dos fatores e identificar padrões recorrentes.
- **Identificação de Tendências e *Insights*:** utilizam a capacidade de zoom e interatividade nos gráficos para aprofundar a análise em categorias específicas de incidentes, identificando tendências e *insights* valiosos.

A implementação do DOC *Analysis* para a análise de incidentes no contexto do projeto revelou-se uma estratégia para compreender e aprimorar os fatores humanos e a resiliência em operações *offshore*. O sistema foi testado apenas no âmbito acadêmico, não sendo disponibilizado ainda para domínio público. Essa ferramenta, embora inicialmente desenvolvida para a indústria de óleo e gás, demonstrou potencial para ser aplicada em diversas áreas e organizações, oferecendo *insights* valiosos para melhorias operacionais e estratégicas.

5.5.1 Considerações para Futuras Implementações

Ao examinar a usabilidade do DOC *Analysis* é crucial realizar testes com usuários para identificar áreas que possam ser aprimoradas. A implementação de instruções mais detalhadas ou tutoriais interativos no sistema pode facilitar a compreensão e a navegação, especialmente para usuários iniciantes. Além disso, introduzir *feedbacks* interativos durante as etapas de criação de estudos e execução de análises contribuirá para uma experiência mais envolvente e informativa. Recomendações para Aprimoramento:

- **Personalização Avançada**: oferecer opções mais avançadas de personalização permitirá que os usuários adaptem a ferramenta de acordo com suas necessidades específicas. A flexibilidade é fundamental para atender a uma variedade de casos de uso.
- **Integração com Fontes Externas**: possibilitar a integração do DOC *Analysis* com fontes externas, como APIs ou bancos de dados, ampliará as opções de análise e proporcionará aos usuários uma visão mais abrangente dos dados.
- **Suporte a Diferentes Formatos de Arquivo**: Ampliar o suporte para diferentes formatos de arquivo é essencial para garantir que o sistema seja capaz de analisar uma ampla variedade de documentos, atendendo às demandas heterogêneas dos usuários.
- **Ferramentas de Exportação**: implementar ferramentas robustas de exportação permitirá que os usuários salvem e compartilhem facilmente os resultados de suas análises em diferentes formatos. Isso promoverá uma colaboração eficaz e a utilização dos dados em outros contextos.
- **Atualizações Contínuas**: comprometer-se com atualizações regulares do sistema é crucial. A incorporação de *feedback* dos usuários, juntamente com a manutenção da relevância em relação às melhores práticas em análise de texto e visualização de dados, garantirá a longevidade do DOC *Analysis*.
- **Utilização de IA Generativa para intepretação dos resultados**: representa uma abordagem inovadora e promissora. A IA generativa, especialmente em sua aplicação na análise de dados visuais, pode oferecer *insights* adicionais e complementares à análise humana. Os algoritmos podem ajudar a identificar tendências emergentes, relações complexas entre variáveis e áreas de interesse que merecem uma investigação mais aprofundada. Além disso, a IA generativa pode auxiliar na detecção de padrões recorrentes ao longo do tempo, permitindo uma análise longitudinal dos dados e uma compreensão mais completa da evolução dos processos e incidentes.

6 CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS

Neste capítulo final, sintetizamos as principais contribuições desta pesquisa, destacando a importância da aplicação de técnicas avançadas de representação do conhecimento, como Taxonomia, TM e NLP, na análise de corpus de documentos textuais. Através do método proposto, foi possível demonstrar a viabilidade e eficácia da extração de termos-chave e a utilização de uma taxonomia para organizar e interpretar esses termos de maneira estruturada. As técnicas de visualização de dados complementaram o processo, oferecendo uma compreensão mais clara e intuitiva dos conteúdos analisados.

6.1 CONSIDERAÇÕES FINAIS

Finalizado o desenvolvimento desta pesquisa é possível destacar a importância e relevância da aplicação de estruturas de representação de conhecimento e técnicas, como Taxonomia, KDT/TM e NLP com ênfase em tarefas de processamento de texto em nível léxico e sintático, assim como a Visualização de Dados na análise de *corpus* de documentos textuais. Através da extração dos principais termos presentes nos textos e da utilização de uma taxonomia para explicar o conteúdo dos documentos, foi possível facilitar a compreensão e interpretação desses textos.

O método proposto permitiu demonstrar a viabilidade prática dessa abordagem. Os resultados obtidos com os testes realizados em um caso de estudo específico demonstraram a eficácia da aplicação na extração dos principais termos presentes nos documentos. A utilização da taxonomia foi fundamental para organizar e relacionar os termos extraídos, proporcionando uma compreensão mais estruturada do conteúdo dos documentos.

A integração de uma Taxonomia com técnicas de KDT, TM, NLP e Visualização de Dados abre possibilidades promissoras para a análise de *corpus* de documentos textuais, auxiliando na extração de conhecimento e na tomada de decisões mais embasadas.

Além disso, é importante ressaltar que este trabalho pode ter impacto direto em aplicações práticas em diversos setores. A capacidade de compreender e interpretar documentos textuais é essencial para profissionais de áreas como jurídica, médica, científicas, entre outras. A aplicação desenvolvida neste trabalho pode fornecer suporte valioso para esses profissionais

na análise e compreensão de documentos complexos, permitindo uma extração mais precisa e eficiente das informações contidas nos textos.

Este trabalho teve como objetivo principal propor e desenvolver um método para apoiar a análise de documentos textuais por meio de técnicas de Mineração de Texto e Taxonomia. Para alcançar esse objetivo, foram definidos objetivos gerais e específicos. O objetivo deste trabalho foi propor e desenvolver um método para apoiar a análise de documentos textuais. Este método baseou-se na identificação de características textuais e na especificação de uma taxonomia que guiou a análise. Além disso, foi aplicado em um cenário de estudo para avaliar o método proposto, utilizando técnicas de Mineração de Texto e Taxonomia.

Os objetivos específicos deste estudo incluíram a identificação das características textuais que possibilitaram o apoio à análise de documentos, a especificação de uma taxonomia para guiar a análise de documentos textuais e a aplicação de técnicas de visualização de dados para apresentar o conhecimento de um domínio específico. Para atingir esses objetivos, foram realizadas diversas etapas, sendo inicialmente identificadas as características textuais relevantes por meio da revisão integrativa da literatura, o que permitiu estabelecer uma base teórica sólida para o desenvolvimento do método.

Posteriormente, uma taxonomia foi especificada para guiar a análise de documentos textuais, fornecendo uma estrutura organizada e sistemática para a interpretação dos textos. Essa taxonomia foi essencial para categorizar os termos e conceitos presentes nos documentos, contribuindo para uma compreensão mais profunda do conteúdo. Além disso, um cenário de estudo foi desenvolvido, focando em plataformas *offshore*, o que permitiu a instanciação do método proposto em um contexto prático e relevante.

Com a instanciação do método, foi possível demonstrar sua relevância na análise de documentos textuais, fornecendo suporte valioso para a compreensão e interpretação desses documentos. Um dos principais benefícios do método proposto é a capacidade de gerar representações visuais intuitivas, como gráficos, que podem auxiliam significativamente no processo de tomada de decisão. Essas visualizações permitem que os usuários identifiquem padrões, tendências e relações entre os tópicos abordados nos documentos de uma maneira clara e compreensível. Ao fornecer uma visão holística e estruturada das informações contidas nos documentos, o sistema DOC *Analysis* assim pode contribuir para uma melhor compreensão do conteúdo e facilita a tomada de decisões fundamentadas.

Vale ressaltar que este trabalho não se encerra aqui, e há oportunidades para futuras melhorias e expansões do método proposto. Entre as possíveis melhorias, pode-se explorar o refinamento das técnicas de processamento de linguagem natural utilizadas, buscando aprimorar a precisão na associação dos documentos à taxonomia definida e na identificação de correspondências entre os tópicos. Além disso, existe a possibilidade de explorar novos cenários de aplicação do método, expandindo seu escopo de atuação para outras áreas, como a análise de relatórios corporativos, a revisão de literatura científica ou a avaliação de documentos em diferentes idiomas. Essas expansões e aprimoramentos futuros podem contribuir ainda mais para a eficácia da análise de documentos textuais e para a tomada de decisões baseadas em informações estruturadas e visuais.

6.2 TRABALHOS FUTUROS

Com base na pesquisa realizada, alguns possíveis trabalhos futuros podem ser explorados:

1. **Refinamento da aplicação:** uma possível melhoria na aplicação desenvolvida seria o refinamento dos algoritmos de TM utilizados, permitindo uma extração ainda mais precisa e eficiente dos termos presentes nos documentos. Isso poderia ser alcançado através do uso de técnicas mais avançadas de processamento de linguagem natural ou algoritmos mais sofisticados de classificação e agrupamento.

2. **Incorporação de outras técnicas:** além do TM, outras técnicas podem ser exploradas para enriquecer a análise dos documentos textuais. Por exemplo, a utilização de técnicas de análise sentimental poderia fornecer *insights* sobre a polaridade ou emoções expressas nos textos analisados.

3. **Exploração de diferentes taxonomias:** a pesquisa propôs o uso de uma taxonomia específica para a aplicação desenvolvida. É possível explorar o uso de diferentes taxonomias em domínios específicos para obter análises mais precisas e personalizadas dos documentos textuais.

4. **Análise de documentos multilíngues:** a aplicação desenvolvida neste trabalho foi focada em documentos em um único idioma. Uma possível expansão seria a análise de documentos multilíngues, incorporando técnicas de tradução automática e adaptando a taxonomia para lidar com diferentes idiomas.

5. **Aplicação em outros domínios:** embora o caso de estudo utilizado tenha sido específico, a aplicação desenvolvida pode ser adaptada e utilizada em diferentes domínios, como saúde, direito, finanças, entre outros. Seria interessante explorar essas possibilidades e avaliar o desempenho da aplicação em diferentes cenários.

6. **Avaliação da eficácia da aplicação:** para garantir a eficácia da aplicação proposta, seria importante realizar estudos adicionais para avaliar sua precisão na extração dos termos e na classificação dos documentos analisados. Isso poderia ser feito comparando os resultados obtidos pela aplicação com avaliação manual ou utilizando métricas específicas de desempenho.

7. **Utilização de Inteligência Artificial Generativa:** explorar a possibilidade de utilizar técnicas de IA Generativa para gerar *insights* adicionais de forma mais natural e abrangente, permitindo uma compreensão mais profunda do conteúdo textual e identificando padrões não óbvios, assim como auxiliando na interpretação dos gráficos obtidos nos resultados dos estudos criados no Doc *Analysis*.

Essas são apenas algumas das possíveis direções para trabalhos futuros, outras oportunidades de estudo incluem a otimização dos algoritmos de extração de termos, a aplicação da tecnologia em diferentes contextos e a integração com outras ferramentas e sistemas existentes. Além disso, é importante considerar o *feedback* dos usuários da aplicação e realizar melhorias com base nos retornos. Dessa forma, as possibilidades de trabalhos futuros podem expandir ainda mais os benefícios da análise de *corpus* de documentos textuais utilizando técnicas avançadas como TM, Representação de Conhecimento por meio de Taxonomias e Visualização de Dados.

6.3 CONTRIBUIÇÕES DA DISSERTAÇÃO

As contribuições deste trabalho se estabelecem em três áreas-chave. Em primeiro lugar, destaca-se a relevância na integração de técnicas avançadas, como TM e NLP, para a análise de documentos textuais. Essa abordagem proporciona uma compreensão mais eficaz e estruturada do conteúdo, facilitando a extração de conhecimento de forma precisa e sistemática. Além disso, o uso do processo de KDT é essencial para organizar e gerenciar o conhecimento extraído, garantindo uma aplicação prática e eficiente dos dados analisados.

Em segundo lugar, o trabalho resultou no desenvolvimento prático do método incorporando técnicas, permitindo aos usuários extrair termos-chave e visualizá-los por meio da taxonomia proposta. Este método demonstra aplicabilidade em diversos setores, incluindo o jurídico, médico e científico, fornecendo uma ferramenta valiosa para a análise e interpretação de documentos textuais por profissionais dessas áreas. A visualização desses termos-chave é especialmente destacada, facilitando a compreensão e navegação pelo conteúdo analisado de maneira intuitiva e eficiente.

Além disso, a dissertação destaca sua contribuição para a Engenharia do Conhecimento, alinhando-se aos princípios desta área ao transformar dados em conhecimento por meio da modelagem e criação de sistemas baseados em conhecimento. Ao combinar diferentes técnicas de processamento, representação e visualização de dados e conhecimento a pesquisa propõe uma abordagem na análise de documentos textuais e provendo uma perspectiva relevante no contexto da Engenharia do Conhecimento.

Por fim, é válido ressaltar que este estudo não se esgota aqui. Existem várias oportunidades de aprimoramento e expansão dessa pesquisa. Por exemplo, a aplicação poderia ser refinada com a inclusão de técnicas adicionais de NLP ou algoritmos mais sofisticados de classificação e agrupamento dos termos extraídos. Além disso, seria interessante explorar a possibilidade de utilizar diferentes taxonomias em diferentes domínios específicos para uma análise mais precisa e personalizada dos documentos textuais. Isso permitiria uma compreensão ainda mais refinada do conteúdo presente nos textos. A pesquisa aponta para o potencial de expansão e aprimoramento, sugerindo o uso de técnicas adicionais de análise textual e diferentes taxonomias em domínios específicos. Essas identificações abrem portas para estudos futuros e o desenvolvimento de soluções mais sofisticadas na análise de documentos textuais, contribuindo para um entendimento mais profundo e abrangente do conteúdo textual em diversos contextos.

REFERÊNCIAS

ABDUL-RAHMAN, A. et al. Rule-based Visual Mappings - With a Case Study on Poetry Visualization. **Computer Graphics Forum**, 2013.

ABEL, Mara; FIORINI, Sandro R. UMA REVISÃO DA ENGENHARIA DO CONHECIMENTO: EVOLUÇÃO, PARADIGMAS E APLICAÇÕES. **International Journal Knowledge Engineering Management**, Florianópolis, v. 2, n. 2, p. 135, mar./maio, 2013. Disponível em: <https://www.researchgate.net/publication/264156426_Uma_revisao_da_Engenharia_do_Conhecimento_Evolucao_Paradigmas_e_Aplicacoes>. Acesso em 17 de out. de 2021.

ADJIN-TETTEY, T. D.; SELORMEY, D.; NKANSAH, H. A. Ubiquitous Technologies and Learning: Exploring Perceived Academic Benefits of Social Media Among Undergraduate Students**. International Journal of Information and Communication Technology Education (IJICTE)**, v. 18, n. 1, p. 1-16, 2022. Disponível em: http://doi.org/10.4018/IJICTE.28675

ADWAN, Omar Y. et al. Twitter sentiment analysis approaches: A survey. International **Journal of Emerging Technologies in Learning,** 2020.

AGGARWAL, Charu C.; ZHAI, ChengXiang. **Mining Text Data**. Boston: Springer, 2012.

ALHUAY-QUISPE, Joel; ESTRADA-CUZCANO, Alonso; BAUTISTA-YNOFUENTE, Lourdes. Analysis and Data Visualization in Bibliometric Studies. **JLIS.it,** 2022.

ALMUTAIRI, Bandar Alhumaidi A. Visualizing patterns of appraisal in texts and corpora. **Text and Talk**, 2013.

ALPAYDIN, E. **Introduction to machine learning**. MIT press, 2010.

ALVAREZ, Guilherme Martins**. Análise de agrupamentos e mineração de opinião como suporte à gestão de ideias**. 2018. Dissertação (Mestrado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, Florianópolis, 2018.

AMADIO, William J.; PROCACCINO, J. Drew. Competitive Analysis of Online Reviews Using Exploratory Text Mining. **Tourism and Hospitality Management**, 2016.

ANDERSON, L. W.; KRATHWOHL, D. R. (Eds.). **A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives**. New York: Longman, 2001.

ANDRADE, R. **Um Modelo para recuperação e comunicação do conhecimento em documentos médicos**. 2011. Tese (Doutorado) - Universidade Federal de Santa Catarina,

Centro Tecnológico, Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, Florianópolis, 2011.

ANDRIANI, Mateus Lohn. **Um método para a construção de taxonomias utilizando a DBpedia**. 2017. Dissertação (Mestrado) – Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, Florianópolis, 2017.

ARAÚJO, Aneide Oliveira; OLIVEIRA, Marcelle Colares. **Tipos de pesquisa**. São Paulo, 1997.

ARTESE, Letícia Silveira. **Modelo de descoberta de conhecimento em texto para detecção de sinais fracos para tecnologias emergentes**. 2023. Tese (Doutorado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, Florianópolis, 2023.

AUCK, Jean Carlo Rossa. **Um Método de aquisição de conhecimento para customização de modelos de capacidade/maturidade de processos de software**. 2011. Tese (Doutorado) – Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, Florianópolis, 2011.

AWS. O que é Java? – Explicação sobre a linguagem de programação Java. Disponível em: https://aws.amazon.com/pt/what-is/java/. Acesso em: 15 mar. 2023.

BABIC, M.; JERMAN-BLAZIC, B. New Cybercrime Taxonomy of Visualization of Data Mining Process**. 2016 39th International Convention on Information and Communication Technology, Electronics and Microelectronics** (MIPRO), 2016.

BAKER, P. **Corpora and Discourse Studies**: Integrating Discourse and Text Linguistics. Cambridge University Press, 2018.

BAKER, P. **Using Corpora in Discourse Analysis**. London: Continuum, 2006.

BAKER, T. et al**. Using a consensus approach to develop a taxonomy for tools that support community engagement in research. Research involvement and engagement**, v. 4, n. 1, p. 1-13, 2018.

BARB, Adrian S.; KILICAY-ERGIN, Nil. Applications of Natural Language Techniques to Enhance Curricular Coherence. **Procedia Computer Science**, v. 168, p. 88-96, 2020.

BARDIN, L. **Análise de conteúdo**. Edições 70, 2016.

BAUMER, Eric P. S.; JASIM, Mahmood; SARVGHAD, Ali; MAHYAR, Narges. Of Course it's Political! A Critical Inquiry into Underemphasized Dimensions in Civic Text Visualization. **Computer Graphics Forum,** 2022.

BAZERMAN, Charles; PRIOR, Paul. **What Writing Does and How It Does It**: An Introduction to Analyzing Texts and Textual Practices. New York: Routledge, 2020.

BECHHOFER, S. et al. Why Linked Data is Not Enough for Scientists. **Future Generation Computer Systems,** v. 26, n. 3, p. 293-305, 2010.

BEIRÃO FILHO, José Alfredo. **Criação e compartilhamento do conhecimento da área de moda em um sistema virtual integrado de informações**. 2011. Tese (Doutorado) - Universidade Federal de Santa Catarina, Florianópolis, 2011.

BEPPLER, Fabiano Duarte. **Um modelo para recuperação e busca de informação baseado em ontologia e no círculo hermenêutico**. 2008. Tese (Doutorado) - Universidade Federal de Santa Catarina, Florianópolis, 2008.

BERMEJO, Paulo Henrique de Souza. **Planejamento estratégico de tecnologia da informação com ênfase em conhecimento**. 2009. Tese (Doutorado) - Universidade Federal de Santa Catarina, Florianópolis, 2009.

BERNERS-LEE, Tim. **HTML - Hypertext Markup Language**. W3C Recommendation 14 December 1995. Disponível em: https://www.w3.org/MarkUp/. Acesso em: 17 mar. 2023.

BERRY, M. J. **Data Mining Techniques**: For Marketing, Sales, and Customer Relationship Management. John Wiley & Sons, 2015.

BETTIO, Raphael Winckler de. **Interrelação das técnicas Term Extration e Query Expansion aplicadas na recuperação de documentos textuais**. 2007. 99 f. Tese (Doutorado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, Florianópolis, 2007. Disponível em: <http://www.bu.ufsc.br/teses/PEGC0030-T.pdf>. Acesso em 18 de out. de 2021.

BIBER, D. Representativeness in Corpus Design. **Literary and Linguistic Computing**, v. 8, n. 4, p. 243-257, 1993.

BIRD, S.; SIMONS, G. (Eds.). **Handbook of Natural Language Processing**. New York: CRC Press, 2009.

BIRD, S.; SIMONS, G. Seven dimensions of portability for language documentation and description. **Language**, 79(3), 557–582, 2003.

BISHOP, C. **Pattern recognition and machine learning**. Springer, 2006.

BISHOP, Christopher M. **Pattern Recognition and Machine Learning**. New York: Springer, 2006.

BIZ, Alexandre Augusto. **Avaliação dos portais turísticos governamentais quanto ao suporte à gestão do conhecimento**. 2009. Tese (Doutorado) – Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-graduação em Engenharia e Gestão do Conhecimento, Florianópolis, 2009.

BLEI, D. M. Probabilistic topic models. **Communications of the ACM**, 55(4), 77-84, 2012.

BLOOM, B. S. et al. **Taxonomy of educational objectives**: The classification of educational goals. New York: David McKay Company, 2001.

BLOOM, B. S. **Taxonomy of Educational Objectives, Handbook I**: The Cognitive Domain. New York: David McKay Co Inc, 1956.

BOEHM, B. W. et al. Um modelo espiral de desenvolvimento e aprimoramento de software. **ACM SIGSOFT Software Engineering Notes**, v. 11, n. 4, p. 14-24, 1986.

BORGMAN, C. L. **Big data, little data, no data**: Scholarship in the networked world. MIT press, 2015.

BORLAND, David; CHRISTOPHERSON, Laura; SCHMITT, Charles. Ontology-Based Interactive Visualization of Patient-Generated Research Questions. **Applied Clinical Informatics**, 2019.

BORTHWICK, A.; STERLING, J.; AGICHTEIN, E. Learning to extract symbolic knowledge from the World Wide Web. In: **Proceedings of the 13th national conference on artificial intelligence.** 1998. p. 509-516.

BORTOLOTTI, D. **O que é MySQL?** 2020. Disponível em: <https://www.hostinger.com.br/tutoriais/o-que-e-mysql>. Acesso em: 17 abr. 2023.

BOVO, A. B. **Um Modelo de descoberta de conhecimento inerente à evolução temporal dos relacionamentos entre elementos textuais**. 2011. Tese (Doutorado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, Florianópolis.

BOYD, Danah; CRAWFORD, Kate. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. Information, **Communication & Society**, v. 15, n. 5, p. 662-679, 2012.

BRACHMAN, R. J. **What is AI?** Fundamentals of Artificial Intelligence. Cambridge: Cambridge University Press, 2006.

BRACHMAN, R. J.; LEVESQUE, H. J. **Knowledge representation and reasoning**. San Francisco: Elsevier, 2004.

BRASIL. Lei Geral de Proteção de Dados Pessoais (LGPD). Lei nº 13.709, de 14 de agosto de 2018.

BREWER, C. A. Color use guidelines for mapping and visualization. In **Cartography and Geographic Information Science**, v. 26, n. 4, p. 261-278, 1999.

BROUGHTON, V. Faceted classification as a basis for knowledge organization in a digital environment: the Bliss Bibliographic Classification as an example. **Journal of Documentation**, v. 62, n. 5, p. 529-542, 2006. Disponível em: https://www.emerald.com/insight/content/doi/10.1108/00220410610692154/full/html. Acesso em: 30 mar. 2023.

BROWN, M. Harnessing the Power of Artificial Intelligence for Document Analysis. **Journal of Information Science**, 45(3), 297-310, 2018.

BROWN, Steve. **React Up and Running**: Building Modern Applications with React. O'Reilly Media, 2019.

BROWN, T. B. et al. Language models are few-shot learners. **arXiv preprint arXiv**: 2005.14165, 2020.

BURKHARDT, Dirk; NAZEMI, Kawa; GINTERS, Egils. Best-Practice Piloting Based on an Integrated Social Media Analysis and Visualization for E-Participation Simulation in Cities. **Procedia Computer Science**, v. 75, p. 66-74, 2015.

BUXTON, B. **Sketching user experiences: getting the design right and the right design**. Morgan Kaufmann, 2007.

BUYYA, R. et al. Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility. **Future Generation Computer Systems**, v. 25, n. 6, p. 599-616, 2009.

CABRAL, R. B. **Concepção, implementação e validação de um enfoque para integração e recuperação de conhecimento distribuído em bases de dados heterogêneas**. 2010. Dissertação (Mestrado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, Florianópolis.

