



UNIVERSIDADE FEDERAL DE SANTA CATARINA
CAMPUS ARARANGUÁ
CENTRO DE CIÊNCIAS, TECNOLOGIA E SAÚDE (CTS)
TECNOLOGIAS DA INFORMAÇÃO E COMUNICAÇÃO

Joel João Medeiros Filho

Análise e Visualização de Metadados de Trabalhos de Conclusão de Curso:
Uma Análise das Produções Acadêmicas

Araranguá
2024

Joel João Medeiros Filho

Análise e Visualização de Metadados de Trabalhos de Conclusão de Curso:
Uma Análise das Produções Acadêmicas

Trabalho de Conclusão de Curso submetido ao curso de Tecnologias da Informação e Comunicação do Centro de Ciência, Tecnologia e Saúde Campus Araranguá da Universidade Federal de Santa Catarina como requisito para a obtenção do título de Bacharel em Tecnologias da Informação e Comunicação.

Orientadora: Profa. Dra. Marina Carradore Sergio.

Araranguá

2024

Medeiros, Joel João

Análise e Visualização de Metadados de Trabalhos de Conclusão de Curso : Uma Análise das Produções Acadêmicas/Medeiros ; orientadora, Marina Carradore Sergio, 2024.

76 p.

Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Santa Catarina, Campus Araranguá, Graduação em Tecnologias da Informação e Comunicação, Araranguá, 2024.

Inclui referências.

1. Tecnologias da Informação e Comunicação. 2. Metadados. 3. Trabalhos de Conclusão de Curso. 4. Visualização de Dados. 5. Clusterização. I. Sergio, Marina Carradore. II. Universidade Federal de Santa Catarina. Graduação em Tecnologias da Informação e Comunicação. III. Título.

Joel João Medeiros Filho

Análise e Visualização de Metadados de Trabalhos de Conclusão de Curso:
Uma Análise das Produções Acadêmicas

Este Trabalho de Conclusão de Curso foi julgado adequado para obtenção do título de bacharel e aprovado em sua forma final pelo Curso de Tecnologias da Informação e Comunicação.

Araranguá, 20 de dezembro de 2024.

Prof. Fabrício Herpich, Dr.
Coordenador do Curso

Banca examinadora

Prof^a. Marina Carradore Sérgio, Dr^a.
Orientadora

Prof. Giovani Mendonça Lunardi, Dr.
Universidade de Federal de Santa Catarina

Prof. Thiago da Silva Fialho, Mestrando
Universidade de Federal de Santa Catarina

Araranguá, 2024.

*Dedico esse trabalho à humanidade,
a todos que acreditaram em mim, e em especial à minha mãe.*

AGRADECIMENTOS

Agradeço imensamente aos meus pais Joel Medeiros e Claudia Goulart que me criaram e custearam minha permanência na universidade. Agradeço também minha orientadora Marina Carradore Sérgio, que diligentemente me orientou e teve a paciência para que este trabalho estivesse pronto.

Também agradeço aos meus primos, que são muitos, meus tios e todo o resto da minha família, em especial a minha avó Tereza Medeiros que fundamentou as bases éticas e morais que possuo, e esteve presente na minha vida apesar de todas as coisas.

Aos meus amigos, que trilharam juntos este caminho que foi a minha presença na universidade, em especial: Raphael Marra e Ricardo Bataglin, aos que estão distantes e compartilhamos nossas vidas apesar disso, agradeço tão somente pela surpresa que foi encontrá-los.

Agradeço a todos os professores, técnicos e colegas que de alguma forma colaboraram com o meu crescimento, principalmente aqueles que nos momentos mais críticos estiveram comigo.

A Universidade Federal de Santa Catarina instituição respeitada e vanguardista na área de Tecnologias da Informação e Comunicação, que foi o palco desse meu crescimento e amadurecimento, seus projetos de pesquisa e monitoria foram peças chaves tanto da minha permanência quanto da minha formação acadêmica.

Agradeço àquela força intrínseca que Espinosa chamou de Deus, manifesto nas leis da natureza e na razão, que se revela na ordem e na beleza do universo, pela inspiração e pela harmonia que guiou este trabalho, minha gratidão por me lembrar da conexão entre todas as coisas e do poder do pensamento.

“A man provided with paper, pencil, and rubber, and subject to strict discipline,
is in effect a universal machine” (Turing, 1912).

RESUMO

Este trabalho analisa e visualiza metadados de Trabalhos de Conclusão de Curso (TCC) da área de Tecnologias da Informação e Comunicação (TIC) da Universidade Federal de Santa Catarina. Utilizando a metodologia Design Science Research (DSRM), foram coletados e tratados dados do repositório institucional para gerar insights sobre a produção acadêmica. Técnicas de processamento de linguagem natural, clusterização e visualização de dados foram aplicadas para identificar padrões, tendências e interseções entre áreas de estudo. Os resultados revelam uma predominância de temas relacionados a dados, inovação tecnológica e educação, com forte contribuição de orientadores específicos e agrupamentos temáticos distintos. A análise evidenciou desafios como a inconsistência de metadados e a necessidade de padronização. As visualizações criadas destacam tendências temporais e áreas de foco acadêmico, contribuindo para a compreensão das direções de pesquisa e para o planejamento estratégico no ensino superior.

Palavras-chave Metadados; Trabalhos de Conclusão de Curso; Visualização de Dados; Tecnologias da Informação e Comunicação; Clusterização.

ABSTRACT

This work analyzes and visualizes metadata from undergraduate thesis projects (TCC) in the field of Information and Communication Technologies (ICT) at the Federal University of Santa Catarina. Using the Design Science Research (DSRM) methodology, data were collected and processed from the institutional repository to generate insights into academic production. Techniques of natural language processing, clustering, and data visualization were applied to identify patterns, trends, and intersections among research areas. The results reveal a predominance of topics related to data, technological innovation, and education, with significant contributions from specific advisors and distinct thematic clusters. The analysis highlighted challenges such as metadata inconsistencies and the need for standardization. The visualizations created emphasize temporal trends and academic focus areas, contributing to the understanding of research directions and strategic planning in higher education.

Keywords: Metadata; Undergraduate Thesis; Data Visualization; Information and Communication Technologies; Clustering.

LISTA DE FIGURAS

Figura 1 - gráfico de barras de exemplo	23
Figura 2 - mapa de calor de exemplo	24
Figura 3 - gráfico de dispersão 2D de exemplo	25
Figura 4 - nuvem de palavras de exemplo.....	26
Figura 5 - exemplo de gráfico de linha.....	27
Figura 6 - exemplo de gráfico de rede	27
Figura 7 - gráfico frequência de números de páginas.....	46
Figura 8 - nuvem de palavras do resumo, título e palavras chaves.....	47
Figura 9 - gráfico das palavras mais frequentes no resumo em português ..	48
Figura 10 - quantidade de trabalhos orientados por orientador	49
Figura 11 - gráfico de distribuição de trabalhos por ano.....	50
Figura 12 – gráfico TCCs por semestre.....	50
Figura 13 - gráfico de quantidade de TCCs acumulados.....	51
Figura 14 - gráfico de rede de orientadores e temas	52
Figura 15 - aproximação do gráfico de rede anterior	53
Figura 16 - gráfico <i>boxplot</i> da distribuição de páginas por cluster.....	54
Figura 17 - dispersão de agrupamentos espaço reduzido PCA	55
Figura 18 -dispersão 3D de agrupamentos espaço reduzido PCA.....	55
Figura 19 - quantidade de TCCs por agrupamento	56
Figura 20 - gráfico de palavras chaves mais frequentes por agrupamento ..	57
Figura 21 - gráfico orientadores por agrupamentos.....	58
Figura 22 - agrupamentos com maior incidência de palavras associadas a educação.....	59
Figura 23 - agrupamentos com maior incidência com desenvolvimento de sistemas e trabalho	60
Figura 24 - agrupamento com predominância de robótica	61
Figura 25 - agrupamento com inovação como foco.....	61
Figura 26 - agrupamento de visualização de dados	62
Figura 27 - agrupamento de redes sociais	63

LISTA DE QUADROS

Quadro 1 - Metodologia DSRM aplicada	36
--	----

LISTA DE TABELAS

Tabela 1 - busca semântica visualização de dados e dinheiro.....	43
Tabela 2 - busca semântica por visualização de dados e chatbot.....	44
Tabela 3 - busca semântica de Inteligência artificial e chatbot.....	45