CAIRO, A. **The functional art: An introduction to information graphics and visualization**. New Riders, 2013.

CAIRO, A. **The Truthful Art**: Data, Charts, and Maps for Communication. New Riders, 2019.

CALVÃO, Leandro Dantas; PIMENTEL, Mariano; FUKS, Hugo. **Do email ao Facebook: uma perspectiva evolucionista sobre os meios de conversação da internet.** Rio de Janeiro: UNIRIO, 2014. Disponível em: https://meiosdeconversacao.uniriotec.br/taxonomia/. Acesso em: 28 set. 2023.

CAMACHO, David et al. **The four dimensions of social network analysis**: An overview of research methods, applications, and software tools. Information Fusion, 2020.

CANTO, Cleunisse Aparecida Rauen De Luca. **Framework conceitual de representação do conhecimento sobre o 'modelo de graduação dual'.** 2022. Tese (Doutorado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, Florianópolis 2022.

CECI, F. **Um Modelo semi-automático para a construção e manutenção de ontologias a partir de bases de documentos não estruturados.** 2012. Dissertação (Mestrado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, Florianópolis.

CHANDOLA, V.; BANERJEE, A.; KUMAR, V. Anomaly detection: A survey. **ACM computing surveys (CSUR)**, v. 41, n. 3, p. 1-58, 2009.

CHEN, L. Data Visualization Techniques. **Journal of Visual Communication**, v. 12, n. 4, p. 189-205, 2018.

CHEN, L.; LIU, H. **Text Mining and Knowledge Discovery**: Theory, Tools, and Applications. Wiley, 2021.

CHEN, M. **Top 10 algorithms for machine learning newbies**. Towards Data Science, 2017.

CHEN, M.; LIU, X. Big data: A survey. **Mobile Networks and Applications,** v. 19, n. 2, p. 171-209, 2014.

CHEN, Siming et al. Supporting Story Synthesis: Bridging the Gap between Visual Analytics and Storytelling. **IEEE Transactions on Visualization and Computer Graphics**, 2020.

CHIANG, M.; LAI, C.; LEE, Y. Data science and its relationship to big data and data-driven decision making. **Big Data Analysis**, v. 4, n. 1, p. 1-15, 2019.

CHIBELUSHI, C. C.; KHAN, M. K. **Knowledge capture**: Issues, techniques and applications. London: Springer, 2002.

COCKBURN, A. **Writing Effective Use Cases**. 1. ed. São Paulo: Bookman, 2000.

CONCEIÇÃO, Cristiano Sena da. **Desenvolvimento de um modelo conceitual da classificação internacional da funcionalidade, incapacidade e saúde baseado na web**. 2007. Dissertação (Mestrado) - Universidade Federal de Santa Catarina, Florianópolis, 2007.

CRESWELL, John W. **Projeto de pesquisa**: métodos qualitativo, quantitativo e misto. 3.ed. Porto Alegre: Artmed, 2010.

CUPANI, Alberto. La peculiaridad del conocimiento tecnológico. **ScientiaeStudia**, São Paulo, v. 4, n. 3, p. 353-71, 2006. Disponível em: <http://www.scielo.br/pdf/ss/v4n3/a01v4n3.pdf>. Acesso em: 06 abr. 2022.

DAVENPORT, T. H.; KIM, J. Big data and managerial decision making. **Academy of Management Journal**, v. 56, n. 2, p. 321-326, 2013.

DENG, X.; ZHANG, X.; XU, Z.; YAO, Y. Hybrid Approach for Automatic Taxonomy Classification. **IEEE Access**, v. 7, p. 123351-123363, 2019.

DEVLIN, J. et al. BERT: Pre-training of deep bidirectional transformers for language understanding. In: **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics** (NAACL), 2019.

DEY, Tathagata. NLP vs NLU vs NLG: Understanding the Differences. **Medium**, 17 maio 2023. Disponível em: https://iamtatha.medium.com/nlp-vs-nlu-vs-nlg-understanding-the-differences-400ad929ed4c. Acesso em: 06 fev. 2024.

DRESCH, Aline; LACERDA, Daniel Pacheco; JUNIOR, Jose Antonio Valle Antunes. **Design Science Research**: Método de Pesquisa para Avanço da Ciência e Tecnologia. Porto Alegre: Bookman, 2015.

DUDYCZ, Helena. Semantics visualization as a user interface in business information searching**. IFIP Advances in Information and Communication Technology**, 2021.

DUDYCZ, Helena. Usability of business information semantic network search visualization. **ACM International Conference Proceeding Series**, 2015.

EICH, Brendan. **JavaScript**: a linguagem de programação da web. 2000.

ELMASRI, R.; NAVATHE, S. B. **Sistemas de banco de dados**. São Paulo: Pearson, 2015.

ETELÄAHO, Anna; SOINI, Jari; JAAKKOLA, Hannu; MATTILA, Anna-Liisa. Visual data mining in software repositories: A survey**. Frontiers in Artificial Intelligence and Applications,** 2018.

FAIRCLOUGH, N. **Analyzing discourse:** Textual analysis for social research. Routledge, 2003.

FARACO, Fernando Melo. **Modelo de conhecimento baseado em tópicos de acórdãos para suporte à análise de petições iniciais**. 2020. 130 p. Dissertação (Mestrado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, Florianópolis, 2020. Disponível em: <http://www.bu.ufsc.br/teses/PEGC0631-D.pdf>. Acesso em 18 de out. de 2021.

FAUST, Richard. **Exploração do espaço de design das interações humano-computador: uma abordagem da gestão do conhecimento ergonômico**. 2013. Tese (Doutorado) – Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, Florianópolis, 2013.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. **AI Magazine**, v. 17, n. 3, p. 37-54, 1996.

FELDMAN, R., SANGER, J. The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. **Cambridge University Press**, 2007.

FELDMAN, S. et al. Taxonomies and their applications. **Journal of Information Science**, v. 30, n. 6, p. 521-549, 2004. Disponível em: https://journals.sagepub.com/doi/abs/10.1177/0165551504047119. Acesso em: 30 mar. 2023.

FENSEL, D. et al. OIL: An ontology infrastructure for the Semantic Web. **IEEE Intelligent Systems**, v. 16, n. 2, p. 38-45, 2001. Disponível em: https://ieeexplore.ieee.org/document/913584. Acesso em: 30 mar. 2023.

FENSEL, D. et al. **The knowledge engineering review**: issues in knowledge representation and reasoning. Cambridge: Cambridge University Press, 2001.

FERNANDES, Roberto Fabiano. **Uma proposta de modelo de aquisição de conhecimento para identificação de oportunidades de negócios nas redes sociais**. 2012. Dissertação (Mestrado) – Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, Florianópolis, 2012.

FEW, S. **Information Dashboard Design**: Displaying Data for At-a-Glance Monitoring. Analytics Press, 2013.

FEW, S. **Now You See It**: Simple Visualization Techniques for Quantitative Analysis. Oakland, CA: Analytics Press, 2009.

FIORIN, J. L. **Elementos de Análise do Discurso**. São Paulo: Contexto, 2016.

FISTER, Iztok; FISTER, Iztok; FISTER, Dušan; PODGORELEC, Vili; SALCEDO-SANZ, Sancho. A comprehensive review of visualization methods for association rule mining: Taxonomy, challenges, open problems and future ideas. **Expert Systems with Applications**, 2023.

FLANAGAN, David. **JavaScript**: The Definitive Guide. 2011.

FLICK, U. **Designing qualitative research**. Sage, 2018.

FLORIDI, L.; TADDEO, M.; TURILLI, M. The ethics of information transparency. **Ethics and Information Technology**, 20(3), p. 199-210, 2018.

FONSECA, F. T. et al. Ontologias: conceitos, ferramentas, aplicações e desafios. **JISTEM**: Journal of Information Systems and Technology Management, v. 5, n. 3, p. 551-574, 2008. Disponível em: https://www.scielo.br/scielo.php?script=sci_arttext&pid=S1807-17752008000300009. Acesso em: 30 mar. 2023.

FOWLER, M. **UML Distilled**: A Brief Guide to the Standard Object Modeling Language. 3. ed. São Paulo: Bookman, 2004.

FU, Siwei et al. VisForum: A visual analysis system for exploring user groups in online forums. **ACM Transactions on Interactive Intelligent Systems**, 2018.

GARCIA, C. Programas de treinamento e capacitação em segurança. In: **Anais do Congresso Brasileiro de Petróleo e Gás**, Rio de Janeiro, 2021.

GÉRON, Aurélien. **Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow**: Concepts, Tools, and Techniques to Build Intelligent Systems. 2. ed. Sebastopol: O'Reilly Media, 2019.

GÓMEZ-RODRÍGUEZ, A. et al. A survey of text mining techniques and applications. **Informatica**, v. 41, n. 4, p. 391-422, 2017.

GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron. **Deep Learning**. Cambridge: MIT Press, 2016.

GREGOR, S.; HART, D.; MARTIN, N. Finding the right issues for design science research. **Communications of the Association for Information Systems**, v. 34, n. 1, p. 49-72, 2013.

GREGOR, S.; JONES, D. The anatomy of a design theory. **Journal of the Association for Information Systems**, v. 8, n. 5, p. 312-335, 2007.

GRUBER, T. R. A translation approach to portable ontology specifications. **Knowledge Acquisition**, v. 5, n. 2, p. 199-220, 1993. Disponível em: https://www.sciencedirect.com/science/article/pii/1042391393900234. Acesso em: 30 mar. 2023.

GUARINO, N. The ontological level. In: MARS, N. (Ed.). **On the Evolution of Information Processing**. Berlin: Springer, 1998. p. 45-68. Disponível em: https://link.springer.com/chapter/10.1007/978-3-642-46825-8_4. Acesso em: 30 mar. 2023.

GUNNING, D. What makes a good feature? or why machine learning fails. **ACM Communications**, v. 60, n. 11, p. 54-63, 2016.

HAMZAH, Muzaffar; VU, Tuong Thuy. A Taxonomy of Twitter Data Analytics Techniques. **Proceedings of the 32nd International Business Information Management Association Conference**, IBIMA 2018 - Vision 2020: Sustainable Economic Development and Application of Innovation Management from Regional Expansion to Global Growth, 2018.

HAN, J.; KAMBER, M**. Data Mining**: Concepts and Techniques. 2. ed. San Francisco: Morgan Kaufmann, 2006.

HAN, Jiawei et al**. Data Mining**: Concepts and Techniques. 3rd ed. Waltham, MA: Morgan Kaufmann, 2011.

HAND, D.; MANNILA, H.; SMYTH, P. **Principles of data mining**. Cambridge, MA: MIT press, 2001.

HARTSON, R.; PYLA, P. **The UX Book**: process and guidelines for ensuring a quality user experience. Morgan Kaufmann, 2012.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. **Springer Series in Statistics**, 2009.

HAVENS, Lucy et al. Beyond Explanation: A case for exploratory text visualizations of non-aggregated, annotated datasets. **1st Workshop on Perspectivist Approaches to Disagreement in NLP, NLPerspectives 2022 as part of Language Resources and Evaluation Conference**, LREC 2022 Workshop, 2022.

HEER, J. & SHNEIDERMAN, B. Interactive dynamics for visual analysis. **Queue**, 10(2), 30, 2012.

HEER, J.; SCHNEIDERMAN, B. Interactive dynamics for visual analysis. **Communications of the ACM**, v. 55, n. 4, p. 45-54, 2012.

HEERY, R. Thesauri and Taxonomies: An Overview of Their Applications in Information Management and Retrieval. **The Electronic Library**, v. 22, n. 4, p. 320-344, 2004.

HEINZLE, R. **Um Modelo de engenharia do conhecimento para sistemas de apoio a decisão com recursos para raciocínio abdutivo**. 2011. Tese (Doutorado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, Florianópolis.

HEVNER, A. R. et al. Design Science Research in Information Systems. **MIS Quarterly**, v. 28, n. 1, p. 75-105, 2004.

HODGE, G**. Classificação, taxonomia e ontologia**. DataGramaZero, Rio de Janeiro, v. 1, n. 5, out. 2000. Disponível em: http://www.dgz.org.br/out00/F_I_art.htm. Acesso em: 30 mar. 2023.

HOLZINGER, Andreas; STOCKER, Christof; DEHMER, Matthias. Big complex biomedical data: Towards a taxonomy of data. **Communications in Computer and Information Science**, 2014.

HOTHO, A.; STAAB, S.; STUMME, G. Ontologies improve text document clustering. In: **Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM 2005)**. IEEE, 2005, p. 541-544.

JACOBSON, I.; CHRISTERSON, M.; JONSSON, P.; ÖVERGAARD, G. **Object-Oriented Software Engineering**: A Use Case Driven Approach. 1. ed. São Paulo: Addison-Wesley, 1992.

JAIN, A. K.; DUBES, R. C**. Algorithms for clustering data**. Englewood Cliffs, NJ: Prentice Hall, 1988.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: a review. **ACM Computing Surveys**, v. 31, n. 3, p. 264-323, 1999.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An introduction to statistical learning**: with applications in R. New York: Springer, 2017.

JÄNICKE, S.; FRANZINI, G.; CHEEMA, M. F.; SCHEUERMANN, G. On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges. **Eurographics Conference on Visualization - State of the Art Reports**, EuroVis-STAR 2015, 2015.

JAVA EE. Introdução a Linguagem de Programação Java. s. d. Disponível em: https://www.devmedia.com.br/iniciando-na-linguagem-java/21136. Acesso em: 15 mar. 2023.

JAVA PLATFORM. Guia Completo de Java: Aprenda a Linguagem de Programação Java. Disponível em: https://www.devmedia.com.br/guia/linguagem-java/38169. Acesso em: 15 mar. 2023.

JOHNSON, A. Data Preprocessing Techniques. **Data Science Journal**, v. 15, n. 3, p. 87-102, 2020.

JOHNSON, L. **Ensuring Reliability and Validity in Qualitative Document Analysis**. Educational Researcher, 48(2), 171-185, 2019.

JOHNSON, Michael. **Mastering JavaScript**: The Ultimate Guide to Modern JavaScript Development. Packt Publishing, 2018.

JOHNSON, Thomas. **MySQL Workbench Data Modeling and Development**. McGraw-Hill Education, 2016.

JONES, A. et al. Choosing Between Document and Text Corpora in Computational Linguistics Research. **Journal of Computational Linguistics,** v. 25, n. 2, p. 45-62, 2020.

JONES, B. Fatores humanos e resiliência na indústria de óleo e gás. **Conferência Internacional de Segurança Industrial**, São Paulo, 2023.

JONES, L. et al. **Understanding Documents**: An Interdisciplinary Approach. Cambridge: Cambridge University Press, 2018.

JONES, R. Document Analysis: An Overview. In **Handbook of Qualitative Research Methods in Entrepreneurship** (pp. 123-140). Edward Elgar Publishing, 2015.

JURAFSKY, D.; MARTIN, J. H. Speech and Language Processing**: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition**. Pearson, 2019.

KALLEY, S. A. Big data, big challenges for evidence-based medicine**. Journal of General Internal Medicine**, v. 29, n. 3, p. 604-607, 2014.

KARAT, J. et al. Métodos de inspeção de usabilidade após 15 anos de pesquisa e prática. In **Anais da conferência SIGCHI sobre fatores humanos em computação**, p. 383-390, 2004.

KASTER, Gerson Bovi. **Framework conceitual baseado em aprendizagem de máquina supervisionada para concepção de sistemas de agentes inteligentes para área judicial**. 2021. 112 p. Dissertação (Mestrado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, Florianópolis, 2021. Disponível em: <https://bu.ufsc.br/teses/PEGC0682-D.pdf>. Acesso em 18 de out. de 2021.

KELLEHER, J. D.; TIERNEY, B. **Data science**: An introduction. Boca Raton: CRC Press, 2018.

KENNEDY, A.; MIKEL-JONES, A. **Corpus Linguistics and Variation in English**: Focus on Non-Native Englishes. Amsterdam: John Benjamins Publishing Company, 2018.

KIEFER, C. et al. Semantic Classification of Heterogeneous Web Content Using a Hybrid Approach. In**: INTERNATIONAL CONFERENCE ON ADVANCED INFORMATION SYSTEMS ENGINEERING**, 21., 2009, Amsterdam. Proceedings… Berlin: Springer, 2009. p. 267-280.

KIMBALL, R.; ROSS, M. **The Data Warehouse Toolkit**: The Definitive Guide to Dimensional Modeling. Hoboken: John Wiley & Sons, 2013.

KITCHIN, R.; LAURIAULT, T. Towards critical data studies: Charting and unpacking data assemblages and their work. **In The Programmable City Working Paper Series**, n. 16, p. 1-19, 2014.

KOCH, I. V. **Linguística Textual**: Introdução e Caminhos. 5. ed. São Paulo: Contexto, 2018.

KONCHADY, M. **Text Mining Application Programming.** Charles River Media, 2001.

KOSARA, R.; MACKINLAY, J. Storytelling: The Next Step for Visualization. In**: IEEE Computer Graphics and Applications**, v. 33, n. 5, p. 44-50, 2016.

KRIPPENDORFF, K. **Content analysis: An introduction to its methodology.** 4th ed. Sage, 2018.

KUCHER, Kostiantyn; ENGSTROM, Nellie; AXELSSON, Wilma; SAVAS, Berkant; KERREN, Andreas. Visualization of Swedish News Articles: A Design Study. Proceedings of **the International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications,** 2024.

LACERDA, Mário Roberto Miranda. **Mapeamento da disposição individual de compartilhar conhecimento a partir dos níveis de consciência informados pela teoria e instrumento de Loevinger**. 2011. Dissertação (Mestrado) - Universidade Federal de Santa Catarina, Florianópolis, 2011.

LAM, W. A Taxonomy of Knowledge Organization Systems. **Knowledge Organization,** v. 45, n. 2, p. 97-110, 2018.

LAUDON, K. C.; LAUDON, J. P. **Sistemas de informação gerenciais**. 14. ed. São Paulo: Pearson Prentice Hall, 2016.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **Nature, 521(7553)**, p. 436-444, 2015.

LEE, Joseph. JavaScript Performance Best Practices. **Smashing Magazine**, 2021.

LEECH, G. et al. **Corpus Annotation**: Linguistic Information from Computer Text Corpora. London: Longman, 1994.

LERDORF, Rasmus. **Personal Home Page Tools (PHP Tools)**. 1995. Disponível em: http://www.php.net/manual/en/history.php.php. Acesso em: 18 mar. 2023.

LI, X. et al. The automatic construction of thesaurus based on network analysis. **Journal of Intelligent & Fuzzy Systems**, v. 35, n. 3, p. 3443-3453, 2018.

LI, Xinyan; WANG, Han; CHEN, Chunyang; GRUNDY, John. An Empirical Study on How Well Do COVID-19 Information Dashboards Service Users' Information Needs. **IEEE Transactions on Services Computing,** 2022.

LI, Y. et al. Data Analysis Methods. **International Journal of Data Science and Analytics**, v. 4, n. 1, p. 56-72, 2021.

LIAO, C. et al. Evaluating the Classification Performance of Hierarchical Taxonomies. **IEEE Access**, v. 7, p. 90187-90195, 2019.

LIDDY, E. D.; TEREDSAI, A. Use of taxonomy in information retrieval. In: BATES, M. J.; MAACK, M. N. (Eds.). **Encyclopedia of library and information sciences.** 2nd ed. New York: Taylor & Francis, 2005. p. 2875-2882. Disponível em: https://www.tandfonline.com/doi/abs/10.1081/E-ELIS2-120043194. Acesso em: 30 mar. 2023.

LIE, Håkon Wium. **Cascading Style Sheets, level 1**. W3C Recommendation 17 Dec 1996. Disponível em: https://www.w3.org/Style/CSS/SAC/Overview.en.html. Acesso em: 17 mar. 2023.

LIMA, Claudio de. **Redes colaborativas como dinâmica de internacionalização da educação superior: um modelo para avaliar o potencial de compartilhamento de conhecimento**. 2023. Dissertação (Mestrado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, Florianópolis, 2023.

LINCOLN, Y. S.; GUBA, E. G. **Naturalistic inquiry**. Sage, 1985.

LIU, B. **Sentiment Analysis and Opinion Mining**. Morgan & Claypool Publishers, 2012.

LIU, B. **Sentiment analysis**: Mining sentiments, opinions, and emotions. Cambridge University Press, 2015. 9.

LIU, M.Y.; LUO, X.W.; WANG, G.B.; LU, W.Z. Intelligent Information Extraction from Government On-Site Inspection Reports of Construction Projects: A Graph-Based Text Mining Approach. **Advanced Engineering Informatics**, 2023.

LIU, Shixia et al. Bridging Text Visualization and Mining: A Task-Driven Survey. **IEEE Transactions on Visualization and Computer Graphics**, 2019.

LIU, T.; AHMED, D. Bangash; BOUALI, F.; VENTURINI, G. Visual and Interactive Exploration of a Large Collection of Open Datasets. **Proceedings of the International Conference on Information Visualisation**, 2013.

LIU, T.; BOUALI, F.; VENTURINI, G. Visual and interactive analysis of a large collection of open data with the relative neighborhood graph. **ACM International Conference Proceeding Series,** 2013.

LIU, Tianyang; BOUALI, Fatma; VENTURINI, Gilles. EXOD: A tool for building and exploring a large graph of open datasets. **Computers and Graphics (Pergamon)**, 2014.

LIU, Y.; LI, H. A survey on approaches to building taxonomies**. Knowledge-Based Systems**, v. 81, p. 35-47, 2015. Disponível em: https://www.sciencedirect.com/science/article/pii/S0950705115001061. Acesso em: 30 mar. 2023.

LOPER, E. & BIRD, S. NLTK: The natural language toolkit. In Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics-Volume 1 (pp. 63-70). **Association for Computational Linguistics**, 2002.

LOPES, L. M.; BARBOSA, V. A. **Desenvolvimento de taxonomias em ambientes digitais**: um estudo de caso. Informação & Sociedade: Estudos, v. 29, n. 2, p. 57-68, 2019.

LUO, T. et al. A Framework for Managing and Querying Multimedia Annotations Using Semantic Web Technologies. **Multimedia Tools and Applications**, v. 74, n. 1, p. 49-71, 2015.

LUZ, Gabriel. A Era do Big Data. **Medium**, 11 de maio de 2019. Disponível em: https://medium.com/gabriel-luz/a-era-do-big-data-64ebad5859f2. Acesso em: 26 set. 2023.

LUZ, Saturnino; SHEEHAN, Shane. Methods and visualization tools for the analysis of medical, political and scientific concepts in Genealogies of Knowledge. **Palgrave Communications**, 2020.

MACHADO, J. A. **WampServer**: instalação e configuração. 2019. Disponível em: <https://www.devmedia.com.br/wampserver-instalacao-e-configuracao/40678>. Acesso em: 16 abr. 2023.

MAI, J. E. **Information and Knowledge Organization**. Routledge, 2015.

MALIK, Sana et al. Cohort comparison of event sequences with balanced integration of visual analytics and statistics. International Conference on Intelligent User Interfaces, **Proceedings IUI**, 2015.

MANICA, Heloise**. Modelo de recuperação e comunicação de conhecimento em emergência médica com utilização de dispositivos portáteis**. 2012. Tese (Doutorado) - Universidade Federal de Santa Catarina, Florianópolis, 2012.

MANNING, C. D. **Foundations of statistical natural language processing**. MIT press, 2018.

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. **Introduction to Information Retrieval**. Cambridge University Press, 2008.

MARCH, S. T.; SMITH, G. F. Design and natural science research on information technology. **Decision support systems**, v. 15, n. 4, p. 251-266, 1995.

MARCH, S. T.; SMITH, G. F.; HINRICHS, K. Design Science Research Methodology: Framework and Outcomes. In: **Innovations in Information Systems Modeling** (pp. 125-143). Springer, 2016.

MARCONDES, C. H.; DEL NERO, A. Construção de taxonomias: uma revisão sistemática de literatura. **Perspectivas em Ciência da Informação**, v. 16, n. 4, p. 3-19, 2011. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-99362011000400002. Acesso em: 30 mar. 2023.

MARCUSCHI, L. A. **Produção Textual, Análise de Gêneros e Compreensão**. São Paulo: Parábola Editorial, 2008.