LISTA DE ABREVIATURAS E SIGLAS

TCC Trabalho de Conclusão de Curso

TIC Tecnologias da Informação e Comunicação

UFSC Universidade Federal de Santa Catarina

LLM *Large Language Models*

HTML *HyperText Markup Language*

HTTP *Hypertext Transfer Protocol*

DOM *Document Object Model*

SUMÁRIO

1	INTRODUÇÃO	16
1.1	OBJETIVOS	17
1.1.1	Objetivo Geral	17
1.1.2	Objetivos Específicos	18
1.2	JUSTIFICATIVA.....	18
1.3	ESTRUTURA E ORGANIZAÇÃO	19
1.4	ADERÊNCIA AO CURSO	20
2	FUNDAMENTAÇÃO TEORICA	21
2.1	VISUALIZAÇÃO DE DADOS	21
2.1.1	A importância da visualização de dados na tomada de decisão	22
2.1.2	Desafios e Limitações	22
2.2	TÉCNICAS DE VISUALIZAÇÕES DE DADOS EMPREGADAS.....	23
2.2.1	Gráfico de Barra	23
2.2.2	Mapa de Calor	24
2.2.3	Gráfico Dispersão	25
2.2.3.1	<i>Gráfico de Dispersão 3D</i>	25
2.2.3.2	<i>Gráfico de Dispersão 2D</i>	25
2.2.4	Nuvem de Palavras	26
2.2.5	Linha	26
2.2.6	Gráfico de rede	27
2.2.7	Gráfico de Caixa (<i>boxplot</i>)	28
2.3	PROCESSAMENTO DE LINGUAGEM NATURAL	28
2.3.1	Aprendizado de Máquina	29
2.3.2	Deep Learning	29
2.3.3	Modelos de Linguagem de Grande Escala (LLMs)	29
2.4	DSPACE	30
2.4.1	Dublin Core	31
2.5	RASPAGEM DE DADOS (WEB SCRAPING).....	32
2.5.1	HTTP (Hypertext Transfer Protocol)	32
2.6	PYTHON	33
3	METODOLOGIA	34

3.1	DESIGN SCIENCE RESEARCH METHODOLOGY	34
3.1.1	Objetivo e Contexto da DSRM.....	35
3.1.2	Metodologia de Design Science Research Methodology	35
3.1.3	Avaliação dos Artefatos na DSRM.....	37
3.2	COLETA E PREPARAÇÃO DOS DADOS	38
3.2.1	Obtenção dos dados.....	38
3.2.2	Preparação dos dados.....	39
3.2.2.1	<i>Colunas deletadas.....</i>	39
3.2.2.2	<i>Dados organizados e adequados categoricamente</i>	39
3.3	ANÁLISE DE DADOS	40
3.4	LIMITAÇÕES DA PESQUISA	41
3.5	RESULTADOS ESPERADOS	41
4	APRESENTAÇÃO E DISCUSSÃO DOS RESULTADOS	43
4.1	BUSCA SEMANTICA.....	43
4.2	VISUALIZAÇÕES REALIZADAS	46
4.2.1	Visualizações e gerais	46
4.2.1.1	<i>Quantidade de Trabalhos por Número de Páginas</i>	46
4.2.1.2	<i>Nuvem de Palavras dos Resumos, Títulos e Palavras-Chave</i>	46
4.2.1.3	<i>Frequência de palavras nos resumos em português.....</i>	47
4.2.1.4	<i>Quantidade de trabalhos por orientador.....</i>	48
4.2.2	Distribuição de TCCs pelo tempo.....	49
4.2.2.1	<i>TCCs por ano.....</i>	49
4.2.2.2	<i>TCCs por semestre</i>	50
4.2.2.3	<i>Evolução temporal acumulada de TCCs por ano</i>	50
4.2.3	Gráficos de rede.....	51
4.2.4	Gráficos de agrupamento (<i>cluster</i>).....	53
4.2.4.1	<i>Páginas de TCCs por agrupamentos Boxplot</i>	53
4.2.4.2	<i>Dispersões dos trabalhos agrupadas</i>	54
4.2.4.3	<i>Distribuição de TCCs por agrupamentos</i>	55
4.2.4.4	<i>Cinco palavras chaves mais relevantes por agrupamento</i>	56
4.2.4.5	<i>Orientadores por agrupamento</i>	57
4.2.4.6	<i>Nuvens de Palavras dos agrupamentos.....</i>	58
4.3	INSIGHTS E INFORMAÇÕES RELEVANTES	63

4.4	LIMITAÇÕES E DESAFIOS	64
5	CONSIDERAÇÕES FINAIS.....	65
	REFERÊNCIAS	67
	APENDICE A – BIBLIOTECAS PYTHON UTILIZADAS.....	71

1 INTRODUÇÃO

Nos últimos anos, a análise e visualização de dados têm se consolidado como ferramentas essenciais para a compreensão de grandes volumes de informações em diversas áreas do conhecimento (Silva, 2019). No contexto acadêmico, os Trabalhos de Conclusão de Curso (TCC) representam uma fonte valiosa de dados sobre a produção científica, refletindo a evolução das áreas de estudo e as principais tendências de pesquisa ao longo do tempo. Essa evolução é particularmente significativa no campo das Tecnologias da Informação e Comunicação (TIC), um setor dinâmico e em constante transformação, onde novas demandas e desafios surgem frequentemente. A análise dos metadados dos TCCs, como títulos, palavras-chave, orientadores e áreas de estudo, pode fornecer *insights* importantes sobre as direções do conhecimento científico e as interseções entre disciplinas.

A visualização de dados destaca-se nesse cenário por facilitar a compreensão de padrões complexos, apresentando-os de maneira acessível e interativa. Conforme Silva (2019), técnicas de visualização de dados não apenas auxiliam na interpretação de informações, mas também potencializam o processo de tomada de decisões ao transformar dados brutos em representações visuais que evidenciam tendências e relações. No entanto, apesar de sua relevância, o uso sistemático de técnicas de visualização aplicadas a metadados acadêmicos ainda é um campo pouco explorado, especialmente no âmbito dos TCCs do curso TIC.

Diante desse contexto, a seguinte pergunta se propõe: “Como técnicas baseadas em análise de agrupamento e busca semântica podem ser aplicadas para analisar documentos textuais em contextos acadêmicos?” para responder essa questão este trabalho propõe investigar como as técnicas de visualização de dados podem ser aplicadas aos metadados dos TCCs do curso TIC da Universidade Federal de Santa Catarina (UFSC), com o objetivo de identificar tendências, padrões e interseções que contribuam para uma melhor compreensão do cenário acadêmico e científico. A delimitação do estudo abrange todos os trabalhos produzidos, permitindo uma análise que capture mudanças significativas nas áreas de pesquisa e temas de interesse.

A relevância desta pesquisa está diretamente relacionada ao potencial de contribuir para a gestão do conhecimento nas instituições de ensino superior, permitindo uma visão mais ampla sobre os focos de estudo e os principais desafios enfrentados por estudantes e orientadores. Conforme destacado por Holanda e Gontijo (2024), o mapeamento de tendências acadêmicas pode subsidiar decisões estratégicas na definição de linhas de pesquisa, priorização de recursos e alinhamento curricular.

Entre os principais desafios identificados neste trabalho, destaca-se a diversidade e heterogeneidade dos dados disponíveis nos TCCs, que podem variar em formato, profundidade e detalhamento. Além disso, a construção de visualizações significativas exige o uso de ferramentas e técnicas avançadas de análise, bem como uma compreensão clara dos objetivos pretendidos. Estudos recentes, como os de Rodrigues e Dias (2017), ressaltam a importância de técnicas como análise de redes e diagramas de dispersão para a identificação de *clusters* de pesquisa e relações entre áreas de estudo.

Por fim, espera-se que os resultados deste trabalho contribuam não apenas para a área de TIC, mas também sirvam como modelo para análises similares em outros campos do conhecimento. Ao integrar técnicas de visualização de dados com a análise de metadados acadêmicos, o estudo oferece uma abordagem inovadora e promissora para o entendimento da produção científica e suas implicações.

1.1 OBJETIVOS

Para melhor compreensão dos objetivos deste projeto, estes foram divididos em objetivo geral e objetivos específicos, delimitados a seguir.

1.1.1 Objetivo Geral

O objetivo deste trabalho é investigar técnicas de visualização de dados aos metadados dos Trabalhos de Conclusão de Curso na área de Tecnologias da Informação e Comunicação da Universidade Federal de Santa Catarina, visando gerar informações que corroborem com o entendimento e indicadores do curso.

1.1.2 Objetivos Específicos

- Analisar os metadados dos TCCs de TIC para identificar tendências e padrões de temas e áreas de pesquisa ao longo do tempo;
- Aplicar técnicas de visualização de dados para representar graficamente as interseções entre as áreas de estudo dos TCCs e suas contribuições para o conhecimento científico;
- Investigar as relações entre os diferentes tópicos de pesquisa através de visualizações que permitam identificar clusters ou agrupamentos significativos.