MARRA, Rafael. **Espiral do Conhecimento**. Key Account Manager. [LinkedIn], 22 de agosto de 2018. Disponível em: https://www.linkedin.com/pulse/espiral-do-conhecimento-rafael-marra/?trackingId=2suvpRbVQDOdPaXP%2FG%2FASA%3D%3D. Acesso em: 26 set. 2023.

MAYER-SCHÖNBERGER, Viktor; CUKIER, Kenneth. **Big Data**: A Revolução Que Transformará Como Vivemos, Trabalhamos e Pensamos. Rio de Janeiro: Elsevier, 2013.

MCAFEE, A.; BRYNJOLFSSON, E. Big data: the management revolution**. Harvard Business Review,** v. 90, n. 10, p. 61-67, 2012.

MCCANDLESS, T.; RASKAR, R.; RAMANI, K. V. Stitching worlds: Scalable infrastructure for interactive visual Analysis. In **IEEE Computer Graphics and Applications**, v. 34, n. 2, p. 20-27, 2014.

McENERY, T.; WILSON, A. **Corpus Linguistics**. Edinburgh: Edinburgh University Press, 1996.

MDN WEB DOCS. **JavaScript**. Disponível em: https://developer.mozilla.org/pt-BR/docs/Web/JavaScript. Acesso em: 16 abr. 2023.

MELGAR SASIETA, Héctor Andrés. **Um Modelo para a visualização de conhecimento baseado em imagens semânticas**. Florianópolis, 2011. 207 p. Dissertação (Mestrado) - Universidade Federal de Santa Catarina, Centro Tecnológico. Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento. Disponível em: <http://www.tede.ufsc.br/teses/PEGC0232-T.pdf>. Acesso em 18 de out. de 2021.

MENCZER, F.; AHN, Y.-Y. Content and Topical Analysis of Weblogs. In: **KOBAYASHI, N.; WU, X. (Eds.). Web Mining and Social Networking**: Techniques and Applications. Hershey: IGI Global, 2011. p. 303-332.

MERRIAM, Sharan B. **Qualitative research**: A guide to design and implementation. John Wiley & Sons, 2009.

MILES, M. B.; HUBERMAN, A. M. **Qualitative Data Analysis**: A Methods Sourcebook. SAGE Publications, 2014.

MIRANDA JUNIOR, J.; NEUMANN POTRICH, L.; LEOMAR TODESCO, J.; SELL, D.; JADER TRIERVEILER, H. Identificando conhecimentos críticos para fortalecer a capacidade de resposta resiliente. **Anais do Congresso Internacional de Conhecimento e Inovação – CIKI**, [S. l.], v. 1, n. 1, 2023. DOI: 10.48090/ciki.v1i1.1299. Disponível em: https://proceeding.ciki.ufsc.br/index.php/ciki/article/view/1299. Acesso em: 19 abr. 2024.

MITCHELL, T. **Machine learning**. McGraw Hill, 1997.

MORAES, R. **Análise textual discursiva**. Unijuí, 2016.

MORAIS, Edison Andrade Martins; AMBRÓSIO, Ana Paula L. **Mineração de Textos**: Relatório Técnico. 2007. Disponível em: <https://ww2.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_005-07.pdf>.

MORENO, P. A.; LUCAS, P. J. F.; OLIVEIRA, M. C. F. Visualização de dados: Evolução, técnicas e desafios. In **Revista de Informática Teórica e Aplicada**, v. 26, n. 2, p. 33-52, 2019.

MOTTA, J. M. C. de G. **Engenharia do conhecimento**: aspectos conceituais. Niterói: Eduff, 1998.

MOURA, Karina. **Ciclo de Vida dos Dados #1**: KDD Process. 2019. Disponível em: https://medium.com/@kvmoura/kdd-process-9b8e3062142. Acesso em: 28 set. 2023.

MURPHY, K. P. **Machine learning**: a probabilistic perspective. Cambridge: MIT Press, 2012.

MYSQL**. Overview of MySQL**. 2021. Disponível em: <https://www.mysql.com/about/>. Acesso em: 17 abr. 2023.

NAPOLI, Marcio. **Aplicação de ontologias para apoiar operações analíticas sobre fontes estruturadas e não estruturadas**. 2011. Dissertação (Mestrado) - Universidade Federal de Santa Catarina, Florianópolis, 2011.

NASCIMENTO, M. A.; GARCIA, A. M. Gestão do conhecimento: uma revisão sistemática de literatura. **Perspectivas em Gestão & Conhecimento**, João Pessoa, v. 4, n. 1, p. 39-57, jan./jun. 2014. Disponível em: https://periodicos.ufpb.br/index.php/pgc/article/view/18904/10446. Acesso em: 30 mar. 2023.

NAUGHTON, Patrick; GOSLING, James. **Java**: História e Principais Conceitos. Disponível em: https://www.devmedia.com.br/java-historia-e-principais-conceitos/25178. Acesso em: 15 mar. 2023.

NEUMANN, R.S.; KUMAR, S.; HAVERKAMP, T.H.A.; SHALCHIAN-TABRIZI, K. BLASTGrabber: A Bioinformatic Tool for Visualization, Analysis and Sequence Selection of Massive BLAST Data. **Bioinformatics**, 2014.

NICOLINI, Aline Torres. **A contribuição da análise do contexto organizacional na concepção de sistemas baseados em conhecimento: tecnologia KMAI®**. 2006. Dissertação (Mestrado) - Universidade Federal de Santa Catarina, Florianópolis, 2006.

NONAKA, I.; TAKEUCHI, H. **Criação de conhecimento na empresa**: como as empresas japonesas geram a dinâmica da inovação. Rio de Janeiro: Elsevier, 1997.

NORAMBUENA, Brian Felipe Keith; MITRA, Tanushree; NORTH, Chris. A survey on event-based news narrative extraction. **ACM Computing Surveys**, 2023.

NOY, N. F.; MCGUINNESS, D. L. Ontology Development 101: A Guide to Creating Your First Ontology. **Stanford Knowledge Systems Laboratory Technical Report** KSL-01-05, 2001. Disponível em: http://protege.stanford.edu/publications/ontology_development/ontology101.pdf. Acesso em: 30 mar. 2023.

OLIVEIRA, E. C. M.; MEDEIROS, B. A. M. Tesauros. In: **ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO**, 13., 2012, Rio de Janeiro. Anais [...]. Rio de Janeiro: ANCIB, 2012. p. 1-17.

OLIVEIRA, L. G. de. **Sistema de recomendação de meios de hospedagem baseado em filtragem colaborativa e informações contextuais**. 2007. Dissertação (Mestrado) - Universidade Federal de Santa Catarina, Centro Tecnológico. Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento.

OLIVEIRA, Thiago Paulo Silva de. **Sistemas baseados em conhecimento e ferramentas colaborativas para a gestão pública: uma proposta ao planejamento público local**. 2009. Dissertação (Mestrado) – Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-graduação em Engenharia e Gestão do Conhecimento, Florianópolis, 2009.

OLIVER, Paulo. **Ontologias**. InfoCon, 2016. Disponível em: https://paulooliver.wixsite.com/infocon/single-post/2016/06/21/ontologias. Acesso em: 26 set. 2023.

ORACLE. **MySQL Workbench**. 2020. Disponível em: https://www.mysql.com/products/workbench/. Acesso em: 15 mar. 2023.

ORTIGOSSA, Evandro S.; DIAS, Fábio Felix; DO NASCIMENTO, Diego Carvalho. Getting over High-Dimensionality: How Multidimensional Projection Methods Can Assist Data Science. **Applied Sciences (Switzerland)**, 2022.

PACHECO, J. A.; ATAIDE, R. C. **Compreensão em Leitura**: Uma abordagem cognitiva. São Paulo: Contexto, 2013.

PACHECO, L. & ATAIDE, M. **Interpretação de textos: teoria e prática**. São Paulo: Contexto, 2013.

PACHECO, Roberto Carlos dos Santos. Dados e Governo Abertos na Sociedade do Conhecimento. **LOD BRASIL Linked Open Data Brasil**. Florianópolis, 2014. Disponível em: <http://www.inf.ufsc.br/~jose.todesco/LODBrasil/Abertura/DadosEGovernoAbertoNaSocConh.pdf>. Acesso em 17 de out. de 2021.

PACHECO, Roberto Carlos dos Santos; SELIG, Paulo Mauricio; KERN, Vinicius Medina. **Seminários de Pesquisa EGC**. Florianópolis: Seminários Egc, 2021, color.

PACHECO, Rosimeri dos Santos; ATAIDE, Antonio Marcio. **Dificuldades de Intepretação de texto na escola – propostas de metodológicas para a superação deste problema**: trabalhando com fábulas e mitos. 2013. Disponível em: < http://www.diaadiaeducacao.pr.gov.br/portals/cadernospde/pdebusca/producoes_pde/2013/2013_unioeste_port_artigo_rosimeri_dos_santos_pacheco.pdf>. Acesso em 16 de out. de 2021.

PANAGIOTIDOU, Georgia; LAMQADDAM, Houda; POBLOME, Jeroen; BROSENS, Koenraad; VERBERT, Katrien; VANDE MOERE, Andrew. Communicating Uncertainty in Digital Humanities Visualization Research. **IEEE Transactions on Visualization and Computer Graphics**, 2023.

PANG, B., LEE, L. Opinion mining and sentiment analysis. **Foundations and Trends in Information Retrieval**, v. 2, n. 1-2, p. 1-135, 2008.

PEFFERS, K. et al. A design science research methodology for information systems research. **Journal of management information systems**, v. 24, n. 3, p. 45-77, 2007.

PEFFERS, K. et al. The design science research process: A model for producing and presenting information systems research. In **Proceedings of the First International Conference on Design Science Research in Information Systems and Technology,** p. 83-106, 2006.

PEFFERS, Ken; TUUNANEN, Tuure; ROTHENBERGER, Marcus A.; CHATTERJEE, Samir. A design science research methodology for information systems research. **Journal of Management Information Systems**, v. 24, n. 3, p. 4577, 2007. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.535.7773&rep=rep1&type=pdf>. Acesso em: 15 out 2021.

PEREIRA, Kariston. **O Raciocínio abdutivo no jogo de xadrez: a contribuição do conhecimento, intuição e consciência da situação para o processo criativo**. 2010. Tese (Doutorado) - Universidade Federal de Santa Catarina, Florianópolis, 2010.

PETROBRAS. **Relatório de sustentabilidade 2020.** Disponível em: https://petrobras.com.br/documents/d/f3a44542-113e-11ee-be56-0242ac120002/relatorio-de-sustentabilidade-2020-petrobras?download=true. Acesso em: 30 jan. 2024.

PHP GROUP. PHP: Hypertext Preprocessor. 2015. Disponível em: https://www.php.net/. Acesso em: 18 mar. 2023.

POLANYI, M. **The tacit dimension**. Gloucester: Peter Smith Publisher Inc., 1967.

PRADO, M. L. et al. Revisão integrativa de literatura: um método de pesquisa para a enfermagem e saúde. **Revista Eletrônica de Enfermagem**, v. 20, p. 1-10, 2018.

PRESSMAN, R. S. **Engenharia de software**: uma abordagem profissional. 8. ed. Porto Alegre: AMGH, 2016.

PRESSMAN, R. S. **Engenharia de software**: uma abordagem profissional. 7ª ed. Porto Alegre: AMGH, 2010.

RADFORD, A.; NARASIMHAN, K.; SALIMANS, T.; SUTSKEVER, I. **Improving language understanding by generative pretraining**. Disponível em: URL: https://s3-us-west-2.amazonaws.com/openai-assets/researchcovers/languageunsupervised/language_understanding_paper.pdf. Acesso em: 06 fev. 2024.

RAMOS JÚNIOR, Hélio Santiago. **Uma ontologia para representação do conhecimento jurídico-penal no contexto dos delitos informáticos**. 2008. Dissertação (Mestrado) – Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Gestão do Conhecimento, Florianópolis, 2008.

RANGANATHAN, Karthika; BARANI, B.; GEETHA, T.V. A Tamil lyrics search and visualization system. **Lecture Notes in Computer Science**, 2013.

RECEPTION THEORY. **How Audiences Make Meaning of Media**. Media Theory, 2023. Disponível em: https://www.mediatheory.net/reception-theory/. Acesso em: 31 mai. 2024.

REIF, K.; MENEZES, N.; SUSIN, P.; ALMEIDA, K.; SANTOS, H. Inovação, aprendizagem e cultura na indústria de óleo e gás: uma análise sociológica das práticas em sistemas sociotécnicos complexos. **Anais do Congresso Internacional de Conhecimento e Inovação – CIKI**, [S. l.], v. 1, n. 1, 2023. DOI: 10.48090/ciki.v1i1.1291. Disponível em: https://proceeding.ciki.ufsc.br/index.php/ciki/article/view/1291. Acesso em: 19 abr. 2024.

REIS, R. **Geração Automática de Texto**: Conceitos e aplicações. Editora Campus, 2020.

RETTIG, M. Prototype fidelity and design complexity: implications for prototype usage and reusability. In **Proceedings of the INTERACT'94 and CHI'94 conference on Human factors in computing systems**, p. 565-571, 1994.

RIBEIRO JUNIOR, D. I. **Modelo de sistema baseado em conhecimento para apoiar processos de tomada de decisão em ciência e tecnologia**. 2010. Tese (Doutorado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, Florianópolis.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. "Why should I trust you?". **Explaining the predictions of any classifier**, 2020.

RIBEIRO, S. F. **Sistema de conhecimento para gestão documental no setor judiciário: uma aplicação no Tribunal Regional Eleitoral de Santa Catarina**. 2010. Dissertação

(Mestrado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, Florianópolis, 2010.

RICCI, F.; ROKACH, L.; SHAPIRA, B. Introduction to recommender systems handbook. **Springer**, 2011.

RODRIGUES, J. P. et al. Revisão integrativa de literatura: conceitos e passos para a sua elaboração. **Research, Society and Development**, v. 10, n. 4, p. 1-23, 2021.

RODRIGUES, L. R. et al. A utilização da revisão integrativa na engenharia de produção: uma análise da produção científica. **Gestão & Produção**, v. 26, n. 2, p. 1-18, 2019.

RODRIGUEZ, D. Resiliência organizacional em ambientes de alto risco. **Revista de Gestão de Riscos**, v. 8, n. 3, p. 70-82, 2020.

ROTHER, Rodrigo Garcia. **Processo para recuperar produtos de inteligência competitiva a partir da memória organizacional**. 2009. Dissertação (Mestrado) – Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-graduação em Engenharia e Gestão do Conhecimento, Florianópolis, 2009.

SALEHEEN, Shibli; LAI, Wei. UIWGViz: An architecture of user interest-based web graph visualization. **Journal of Visual Languages and Computing**, 2018.

SALLES, Bertholdo Werner. Desenvolvimento de uma base de conhecimento de casos clínicos de pacientes portadores de desordem temporomandibular, como forma de organização do conhecimento e auxilio no diagnóstico. 2009. Tese (Doutorado) – Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, Florianópolis, 2009.

SALLOUM, Said A.; AL-EMRAN, Mostafa; ABDEL MONEM, Azza; SHAALAN, Khaled. Using Text Mining Techniques for Extracting Information from Research Articles. In: **Intelligent Natural Language Processing**: Trends and Applications, 2017.

SALM JUNIOR, José Francisco. **Padrão de projeto de ontologias para inclusão de referências do novo serviço público em plataformas de governo aberto**. 2012. Tese (Doutorado) – Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, Florianópolis, 2012.

SANTOS, Cristina Souza. **O acesso ao conhecimento em sistemas inteligentes de gestão e análise estratégicas: uma aplicação na segurança pública**. 2006. Dissertação (Mestrado) - Universidade Federal de Santa Catarina, Florianópolis, 2006.

SANTOS, V. M. **Leitura e Compreensão Textual**: Uma Nova Abordagem. Rio de Janeiro: PUC-Rio, 2019.

SCHWEITZER, Fernanda. **Produção científica em área de construção interdisciplinar: educação a distância no Brasil**. 2010. Dissertação (Mestrado) – Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, Florianópolis, 2010.

SEBASTIANI, F. Machine learning in automated text categorization. **ACM computing surveys (CSUR)**, v. 34, n. 1, p. 1-47, 2002.

SEGEL, E.; HEER, J. Narrative Visualization: Telling Stories with Data. In: **IEEE Transactions on Visualization and Computer Graphics**, v. 16, n. 6, p. 1139-1148, 2010.

SHURKHOVETSKYY, G.; ANDRIENKO, N.; ANDRIENKO, G.; FUCHS, G. Data abstraction for visualizing large time series. **Computer Graphics Forum**, 2018.

SHUTTERSTOCK. Text mining concept with icons. Written resources, data mining, text extraction, algorithm, analysis, information, database, forecasting icons. Web vector infographic in minimal flat line style. 2023. Disponível em: https://www.shutterstock.com/pt/image-vector/text-mining-concept-icons-written-resources-2152022023/edit?chatId=db96996139ae41499a391ad02a9de9f0. Acesso em: 28 set. 2023.

SILVA, A. B. Avanços em Processamento de Linguagem Natural: Uma revisão crítica. **Revista de Inteligência Artificial,** 12(3), 45-62, 2021.

SILVA, Dhiogo Cardoso da. **Uma Arquitetura de business intelligence para processamento analítico baseado em tecnologias semânticas e em linguagem natural**. 2011. Dissertação (Mestrado) - Universidade Federal de Santa Catarina, Florianópolis, 2011.

SILVA, Humberto; SIQUEIRA, Alexandre de Oliveira; ARAUJO, Marcus; DORNELAS, Jairo. **Sejamos Pragmáticos**: Pesquisas em Sistemas de Informação com Relevância e Rigor. 2017. Disponível em: <https://www.researchgate.net/publication/322042557_Sejamos_Pragmaticos_Pesquisas_em_Sistemas_de_Informacao_com_Relevancia_e_Rigor>. Acesso em 28 out. 2021

SILVA, Madalena Pereira da. **Um modelo de gerenciamento da qualidade de experiência para a provisão de serviços cientes de contexto**. 2017. Tese (Doutorado) – Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, Florianópolis, 2017.

SILVA, P. R.; MELO, J. F. **Interpretação e Análise de Documentos**. Porto Alegre: Editora Sulina, 2020.

SILVERMAN, D. **Doing qualitative research**. Sage, 2013.

SINCLAIR, J. **Corpus, Concordance, Collocation**. Oxford: Oxford University Press, 1991.

SINDELAR, Tim; ZUCCHERATO, Allan. **The Definitive Guide to Symfony**. Apress, 2013.

SKLIAROVA, Iryna; KRAVTSOV, Oleg; HOLUBEV, Yuriy. **Web Application Development with PHP 4.0**. Wrox Press, 2007.

SMITH, J. **Corpora in Linguistics**: An Overview. Cambridge University Press, 2018

SMITH, J. Data Collection Methods. **Journal of Data Science**, v. 7, n. 2, p. 245-261, 2019.

SMITH, J. **Qualitative research methods in social sciences**. Sage Publications, 2010.

SMITH, J.; JOHNSON, A. **Advances in Knowledge Discovery in Text**: Techniques and Applications. Springer, 2020.

SMITH, J.; JONES, K.; BROWN, L. Taxonomy in Information Science. **Journal of Information Science**, v. 45, n. 2, p. 256-273, 2019.

SMITH, John. **Mastering MySQL Workbench**. Packt Publishing, 2018.

SMITH, R. S. **Writing a Requirements Document**. 2023.

SMITH, T.; JOHNSON, M. **Document Design for Technical Readers**. New York: Oxford University Press, 2015.

SNYDER, C.; WOZNIAK, J. **Rapid contextual design**: a how-to guide to key techniques for user-centered design. Morgan Kaufmann, 2010.

SOARES, L. F. **Processamento de Linguagem Natural**: Uma abordagem prática. Editora Novatec, 2018.

SOUZA, Luiz Fernando Spillere de. **Modelo de mineração de ideias utilizando técnicas de engenharia do conhecimento**. 2021. Tese (Doutorado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, Florianópolis, 2021.

SSM (Superintendência de Segurança Operacional e Meio Ambiente). Relatório de Investigação do Incidente de Explosão no FPSO Cidade de São Mateus em 11/02/2015. Disponível em: https://www.gov.br/anp/pt-br/assuntos/exploracao-e-producao-de-oleo-e-gas/seguranca-operacional/incidentes/relatorios-de-investigacao-de-incidentes-1/arquivos-relatorios-de-investigacao-de-incidentes/fpso-cidade-de-sao-mateus/relatorio-de-investigacao-fpso-cidade-de-sao-mateus.pdf. Acesso em: 12 mai. 2023.

STACK OVERFLOW. Java: o que é, linguagem e um Guia para iniciar na tecnologia. Disponível em: https://www.alura.com.br/artigos/java. Acesso em: 15 mar. 2023.

STACK OVERFLOW. WampServer. 2021. Disponível em: <https://stackoverflow.com/questions/tagged/wampserver>. Acesso em: 16 abr. 2023.

STOJANOVIC, L. et al. Knowledge engineering for the Semantic Web. **IEEE Intelligent Systems**, v. 16, n. 2, p. 26-34, 2001. Disponível em: https://ieeexplore.ieee.org/document/913582. Acesso em: 30 mar. 2023.

STRADIOTTO, César Ramirez Kejelin. **Método de construção de ontologias multilingues com associação de conceitos a objetos em espaço 3D**. 2011. Tese (Doutorado) – Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, Florianópolis, 2011.

STUBBS, M. **Text and corpus analysis**: Computer-assisted studies of language and culture. Blackwell, 1996.

STUBBS, M. **Words and phrases**: Corpus studies of lexical semantics. Blackwell, 2001.

STUDER, R. et al. Knowledge engineering: principles and methods. **Data & Knowledge Engineering**, v. 25, n. 1-2, p. 161-197, 1998. Disponível em: https://www.sciencedirect.com/science/article/pii/S0169023X98000487. Acesso em: 30 mar. 2023.

SUTTON, R. S.; BARTO, A. G. **Reinforcement learning**: An introduction. MIT press, 2018.

SUZEN, Neslihan; GORBAN, Alexander; LEVESLEY, Jeremy; MIRKES, Evgeny. **Semantic Analysis for Automated Evaluation of the Potential Impact of Research Articles**. 2021. 36 p. Disponível em: https://doi.org/10.48550/arXiv.2104.12869. Acesso em: 26 jan. 2024.

SVENONIUS, E. The Intellectual Foundation of Information Organization. **The MIT Press**, 2000.

SWIETLICKI, Laura; CUBAUD, Pierre. Overview Visualizations for Large Digitized Correspondence Collections: A Design Study. **Lecture Notes in Computer Science**, 2022.

TAN, P.; STEINBACH, M.; KUMAR, V. **Introdução ao Data Mining**. Porto Alegre: Bookman, 2005.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introdução à Mineração de Dados**. Pearson, 2017.

TAO, C. et al. An Ontology-driven Framework for Information Retrieval in Disaster Management. **Journal of Information Science**, v. 39, n. 2, p. 204-220, 2013.

TERRA, José Cláudio; GLINA, Deise Maria; FERRAZ, Maria Lúcia do Carmo Cruz. **Saúde mental e trabalho**: guia para os profissionais de saúde. São Paulo: Roca, 2012.

TERRA, Rodrigo. **O que é Análise de Dados?** Makerzine, 11/05/2023. Disponível em: https://www.makerzine.com.br/programacao/o-que-e-analise-de-dados/. Acesso em: 29 set. 2023.

TERVEEN, L.; HILL, W. C. E. A taxonomy of web search. **ACM SIGIR Forum**, v. 35, n. 2, p. 14-18, 2001. Disponível em: https://dl.acm.org/doi/abs/10.1145/384694.384698. Acesso em: 30 mar. 2023.

THOMPSON, P. The Cambridge Handbook of English Corpus Linguistics. **Cambridge University Press**, 2020.

TIBSHIRANI, R.; WAINWRIGHT, M.; HASTIE, T. Statistical learning with sparsity: the lasso and generalizations. **Chapman and Hall/CRC**, 2013.

TJPI. Tribunal de Justiça do Piauí. Lista de tipos de documentos. [s. d.]. Disponível em: <https://www.tjpi.jus.br/e-tjpi/lista_tipos_documentos.php>. Acesso em 16 de out. de 2021.

TORRACO, R. J. Writing integrative literature reviews: Guidelines and examples. **Sage publications**, 2010.