1.2 JUSTIFICATIVA

A aplicação de técnicas de visualização de dados aos metadados de TCC na área de Tecnologias da Informação e Comunicação é fundamental para compreender a evolução do conhecimento científico e as interseções entre suas áreas de estudo. Segundo Santos e Perry (2023), os painéis de visualização educacional têm sido amplamente adotados para oferecer *insights* sobre o desempenho acadêmico e padrões de pesquisa em diferentes níveis de ensino, proporcionando informações acessíveis e organizadas a partir de grandes volumes de dados. Isso demonstra a relevância de tecnologias que integrem análises visuais no contexto educacional.

A integração de ferramentas de análise visual no ambiente acadêmico tem o potencial de melhorar a compreensão de padrões em produções científicas. Linhares (2020) destaca que ferramentas de visualização educacional auxiliam tanto gestores quanto pesquisadores ao identificar tendências de pesquisa e lacunas no conhecimento, por meio de representações gráficas de dados qualitativos e quantitativos. Tais técnicas contribuem para uma análise mais eficaz e precisa, gerando relatórios dinâmicos que podem orientar políticas educacionais e estratégias de ensino.

Além disso, Stoll et al. (2019) demonstram que as tecnologias de visualização de dados acadêmicos desempenham um papel essencial na análise preditiva e recomendação de soluções no ensino superior. Combinando modelos preditivos e visualizações, sua aplicação permite identificar padrões significativos de aprendizagem e desempenho, otimizando o acompanhamento de estudantes e facilitando a tomada de decisões com base em dados confiáveis e visualmente representados.

No contexto das pesquisas de TCC em TIC, o uso de técnicas de visualização de dados para explorar os metadados pode revelar tendências e agrupamentos significativos entre os temas abordados. Estudos recentes como o de Santos e Perry (2023) evidenciam que essas abordagens não apenas aumentam a acessibilidade dos dados, mas também oferecem uma compreensão mais profunda das interseções temáticas entre diferentes áreas do conhecimento.

1.3 ESTRUTURA E ORGANIZAÇÃO

Este trabalho está organizado da seguinte forma:

- **Capítulo 1: Introdução:** Neste capítulo, são apresentados o tema, a justificativa, o objetivo geral e os objetivos específicos da pesquisa. Além disso, são discutidos os principais conceitos e a relevância da visualização de dados no contexto acadêmico, especialmente em Trabalhos de Conclusão de Curso (TCC);
- **Capítulo 2: Fundamentação Teórica:** O capítulo dedica-se à revisão de literatura sobre visualização de dados, abordando suas técnicas, aplicações e impacto no campo acadêmico. Também são discutidos os conceitos de metadados, especialmente no contexto de TCCs, e as técnicas usadas para análise de dados acadêmicos;
- **Capítulo 3: Metodologia:** Apresenta a metodologia adotada para a coleta e análise dos metadados dos TCCs, descrevendo as fontes de dados, as ferramentas de visualização utilizadas e o processo de análise dos dados;

- **Capítulo 4: Resultados e Discussão:** Aqui são apresentados os resultados da análise dos metadados dos TCCs, acompanhados das visualizações geradas. São discutidos os padrões identificados, as interseccionalidades entre as áreas de pesquisa e as implicações desses resultados para o avanço do conhecimento em TIC;
- **Capítulo 5: Conclusão:** Este capítulo sintetiza as principais descobertas da pesquisa, destacando a contribuição da visualização de dados na compreensão dos TCCs. Além disso, são feitas sugestões para trabalhos futuros;
- **Referências:** Apresenta as referências bibliográficas utilizadas ao longo do trabalho.

1.4 ADERÊNCIA AO CURSO

A análise dos metadados de trabalhos acadêmicos do curso de TIC é uma possibilidade ímpar de entender mais sobre o próprio curso, o que ressoa com o caráter de pesquisa e inovação sendo uma nova perspectiva dos TCCs do curso.

O seguinte trabalho enquadra-se na área de visualização de dados, o repositório específico do curso resulta 26 trabalhos de conclusão de curso nessa área. Após análise dos resumos de todos os trabalhos, identificou-se apenas 3 trabalhos que também estão vinculados a processamento de texto, sendo eles: uma arquitetura para descoberta de conhecimento a partir de bases textuais (Silva, 2012), um sistema voltado à indexação e recuperação de informação integrado à ontologia (Bitencourt, 2013) e uma arquitetura de descoberta de conhecimento baseada na correlação e associação temporal de padrões textuais (Sérgio, 2013).

Apesar disso nenhum trabalho em específico trata de trabalhos de conclusão de curso ou trabalhos acadêmicos, essa pesquisa demonstra a falta de análises de produções acadêmicas dentro do TIC, e como esse é um nicho amplo para as pesquisas.

2 FUNDAMENTAÇÃO TEORICA

Nesse capítulo são explorados os termos e conceitos utilizados durante o trabalho para fundamentar o entendimento dele.

2.1 VISUALIZAÇÃO DE DADOS

A visualização de dados é uma ferramenta essencial no contexto atual, onde há uma crescente geração de grandes volumes de informações. Consiste na representação gráfica de dados abstratos e estruturados com o objetivo de facilitar a compreensão, análise e comunicação de informações complexas. Esta área interdisciplinar combina conceitos de estatística, ciência da computação, *design* e psicologia cognitiva, oferecendo métodos eficazes para explorar, interpretar e comunicar grandes conjuntos de dados de maneira intuitiva e acessível.

De acordo com Card, Mackinlay e Shneiderman (1999), a visualização de dados transforma informações abstratas em representações visuais interativas, permitindo que padrões, tendências e relações sejam mais facilmente identificados. Embora sua relevância tenha se intensificado com a explosão de dados digitais, suas raízes remontam a práticas históricas. Com o tempo, a evolução da tecnologia e dos métodos computacionais expandiu as ferramentas e técnicas disponíveis, tornando a visualização de dados mais sofisticada e acessível.

A capacidade da visualização de dados de transmitir informações complexas de maneira clara e acessível é uma de suas características mais notáveis. Como argumenta Munzner (2014), gráficos bem projetados podem reduzir drasticamente a carga cognitiva, sintetizando informações complexas em imagens simples e compreensíveis.

Um exemplo significativo dessa aplicação é o uso de *dashboards* para monitoramento em tempo real, como os utilizados amplamente durante a pandemia de COVID-19. Esses *dashboards* permitiram que governos e organizações comunicassem dados de forma clara, acessível e visualmente atraente, promovendo respostas rápidas e informadas em um cenário de crise global (Bohnert; Pavão; Silva, 2022).

A visualização de dados não apenas facilita a comunicação de informações, mas também contribui para a tomada de decisões, sendo fundamental em diversas áreas, como saúde, negócios, educação e governança. Ao fornecer *insights* visuais de forma eficaz, ela transforma dados brutos em conhecimento acionável, permitindo que os tomadores de decisão compreendam o panorama completo e ajam com base em informações precisas e claras (Regly; Souza, 2023).

2.1.1 A importância da visualização de dados na tomada de decisão

A visualização de dados vai além de simplesmente apresentar informações, pois desempenha um papel crucial no processo de tomada de decisão (Silva, 2007). Ferramentas modernas como *Tableau* e *Power BI* permitem que analistas explorem diferentes perspectivas de um conjunto de dados e identifiquem *insights*. Pesquisas indicam que visualizações interativas são especialmente eficazes para detectar anomalias e analisar tendências históricas. De acordo com Heer, Bostock e Ogievetsky (2010), ferramentas interativas oferecem uma exploração mais profunda dos dados, incentivando a descoberta de padrões que não seriam facilmente identificados em representações estáticas.

No contexto corporativo, *dashboards* desempenham um papel fundamental na análise de desempenho, permitindo que gestores monitorem métricas-chave em tempo real. Essas ferramentas não apenas promovem a transparência, mas também permitem decisões mais ágeis e baseadas em dados, que são essenciais para a competitividade no ambiente empresarial atual.

2.1.2 Desafios e Limitações

Apesar de seu vasto potencial, a visualização de dados apresenta desafios significativos. O primeiro é a alfabetização em dados: muitos usuários não possuem o conhecimento necessário para interpretar gráficos complexos de forma precisa, o que pode levar a erros de interpretação. Além disso, o excesso de informações visuais ou o uso inadequado de gráficos pode comprometer a clareza da mensagem, um fenômeno conhecido como "sobrecarga de dados" (FEW, 2012).

Outro desafio é a representação de grandes volumes de dados. Embora ferramentas modernas ofereçam recursos avançados, a criação de visualizações que sejam simultaneamente informativas e visualmente limpas requer um equilíbrio cuidadoso entre *design* e funcionalidade. Como observam Yau (2013) e Cairo (2016), é fundamental que os designers de visualizações compreendam não apenas os dados, mas também o público-alvo e o contexto em que as informações serão consumidas.

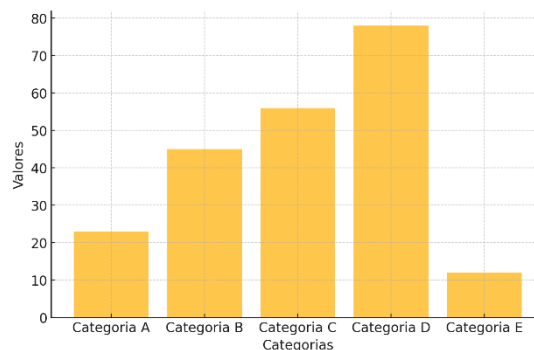
2.2 TÉCNICAS DE VISUALIZAÇÕES DE DADOS EMPREGADAS

Dentre as várias técnicas de visualizações de dados, pode-se citar o: gráfico de barra, mapa de calor (*heatmap*), gráfico de dispersão de 2D e 3D, nuvens de palavras, gráficos de linhas, de caixa e gráfico de redes, estão descritas de forma detalhada abaixo.

2.2.1 Gráfico de Barra

Os gráficos de barra são amplamente utilizados para comparar categorias distintas em termos de valores quantitativos. Eles são eficazes quando se deseja exibir dados categóricos, permitindo que as diferenças entre as categorias sejam visualizadas de forma clara. Segundo Few (2012) gráficos de barra são uma das

Figura 1 - gráfico de barras de exemplo



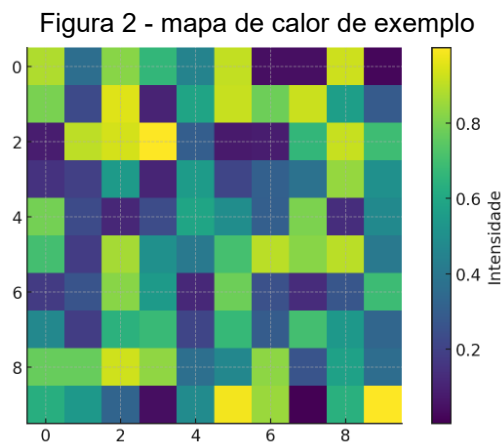
Fonte: elaborado pelo autor 1

formas mais intuitivas de visualizar relações quantitativas, pois a altura ou comprimento das barras facilita a comparação de valores entre categorias distintas. Além disso, os gráficos de barra podem ser dispostos horizontal ou verticalmente, dependendo do espaço disponível e do objetivo da análise.

Outra vantagem dos gráficos de barra é sua capacidade de comunicar de forma precisa e eficiente, quando bem projetados. É crucial evitar elementos desnecessários, como efeitos 3D, que podem confundir a interpretação dos dados Few (2012). Quando utilizados corretamente, eles são ideais para destacar discrepâncias ou tendências entre categorias, sendo ferramentas indispensáveis em análises quantitativas simples.

2.2.2 Mapa de Calor

Heatmaps, ou mapas de calor, são representações gráficas que utilizam escalas de cores para exibir dados quantitativos em uma matriz bidimensional. Eles são particularmente eficazes para identificar padrões e anomalias em grandes conjuntos de dados. Few (2012) destaca que a aplicação cuidadosa de cores em



Fonte: elaborado pelo autor

heatmaps permite que o espectador compreenda rapidamente as relações e os destaques nos dados (p. 61). Por exemplo, em análises de correlação, a intensidade das cores ajuda a ilustrar a força e a direção da relação entre variáveis. Além de sua capacidade exploratória, os *heatmaps* são uma ferramenta visual poderosa para apresentações e relatórios, ajudando a simplificar a comunicação de insights complexos.

2.2.3 Gráfico Dispersão

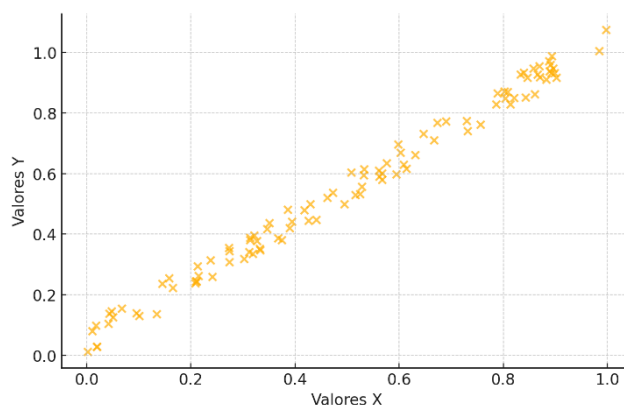
2.2.3.1 Gráfico de Dispersão 3D

Os gráficos de dispersão em três dimensões são usados quando há a necessidade de explorar relações entre três variáveis quantitativas. Embora possam ser úteis em contextos específicos, Few (2012) alerta que a adição de uma terceira dimensão frequentemente compromete a clareza e dificulta a interpretação dos dados (p. 197). Portanto, seu uso deve ser limitado a situações em que o benefício da visualização tridimensional supera os desafios impostos pela complexidade adicional.

2.2.3.2 Gráfico de Dispersão 2D

Os gráficos de dispersão em duas dimensões são usados para ilustrar a relação entre duas variáveis quantitativas. Few (2012) ressalta que esses gráficos são ideais para identificar padrões, tendências ou agrupamentos nos dados, uma vez que cada ponto no gráfico representa um par de valores. Por exemplo, ao analisar a relação entre despesas de marketing e vendas, um gráfico de dispersão pode ajudar a revelar correlações ou comportamentos anômalos.

Figura 3 - gráfico de dispersão 2D de exemplo



Fonte: elaborado pelo autor

É importante projetar gráficos de dispersão com clareza, garantindo que os eixos estejam bem definidos e que os pontos sejam representados de forma legível (Few (2012)). A simplicidade no design é essencial para que os espectadores possam identificar facilmente as informações importantes e tirar conclusões precisas.

2.2.4 Nuvem de Palavras

Word clouds, ou nuvens de palavras, são ferramentas visuais amplamente empregadas para resumir textos de forma visual e intuitiva, destacando as palavras mais frequentes ou relevantes. De acordo com Lohmann et al. (2015), elas utilizam o tamanho das fontes para representar a frequência ou relevância de termos, enquanto outras propriedades visuais, como cores e posicionamento, podem ser usadas para enriquecer a apresentação gráfica. Essa abordagem facilita uma compreensão inicial do conteúdo textual, permitindo que o leitor identifique rapidamente os principais tópicos ou padrões presentes em um conjunto de dados textuais.

Figura 4 - nuvem de palavras de exemplo

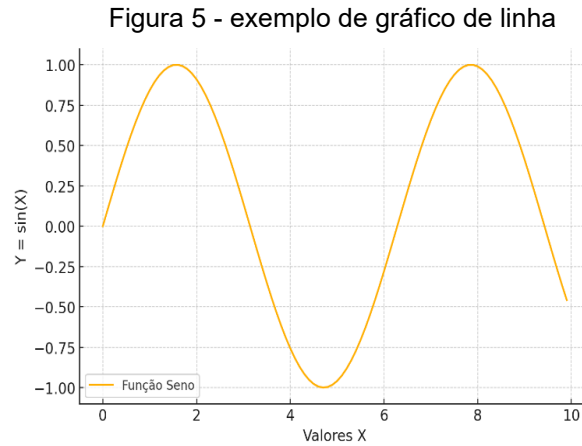


Fonte: elaborado pelo autor

Apesar de serem úteis para visualização inicial, as *word clouds* apresentam limitações quando aplicadas em análises comparativas de múltiplos textos. Como destaca Lohmann et al. (2015), essas limitações incluem a dificuldade em diferenciar fontes textuais e em capturar nuances de relações entre palavras em diferentes contextos. Entretanto, sua aplicação continua a ser uma escolha valiosa em estudos exploratórios e em apresentações visuais, por sua capacidade de condensar e comunicar informações complexas de maneira visualmente acessível e de fácil compreensão.