TUFTE, E. **The Visual Display of Quantitative Information**. Cheshire, Connecticut: Graphics Press LLC, 2001.

UFS-GISI. Definição das Ferramentas da Engenharia do Conhecimento. 2015. Disponível em: https://ecufs-gisi.blogspot.com/2015/01/definicao-das-ferramentas-da-engenharia.html. Acesso em: 28 set. 2023.

URIARTE, Flavia Maia da Nova. **Portal corporativo como canal para gestão do conhecimento**. 2006. Dissertação (Mestrado) - Universidade Federal de Santa Catarina, Florianópolis, 2006.

USCHOLD, M.; GRUNINGER, M. Ontologies and semantics for seamless connectivity. **ACM SIGMOD Record**, v. 33, n. 4, p. 15-18, 2004.

USCHOLD, M.; GRUNINGER, M. Ontologies: principles, methods and applications. **Knowledge Engineering Review**, v. 11, n. 2, p. 93-136, 1996. Disponível em: https://www.cambridge.org/core/journals/knowledge-engineering-review/article/ontologies-principles-methods-and-applications/407F44BD567048193ECC15C56DAB0A16. Acesso em: 30 mar. 2023.

VASWANI, A. et al. Attention is all you need. In: **Advances in neural information processing systems**, p. 5998-6008, 2017.

WAMPSERVER. Home. 2021. Disponível em: <https://www.wampserver.com/en/>. Acesso em: 16 abr. 2023.

WANG, J.; YU, L.; YU, L.; HAN, Q.; SHEN. Deep learning for content-based image retrieval: A comprehensive study. **ACM Computing Surveys** (CSUR), 53(2), p. 1-42, 2020.

WANG, K.; ZHANG, C.; CHEN, H.; YUE, Y.; ZHANG, W.; ZHANG, M.; QI, X.; FU, Z. Karst landscapes of China: patterns, ecosystem processes and services. **Landscape Ecology**, v. 34, p. 2743-2763, 2019.

WANG, L. et al. An ontology-based knowledge management system for product development. **Journal of Intelligent Manufacturing**, v. 29, n. 7, p. 1493-1505, 2018.

WANG, S. et al. A Domain-specific Taxonomy-based Framework for Enterprise Information Integration. **Journal of Systems and Software**, v. 80, n. 10, p. 1712-1727, 2007.

WANG, Y. et al. Taxonomy creation by domain experts and its applications: An approach based on the similarity measure. In: **2019 IEEE International Conference on Big Data (Big Data)**. IEEE, 2019. p. 3126-3133.

WANG, Y.; ZHANG, Q. Knowledge Discovery in Text: Methods, **Algorithms, and Applications**. Elsevier, 2022.

WATZLAWICK, P. et al. **Pragmática da Comunicação Humana**: Um Estudo dos Padrões, Patologias e Paradoxos da Interação. São Paulo: Cultrix, 2006.

WELLING, Luke; THOMSON, Laura. **PHP and MySQL Web Development**. Addison-Wesley, 2016.

WELTER, Márcio. **Método de identificação de padrões em discurso político a partir da descoberta de conhecimento**. 2021. Tese (Doutorado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, Florianópolis, 2021.

WHITTEMORE, R.; KNAFL, K. The integrative review: an updated methodology. **Journal of advanced nursing**, v. 52, n. 5, p. 546-552, 2005.

WHITTEMORE, R.; KNAFL, K. The integrative review: Updated methodology. **Journal of advanced nursing**, v. 52, n. 5, p. 541-552, 2005.

WICKHAM, Hadley. A Layered Grammar of Graphics. **Journal of Computational and Graphical Statistic**s, vol. 19, no. 1, 2010, p. 3-28

WILGES, Beatriz. **Um modelo para organização de documentos no contexto da memória organizacional.** 2014. 125 p. Tese (Doutorado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, Florianópolis, 2014. Disponível em: <https://bu.ufsc.br/teses/PEGC0346-T.pdf>. Acesso em 18 de out. de 2021.

WITTEN, Ian H.; FRANK, Eibe; HALL, Mark A. **Data Mining**: Practical Machine Learning Tools and Techniques. 4th ed. San Francisco: Morgan Kaufmann, 2016.

WONG, B.; THOMAS, J. Visual Analysis. In **Proceedings of the 2004 IEEE Symposium on Information Visualization**, p. 19-25, 2004.

WONG, D. The Wall Street Journal Guide to Information Graphics: The Dos and Don'ts of Presenting Data, Facts, and Figures. **W. W. Norton & Company**, 2019.

WU, S.; LI, Y.; LI, Y. Knowledge discovery in social media data: a survey. **Social Network Analysis and Mining**, v. 4, n. 1, p. 1-13, 2014.

YARMOHAMMADI, Saman; POURABOLGHASEM, Reza; CASTRO-LACOUTURE, Daniel. Mining implicit 3D modeling patterns from unstructured temporal BIM log text data. **Automation in Construction**, 2017.

YUAN, Jun; CHEN, Changjian; YANG, Weikai; LIU, Mengchen; XIA, Jiazhi; LIU, Shixia. A Survey of Visual Analytics Techniques for Machine Learning. **Computational Visual Media**, 2020.

ZAHAIRA, M. et al. Spark: Cluster computing with working sets. **HotCloud**, v. 10, p. 10-16, 2010.

ZEND. Zend Framework. 2010. Disponível em: https://framework.zend.com/. Acesso em: 18 mar. 2023.

ZHANG, Q.; GAO, Y. The Role of Taxonomies in Information Retrieval Systems. **Information Processing & Management**, v. 57, n. 1, p. 102075, 2020.

ZHANG, W.; DONG, Y.; DING, Y. Building Taxonomies Automatically from Keywords. In: **INTERNATIONAL CONFERENCE ON WORLD WIDE WEB**, 16., 2007, Banff. Proceedings. New York: ACM, 2007. p. 1053-1054.

APÊNDICE A – Protocolo de Revisão Integrativa

UNIVERSIDADE FEDERAL DE SANTA CATARINA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA E GESTÃO DO CONHECIMENTO

Protocolo para Revisão Integrativa[14]

1. Data

26 de novembro de 2022

2. Nomes dos pesquisadores / instituições

Pesquisadora: Tatiana Tozzi

3. Fundamentos teóricos da pesquisa:

Text mining é a área da ciência da computação que se dedica a extrair informações relevantes de textos não estruturados, como e-mails, artigos, postagens em redes sociais, entre outros. Essas informações podem ser relacionadas a sentimentos, opiniões, tópicos mais discutidos, entre outros aspectos que podem ser úteis para empresas e pesquisadores.

Taxonomy, o termo "taxonomia" refere-se à classificação e organização sistemática de elementos ou conceitos dentro de um determinado domínio. É uma abordagem para organizar informações de forma hierárquica, onde os elementos são agrupados com base em características comuns. A taxonomia é frequentemente utilizada para criar uma estrutura que facilite a compreensão, a organização e a busca de informações.

Information visualization é a área que se dedica a representar informações e dados de forma visual, por meio de gráficos, mapas, diagramas, entre outras representações visuais. Essa área é importante para a compreensão de grandes conjuntos de dados e para a

[14] Adaptado de KITCHENHAM *Protocol for Systematic Review*, 2006.

comunicação de informações complexas de forma clara e objetiva. Além disso, a informação visualização é utilizada em diversas áreas, como jornalismo, publicidade, ciência, entre outras.

4. **Questão de pesquisa:**

O que tem sido publicado sobre *text mining*, *taxonomy* e *information visualization*, no período de 2013 a 2024?

5. Bases de dados consultadas: liste as bases de dados que serão pesquisadas. Liste revistas ou websites que serão pesquisados.

Base de dados	**Site**
IEEEXplore	https://ieeexplore.ieee.org/Xplore/home.jsp
SpringerLink	https://link.springer.com/
WebScience	https://www.webofscience.com/wos/woscc/basic-search
Scopus	https://www.scopus.com/home.uri
Science Direct	https://www.sciencedirect.com/

6. Critérios de inclusão e exclusão:

Critérios de inclusão	**Critérios de exclusão**
- Artigos originais publicados nas bases de dados escolhidas neste estudo; - Artigos publicados nos idiomas português e inglês; - Artigos conforme a combinação dos descritores ou palavras chaves selecionadas para este estudo, no resumo e/ou título; - Estudos publicados no período de 2013 a 2025.	- Estudos publicados em outros meios de comunicação que não sejam periódicos científicos; - Artigos duplicados em bases; - Artigos que não estão disponibilizados no formato completo para análise e estudos que não respondam à questão de pesquisa; - Artigos que não se enquadram aos descritos e/ou palavras chave; - Artigos publicados anteriormente ao período escolhido.

7. Estratégias de busca:

Base de dados	**Estratégia de busca**	**Qtde.**
IEEEXplore	("Document Title":"text mining") AND ("Document Title":taxonomy) AND ("Document Title":"data visualization") AND ("Document Title":"information visualization") AND ("Abstract":"text mining") AND ("Abstract":taxonomy) AND ("Abstract":"data visualization”) AND ("Abstract":"information visualization") e ("All Metadata":"text mining") AND ("All Metadata":taxonomy) AND ("All Metadata":"data visualization") AND ("All Metadata":"information visualization")	0
SpringerLink	*"text mining" AND taxonomy AND "data visualization" AND "information visualization"- 2013 - 2023*	*10*
WebScience	*(((ALL=(text mining)) AND ALL=(taxonomy)) AND ALL=(data visualization)) AND ALL=(information visualization) - Timespan: 2013-01-01 to 2023-12-31 (Publication Date)*	*4*
Scopus	*"text mining" AND taxonomy AND "data visualization" AND "information visualization" AND PUBYEAR > 2013 AND PUBYEAR < 2024 AND (LIMIT-TO (DOCTYPE , "ar") OR LIMIT-TO (DOCTYPE , "cp"))*	*61*
Science Direct	*"text mining" AND taxonomy AND "data visualization" AND "information visualization" – 2013-2024*	*6*
	Total	*81*

8. Critérios de qualidade para seleção dos artigos

- Artigos de acesso livre;
- Títulos dos artigos dentro dos critérios de busca;
- Resumo dos artigos dentro dos critérios de busca;
- Artigos cujo objetivo geral e/ou específico refere-se aos objetos deste estudo.

9. Estratégias de extração dos dados:

Os dados dos artigos serão exportados utilizando o formato .ris, comumente utilizado para ser utilizados no Gerenciador de Bibliografia EndNote.

10. Estratégias de análise dos dados:

Os dados serão classificados, conforme:

Seleção	Exclusão
Resumo Aderente	Não Aderente
Título Aderente	

E classificados utilizando *Tags* (etiquetas) para facilitar a localização das temáticas dos artigos.

11. Estratégia de disseminação do conhecimento:

Esperamos que os resultados desta Revisão Integrativa possam resultar na publicação de um artigo em um evento científico ou em uma revista especializada.

12. Cronograma das atividades

	Período	
Atividade	Outubro-	Janeiro-

	Dezembro/2022			Fevereiro/2023		
Pré-projeto						
Elaboração do protocolo						
Busca dos estudos						
Seleção dos estudos						
Organização dos estudos						
Avaliação crítica dos estudos						
Discussão e Conclusão						

APÊNDICE B – Desenvolvimento da Revisão Integrativa

Identificação do problema

Na primeira fase foi identificado o problema, o qual foi percebido por meio de uma pesquisa em algumas bases de dados, assim observamos como os temas de interseção entre *Text mining*, *machine learning*, *information visualization* e *taxonomy* são pouco abordados em pesquisas e publicações. Temos como tema a interseção entre três áreas de pesquisa: *Text Mining*, *Machine Learning* e *Information Visualization*. Portanto, o tema geral é a integração dessas três disciplinas e suas aplicações conjuntas no período de 2017 a 2022. Logo a questão de pesquisa é: "O que tem sido publicado sobre *text mining*, e *information visualization*, no período de 2013 a 2024?"

Busca na literatura

Em seguida, foi iniciada uma busca na literatura a qual foi realizada inicialmente selecionado as quatro bases de publicações para a recuperação dos artigos, foram utilizadas nesta pesquisa as bases: IEEEXplore, SpringerLink, WebScience, Scopus e Science Direct.

Estas bases foram escolhidas por dois motivos:

1. A facilidade que as bases possibilitam de customizar as *strings* de busca, podendo assim obter resultados mais precisos e relevantes para responder à questão de pesquisa.

2. A grande quantidade de publicações presentes em cada base, resultando que tivéssemos mais resultados.

As palavras-chaves (*strings*) utilizadas "*text mining*", *taxonomy*, "*data visualization*", "*information visualization*". Para um artigo ser capturado pelo filtro de busca ele deveria conter as duas *strings*, presentes no título, resumo ou palavras-chave, no Quadro 13 são apresentadas as *strings* de busca utilizadas em cada busca.

Quadro 13 - *Strings* de busca utilizadas

Base de dados	Estratégia de busca
IEEEXplore	("Document Title":"text mining") AND ("Document Title":taxonomy) AND ("Document Title":"data visualization") AND ("Document Title":"information visualization") AND ("Abstract":"text mining") AND

Base de dados	Estratégia de busca
	("Abstract":taxonomy) AND ("Abstract":"data visualization") AND ("Abstract":"information visualization") e ("All Metadata":"text mining") AND ("All Metadata":taxonomy) AND ("All Metadata":"data visualization") AND ("All Metadata":"information visualization")
SpringerLink	"text mining" AND taxonomy AND "data visualization" AND "information visualization"- 2013 – 2023
WebScience	**(((ALL=(text mining)) AND ALL=(taxonomy)) AND ALL=(data visualization)) AND ALL=(information visualization) - Timespan: 2013-01-01 to 2023-12-31 (Publication Date)**
Scopus	"text mining" AND taxonomy AND "data visualization" AND "information visualization" AND PUBYEAR > 2013 AND PUBYEAR < 2024 AND (LIMIT-TO (DOCTYPE , "ar") OR LIMIT-TO (DOCTYPE , "cp"))
	"text mining" AND taxonomy AND "data visualization" AND "information visualization" Published from 2017 - to Present– 2013-2024
Science Direct	("Document Title":"text mining") AND ("Document Title":taxonomy) AND ("Document Title":"data visualization") AND ("Document Title":"information visualization") AND ("Abstract":"text mining") AND ("Abstract":taxonomy) AND ("Abstract":"data visualization") AND ("Abstract":"information visualization") e ("All Metadata":"text mining") AND ("All Metadata":taxonomy) AND ("All Metadata":"data visualization") AND ("All Metadata":"information visualization")
	"text mining" AND taxonomy AND "data visualization" AND "information visualization"- 2013 - 2023

Fonte: Elaborado pela autora.

Na base IEEEXplore mesmo utilizando a diferentes *strings* e sem a limitação do limite de tempo não foram encontradas publicações. A análise se limitou a buscar artigos publicados em revistas científicas, em inglês, português ou espanhol, no período contido entre 2013 e 2024. Ao todo foram recuperados 81 (oitenta e um) artigos das bases de dados. A contribuição de cada base é detalhada na Tabela 1.

Tabela 1 - Contribuição de cada Base de dados

Base de dados	**Número de artigos**
IEEEXplore	0
SpringerLink	10
WebScience	4
Scopus	61
Science Direct	6
Total	81

Fonte: Elaborado pela autora.

Avaliação dos dados

Uma vez realizadas as pesquisas nas bases, os resultados foram selecionados e exportados utilizando o formato de arquivo .RIS (*Information Systems Research*), os quais foram armazenados no Mega[15], na sequência os arquivos das bases foram importados individualmente no sistema Web Academical[16]. O Academical é um sistema *online*, que auxilia no desenvolvimento de pesquisas integrativas e bibliométricas e possibilita que múltiplos pesquisadores participem simultaneamente na análise dos artigos, classificação, etiquetação e exploração dos mesmos (Academical, [s.d.]). Foram criadas duas etapas para classificação e exclusão dos artigos:

- **1ª Etapa:** depois que todos os artigos foram carregados no sistema, a primeira fase foi aplicada os critérios de classificação e exclusão buscando identificar apenas os artigos que continham de fatos ambos termos presentes no título ou resumo.
- **2ª Etapa:** foi realizado o refinamento da pesquisa, reanalisados dos artigos que tínhamos dúvidas, e separados dos artigos que apresentavam apenas um dos termos.

No sistema foram criados os seguintes critérios para a classificação e exclusão de artigos:

[15] https://mega.nz/
[16] https://app.academical.com.br/

Classificação (critérios de inclusão):

- Título e resumo aderentes: quando o título e o resumo indicavam uma linha de pesquisa aderente ao tema de interesse desta pesquisa.
- Analisar: presente na primeira etapa, quando o artigo deveria ser classificado, porém realizado na segunda etapa.

Exclusão (critérios de exclusão):

- Anterior ao período: quando o artigo estava fora do período (critério não utilizado).
- Duplicado: quando o artigo estava presente em duas bases simultâneas.
- Não aderente: quando nem título nem o resumo indicavam uma linha de pesquisa aderente ao tema de interesse deste artigo.

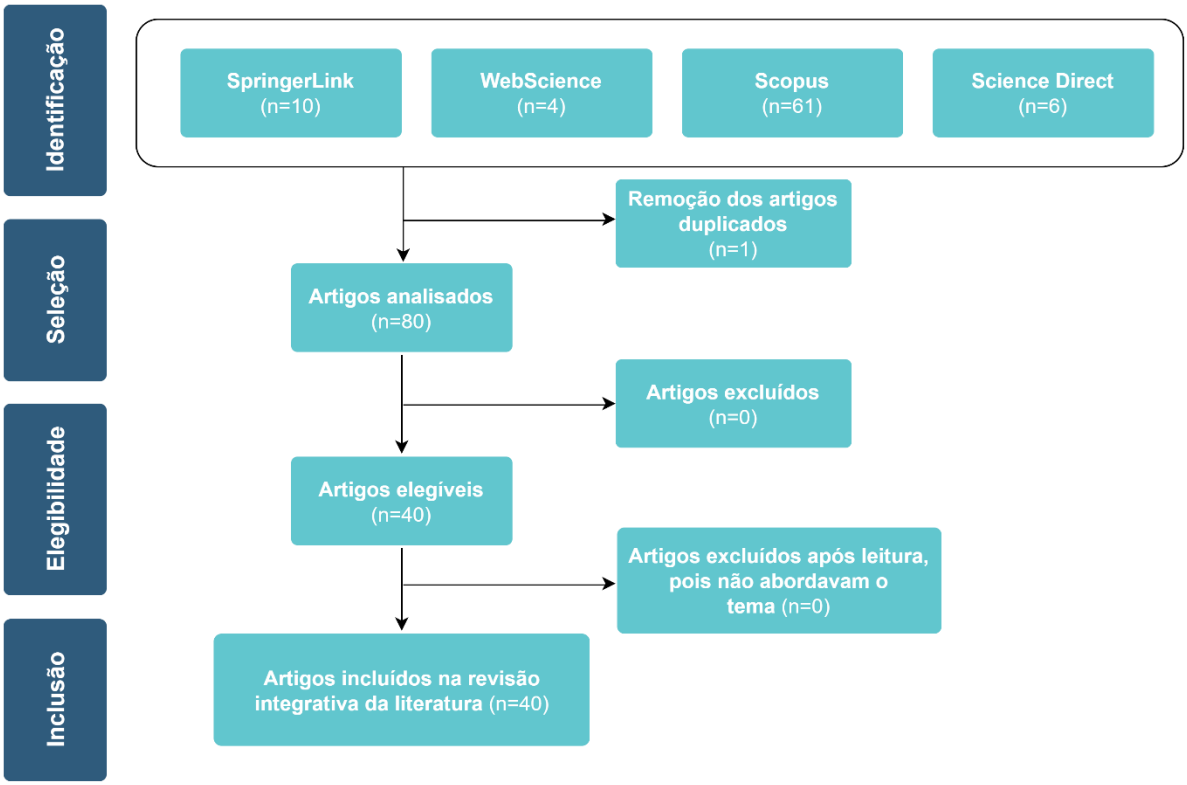
A seleção final dos artigos recuperados na pesquisa e considerados aderentes ao escopo deste trabalho resultou em 40 (quarenta) artigos, presentes nas bases de dados selecionada os quais foram acessados utilizando a Rede CAFe[17] (Comunidade Acadêmica Federada).

Para extrair as informações do título, autores, local de publicação, resumo e palavras-chave, foi utilizada a funcionalidade Análise Bibliométrica presente no sistema Academical, o qual foi salvo no Mega[18] (serviço de armazenamento de dados e transferências em nuvem) da pesquisa.

Os artigos elegíveis foram analisados, buscando identificar quais ocorriam de fato a interseção entre *Text mining*, *information visualization* e *taxonomy*, a Figura 11, apresenta o fluxograma referente ao processo de seleção dos artigos que compõem este estudo, conforme recomenda o PRISMA (*Preferred Reporting Items for Systematic Reviews and Meta-Analyses*).

[17] https://www.gov.br/inpe/pt-br/area-conhecimento/biblioteca/acervo-digital/acesso-cafe
[18] https://mega.io/

Figura 38 - Fluxograma de amostragem da revisão integrativa



Fonte: Elaborado pela autora.

Os artigos que foram analisados nesta pesquisa foram publicados em 40 (quarenta) revistas diferentes, sendo que no resultado geral dos artigos, ou seja, na análise das etapas dos 81 artigos essas revistas tiveram múltiplas publicações. A Tabela 2, elenca os periódicos dos artigos considerados nesta pesquisa.