2.2.5 Linha

Os gráficos de linha são representações visuais que conectam pontos de dados por meio de linhas, facilitando a análise de tendências e variações ao longo do tempo ou em séries contínuas. De acordo com Few (2012), eles são ideais para



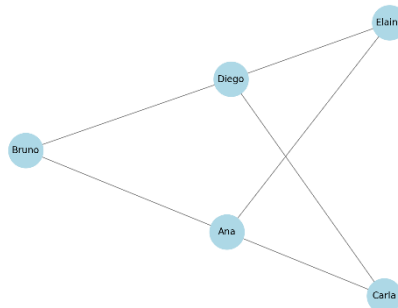
Fonte: elaborado pelo autor

sintetizar dados complexos e oferecer uma interpretação clara de fenômenos temporais ou relacionais. Sua eficácia está ligada à clareza na escolha de escalas, rótulos nos eixos e precisão na representação dos pontos, garantindo que as informações sejam transmitidas com objetividade (Rosen; Quadri, 2020).

2.2.6 Gráfico de rede

Os gráficos de rede, ou grafos, são representações matemáticas que modelam sistemas compostos por entidades interconectadas, denominadas nós, e suas interações, chamadas arestas. Essas estruturas são fundamentais para a análise de sistemas complexos em diversas áreas, como sociologia, biologia e ciência da computação, permitindo a visualização e compreensão das relações e padrões subjacentes entre os componentes do sistema (Barabási; Pósfai, 2016).

Figura 6 - exemplo de gráfico de rede



Fonte: elaborado pelo autor

Por exemplo, na sociologia, os grafos são utilizados para mapear e estudar as interações sociais, identificando padrões de comportamento e estruturas de poder dentro de uma comunidade (Scott, 2017).

2.2.7 Grafico de Caixa (*boxplot*)

O *boxplot*, ou gráfico de caixa, é uma ferramenta estatística eficaz para resumir e visualizar distribuições de dados quantitativos. Ele exibe medidas de tendência central, como a mediana, e de dispersão, como os quartis, além de identificar possíveis valores atípicos (*outliers*). A estrutura básica do *boxplot* inclui uma caixa que se estende do primeiro ao terceiro quartil, com uma linha interna representando a mediana. Linhas, conhecidas como "bigodes", se estendem dos quartis até os valores mínimos e máximos dentro de um intervalo específico, geralmente 1,5 vezes o intervalo interquartil. Valores fora desse intervalo são considerados outliers e são plotados individualmente. Essa representação gráfica permite uma análise rápida da distribuição dos dados, facilitando a identificação de assimetrias, dispersões e valores extremos, sendo amplamente utilizada em pesquisas científicas para a interpretação de dados quantitativo (Neto, 2017).

2.3 PROCESSAMENTO DE LINGUAGEM NATURAL

Jurafsky e Martin (2009) definem que o processamento de Linguagem Natural (PLN) é uma área da inteligência artificial que se dedica ao estudo das interações entre computadores e linguagens humanas naturais. Ele combina técnicas de linguística computacional e aprendizado de máquina para permitir que sistemas computacionais compreendam, interpretem e produzam texto ou fala de maneira semelhante aos seres humanos. O PLN abrange tarefas como análise de sentimentos, tradução automática e extração de informações. Sua evolução tem sido impulsionada pela aplicação de algoritmos de aprendizado profundo e modelos de linguagem avançados, que oferecem maior precisão na interpretação semântica (Jurafsky e Martin (2009)).

2.3.1 Aprendizado de Máquina

Aprendizado de máquina mais conhecido como *Machine Learning* (ML) é uma subárea da inteligência artificial que utiliza algoritmos para construir modelos capazes de generalizar padrões a partir de dados (Ludermir, 2021). Esses algoritmos podem ser supervisionados, não supervisionados ou por reforço, dependendo da natureza do problema. No contexto do PLN, o aprendizado de máquina é frequentemente utilizado para tarefas como categorização de texto, classificação de documentos e análise de tópicos. Modelos clássicos, como regressão logística e máquinas de vetores de suporte, foram amplamente aplicados antes do advento das técnicas de aprendizado profundo (Murphy 2014).

2.3.2 Deep Learning

O *Deep Learning* (aprendizado profundo) é uma subárea do aprendizado de máquina que utiliza redes neurais artificiais para modelar dados complexos (Lecun; Bengio; Hinton, 2015). Essa abordagem tem sido particularmente eficaz em tarefas de PLN, como tradução automática, reconhecimento de fala e geração de texto. Redes neurais profundas são compostas por várias camadas de unidades computacionais, que permitem que o modelo aprenda representações hierárquicas e não-lineares dos dados. A introdução de redes neurais convolucionais (CNNs) e redes neurais recorrentes (RNNs) tem permitido avanços significativos no entendimento e na geração de linguagem natural (Lecun; Bengio; Hinton, 2015). O uso de *deep learning* em PLN resultou em modelos mais robustos e precisos, que superaram abordagens tradicionais em uma série de tarefas, especialmente em modelos de grande escala que podem entender nuances complexas da linguagem.

2.3.3 Modelos de Linguagem de Grande Escala (LLMs)

Os Modelos de Linguagem de Grande Escala (LLMs, do inglês *Large Language Models*) são redes neurais profundas treinadas em grandes volumes de texto para compreender, gerar e manipular linguagem natural (Moraes et al., 2024). Esses modelos, como o GPT-3, BERT e T5, têm sido aplicados em uma ampla

gama de tarefas de PLN, desde a geração de texto até a tradução e a análise de sentimentos (Brown et al., 2020). A principal vantagem dos LLMs é sua capacidade de aprender representações contextuais da linguagem, o que permite um desempenho superior em tarefas que envolvem ambiguidade ou contexto. Esses modelos são treinados utilizando grandes quantidades de dados não estruturados, o que possibilita uma compreensão profunda e sofisticada da linguagem natural. O uso de LLMs tem revolucionado a forma como sistemas de PLN são aplicados, oferecendo soluções mais eficazes e precisas, especialmente em áreas como o processamento de grandes volumes de texto acadêmico, como no caso da análise de metadados de trabalhos de conclusão de curso (Vaswani et al., 2023).

2.4 DSPACE

O *DSpace* é um sistema de repositório digital de código aberto desenvolvido pelo *MIT Libraries* e *Hewlett-Packard Labs*, utilizado pela Universidade Federal de Santa Catarina (UFSC) para o armazenamento de seus trabalhos acadêmicos, como monografias, teses e dissertações. Este sistema foi projetado para armazenar, gerenciar, indexar e distribuir materiais de pesquisa e publicações acadêmicas em formatos digitais, sendo uma ferramenta crucial para instituições de pesquisa multidisciplinares. Ele possibilita que diferentes departamentos e centros de pesquisa gerenciem suas próprias submissões de maneira personalizada, atendendo às necessidades específicas de cada área (Smith et al., 2003).

O *DSpace* utiliza um modelo de metadados baseado no *Dublin Core*, que qualifica e organiza a descrição dos itens, facilitando a busca e a recuperação de informações. A interface do usuário é acessível via web e oferece diferentes níveis de acesso, com funções distintas para depositantes, usuários finais e administradores do sistema. O sistema também suporta a preservação digital, assegurando que os materiais arquivados permaneçam acessíveis e utilizáveis ao longo do tempo, mesmo com a evolução das tecnologias (Smith et al., 2003).

DSpace foi projetado para ser adaptável e expansível, permitindo que outras instituições possam utilizá-lo e modificá-lo conforme suas necessidades locais. A *DSpace Federation*, que fomenta a colaboração entre diferentes instituições,

promove a interoperabilidade e o desenvolvimento contínuo do sistema, tornando-o uma plataforma dinâmica e em constante aprimoramento (Smith et al., 2003).

2.4.1 Dublin Core

O Dublin Core é um padrão internacional de metadados utilizado para a descrição e organização de recursos digitais, facilitando sua recuperação e acessibilidade. Esse padrão é composto por 15 elementos essenciais, que são utilizados para fornecer informações detalhadas e estruturadas sobre os recursos, promovendo sua identificação e facilitando o processo de pesquisa e recuperação de dados. Esses elementos são:

1. Título: Representa o nome dado ao recurso.
2. Criador: Indica a entidade responsável pela criação do conteúdo.
3. Assunto: Descreve o tema ou os tópicos abordados no recurso.
4. Descrição: Oferece um resumo ou detalhes sobre o conteúdo do recurso.
5. Publicador: Refere-se à entidade responsável por tornar o recurso disponível.
6. Colaborador: Menciona outras contribuições intelectuais ao conteúdo.
7. Data: Especifica a data associada ao recurso.
8. Tipo: Define a natureza ou o gênero do conteúdo.
9. Formato: Descreve a manifestação física ou digital do recurso.
10. Identificador: Representa uma referência única para o recurso.
11. Fonte: Indica o recurso de onde o presente foi derivado.
12. Idioma: Especifica a língua do conteúdo.
13. Relação: Descreve a relação do recurso com outros recursos.
14. Cobertura: Determina o escopo ou a extensão do recurso, seja no aspecto temporal ou espacial.
15. Direitos autorais: Informa sobre os direitos atribuídos ao recurso.