Tabela 2 - Periódicos que contém os artigos selecionados

Título do artigo	**Periódico**
Best-Practice Piloting Based on an Integrated Social Media Analysis and Visualization for E-Participation Simulation	*Procedia Computer Science, Volume 75, 2015, Pages 66-74*
Applications of Natural Language Techniques to Enhance Curricular Coherence	*Procedia Computer Science, Volume 168, 2020, Pages 88-96*
On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges	*Eurographics Conference on Visualization - State of the Art Reports, EuroVis-STAR 2015*
A taxonomy of Twitter data analytics techniques	*Proceedings of the 32nd International*

	Business Information Management Association Conference, IBIMA 2018 - Vision 2020
Visualization of Swedish News Articles: A Design Study	*Proceedings of the International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*
Mining implicit 3D modeling patterns from unstructured temporal BIM log text data	*Automation in Construction*
Supporting Story Synthesis: Bridging the Gap between Visual Analytics and Storytelling	*IEEE Transactions on Visualization and Computer Graphics*
Getting over High-Dimensionality: How Multidimensional Projection Methods Can Assist Data Science	*Applied Sciences (Switzerland)*
Tools and Techniques forBridging Text Visualization and Mining: A Task-Driven Survey	*IEEE Transactions on Visualization and Computer Graphics*
Rule-based Visual Mappings - With a Case Study on Poetry Visualization	*Computer Graphics Forum*
Big complex biomedical data: Towards a taxonomy of data	*Communications in Computer and Information Science*
Visual and Interactive Exploration of a Large Collection of Open Datasets	*Proceedings of the International Conference on Information Visualisation*
Using Text Mining Techniques for Extracting Information from Research ArticlesA comprehensive review of visualization methods for association rule mining: Taxonomy, challenges, open problems	*Expert Systems with Applications*
Visualizing patterns of appraisal in texts and corpora	*Text and Talk*
Methods and visualization tools for the analysis of medical, political and scientific concepts in Genealogies of Knowledge	*Palgrave Communications*
VisForum: A visual analysis system for exploring user groups in online forums	*ACM Transactions on Interactive Intelligent Systems*
Visual and interactive analysis of a large collection of open data with the relative neighborhood graph	*ACM International Conference Proceeding Series*
Cohort comparison of event sequences with balanced integration of visual analytics and statistics	*International Conference on Intelligent User Interfaces, Proceedings IUI*
Twitter sentiment analysis approaches: A survey	*International Journal of Emerging*

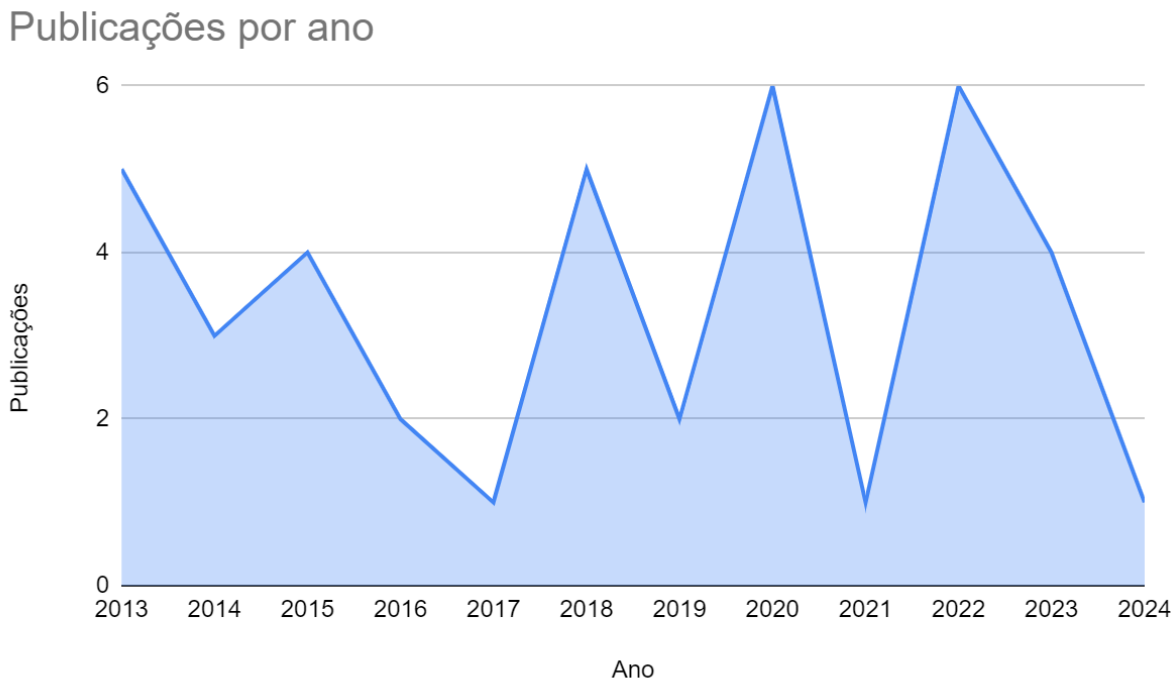
	Technologies in Learning
Beyond Explanation: A Case for Exploratory Text Visualizations of Non-Aggregated, Annotated Datasets	*1st Workshop on Perspectivist Approaches to Disagreement in NLP, NLPerspectives 2022 as part of Language Resources and Evaluation Conference, LREC 2022 Workshop*
Visual data mining in software repositories: A survey	*Frontiers in Artificial Intelligence and Applications*
A Survey on Event-Based News Narrative Extraction	*ACM Computing Surveys*
Semantics Visualization as a User Interface in Business Information Searching	*IFIP Advances in Information and Communication Technology*
The four dimensions of social network analysis: An overview of research methods, applications, and software tools	*Information Fusion*
EXOD: A tool for building and exploring a large graph of open datasets	*Computers and Graphics (Pergamon)*
An Empirical Study on How Well Do COVID-19 Information Dashboards Service Users' Information Needs	*IEEE Transactions on Services Computing*
Overview Visualizations for Large Digitized Correspondence Collections: A Design Study	*Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*
Usability of business information semantic network search visualization	*ACM International Conference Proceeding Series*
A Tamil lyrics search and visualization system	*Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*
Data abstraction for visualizing large time series	*Computer Graphics Forum*
Of Course it's Political! A Critical Inquiry into Underemphasized Dimensions in Civic Text Visualization	*Computer Graphics Forum*
Communicating Uncertainty in Digital Humanities Visualization Research	*IEEE Transactions on Visualization and Computer Graphics*
UIWGViz: An architecture of user interest-based web graph visualization	*Journal of Visual Languages and Computing*
Ontology-Based Interactive Visualization of Patient-Generated Research Questions	*Applied Clinical Informatics*
Analysis and data visualization in bibliometric studies; [Análisis y visualización de datos en	*JLIS.it*

estudios bibliométricos]	
Competitive analysis of online reviews using exploratory text mining	*Tourism and Hospitality Management*
A survey of visual analytics techniques for machine learning	*Computational Visual Media - Volume 7, pages 3–36, (2021)*
New Cybercrime Taxonomy of Visualization of Data Mining Process	*2016 39TH INTERNATIONAL CONVENTION ON INFORMATION AND COMMUNICATION TECHNOLOGY, ELECTRONICS AND MICROELECTRONICS (MIPRO) - Page349-351*
BLASTGrabber: a bioinformatic tool for visualization, analysis and sequence selection of massive BLAST data	*BIOINFORMATICS*
Intelligent information extraction from government on-site inspection reports of construction projects: A graph-based text mining approach	*ADVANCED ENGINEERING INFORMATICS*

Fonte: Elaborado pela autora.

Em relação ao período analisado, inicialmente esperávamos encontrar um maior número de publicações, por isso definimos o período temporal entre 2013 a 2024, porém no decorrer das buscas nas bases, observamos que poucos autores abordam todos os temas em conjunto, resultado assim a realização da pesquisa usando a *strings* simples para encontrar mais resultados relevantes a esta pesquisa. A Figura 39 apresenta como ocorreu a distribuição ao longo dos anos avaliados nesta pesquisa.

Figura 39 - Distribuição de Publicações por ano



Fonte: Elaborado pela autora.

Num recorte de autores, 127 autores que contribuíram para o objetivo das pesquisas. A Tabela 3, apresenta os autores de cada um dos artigos e o ano de publicação dos mesmos.

Tabela 3 - Autores dos artigos

Título do artigo	**Autores**	**Ano de publicação**
Best-Practice Piloting Based on an Integrated Social Media Analysis and Visualization for E-Participation Simulation in Cities,	Dirk Burkhardt, Kawa Nazemi, Egils Ginters	2015
Applications of Natural Language Techniques to Enhance Curricular Coherence	Adrian S. Barb, Nil Kilicay-Ergin	2020
On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges	Jänicke, S., Franzini, G., Cheema, M.F., Scheuermann, G.	2015
A taxonomy of Twitter data analytics techniques	Hamzah, Muzaffar, Vu, Tuong Thuy	2018
Visualization of Swedish News Articles: A	Kucher, Kostiantyn,	2024

Título do artigo	Autores	Ano de publicação
Design Study	Engstrom, Nellie, Axelsson, Wilma, Savas, Berkant, Kerren, Andreas	
Mining implicit 3D modeling patterns from unstructured temporal BIM log text data	Yarmohammadi, Saman, Pourabolghasem, Reza, Castro-Lacouture, Daniel	2017
Examining data visualization pitfalls in scientific publicationsSupporting Story Synthesis: Bridging the Gap between Visual Analytics and Storytelling	Chen, Siming, Li, Jie, Andrienko, Gennady, Andrienko, Natalia, Wang, Yun, Nguyen, Phong H., Turkay, Cagatay	2020
Getting over High-Dimensionality: How Multidimensional Projection Methods Can Assist Data Science	Ortigossa, Evandro S., Dias, Fábio Felix, Do Nascimento, Diego Carvalho	2022
Tools and Techniques forBridging Text Visualization and Mining: A Task-Driven Survey	Liu, Shixia, Wang, Xiting, Collins, Christopher, Dou, Wenwen, Ouyang, Fangxin, El-Assady, Mennatallah, Jiang, Liu, Keim, Daniel A.	2019
Rule-based Visual Mappings - With a Case Study on Poetry Visualization	Abdul-Rahman, A., Lein, J., Coles, K., Maguire, E., Meyer, M., Wynne, M., Johnson, C.R., Trefethen, A., Chen, M.	2013
Big complex biomedical data: Towards a taxonomy of data	Holzinger, Andreas, Stocker, Christof, Dehmer, Matthias	2014
Visual and Interactive Exploration of a Large Collection of Open Datasets	Liu, T., Ahmed, D. Bangash, Bouali, F., Venturini, G.	2013
Using Text Mining Techniques for Extracting Information from Research ArticlesA comprehensive review of visualization methods for association rule mining: Taxonomy, challenges, open problems	Fister, Iztok, Fister, Iztok, Fister, Dušan, Podgorelec, Vili, Salcedo-Sanz, Sancho	2023
Visualizing patterns of appraisal in texts and corpora	Almutairi, Bandar Alhumaidi A.	2013
Effective Natural Language Processing and Interpretable Machine Learning for Structuring CT Liver-Tumor	Luz, Saturnino, Sheehan, Shane	2020

Título do artigo	Autores	Ano de publicação
ReportsMethods and visualization tools for the analysis of medical, political and scientific concepts in Genealogies of Knowledge		
VisForum: A visual analysis system for exploring user groups in online forums	Fu, Siwei, Wang, Yong, Yang, Yi, Bi, Qingqing, Guo, Fangzhou, Qu, Huamin	2018
Visual and interactive analysis of a large collection of open data with the relative neighborhood graph	Liu, T., Bouali, F., Venturini, G.	2013
Cohort comparison of event sequences with balanced integration of visual analytics and statistics	Malik, Sana, Du, Fan, Monroe, Megan, Onukwugha, Eberechukwu, Plaisant, Catherine, Shneiderman, Ben	2015
Twitter sentiment analysis approaches: A survey	Adwan, Omar Y., Al-Tawil, Marwan, Huneiti, Ammar M., Shahin, Rawan A., Abu Zayed, Abeer A., Al-Dibsi, Razan H.	2020
Beyond Explanation: A Case for Exploratory Text Visualizations of Non-Aggregated, Annotated Datasets	Havens, Lucy, Bach, Benjamin, Terras, Melissa, Alex, Beatrice	2022
Visual data mining in software repositories: A survey	Eteläaho, Anna, Soini, Jari, Jaakkola, Hannu, Mattila, Anna-Liisa	2018
A Survey on Event-Based News Narrative Extraction	Keith Norambuena, Brian Felipe, Mitra, Tanushree, North, Chris	2023
Semantics Visualization as a User Interface in Business Information Searching	Dudycz, Helena	2021
The four dimensions of social network analysis: An overview of research methods, applications, and software tools	Camacho, David, Panizo-LLedot, Ángel, Bello-Orgaz, Gema, Gonzalez-Pardo, Antonio, Cambria, Erik	2020
EXOD: A tool for building and exploring a large graph of open datasets	Liu, Tianyang, Bouali, Fatma, Venturini, Gilles	2014
An Empirical Study on How Well Do COVID-19 Information Dashboards Service Users' Information Needs	Li, Xinyan, Wang, Han, Chen, Chunyang, Grundy, John	2022

Título do artigo	Autores	Ano de publicação
Overview Visualizations for Large Digitized Correspondence Collections: A Design Study	Swietlicki, Laura, Cubaud, Pierre	2022
Usability of business information semantic network search visualization	Dudycz, Helena	2015
A Tamil lyrics search and visualization system	Ranganathan, Karthika, Barani, B., Geetha, T.V.	2013
Data abstraction for visualizing large time series	Shurkhovetskyy, G., Andrienko, N., Andrienko, G., Fuchs, G.	2018
Of Course it's Political! A Critical Inquiry into Underemphasized Dimensions in Civic Text Visualization	Baumer, Eric P. S., Jasim, Mahmood, Sarvghad, Ali, Mahyar, Narges	2022
Communicating Uncertainty in Digital Humanities Visualization Research	Panagiotidou, Georgia, Lamqaddam, Houda, Poblome, Jeroen, Brosens, Koen	2023
UIWGViz: An architecture of user interest-based web graph visualization	Saleheen, Shibli; Lai, Wei	2018
Ontology-Based Interactive Visualization of Patient-Generated Research Questions	Borland, David; Christopherson, Laura; Schmitt, Charles	2019
Analysis and data visualization in bibliometric studies	Alhuay-Quispe, Joel; Estrada-Cuzcano, Alonso; Bautista-Ynofuente, Lourdes	2022
Competitive analysis of online reviews using exploratory text mining	Amadio, William J.; Procaccino, J. Drew	2016
A survey of visual analytics techniques for machine learning	Jun Yuan, Changjian Chen, Weikai Yang, Mengchen Liu, Jiazhi Xia & Shixia Liu	2020
New Cybercrime Taxonomy of Visualization of Data Mining Process	Babic, M.; Jerman-Blazic, B.	2016
BLASTGrabber: a bioinformatic tool for visualization, analysis and sequence selection of massive BLAST data	Neumann, RS; Kumar, S; Haverkamp, THA; Shalchian-Tabrizi, K	2014
Intelligent information extraction from government on-site inspection reports of construction projects: A graph-based text mining approach	Liu, MY; Luo, XW; Wang, GB; Lu, WZ	2023

Fonte: Elaborado pela autora.

APÊNDICE C – Tecnologias adotadas

Para o desenvolvimento do sistema DOC *Analysis* foram utilizadas diversas tecnologias, nas seções a seguir são apresentadas as tecnologias adotadas.

JAVA

A linguagem de programação Java, cuja história remonta ao Projeto Green em 1991, conduzido por Patrick Naughton e James Gosling, teve como propósito original o desenvolvimento de tecnologias inovadoras para a interconexão de dispositivos eletrônicos de consumo. Essa iniciativa culminou na criação do sistema operacional GreenOS, demonstrando a capacidade de avanços na comunicação entre dispositivos (NAUGHTON; GOSLING, 1991). Em um curioso acontecimento, a linguagem, inicialmente batizada de Oak, foi renomeada para Java devido à influência da cafeteria local onde a equipe frequentemente tomava café, servindo uma marca originária da região de Java (NAUGHTON; GOSLING, 1991). Esse nome não apenas reflete a jornada da linguagem, mas também simboliza sua interconexão e propósito.

O Java não se limita apenas a dispositivos eletrônicos, tornando-se uma plataforma essencial para o desenvolvimento de aplicações globais. De acordo com uma pesquisa recente, a linguagem Java é a sexta mais utilizada por desenvolvedores, representando cerca de 33% das preferências (STACKOVERFLOW, 2022). Sua estrutura orientada a objetos e a presença da Máquina Virtual Java (JVM) permitem a portabilidade dos programas, convertendo-os em bytecode para execução na máquina virtual (STACKOVERFLOW, 2022). A plataforma Java é organizada em diferentes edições, como Java SE (*Java Standard Edition*), Java ME (*Java Micro Edition*) e Java EE (*Java Enterprise Edition*), cada uma destinada a atender a diferentes necessidades de desenvolvimento (JAVA PLATFORM, 2023).

A independência de plataforma é uma característica marcante do Java, permitindo que programas desenvolvidos na linguagem sejam executados em diversos sistemas operacionais e hardwares, tornando-o uma escolha versátil para projetos multiplataforma (JAVA PLATFORM, 2023; JAVA EE, s.d.). Além disso, sua biblioteca padrão e as ferramentas disponíveis facilitam a distribuição e o desenvolvimento de programas, enquanto sua robustez e segurança garantem um ambiente propício para a criação de aplicações confiáveis (JAVA EE, s.d.; AWS, 2023).

PHP

PHP (*Hypertext Preprocessor*), é uma linguagem de programação amplamente utilizada para o desenvolvimento *web*, tem desempenhado um papel crucial na criação de aplicações dinâmicas e interativas. Desde sua criação por Rasmus Lerdorf em 1994, o PHP passou por várias evoluções e versões, tornando-se uma escolha popular para desenvolvedores de todo o mundo (LERDORF, 1995). Com sua sintaxe simples e flexibilidade, o PHP permite aos desenvolvedores criar sites dinâmicos que interagem com bancos de dados e oferecem uma experiência personalizada aos usuários. Além disso, a ampla comunidade de desenvolvedores contribuiu para a criação de inúmeras bibliotecas e *frameworks* que facilitam o desenvolvimento eficiente de projetos *web* (ZEND, 2010).

Um dos principais pontos fortes do PHP é sua compatibilidade com uma variedade de sistemas operacionais e servidores *web*, como o Apache e o Nginx, o que o torna uma escolha versátil para diferentes ambientes de hospedagem (WELLING; THOMSON, 2016). Além disso, o PHP também oferece suporte a uma variedade de bancos de dados, incluindo MySQL e PostgreSQL, permitindo que os desenvolvedores escolham a opção que melhor se adapte às necessidades do projeto (SKLIAROVA *et al*., 2007). A natureza de código aberto do PHP tem incentivado a colaboração e a melhoria contínua da linguagem ao longo dos anos, resultando em um ecossistema rico de recursos e funcionalidades.

O PHP também enfrentou críticas ao longo do tempo, principalmente devido a práticas de programação inconsistentes e vulnerabilidades de segurança ocasionais (SINDELAR; ZUCCHERATO, 2013). A introdução do PHP 7 trouxe melhorias significativas no desempenho e na segurança, abordando muitas dessas preocupações e permitindo que o PHP permaneça uma escolha relevante no cenário de desenvolvimento atual (PHP Group, 2015). Com a ascensão de outras linguagens e *frameworks*, como Python e Node.js, o PHP continua a evoluir para atender às demandas em constante mudança da *web* moderna.

HTML E CSS

O desenvolvimento *web* revolucionou a maneira como interagimos com a informação e como as organizações e pessoas se comunicam. Nesse cenário, Linguagem de marcação de hipertexto (do inglês *HyperText Markup Language* - HTML) e Folhas de estilo em cascata (do

inglês *Cascading Style Sheets* - CSS) desempenham papéis cruciais na criação de *websites* funcionais e visualmente atraentes.

O HTML, introduzido por Tim Berners-Lee em 1991 (BERNERS-LEE, 1995), é uma linguagem de marcação que estrutura o conteúdo de uma página *web*. Ele utiliza *tags* para definir elementos como cabeçalhos, parágrafos, *links* e imagens. Já o CSS, proposto por Håkon Wium Lie em 1994 (LIE, 1996), é responsável pela apresentação visual, permitindo o controle de cores, fontes, layouts e efeitos.

A integração harmoniosa de HTML e CSS permite a criação de páginas *web* bem organizadas e estilizadas. O HTML define a estrutura, enquanto o CSS proporciona o estilo, garantindo a separação de preocupações e a reutilização de código. Com a evolução constante, o HTML5, lançado em 2014, introduziu elementos semânticos e recursos multimídia, enquanto o CSS3 trouxe animações e efeitos visuais avançados.

JAVASCRIPT

JavaScript, uma linguagem de programação amplamente utilizada no desenvolvimento *web* e *mobile*, tem uma rica história que remonta a sua criação por Brendan Eich em 1995 (EICH, 2000). Sua importância é inegável, sendo uma das principais linguagens do mercado e contando com uma comunidade de desenvolvedores vibrante. A habilidade do JavaScript em interagir com HTML e CSS (FLANAGAN, 2011) é um dos fatores-chave para tornar páginas mais interativas e responsivas, proporcionando experiências mais envolventes aos usuários.

A versatilidade do JavaScript é notável em sua capacidade de operar tanto no lado do cliente quanto no lado do servidor. No ambiente do cliente, ele é executado diretamente nos navegadores dos usuários, permitindo a criação de ações sem recarregar a página, o que é fundamental para a fluidez dos Aplicativos de página única (do inglês *Single Page Applications* - SPAs) (JOHNSON, 2018). No servidor, o JavaScript também encontra sua aplicação, desempenhando um papel crucial no desenvolvimento de aplicações web completas.

O ecossistema JavaScript é rico em bibliotecas e *frameworks*, e um destaque notável é o *React*, desenvolvido pelo Facebook (BROWN, 2019). O *React* revolucionou a forma como os desenvolvedores constroem interfaces de usuário, ao possibilitar a criação de componentes reutilizáveis que aceleram o processo de desenvolvimento. O percurso do JavaScript também teve seus desafios, incluindo questões de segurança e desempenho. À medida que a linguagem

evolui, desenvolvedores devem estar atentos às melhores práticas e tendências atuais (Lee, 2021) para garantir que suas aplicações sejam seguras e eficientes.

A trajetória do JavaScript é uma história de inovação contínua. Desde suas origens nos anos 90 até os dias atuais, a linguagem tem desempenhado um papel fundamental na revolução da *web*, capacitando desenvolvedores a criar experiências dinâmicas e interativas. A constante evolução e a diversidade de aplicações fazem do JavaScript uma ferramenta indispensável no kit de ferramentas de qualquer desenvolvedor *web* moderno.

MYSQL WORKBENCH

No cenário contemporâneo de desenvolvimento de software e análise de dados, a utilização de sistemas de gerenciamento de bancos de dados (SGBDs) desempenha um papel crucial. Dentre as ferramentas disponíveis, destaca-se o MySQL Workbench, uma aplicação desenvolvida pela Oracle Corporation que oferece um ambiente integrado para criação, modelagem, administração e manutenção de bancos de dados MySQL. Desde seu lançamento em 2003, o MySQL Workbench tem se mostrado uma solução poderosa e versátil, proporcionando aos desenvolvedores e administradores um conjunto de recursos eficientes para lidar com a complexidade dos bancos de dados modernos (Oracle, 2020).

Uma das características distintivas do MySQL Workbench é sua interface intuitiva e de fácil utilização. Através de uma abordagem visual, os usuários podem projetar esquemas de bancos de dados, criar tabelas, definir relações e estabelecer chaves estrangeiras, tudo isso de maneira simples e interativa. Além disso, a ferramenta oferece uma ampla gama de recursos para otimização de consultas, permitindo aos desenvolvedores a análise de desempenho e a identificação de gargalos em seus sistemas (Smith, 2018).

Outro aspecto notável do MySQL Workbench é sua capacidade de engenharia reversa, que permite aos usuários importar bancos de dados existentes para o ambiente de trabalho. Essa funcionalidade é especialmente valiosa para equipes de desenvolvimento que herdam projetos legados ou precisam colaborar em sistemas já em produção. Através da engenharia reversa, é possível obter um modelo visual da estrutura do banco de dados, facilitando a compreensão e a documentação do sistema (Johnson, 2016).

WAMPSERVER

De acordo com Instituto Brasileiro de Geografia e Estatística (1993), tabela é uma forma não discursiva de apresentar informações em que os números representam a informação central. O WampServer é um ambiente de desenvolvimento *web* que permite a criação de aplicações PHP, MySQL e Apache em um único pacote. Essa ferramenta é amplamente utilizada por desenvolvedores *web* devido à sua facilidade de instalação e configuração (WAMPSERVER, 2021). Além disso, o WampServer possui uma interface amigável que permite gerenciar facilmente as configurações do servidor.

Uma das principais vantagens do WampServer é a sua capacidade de suportar diferentes versões do PHP e do MySQL. Isso permite que os desenvolvedores testem suas aplicações em diferentes ambientes antes de colocá-las em produção (MACHADO, 2019). Além disso, o WampServer possui recursos avançados que facilitam o desenvolvimento, como a possibilidade de criar virtual hosts para testar diferentes projetos simultaneamente.

Outra vantagem do WampServer é a sua comunidade ativa e engajada. Existem diversos fóruns e grupos de discussão onde os desenvolvedores podem compartilhar conhecimentos e solucionar dúvidas (STACK OVERFLOW, 2021). Além disso, existem diversos plugins e extensões disponíveis para o WampServer que podem ser utilizados para melhorar a produtividade e a eficiência no desenvolvimento.