Esses elementos são utilizados para organizar e classificar recursos digitais de maneira padronizada, permitindo sua fácil busca e recuperação. O Dublin Core é adotado em repositórios digitais, bibliotecas, arquivos e outros sistemas de gestão

de informações digitais, por sua simplicidade e aplicabilidade em diferentes contextos (Alves; Souza, 2007).

2.5 RASPAGEM DE DADOS (WEB SCRAPING)

Raspagem de dados mais conhecido como *Web Scrapping* é definido no contexto de sistemas de dados como o processo de extração automatizada de dados a partir de páginas *HTML* acessíveis por *HTTP* (Bressoud e White, 2020, p. 681-693). Essa técnica envolve a aplicação de métodos programáticos para acessar, navegar e coletar informações de sites, muitas vezes com o objetivo de organizar os dados extraídos em formatos estruturados, como tabelas ou bancos de dados. Segundo Bressoud e White (2020), essa prática combina conceitos de modelos de dados hierárquicos com operações de rede e *HTTP*, permitindo a integração de dados não estruturados de páginas web com sistemas de dados tradicionais.

Embora o *web scraping* seja uma prática comum e amplamente utilizada em diversas áreas, muitos países, incluindo o Brasil, ainda carecem de uma legislação clara que regule essa atividade. Como observa Pedrina (2024), a situação torna-se ainda mais complexa quando grandes empresas utilizam essa técnica para obter vantagens competitivas ou realizar o rastreamento de dados pessoais, levantando sérias questões éticas. Esse cenário nos impõe a necessidade urgente de discutir o equilíbrio entre a inovação tecnológica e as implicações éticas de seu uso, especialmente quando se trata do impacto da coleta de dados sem o devido consentimento e os riscos associados à privacidade e à segurança da informação.

2.5.1 HTTP (Hypertext Transfer Protocol)

O *Hypertext Transfer Protocol (HTTP)* é um protocolo de comunicação utilizado para a transferência de dados entre clientes e servidores na web. Ele opera como a base para a troca de informações na *World Wide Web*, possibilitando a transmissão de hipertextos, imagens, vídeos e outros tipos de dados. Desenvolvido inicialmente por Tim Berners-Lee no final dos anos 1980, o *HTTP* evoluiu para

atender às crescentes demandas de desempenho e segurança da internet moderna (Tim Berners-Lee et al., 1999)

O funcionamento do *HTTP* é baseado em um modelo de requisição e resposta. O cliente, geralmente um navegador web, envia uma requisição ao servidor solicitando recursos, como páginas *HTML* ou arquivos. Em resposta, o servidor retorna o recurso solicitado ou uma mensagem de erro, caso o recurso não esteja disponível. Cada troca de mensagens no *HTTP* segue um formato padronizado, composto por um cabeçalho, que contém informações sobre a requisição ou resposta, e, opcionalmente, um corpo que transporta os dados requisitados (Tim Berners-Lee et al., 1999).

O protocolo desempenha um papel essencial na comunicação web moderna, suportando aplicações desde simples páginas estáticas até complexos sistemas interativos. Sua evolução contínua reflete os esforços para atender às demandas por maior velocidade, segurança e flexibilidade no ambiente digital.

2.6 PYTHON

De acordo com a Python Software Foundation ([s.d.]) Python é uma linguagem de programação interpretada, interativa e orientada a objetos, que incorpora módulos, exceções, tipagem dinâmica e tipos de dados de alto nível. Ela suporta múltiplos paradigmas de programação, incluindo a programação procedural, orientada a objetos e funcional, oferecendo ao desenvolvedor uma sintaxe clara e simples, além de interfaces para diversas chamadas e bibliotecas do sistema, bem como para múltiplos sistemas de janelas. Além disso, o Python é extensível por meio de linguagens como C ou C++, sendo também utilizado como linguagem de extensão para aplicações que necessitam de uma interface programável. Sua portabilidade permite a execução em diversas variantes do Unix, incluindo Linux e macOS, além do Windows. Nesse trabalho empregam-se várias bibliotecas de recursos que se encontram listadas e definidas no “APENDICE A – Bibliotecas Python Utilizadas”, já que o foco do disposto trabalho não está na codificação.

3 METODOLOGIA

Este capítulo tem como objetivo fornecer uma descrição dos métodos e técnicas empregados na condução da pesquisa. Como metodologia da elaboração do trabalho, a metodologia escolhida foi a *Design Science Research Methodology (DSRM)*, visto que a *Design Science Research (DSR)* visa compreender os fenômenos dos sistemas de informação.

3.1 DESIGN SCIENCE RESEARCH METHODOLOGY

A *DSRM* proposta por Peffers *et. all* (2007) é um modelo conceitual para realizar pesquisas em ciência do design (DS) no campo de sistemas de informação (IS).

De acordo com Geerts (2011), as pesquisas nas ciências naturais seguem uma estrutura padrão composta por etapas como definição de problema, revisão de literatura, desenvolvimento de hipóteses, coleta de dados, análise, apresentação de resultados e discussão. Em contrapartida, as ciências do “artificial”, como as define Simon (2019), concentram-se na criação de artefatos projetados para alcançar objetivos específicos. O propósito desses projetos é transformar situações existentes em preferenciais (Simon, 1996).

Na abordagem de *Design Science Research Methodology (DSRM)*, os artefatos têm um papel central. São criações humanas voltadas para propósitos práticos, com o objetivo de resolver problemas e melhorar práticas existentes Geerts (2011). Segundo March e Smith (1995), os artefatos podem ser classificados em quatro tipos principais:

1. **Conceitos:** Vocabulário do domínio, que descreve os termos e as tarefas associadas ao problema.
2. **Modelos:** Relações entre conceitos que representam objetos ou fenômenos do mundo real.
3. **Métodos:** Conjunto de passos sistemáticos para executar tarefas específicas dentro de um contexto determinado.
4. **Instâncias:** Concretizações dos artefatos em seus ambientes, que demonstram a viabilidade e eficácia da solução proposta.

Geerts (2011) destaca que, para que esses artefatos sejam eficazes, eles devem possuir características de relevância e novidade. De acordo com Hevner et al. (2004), a *DSR* deve focar na resolução de problemas não resolvidos ou na melhoria significativa de soluções existentes, sempre buscando inovação no processo de criação dos artefatos. Esse enfoque prático e inovador permite que a *DSR* faça contribuições significativas para o avanço do conhecimento aplicado e da solução de problemas no mundo real.

3.1.1 Objetivo e Contexto da DSRM

O objetivo da *DSRM* é aprimorar a compreensão dos fenômenos relacionados a sistemas de informação, por meio da criação de artefatos tecnológicos que incorporam soluções para problemas previamente identificados (Hevner; Chatterjee, 2010). Em um mundo globalizado, esses artefatos devem ser adaptáveis a diversos contextos culturais (Lawrence; Tuunanen; Myers, 2010).

3.1.2 Metodologia de Design Science Research Methodology

A metodologia *Design Science Research (DSRM)*, proposta por Peffers et al. (2007), é amplamente adotada em projetos de sistemas de informação (Geerts (2011)). Ela é composta por seis etapas, que podem ser adaptadas conforme a necessidade do projeto, sendo elas:

1. Identificação do problema e motivação: Esta etapa envolve a definição clara do problema a ser resolvido e a justificativa de sua relevância. Também é crucial a revisão do estado da arte, para situar a pesquisa no contexto atual, destacando a importância da solução proposta para o avanço do conhecimento ou para a resolução de desafios práticos.
2. Definição dos resultados esperados: Nessa fase, são estabelecidos objetivos específicos e mensuráveis para a solução. Além disso, é importante considerar a viabilidade da solução e explorar alternativas ou soluções previamente apresentadas, garantindo que os resultados atendam às necessidades identificadas na fase inicial.

3. Projeto e desenvolvimento: Aqui, o artefato é projetado e desenvolvido. Esta etapa envolve a definição das funcionalidades essenciais, bem como a arquitetura do sistema ou solução proposta. A construção do artefato deve ser orientada pelas necessidades identificadas nas etapas anteriores, com o objetivo de resolver de maneira eficaz o problema delineado.
4. Demonstração: Após o desenvolvimento, é fundamental demonstrar a utilidade e a eficácia do artefato. Isso pode ser feito por meio de experimentos, simulações, protótipos ou estudos de caso, que evidenciem como o artefato pode ser aplicado no contexto real e como ele resolve o problema proposto.
5. Avaliação: A avaliação analisa o desempenho do artefato em relação aos objetivos definidos e como ele resolve o problema identificado. Essa etapa pode levar à necessidade de ajustes ou melhorias no artefato, que serão feitas em ciclos de refinamento, caso necessário.
6. Comunicação: Por fim, a etapa de comunicação visa divulgar os resultados obtidos, destacando o problema abordado, a relevância da solução proposta e o artefato desenvolvido. A comunicação é essencial para disseminar o conhecimento gerado e assegurar que a contribuição seja reconhecida tanto no ambiente acadêmico quanto no profissional.