APÊNDICE D – SQL – *Back-end*

```sql
-- MySQL Workbench Forward Engineering

SET @OLD_UNIQUE_CHECKS=@@UNIQUE_CHECKS, UNIQUE_CHECKS=0;
SET @OLD_FOREIGN_KEY_CHECKS=@@FOREIGN_KEY_CHECKS, FOREIGN_KEY_CHECKS=0;
SET @OLD_SQL_MODE=@@SQL_MODE,
SQL_MODE='ONLY_FULL_GROUP_BY,STRICT_TRANS_TABLES,NO_ZERO_IN_DATE,NO_ZERO_DATE,ERROR_FOR_DIVISION_BY_ZERO,NO_ENGINE_S
UBSTITUTION';

-- -----------------------------------------------------
-- Schema accident_investigation
-- -----------------------------------------------------

-- -----------------------------------------------------
-- Schema accident_investigation
-- -----------------------------------------------------
CREATE SCHEMA IF NOT EXISTS `accident_investigation` DEFAULT CHARACTER SET utf8mb4 COLLATE utf8mb4_0900_ai_ci ;
USE `accident_investigation` ;

-- -----------------------------------------------------
-- Table `accident_investigation`.`study`
-- -----------------------------------------------------
CREATE TABLE IF NOT EXISTS `accident_investigation`.`study` (
  `study_id` INT NOT NULL,
  `description` VARCHAR(100) NOT NULL,
  `taxonomy_file` VARCHAR(255) NOT NULL,
  `report_file` VARCHAR(255) NOT NULL,
  `annotated_report` TEXT NULL,
  `date` TIMESTAMP NOT NULL DEFAULT now(),
  PRIMARY KEY (`study_id`))
ENGINE = InnoDB;


-- -----------------------------------------------------
-- Table `accident_investigation`.`dimension`
-- -----------------------------------------------------
CREATE TABLE IF NOT EXISTS `accident_investigation`.`dimension` (
  `dimension_id` INT NOT NULL,
  `study_id` INT NOT NULL,
  `description` VARCHAR(100) NOT NULL,
  PRIMARY KEY (`dimension_id`),
  INDEX `fk_dimension_study_idx` (`study_id` ASC) VISIBLE,
  CONSTRAINT `fk_dimension_study`
    FOREIGN KEY (`study_id`)
    REFERENCES `accident_investigation`.`study` (`study_id`)
    ON DELETE CASCADE
    ON UPDATE NO ACTION)
ENGINE = InnoDB;


-- -----------------------------------------------------
-- Table `accident_investigation`.`factor`
-- -----------------------------------------------------
CREATE TABLE IF NOT EXISTS `accident_investigation`.`factor` (
  `factor_id` INT NOT NULL AUTO_INCREMENT,
  `dimension_id` INT NOT NULL,
  `description` VARCHAR(100) NOT NULL,
  PRIMARY KEY (`factor_id`),
  INDEX `fk_factor_dimension_idx` (`dimension_id` ASC) VISIBLE,
  CONSTRAINT `fk_factor_dimension`
    FOREIGN KEY (`dimension_id`)
    REFERENCES `accident_investigation`.`dimension` (`dimension_id`)
    ON DELETE CASCADE
    ON UPDATE NO ACTION)
ENGINE = InnoDB;

```

```
64
65 -- -----------------------------------------------------
66 -- Table `accident_investigation`.`term`
67 -- -----------------------------------------------------
68 CREATE TABLE IF NOT EXISTS `accident_investigation`.`term` (
69   `term_id` INT NOT NULL AUTO_INCREMENT,
70   `factor_id` INT NOT NULL,
71   `preferred_term` VARCHAR(200) NOT NULL,
72   `search_term` VARCHAR(200) NOT NULL,
73   `frequency` INT NOT NULL,
74   PRIMARY KEY (`term_id`),
75   INDEX `fk_term_factor_idx` (`factor_id` ASC) VISIBLE,
76   CONSTRAINT `fk_term_factor`
77     FOREIGN KEY (`factor_id`)
78     REFERENCES `accident_investigation`.`factor` (`factor_id`)
79     ON DELETE CASCADE
80     ON UPDATE NO ACTION)
81 ENGINE = InnoDB
82 DEFAULT CHARACTER SET = utf8mb4
83 COLLATE = utf8mb4_0900_ai_ci;
84
85
86 -- -----------------------------------------------------
87 -- Table `accident_investigation`.`relation`
88 -- -----------------------------------------------------
89 CREATE TABLE IF NOT EXISTS `accident_investigation`.`relation` (
90   `source_term_id` INT NOT NULL,
91   `target_term_id` INT NOT NULL,
92   `value` INT NOT NULL,
93   INDEX `fk_relation_term_1_idx` (`source_term_id` ASC) VISIBLE,
94   INDEX `fk_relation_term_2_idx` (`target_term_id` ASC) VISIBLE,
95   PRIMARY KEY (`source_term_id`, `target_term_id`),
96   CONSTRAINT `fk_relation_term`
97     FOREIGN KEY (`source_term_id`)
98     REFERENCES `accident_investigation`.`term` (`term_id`)
99     ON DELETE CASCADE
100    ON UPDATE NO ACTION,
101  CONSTRAINT `fk_relation_term2`
102    FOREIGN KEY (`target_term_id`)
103    REFERENCES `accident_investigation`.`term` (`term_id`)
104    ON DELETE CASCADE
105    ON UPDATE NO ACTION)
106 ENGINE = InnoDB
107 DEFAULT CHARACTER SET = utf8mb4
108 COLLATE = utf8mb4_0900_ai_ci;
109
110
111 -- -----------------------------------------------------
112 -- Table `accident_investigation`.`paragraph`
113 -- -----------------------------------------------------
114 CREATE TABLE IF NOT EXISTS `accident_investigation`.`paragraph` (
115   `paragraph_id` INT NOT NULL AUTO_INCREMENT,
116   `study_id` INT NOT NULL,
117   `text` TEXT NOT NULL,
118   PRIMARY KEY (`paragraph_id`),
119   INDEX `fk_paragraph_study_idx` (`study_id` ASC) VISIBLE,
120   CONSTRAINT `fk_paragraph_study`
121     FOREIGN KEY (`study_id`)
122     REFERENCES `accident_investigation`.`study` (`study_id`)
123     ON DELETE CASCADE
124     ON UPDATE NO ACTION)
125 ENGINE = InnoDB;
126
127
128 -- -----------------------------------------------------
129 -- Table `accident_investigation`.`paragraph_term`
130 -- -----------------------------------------------------
131 CREATE TABLE IF NOT EXISTS `accident_investigation`.`paragraph_term` (
132   `paragraph_id` INT NOT NULL,
133   `term_id` INT NOT NULL,
134   `offsets` VARCHAR(100) NOT NULL,
135   PRIMARY KEY (`paragraph_id`, `term_id`),
136   INDEX `fk_paragraph_term_paragraph_idx` (`paragraph_id` ASC) VISIBLE,
137   INDEX `fk_paragraph_term_term_idx` (`term_id` ASC) VISIBLE,
138   CONSTRAINT `fk_term_has_paragraph_paragraph`
139     FOREIGN KEY (`paragraph_id`)
140     REFERENCES `accident_investigation`.`paragraph` (`paragraph_id`)
141     ON DELETE CASCADE
142     ON UPDATE NO ACTION,
143   CONSTRAINT `fk_paragraph_content_term`
144     FOREIGN KEY (`term_id`)
145     REFERENCES `accident_investigation`.`term` (`term_id`)
```

```
    ON DELETE CASCADE
    ON UPDATE NO ACTION)
ENGINE = InnoDB
DEFAULT CHARACTER SET = utf8mb4
COLLATE = utf8mb4_0900_ai_ci;


-- -----------------------------------------------------
-- Table `accident_investigation`.`factor_term_relation`
-- -----------------------------------------------------
CREATE TABLE IF NOT EXISTS `accident_investigation`.`factor_term_relation` (
  `factor_id` INT NULL,
  `source_term_id` INT NULL,
  `target_term_id` INT NULL,
  `value` INT NOT NULL,
  INDEX `fk_table1_factor1_idx` (`factor_id` ASC) VISIBLE,
  INDEX `fk_table1_term1_idx` (`source_term_id` ASC) VISIBLE,
  INDEX `fk_table1_term2_idx` (`target_term_id` ASC) VISIBLE,
  CONSTRAINT `fk_factor_term_relation_factor`
    FOREIGN KEY (`factor_id`)
    REFERENCES `accident_investigation`.`factor` (`factor_id`)
    ON DELETE CASCADE
    ON UPDATE NO ACTION,
  CONSTRAINT `fk_factor_term_relation_term_1`
    FOREIGN KEY (`source_term_id`)
    REFERENCES `accident_investigation`.`term` (`term_id`)
    ON DELETE CASCADE
    ON UPDATE NO ACTION,
  CONSTRAINT `fk_factor_term_relation_term_2`
    FOREIGN KEY (`target_term_id`)
    REFERENCES `accident_investigation`.`term` (`term_id`)
    ON DELETE CASCADE
    ON UPDATE NO ACTION)
ENGINE = InnoDB;


-- -----------------------------------------------------
-- Table `accident_investigation`.`visualization`
-- -----------------------------------------------------
CREATE TABLE IF NOT EXISTS `accident_investigation`.`visualization` (
  `study_id` INT NOT NULL,
  `visualization_id` INT NOT NULL,
  `text` TEXT NOT NULL,
  PRIMARY KEY (`study_id`, `visualization_id`),
  CONSTRAINT `fk_visualization_study1`
    FOREIGN KEY (`study_id`)
    REFERENCES `accident_investigation`.`study` (`study_id`)
    ON DELETE CASCADE
    ON UPDATE NO ACTION)
ENGINE = InnoDB;


SET SQL_MODE=@OLD_SQL_MODE;
SET FOREIGN_KEY_CHECKS=@OLD_FOREIGN_KEY_CHECKS;
SET UNIQUE_CHECKS=@OLD_UNIQUE_CHECKS;

```

APÊNDICE E – SQL – *Front-end*

```sql
-- phpMyAdmin SQL Dump
-- version 4.9.7
-- https://www.phpmyadmin.net/
--
-- Host: 127.0.0.1:3306
-- Tempo de geração: 01-Jun-2023 às 18:18
-- Versão do servidor: 5.7.36
-- versão do PHP: 5.6.40

SET SQL_MODE = "NO_AUTO_VALUE_ON_ZERO";
SET AUTOCOMMIT = 0;
START TRANSACTION;
SET time_zone = "+00:00";


/*!40101 SET @OLD_CHARACTER_SET_CLIENT=@@CHARACTER_SET_CLIENT */;
/*!40101 SET @OLD_CHARACTER_SET_RESULTS=@@CHARACTER_SET_RESULTS */;
/*!40101 SET @OLD_COLLATION_CONNECTION=@@COLLATION_CONNECTION */;
/*!40101 SET NAMES utf8mb4 */;

--
-- Banco de dados: `bd_estudos`
--

-- --------------------------------------------------------

--
-- Estrutura da tabela `estudos`
--

DROP TABLE IF EXISTS `estudos`;
CREATE TABLE IF NOT EXISTS `estudos` (
  `id` int(11) NOT NULL AUTO_INCREMENT,
  `nome_do_estudo` varchar(255) NOT NULL,
  `arquivo_1` longblob,
  `arquivo_2` longblob,
  `anotacoes` text,
  PRIMARY KEY (`id`)
) ENGINE=MyISAM AUTO_INCREMENT=28 DEFAULT CHARSET=latin1;

--
-- Extraindo dados da tabela `estudos`
--

INSERT INTO `estudos` (`id`, `nome_do_estudo`, `arquivo_1`, `arquivo_2`, `anotacoes`) VALUES
(21, '', 0x7461786f6e6f6d792f, 0x7461786f6e6f6d792f, '');
COMMIT;

/*!40101 SET CHARACTER_SET_CLIENT=@OLD_CHARACTER_SET_CLIENT */;
/*!40101 SET CHARACTER_SET_RESULTS=@OLD_CHARACTER_SET_RESULTS */;
/*!40101 SET COLLATION_CONNECTION=@OLD_COLLATION_CONNECTION */;

```

APÊNDICE F – Documentação do sistema

ESCOPO DO SISTEMA

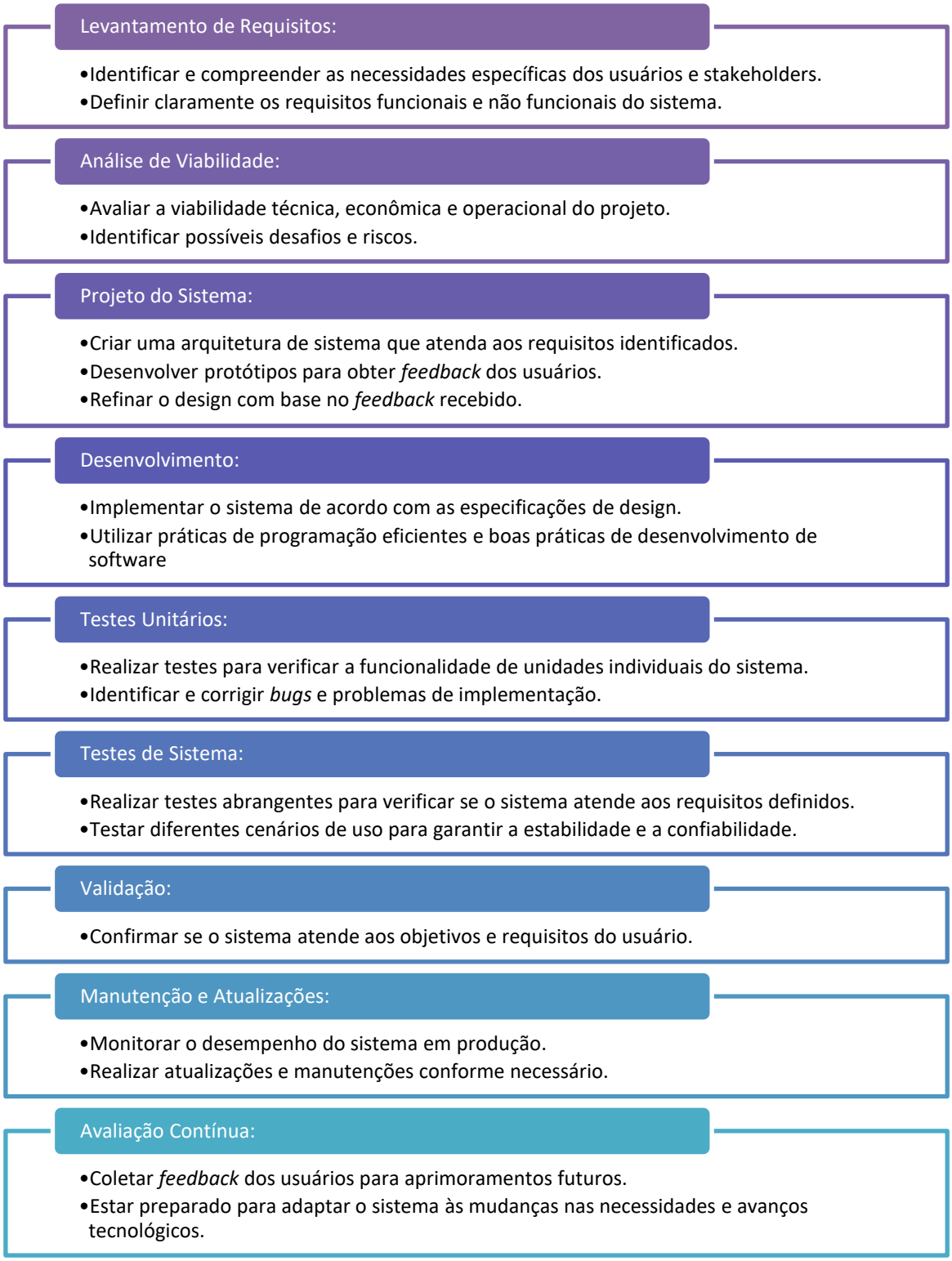
O sistema DOC *Analysis* visa proporcionar uma solução abrangente para a análise de documentos de texto e a visualização de tópicos relevantes em conformidade com taxonomias definidas pelos usuários. A plataforma permite que os usuários realizem o upload de documentos, associando-os a uma taxonomia específica, além de incluir informações adicionais, como título do estudo e comentários. O cerne do sistema reside na capacidade de analisar o conteúdo do documento em relação à taxonomia, identificando correspondências entre tópicos abordados e os termos definidos na estrutura taxonômica.

A geração de gráficos visuais intuitivos e enriquecidos é uma característica central do DOC Analysis, destacando visualmente as relações entre os tópicos do documento e os termos da taxonomia. Esses gráficos são armazenados no banco de dados, associados ao documento e à taxonomia utilizada, proporcionando aos usuários a capacidade de revisitar estudos anteriores de maneira eficiente. Além disso, o sistema se compromete a garantir a segurança e privacidade dos dados do usuário, implementando medidas robustas para proteção das informações. O acesso ao sistema é viabilizado por meio de navegadores *web* modernos, assegurando uma experiência consistente e intuitiva ao usuário, desde o upload até a visualização dos resultados analíticos. No âmbito dos requisitos não funcionais, o DOC *Analysis* prioriza o desempenho responsivo, escalabilidade para lidar com volumes crescentes de estudos armazenados, e compatibilidade com diversos navegadores *web*. A segurança de dados, a disponibilidade do sistema e a acessibilidade são fatores críticos considerados no desenvolvimento, assegurando que o sistema seja eficiente, confiável e amplamente acessível. A interface do usuário é projetada para ser intuitiva, demandando mínimo treinamento, enquanto a consistência visual e a integração eficiente com o banco de dados são elementos essenciais para uma experiência coesa e eficaz. Em consonância com as práticas contemporâneas, o DOC *Analysis* utiliza tecnologias *web* atualizadas para garantir compatibilidade, desempenho e segurança de ponta.

ETAPAS DE DESENVOLVIMENTO

O desenvolvimento de sistemas para pesquisa científica envolve várias etapas, desde a concepção até a implementação. A Figura 40, apresentas as principais etapas desse processo:

Figura 40 - Etapas de desenvolvimento de sistemas



Fonte: Elaborado pela autora.

No final espera-se garantir o desenvolvimento eficiente e eficaz de sistemas para pesquisa científica, proporcionando resultados confiáveis e contribuindo para o progresso do desenvolvimento.

REQUISITOS FUNCIONAIS

Apresenta-se o conjunto de requisitos funcionais (RF) do sistema *DOC Analysis*:

RF 1. O usuário deve ser capaz de fazer o upload de um documento de texto e uma taxonomia associada através da interface do sistema.

RF 2. O sistema deve permitir a inclusão do título do estudo e de comentários.

RF 3. O sistema deve analisar o conteúdo do documento de texto com base na taxonomia fornecida pelo usuário, identificando as correspondências entre os tópicos do documento e os termos da taxonomia.

RF 4. Com base na análise realizada, o sistema deve gerar gráficos visuais que representem os tópicos abordados no documento e como eles se relacionam com a taxonomia. Os gráficos devem ser intuitivos e compreensíveis.

RF 5. Os gráficos gerados devem ser enriquecidos com recursos visuais, como cores distintas, ícones ou formas geométricas, para facilitar a identificação dos diferentes elementos e relações.

RF 6. Após a análise, os resultados e os gráficos gerados devem ser armazenados no banco de dados do sistema, associados ao documento de texto e à taxonomia utilizada.

RF 7. O usuário deve ser capaz de visualizar os estudos anteriores armazenados no banco de dados, incluindo os documentos, as taxonomias, os resultados da análise e os gráficos gerados.

RF 8. O sistema deve permitir que o usuário exclua estudos anteriores armazenados, caso deseje remover informações específicas do banco de dados.

RF 9. O sistema deve ser acessível através de navegadores *web* modernos, garantindo uma experiência consistente em diferentes plataformas.

RF 10. A interface do sistema deve ser intuitiva e amigável, orientando o usuário através do processo de envio de documentos, escolha de taxonomias, análise, geração de gráficos e visualização de resultados.

RF 11. O sistema deve garantir a segurança e privacidade dos dados do usuário, implementando medidas adequadas para proteger as informações armazenadas e transmitidas.

É importante ressaltar que na maioria de sistemas utilizados para análise, conversão e tradução encontrados hoje na *web*, poucos são de uso local o que pode acarretar no vazamento de dados enviados. Ao utilizar um sistema local, sua execução e armazenamento de dados utilizados têm chances menores de vazamento de dados e cópias não autorizadas.

REQUISITOS NÃO FUNCIONAIS

Os requisitos não funcionais (RNF) do sistema *DOC Analysis* são apresentados na sequência:

RNF 1. O sistema deve ter um desempenho responsivo, garantindo tempos de resposta rápidos ao realizar a análise e gerar os gráficos, mesmo com grandes volumes de dados.

RNF 2. O sistema deve ser capaz de lidar com um aumento gradual no número de estudos armazenados, mantendo a performance e a capacidade de resposta.

RNF 3. O sistema deve ser compatível com uma ampla gama de navegadores *web* modernos, assegurando uma experiência consistente para os usuários.

RNF 4. Os dados dos usuários, incluindo documentos enviados e resultados analíticos, devem ser protegidos por medidas de segurança.

RNF 5. O sistema deve estar disponível para acesso e utilização pelos usuários durante a maior parte do tempo, com tempos de inatividade planejados mínimos para manutenção.

RNF 6. A interface do usuário deve ser intuitiva e de fácil utilização, exigindo pouco ou nenhum treinamento para que os usuários possam aproveitar todas as funcionalidades.

RNF 7. O sistema deve ser projetado de acordo com as diretrizes de acessibilidade, garantindo que pessoas com deficiências possam interagir com ele de forma eficaz.

RNF 8. A aparência visual do sistema, incluindo cores, ícones e layouts, deve ser consistente em todas as páginas e elementos, proporcionando uma experiência coesa.

RNF 9. O sistema deve ser capaz de se integrar ao banco de dados de forma eficiente, garantindo a recuperação rápida e precisa dos estudos armazenados.

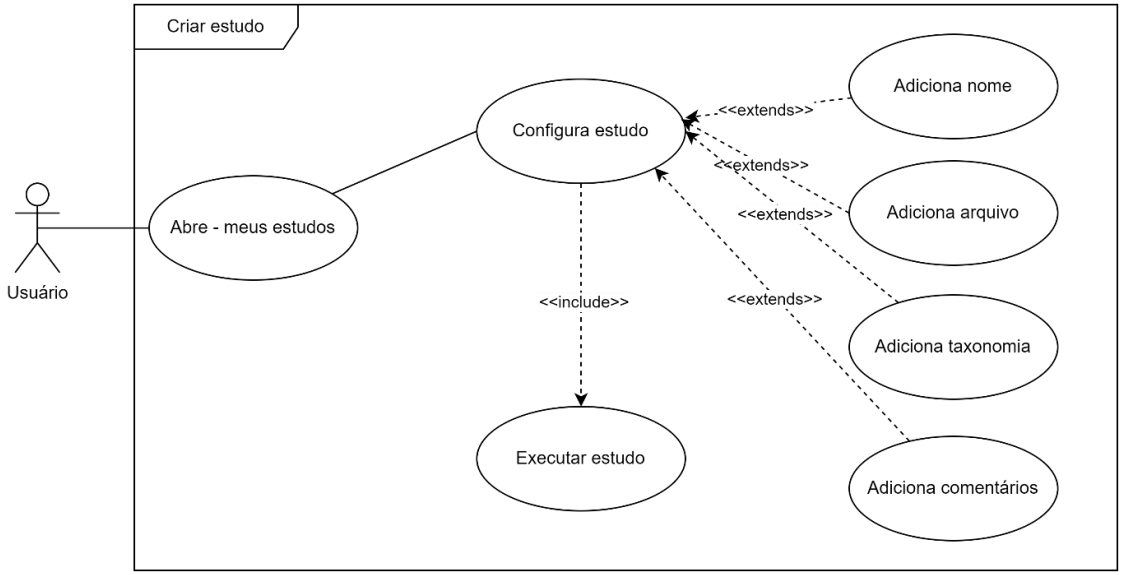
RNF 10. Os gráficos gerados devem ser criados de maneira eficiente, evitando atrasos significativos na geração e exibição dos resultados.

RNF 11. O sistema deve utilizar tecnologias *web* atualizadas e de ponta para garantir a compatibilidade, desempenho e segurança.

CASOS DE USO

Os casos de uso auxiliam na visualização das funcionalidades do sistema desenvolvido, a Figura 41 demonstra o caso de uso de criar novo estudo, onde o usuário acessa por meio cabeçalho ou página inicial.

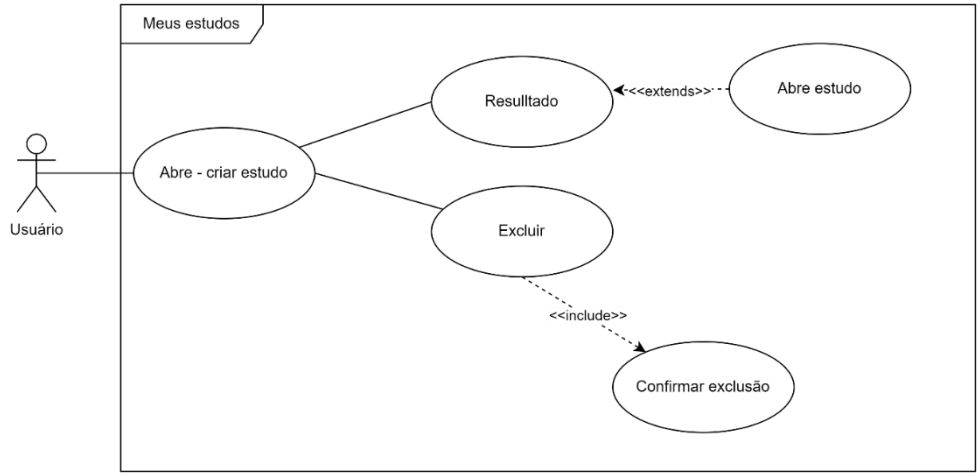
Figura 41 - Caso de uso: criar novo estudo



Fonte: Elaborado pela autora

A Figura 42, demonstra o caso de uso quando o usuário irá abrir um estudo anteriormente criado, assim o usuário poderá visualizar o resultado ou excluir o estudo.

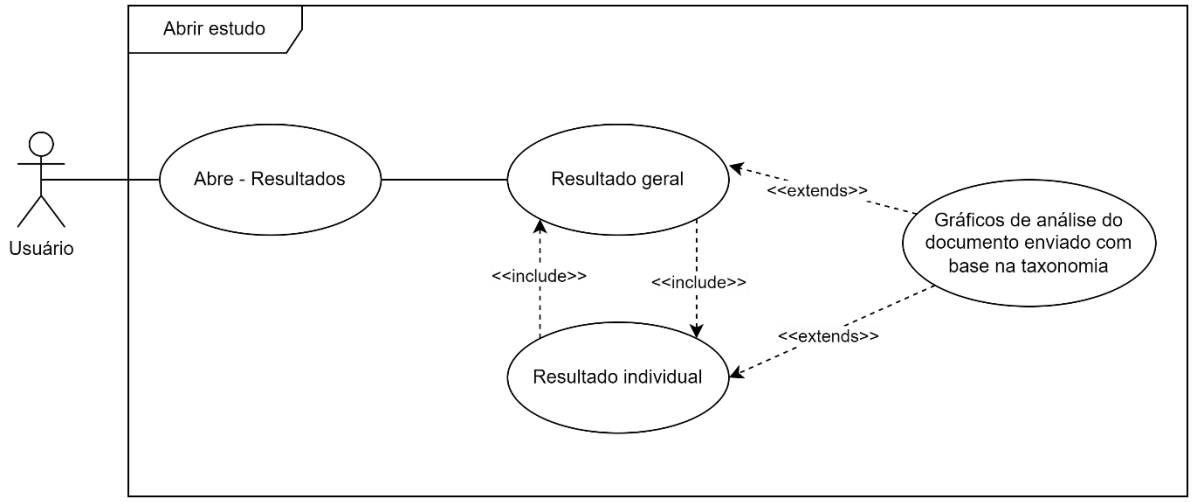
Figura 42 - Caso de uso: Criar estudo



Fonte: Elaborado pela autora

Na Figura 43, demonstra o caso de uso dos Resultados, o usuário pode optar em qual tipo de visualização do resultado geral ou individual.

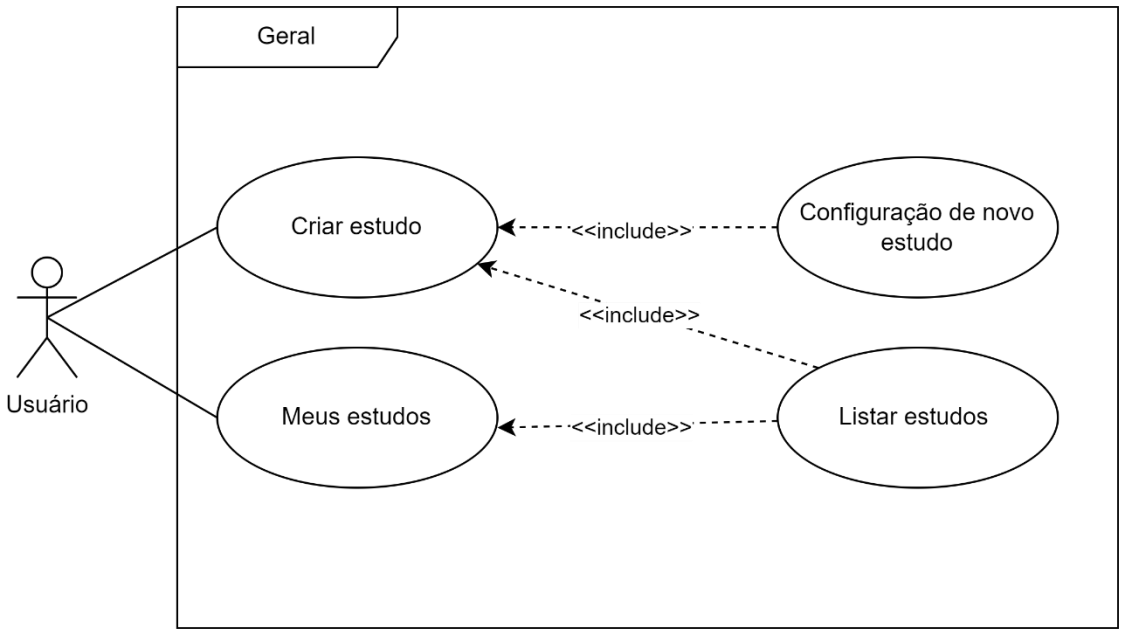
Figura 43 - Caso de uso: Resultados



Fonte: Elaborado pela autora

A Figura 44 representa o caso de uso geral do sistema, quando o usuário acessa a página inicial, o mesmo poderá criar estudos ou visualizar todos os estudos já criados.

Figura 44 - Caso de uso: Geral



Fonte: Elaborado pela autora

ARQUITETURA

A arquitetura do sistema DOC *Analysis* foi cuidadosamente planejada para proporcionar eficiência, escalabilidade e segurança. A estrutura delineada é essencial para garantir que o sistema atenda aos requisitos funcionais e não funcionais estabelecidos, promovendo um ambiente robusto e confiável para a análise de documentos de texto e visualização relevantes.

A arquitetura do DOC *Analysis* é composta por diversos componentes inter-relacionados:

Interface do Usuário (UI): A camada de interface do usuário oferece a experiência visual e interativa para os usuários. É por meio dessa interface que os documentos são carregados, taxonomias são associadas e os resultados são visualizados.

Servidor de Aplicação: O servidor de aplicação é responsável por processar as solicitações dos usuários, coordenando as operações do sistema. Ele gerencia a lógica de negócios, comunica-se com o banco de dados e orquestra a geração de gráficos.

Banco de Dados: O MySQL é o sistema de gerenciamento de banco de dados escolhido para armazenar informações críticas, incluindo documentos de texto, taxonomias, resultados de análises e gráficos gerados. A estrutura do banco de dados é projetada para otimizar a recuperação e armazenamento eficientes desses dados.

Módulo de Análise de Texto: Este módulo é responsável por analisar o conteúdo dos documentos de texto em relação às taxonomias definidas pelos usuários. Ele identifica correspondências entre tópicos abordados e termos da estrutura taxonômica.

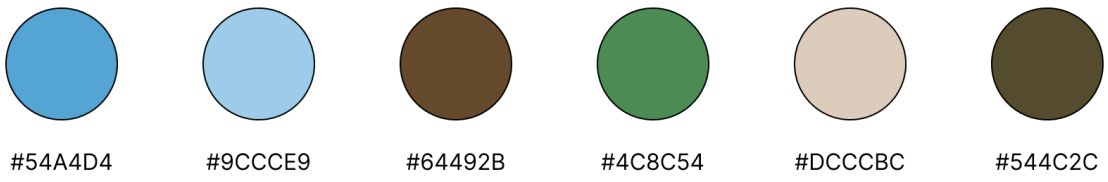
Gerador de Gráficos: O componente de geração de gráficos é acionado pelo módulo de análise de texto e é responsável por criar visualizações gráficas intuitivas que destacam as relações entre o documento e os termos da taxonomia.

IDENTIDADE VISUAL

A identidade visual do DOC *Analysis* foi cuidadosamente elaborada para refletir os princípios fundamentais de clareza, eficiência e modernidade. O design system desempenhou um papel crucial na criação de uma experiência de usuário consistente e atraente. Os principais elementos da identidade visual são:

- **Logotipo:** O logotipo do DOC *Analysis* incorpora elementos que simbolizam a análise de documentos e a visualização de informações. Cores sóbrias foram escolhidas para transmitir profissionalismo e confiabilidade.
- **Paleta de Cores:** Uma paleta de cores foi selecionada (Figura 45), combinando tons que promovem uma atmosfera profissional e acessível. A consistência nas escolhas cromáticas foi mantida em todo o sistema, contribuindo para uma identidade visual coesa.

Figura 45 - Paleta de cores



- Tipografia: A seleção de fontes foi feita com foco na legibilidade e modernidade. Tipos de letra limpos e contemporâneos foram escolhidos para garantir uma comunicação clara e eficaz em toda a plataforma.

DESENVOLVIMENTO DO *FRONT-END*

DESIGN SYSTEM

O design system do DOC *Analysis* é uma abordagem sistemática para o design de interfaces e experiências do usuário, e para a sua implementação, optou-se pela integração do *framework Bootstrap*[19]. Este design system abrange componentes reutilizáveis, diretrizes de design e princípios que asseguram uma coesão visual em toda a aplicação. Elementos do *Design system*:

- **Componentes de Interface com *Bootstrap*:** botões, campos de formulário, barras de navegação e outros componentes foram padronizados utilizando os estilos do *Bootstrap*. Isso cria uma experiência de usuário consistente, melhorando a usabilidade e acelerando o desenvolvimento.
- **Diretrizes de *Layout* Claras:** foram estabelecidas orientações claras de layout, alinhadas com as convenções do *Bootstrap*, garantindo uma disposição intuitiva e eficiente dos elementos na interface. Isso facilita a navegação do usuário e contribui para uma experiência coesa.

PROTÓTIPOS

O desenvolvimento do DOC *Analysis* envolveu a criação de protótipos em diferentes níveis de fidelidade, desde protótipos de baixa fidelidade até protótipos de alta fidelidade. Cada etapa do processo de prototipagem teve como objetivo refinar a experiência do usuário e validar as funcionalidades do sistema.

PROTÓTIPOS DE BAIXA FIDELIDADE

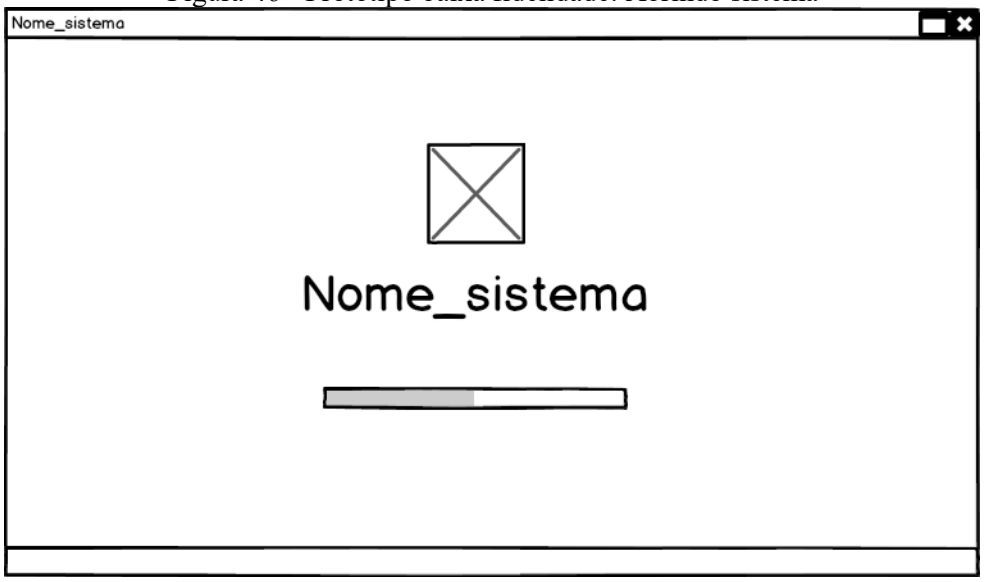
Os protótipos de baixa fidelidade foram concebidos como esboços iniciais que delinearam a estrutura geral e o fluxo de interação do DOC *Analysis*. Esses protótipos foram

[19] https://getbootstrap.com/

úteis para explorar conceitos de layout, identificar os principais elementos de interface e receber *feedback* inicial de *stakeholders*.

A Figura 46, ilustra o carregamento do sistema, uma vez que estes protótipos representam uma ideia da versão inicial do sistema.

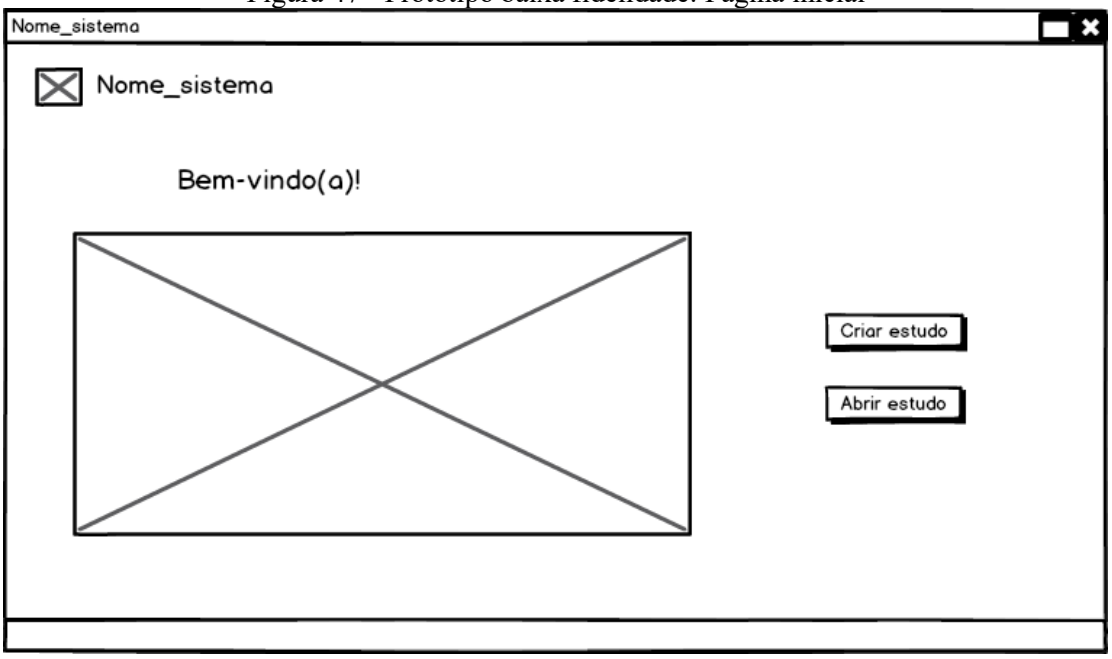
Figura 46 - Protótipo baixa fidelidade: Abrindo sistema



Fonte: Elaborado pela autora

A Figura 47, representa o protótipo de baixa fidelidade da página inicial do sistema, na qual o usuário poderia optar entre criar ou abrir um estudo.

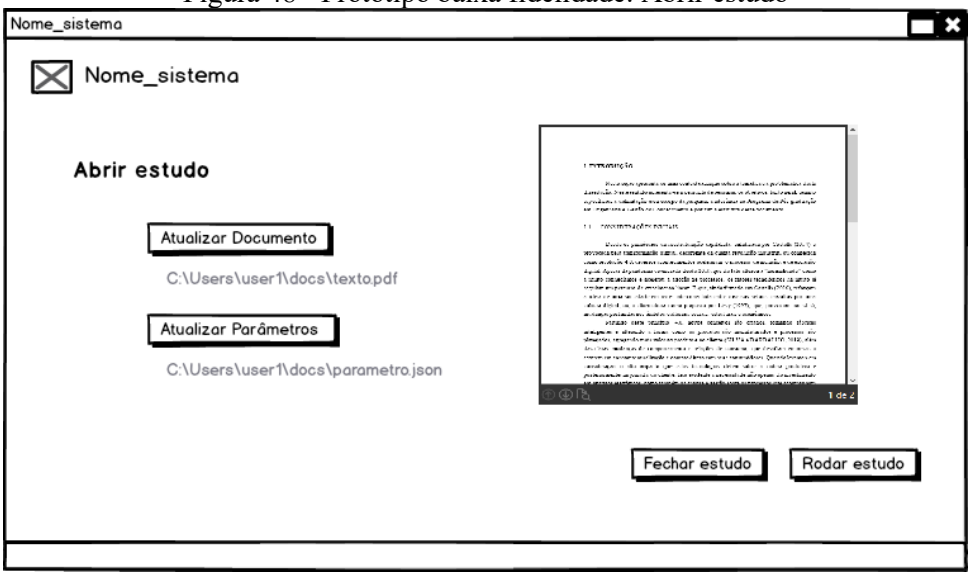
Figura 47 - Protótipo baixa fidelidade: Página inicial



Fonte: Elaborado pela autora

A Figura 48, ilustra o protótipo da página de Abrir estudo, onde o usuário poderia fechar ou rodar o estudo.

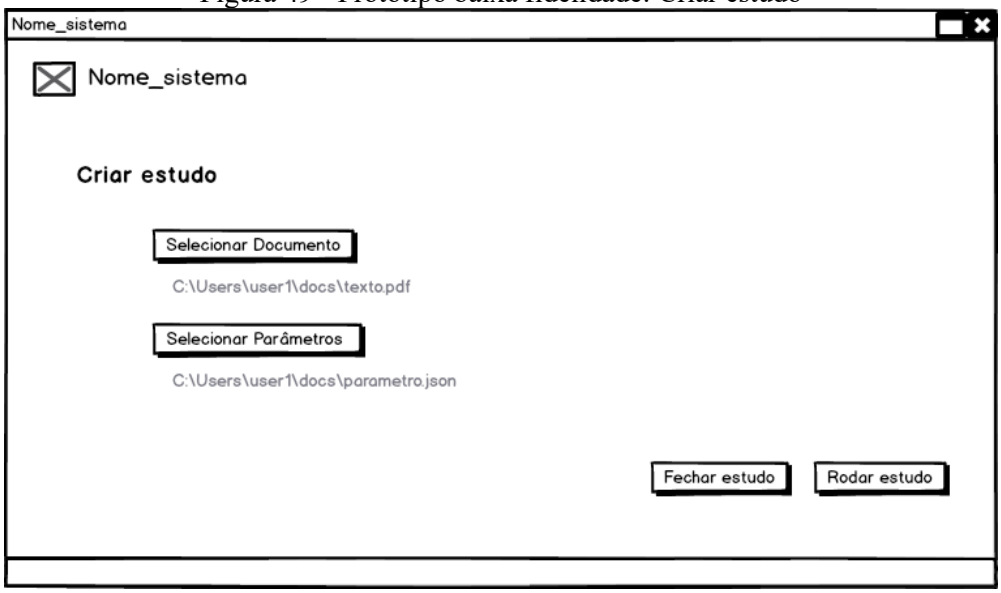
Figura 48 - Protótipo baixa fidelidade: Abrir estudo



Fonte: Elaborado pela autora

Na Figura 49 é ilustrada a página de criar estudo, o usuário deve selecionar o documento de texto e a taxonomia e na sequência rodar o estudo.

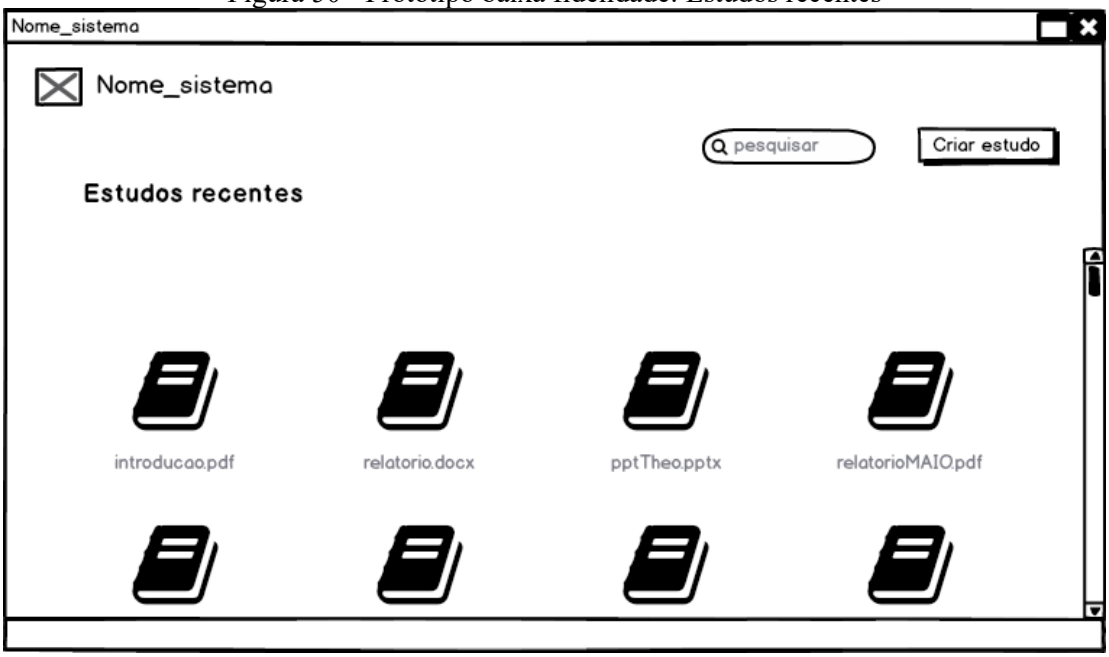
Figura 49 - Protótipo baixa fidelidade: Criar estudo



Fonte: Elaborado pela autora

A Figura 50, demonstra a página de estudo recentes, responsável por listar todos os estudos já criados.

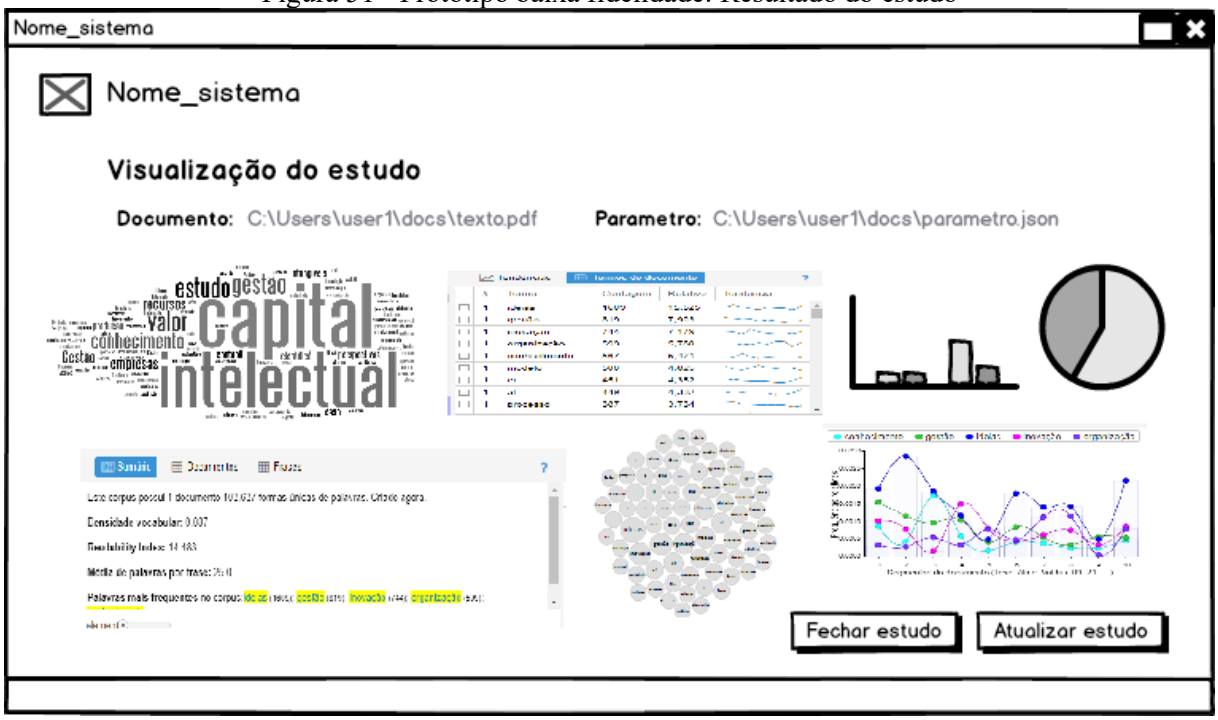
Figura 50 - Protótipo baixa fidelidade: Estudos recentes



Fonte: Elaborado pela autora

Uma vez clicado em ‘rodar estudo’ o resultado é exibido como o ilustrado pela Figura 51.