Essa metodologia permite uma abordagem estruturada e iterativa para o desenvolvimento de soluções inovadoras, combinando a teoria com a prática para resolver problemas reais de maneira eficaz e eficiente. Abaixo o quadro preenchido com as etapas da metodologia.

Quadro 1 - Metodologia DSRM aplicada

Identificação do Problema	Análise dos trabalhos de conclusão de curso (TCCs) disponíveis no repositório da Universidade Federal de Santa Catarina (UFSC), com o objetivo de identificar <i>insights</i> , oportunidades e tendências relevantes.
Definição dos Requisitos	Estabelecimento de critérios para selecionar, processar e analisar os TCCs.
Design da Solução	Realização de diversas visualizações dos dados.

Desenvolvimento e implementação	Emprego de: <i>scraping</i> de dados, aplicação de técnicas de agrupamento, visualização, processamento de linguagem natural, aproximação por uso de algoritmo <i>fuzzy</i> , aplicação de algoritmos de aproximação <i>K-means</i> . Com o intuito de gerar visualizações e dados que facilitem a interpretação de dados.
Avaliação e Confirmação	Análise dos agrupamentos gerados para avaliar a precisão e relevância das agrupações e validação da qualidade das informações.
Comunicação dos resultados	Apresentação dos <i>insights</i> e dados obtidos bem como suas interpretações derivadas, enfatizando as tendências e interseccionalidades.

Fonte: Elaborado pelo autor.

3.1.3 Avaliação dos Artefatos na DSRM

A etapa de avaliação é essencial para garantir a confiabilidade dos resultados. Segundo Lacerda et al. (2013), cuidados e procedimentos específicos são necessários para esse processo. Os métodos de avaliação incluem:

- **Observacional:** Estudos de caso ou de campo analisam o artefato em profundidade ou em múltiplos projetos.
- **Analítico:** Análises estatísticas, de arquitetura, otimização ou dinâmicas avaliam as propriedades do artefato.
- **Experimental:** Experimentos controlados ou simulações testam o artefato em ambientes artificiais.
- **Teste:** Testes funcionais (caixa preta) e estruturais (caixa branca) identificam falhas.
- **Descritivo:** Argumentos informados ou cenários demonstram a utilidade do artefato com base em pesquisas existentes.

Esses métodos asseguram a eficácia e a aplicabilidade do artefato, contribuindo para a relevância da pesquisa. Esta abordagem metodológica orientada pela *DSR* se torna uma ferramenta poderosa para a criação e avaliação de soluções

inovadoras em sistemas de informação, promovendo avanços significativos em diferentes contextos acadêmicos e práticos.

3.2 COLETA E PREPARAÇÃO DOS DADOS

Os metadados dos TCCs foram coletados por meio da técnica de Web Scraping, aplicados na plataforma de repositório institucional da UFSC, que utiliza o software DSpace como solução para gerenciar e disponibilizar os conteúdos do repositório institucional (UFSC, 2015). A coleção selecionada para o estudo foi intitulada "TCC Tecnologias da Informação e Comunicação (Araranguá)", e os dados foram coletados em 1º de novembro de 2024.

3.2.1 Obtenção dos dados

A extração de dados foi realizada utilizando bibliotecas da linguagem Python, com destaque para a *Requests*, responsável pela obtenção do conteúdo *HTML* das páginas, e a *BeautifulSoup*, empregada para a seleção de elementos a partir da árvore *DOM (Document Object Model)* da página. Inicialmente, aproveitou-se a numeração sequencial utilizada pela página para navegação entre as coleções, permitindo a extração sistemática de todos os *links* referentes aos registros individuais de TCCs, os quais foram armazenados em uma lista de *URLs (Uniform Resource Locator)*.

Com a lista de *URLs* preparada, foi empregado o recurso *ThreadPoolExecutor* do Python para realizar requisições assíncronas aos registros individuais. Essa abordagem reduziu significativamente o tempo de processamento, pois as *threads* executam de forma paralela, sem depender umas das outras para concluir o programa. Para assegurar a obtenção completa e precisa dos dados, implementou-se uma lógica de repetição de requisições em casos de respostas do servidor distintas de sucesso.

Após o acesso a cada URL, os dados foram extraídos diretamente do *HTML* utilizando a biblioteca *Pandas*, que possibilita a leitura de tabelas estruturadas contidas nos documentos. Nesse processo, a primeira tabela foi selecionada,

correspondendo aos metadados de cada registro de TCC. Essa metodologia garantiu a eficiência e a precisão na coleta das informações desejadas.

3.2.2 Preparação dos dados

Os dados coletados foram organizados em uma tabela utilizando a biblioteca Pandas, com o objetivo de consolidar e normalizar a estrutura. Durante o processo, foram identificadas múltiplas colunas com o mesmo nome, como, por exemplo, *subject* (assunto). Para evitar redundâncias, os valores dessas colunas duplicadas foram combinados em uma única coluna, estruturada como uma lista contendo todos os valores associados. Essa abordagem garantiu maior uniformidade e consistência no conjunto de dados, facilitando etapas subsequentes de análise e processamento.

3.2.2.1 Colunas deletadas

Durante o processo de preparação dos dados, foram excluídas todas as colunas que não apresentavam relevância ou continham dados excessivamente esparsos, considerando a limitação do conjunto de dados e o escopo bem definido da análise. Exemplos incluem colunas com informações uniformes, como o tipo de trabalho (TCC da graduação), direitos autorais (*Open Access*), instituição responsável (UFSC), local de publicação (Araranguá), idioma do documento (Português Brasileiro) e descrição do documento (TCCs do TIC). Essas colunas, por possuírem valores constantes ou nulos, foram consideradas de baixo valor analítico.

Adicionalmente, Outras datas, preenchidas manualmente, foram descartadas por não apresentarem consistência ou relevância significativa para a análise.

3.2.2.2 Dados organizados e adequados categoricamente

O processo de tratamento dos resumos envolveu a separação dos textos, inicialmente agrupados em português e inglês, utilizando a biblioteca *langdetect*, que identifica o idioma com base em um modelo de aprendizado de máquina. Essa

abordagem permitiu categorizar os textos corretamente por idioma, facilitando análises posteriores.

Os nomes dos orientadores, inseridos manualmente pelos alunos, apresentavam múltiplas variações, o que dificultava a quantificação e análise das orientações. Para solucionar esse problema, aplicou-se a biblioteca *fuzzywuzzy*, que utiliza o algoritmo de *Levenshtein* para calcular a similaridade entre textos. Essa técnica permitiu consolidar variações de um mesmo nome em uma única representação padronizada, garantindo maior consistência nos dados.

No caso das palavras-chave, utilizou-se a mesma biblioteca, mas com um nível de similaridade mais rigoroso para evitar agrupamentos inadequados. Assim, termos como "visualização gráfica" e "visualização de dados" foram tratados como distintos, preservando a especificidade dos conceitos.

Para a elaboração de nuvens de palavras e etapas subsequentes, realizou-se uma limpeza prévia dos dados, eliminando palavras de parada, caracteres especiais, uniformizando todas as letras para minúsculas e consolidando os dados dos resumos (em português e inglês), títulos e palavras-chave em uma única coluna. Esses dados foram preparados para serem transformados em vetores *embedding* a serem usados na busca semântica e nos processos de clusterização.

A coluna que continha a quantidade de páginas do trabalho também foi tratada, uma vez que os dados eram frequentemente inseridos de forma inconsistente, incluindo informações irrelevantes. Foram removidos todos os caracteres não numéricos, e valores que ultrapassavam 1000 páginas foram considerados nulos, pois representavam erros, como a inserção equivocada do ano de publicação. Essa etapa garantiu maior precisão e confiabilidade nas informações relacionadas ao número de páginas.

3.3 ANÁLISE DE DADOS

A análise de dados foi realizada com base nos metadados e informações textuais coletadas e tratados previamente. Primeiramente, aplicou-se técnicas de visualização para compreender a distribuição dos dados categóricos e numéricos.

Gráficos foram gerados para observar padrões nas variáveis, como anos de publicação, palavras-chave mais frequentes e áreas temáticas predominantes.