Figura 51 - Protótipo baixa fidelidade: Resultado do estudo

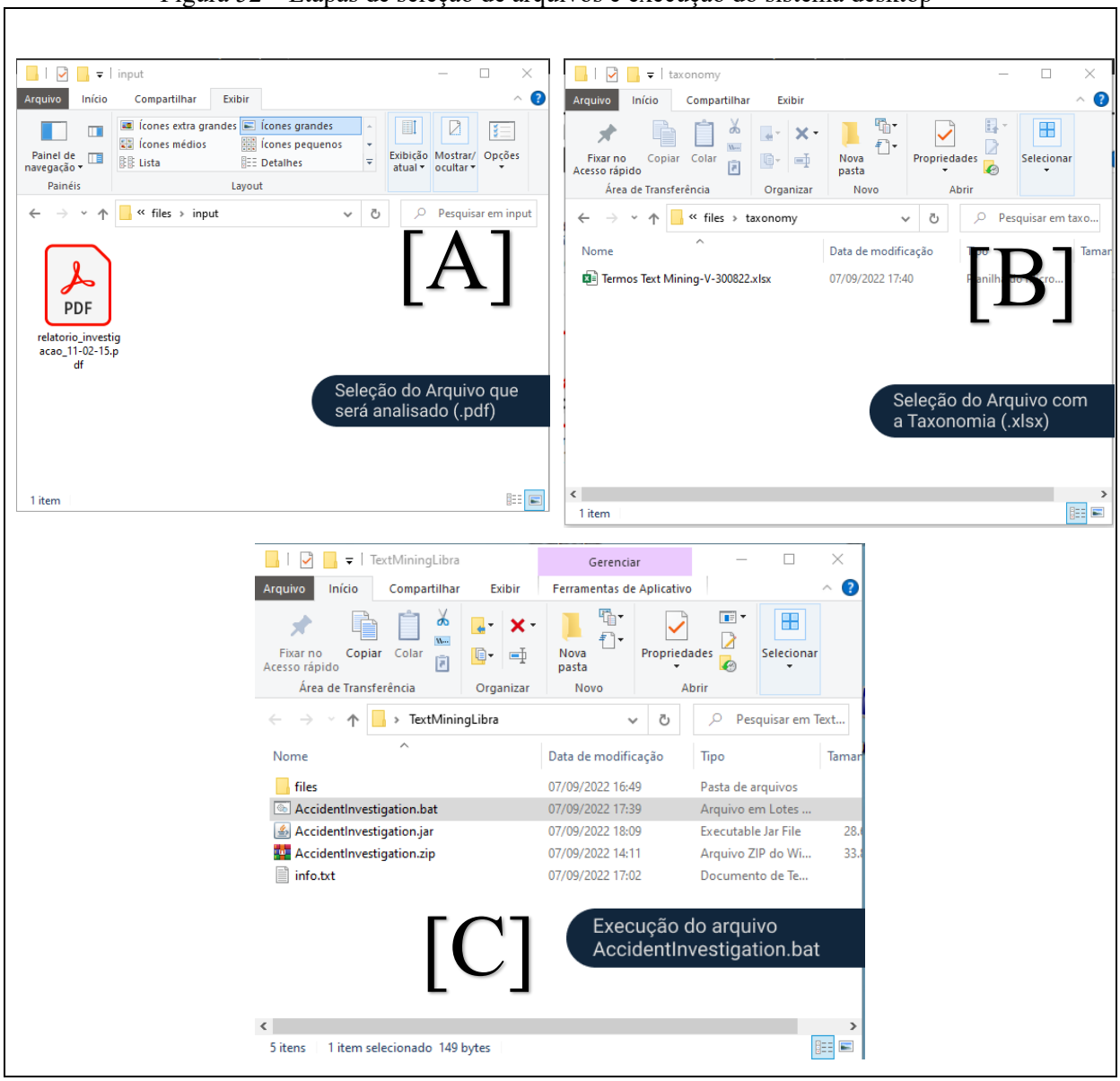


Fonte: Elaborado pela autora

PROTÓTIPOS DE MÉDIA FIDELIDADE

Os protótipos de média fidelidade refinaram os detalhes visuais e de interação, incorporando *feedback* coletado dos protótipos de baixa fidelidade. Elementos de design, como cores, tipografia e iconografia, foram mais definidos nessa etapa para criar uma representação mais próxima do produto final. A Figura 52 demonstra a execução do módulo *back-end*, [A] inicialmente é inserido o arquivo de texto a ser analisado na pasta ‘input’, [B] na pasta ‘taxonomy’ deveria ser inserido o arquivo com a taxonomia, [C] executar o módulo.

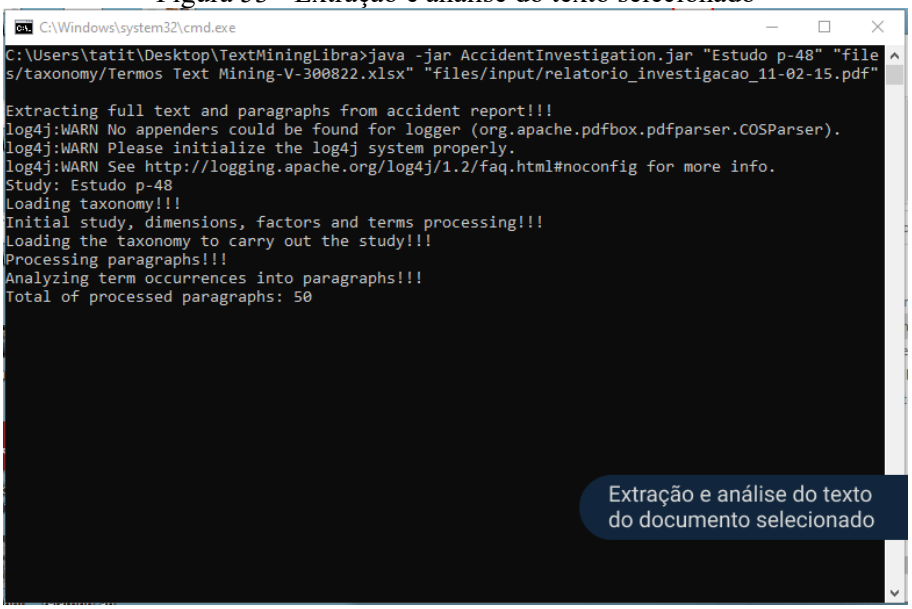
Figura 52 – Etapas de seleção de arquivos e execução do sistema desktop



Fonte: Elaborado pela autora

Após iniciado o módulo *back-end*, é iniciado o processo de extração do texto (Figura 53) completo e dos parágrafos do arquivo de texto e realizada a análise do texto conforme taxonomia. Ao final, os arquivos JSON e CSV são gerados, os quais são utilizados para exibição dos dados nos gráficos.

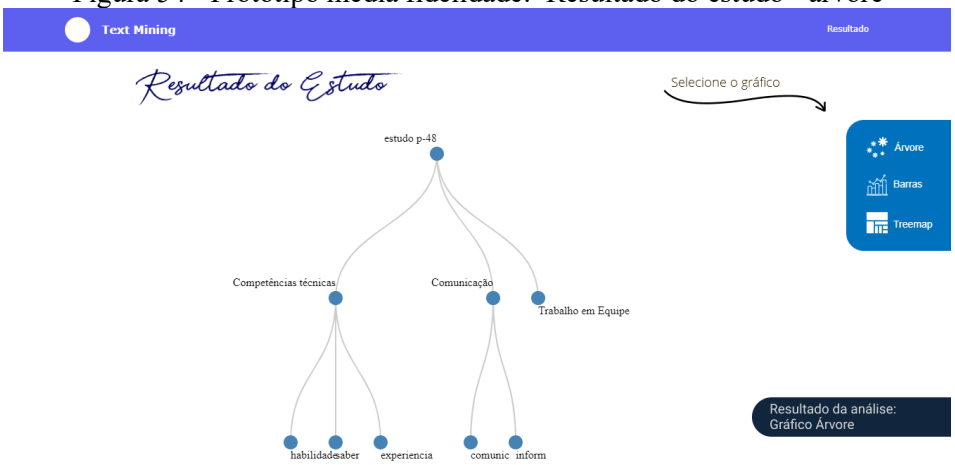
Figura 53 - Extração e análise do texto selecionado



Fonte: Elaborado pela autora

Os gráficos eram exibidos utilizado um navegador *web*, a Figura 54 representa os dados resultados da extração e análise do texto em um gráfico de árvore.

Figura 54 - Protótipo média fidelidade: Resultado do estudo - árvore

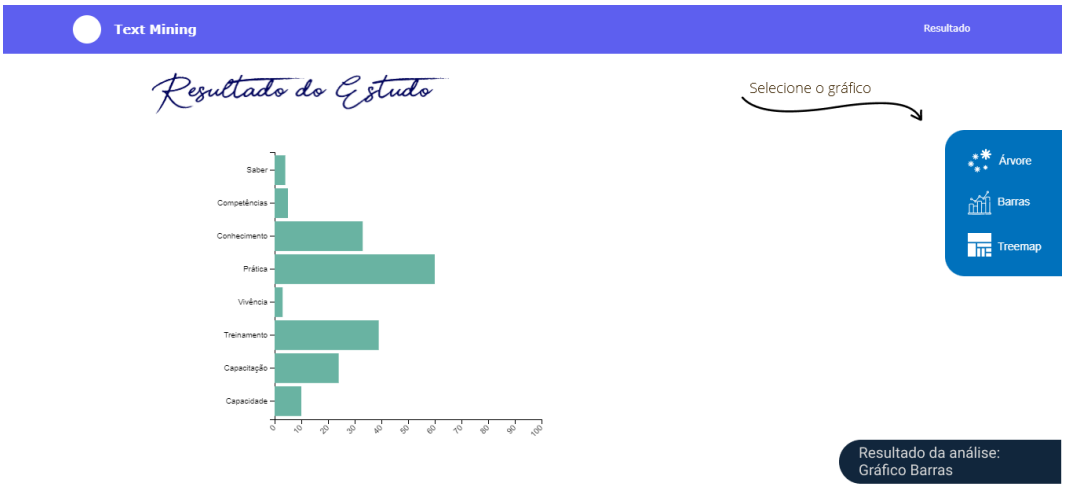


Fonte: Elaborado pela autora

Já a

Figura 55, demonstra o resultado do estudo utilizando um gráfico de barras horizontais e a Figura 56 usando um gráfico *treemap*.

Figura 55 - Protótipo média fidelidade: Resultado do estudo - barras horizontais



Fonte: Elaborado pela autora

Figura 56 - Protótipo média fidelidade: Resultado do estudo - treemap

Fonte: Elaborado pela autora

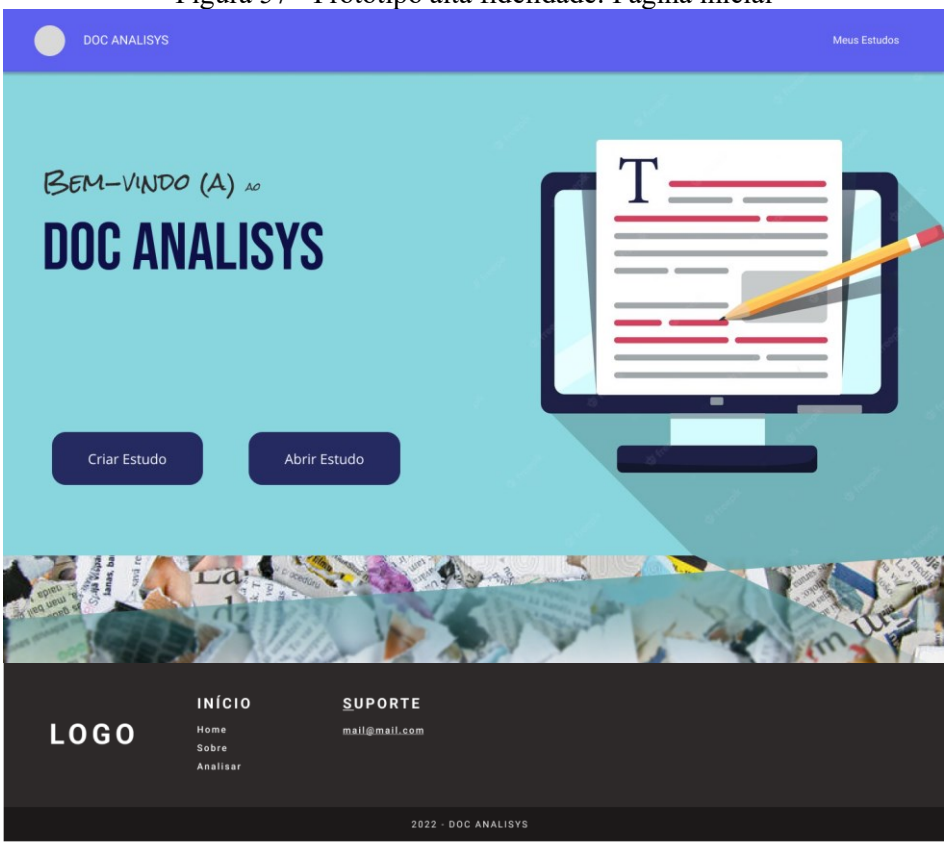
Nos protótipos de média fidelidade, é evidente uma notável progressão em comparação com os protótipos de baixa fidelidade, especialmente no que diz respeito à concepção de um sistema *web*.

PROTÓTIPOS DE ALTA FIDELIDADE

Os protótipos de alta fidelidade representam a versão mais próxima do design final do DOC *Analysis*. Nessa parte, detalhes finos foram refinados, e a interatividade completa da interface foi incorporada. Esses protótipos serviram como uma ferramenta valiosa para validar o design.

A página inicial, representada pela Figura 57, demonstra ações que o usuário poderá realizar, como Criar ou Abrir estudos, assim como visualizar todos os estudos criados, quando acionado o botão ‘Meus estudos’.

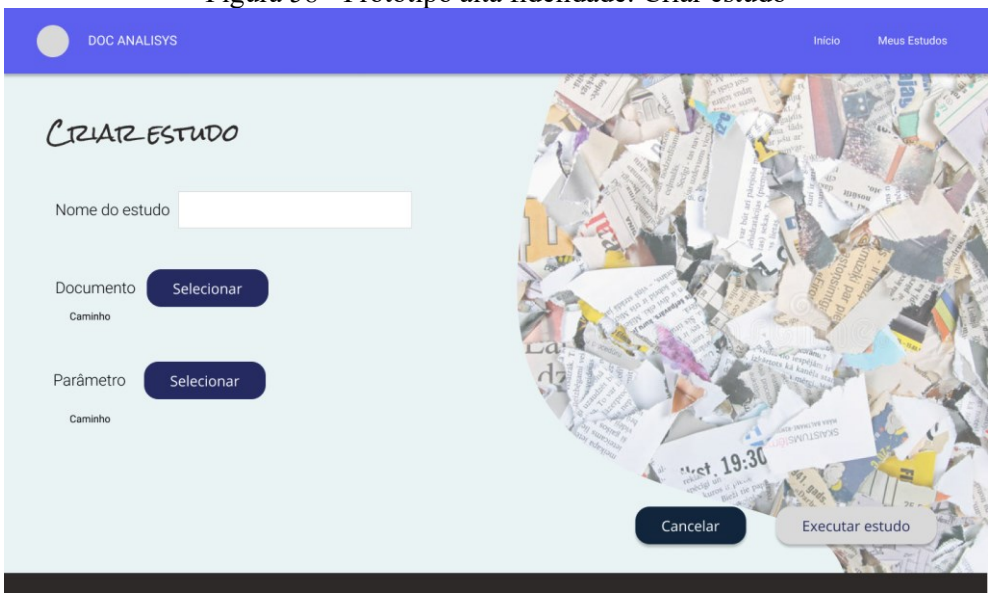
Figura 57 - Protótipo alta fidelidade: Página inicial



Fonte: Elaborado pela autora

Por meio da Figura 58, é possível ver a página de ‘Criar estudo’ do protótipo de alta fidelidade.

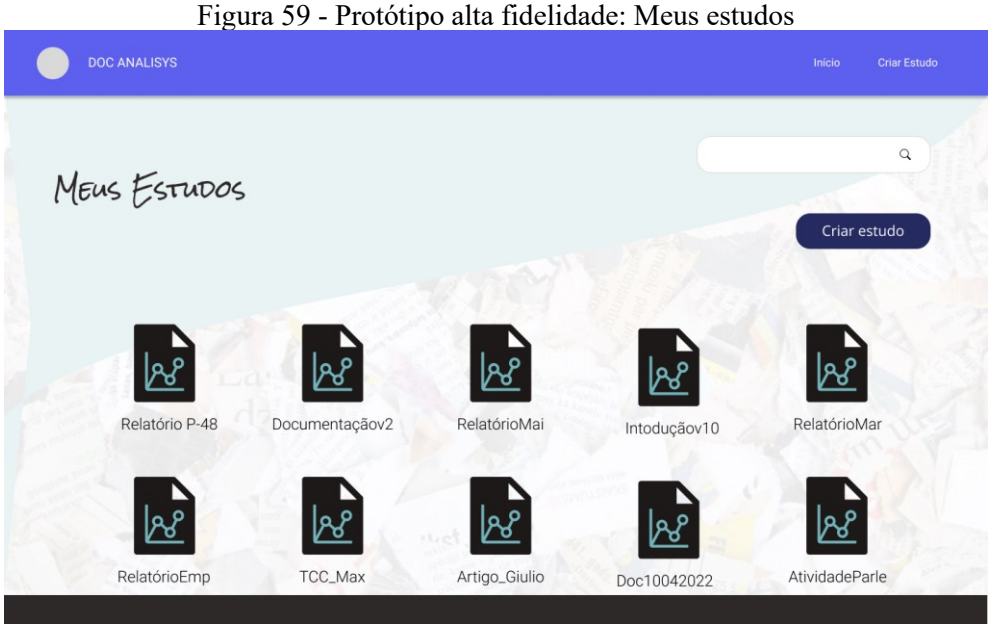
Figura 58 - Protótipo alta fidelidade: Criar estudo



Fonte: Elaborado pela autora

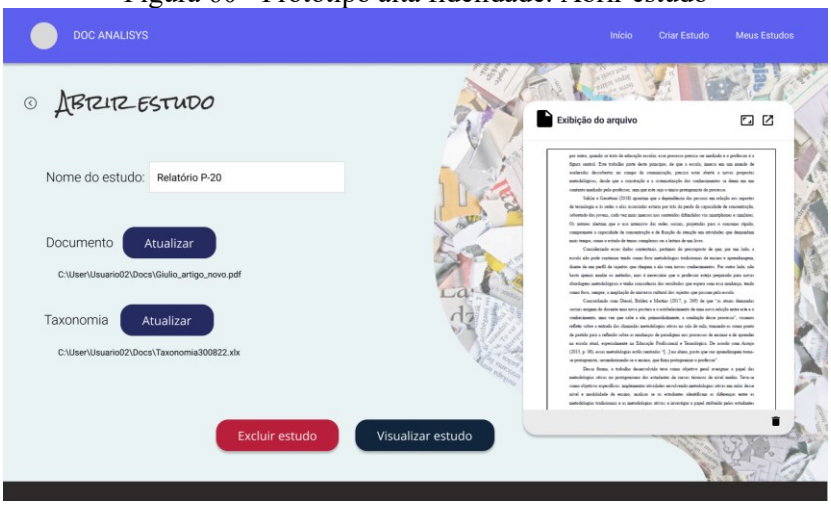
A Figura 59, apresenta a página Meus estudos, aonde são listados todos os estudos criados no sistema. Já a Figura 60, o protótipo de alta fidelidade da página Abri estudo.

Figura 59 - Protótipo alta fidelidade: Meus estudos



Fonte: Elaborado pela autora

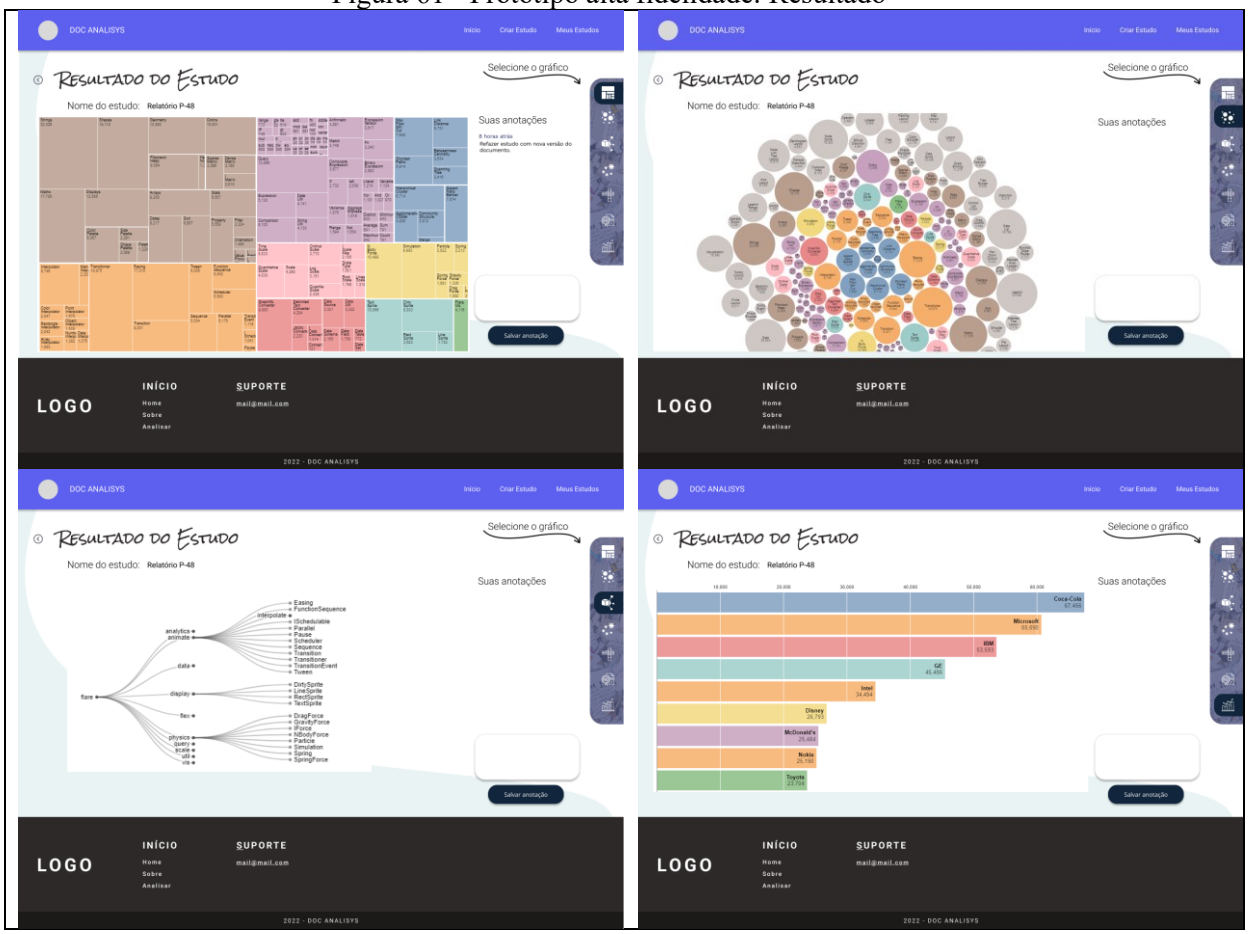
Figura 60 - Protótipo alta fidelidade: Abrir estudo



Fonte: Elaborado pela autora

Os resultados do estudo realizado são apresentados por meio da Figura 61, utilizado os gráficos *treemap*, *zoomable circle*, árvore e barras horizontais.

Figura 61 - Protótipo alta fidelidade: Resultado



Fonte: Elaborado pela autora