Para os dados textuais, foram gerados vetores *embedding* a partir de resumos, títulos e palavras-chave, utilizando o modelo de processamento de linguagem natural *all-MiniLM-L6-v2*, uma das referências para aplicações como esta, para similaridades semânticas. Esses vetores foram empregados em análises de clusterização, com algoritmos como *K-Means*, a fim de identificar agrupamentos temáticos entre os trabalhos.

Também foram criadas nuvens de palavras para identificar tendências e termos mais recorrentes nos textos. Além disso, foi realizada uma análise de concorrência de palavras-chave para explorar a relação entre diferentes temas abordados nos TCCs.

3.4 LIMITAÇÕES DA PESQUISA

Dentre as limitações pode-se citar a inconsistência nos dados de entrada, pois informações inseridas manualmente, como nomes de orientadores e palavras-chave, apresentaram variações que exigiram esforços significativos de padronização, podendo ainda haver casos não identificados de duplicidade ou erro.

A base de dados restrita devido a análise foi limitada aos trabalhos de conclusão de curso (TCCs) disponíveis na coleção específica do repositório institucional da UFSC. Isso restringe as conclusões apenas ao contexto do curso de Tecnologias da Informação e Comunicação, impossibilitando generalizações para outros cursos ou instituições.

Também pode-se citar os desafios técnicos da utilização de técnicas modernas, como *embeddings* e algoritmos de clusterização, a análise semântica pode ter limitações ao capturar nuances mais complexas do texto, como ironia ou ambiguidades.

3.5 RESULTADOS ESPERADOS

Espera-se que a análise permita identificar padrões e tendências significativas na produção acadêmica do curso de Tecnologias da Informação e Comunicação. Entre os resultados esperados estão:

- Temas recorrentes: Determinar quais áreas de pesquisa têm maior representatividade nos TCCs, possibilitando um mapeamento das linhas de interesse predominantes no curso;
- Orientadores influentes: Identificar os orientadores mais frequentes e os temas com os quais mais trabalham, oferecendo subsídios para análise de orientação acadêmica;
- *Insights* semânticos: A partir da clusterização de dados textuais, espera-se agrupar os trabalhos em tópicos, auxiliando na criação de taxonomias para estudos futuros;
- Impacto visual: Por meio de visualizações de dados, como gráficos interativos e nuvens de palavras, apresentar *insights* claros e acessíveis, que auxiliem docentes e discentes no planejamento de futuras pesquisas.

Em suma, os resultados esperados buscam não apenas explorar o panorama acadêmico atual do curso, mas também propor direções estratégicas para potencializar a produção científica na área.

4 APRESENTAÇÃO E DISCUSSÃO DOS RESULTADOS

A análise de dados forneceu *insights* relevantes sobre a produção acadêmica do curso de Tecnologias da Informação e Comunicação (TIC), permitindo a identificação de padrões, tendências e características específicas dos trabalhos de conclusão de curso (TCCs).

4.1 BUSCA SEMANTICA

Agora apresentam-se os resultados obtidos por meio da aplicação de técnicas de busca semântica para identificar trabalhos relevantes relacionados a tópicos específicos no contexto da visualização de dados, inteligência artificial e chatbots. As tabelas a seguir mostram os títulos dos trabalhos, as palavras-chave associadas e os respectivos scores de relevância. O score é calculado com base na similaridade semântica entre as consultas e os documentos analisados.

Tabela 1 - busca semântica visualização de dados e dinheiro

Score	Título	Palavras chaves
0.7050	Visualização de Dados Aplicada a Análise de Dados Contábeis	Visualização de dados, Contabilidade, Dashboards, micro e pequenas empresas, tomada de decisão
0.7034	Visualização de Dados Abertos Educacionais do PROUNI com ChartJS	visualização de dados, Dados Abertos, ChartJS
0.6826	Visualização de Dados Financeiros: Uma Abordagem Prática na Bolsa de Valores Brasileira	visualização de dados, mercado financeiro, tomada de decisão, ações brasileiras
0.6767	Explorando a Integração entre Visualização de Dados e a Computação Quântica: Uma Revisão Sistemática	Tecnologias das Informação e Comunicação, Computação Quântica, visualização de dados
0.6759	View.gov: visualização de dados de endereços eletrônicos do governo com a biblioteca D3	visualização de dados, Dados Abertos, Php, D3, Tecnologias das Informação e Comunicação

Fonte: elaborado pelo autor

A Tabela 1 apresenta os resultados da busca semântica com o foco em visualização de dados e dinheiro. Os trabalhos com maior relevância incluem estudos que exploram a aplicação da visualização de dados no contexto contábil e financeiro, como em micro e pequenas empresas e no mercado de ações. Esse conjunto evidencia a relação direta entre a utilização de ferramentas de visualização de dados e a tomada de decisão em contextos econômicos e financeiros.

Já a Tabela 2 foca na busca semântica por trabalhos relacionados à visualização de dados e *chatbots*. Neste caso, observou-se uma sobreposição com a Tabela 1, pois trabalhos que incluem a visualização de dados em áreas mais amplas também aparecem nesta consulta. Destacam-se estudos que utilizam bibliotecas específicas, como *ChartJS* e *D3*, para trabalhar com dados abertos, indicando a relevância de tecnologias de código aberto em contextos variados de visualização de informações.

Tabela 2 - busca semântica por visualização de dados e chatbot

Score	Título	Palavras chaves
0.6045	Análise da Evolução Histórica dos Chatbots: uma comparação moderna	<i>Chatbots</i> , <i>Chatbots</i> , Inteligência artificial
0.5773	Explorando a Integração entre Visualização de Dados e a Computação Quântica: Uma Revisão Sistemática	Tecnologias das Informação e Comunicação, Computação Quântica, visualização de dados
0.5684	View.gov: visualização de dados de endereços eletrônicos do governo com a biblioteca D3	visualização de dados, Dados Abertos, <i>Php</i> , <i>D3</i> , Tecnologias das Informação e Comunicação
0.5620	Visualização de Dados Abertos Educacionais do PROUNI com ChartJS	visualização de dados, Dados Abertos, <i>ChartJS</i>
0.5546	VisdadosENEM: visualização de dados do ENEM com a biblioteca D3	Análise de Dados, visualização de dados, Novas tecnologias da informação e comunicação, visualização de dados

Fonte: elaborado pelo autor

A Tabela 2 revelou uma conexão relevante entre visualização de dados e chatbots, mas de forma menos direta em comparação à Tabela 1. A inclusão de estudos relacionados a novas tecnologias de informação e comunicação reflete a

diversidade de aplicações da visualização de dados. A utilização de ferramentas específicas, como a biblioteca *D3*, aparece como um ponto comum entre várias abordagens.

Na Tabela 3, a busca semântica foi ajustada para explorar diretamente a relação entre inteligência artificial e *chatbots*. Os resultados demonstram um foco maior em estudos sobre a evolução histórica dos *chatbots* e em como a inteligência artificial é aplicada em diferentes contextos, como jogos computacionais e acessibilidade. Apesar de a relevância geral ser menor, observa-se uma ligação entre inteligência artificial, automação e *chatbots*, o que complementa os achados da Tabela 2.

Tabela 3 - busca semântica de Inteligência artificial e chatbot

Score	Título	Palavras chaves
0.5610	Análise da Evolução Histórica dos Chatbots: uma comparação moderna	<i>Chatbots</i> , <i>Chatbots</i> , Inteligência artificial
0.4257	Um estudo sobre a interação entre usuários e chatterbots	
0.3824	Análise das técnicas de inteligência artificial aplicadas em jogos computacionais	Inteligência artificial, Jogos computacionais, Estratégias táticas, Artificial intelligence, <i>games</i> , <i>Strategies tactics</i>
0.3824	Desenvolvimento de um protótipo de sistema web para facilitar a comunicação entre surdos e ouvintes	<i>Chatbots</i> , Sistema Web, Acessibilidade, Surdos
0.3365	Aplicação da automação robótica de processos em atividades administrativas da Universidade Federal de Santa Catarina	Robotic Process Automation, University, Planning and Monitoring of Teaching Activities, Planejamento, Gestão Universitária, Processos

Fonte: elaborado pelo autor

As tabelas apresentadas fornecem uma visão abrangente da conexão entre visualização de dados, inteligência artificial e *chatbots*. A análise semântica evidencia como essas áreas podem se interseccionar em aplicações práticas, como a tomada de decisão em contextos financeiros, o uso de tecnologias de código aberto e o desenvolvimento de sistemas inteligentes e acessíveis. Essa abordagem permite identificar lacunas e tendências nas áreas estudadas, contribuindo para o avanço do conhecimento no campo.

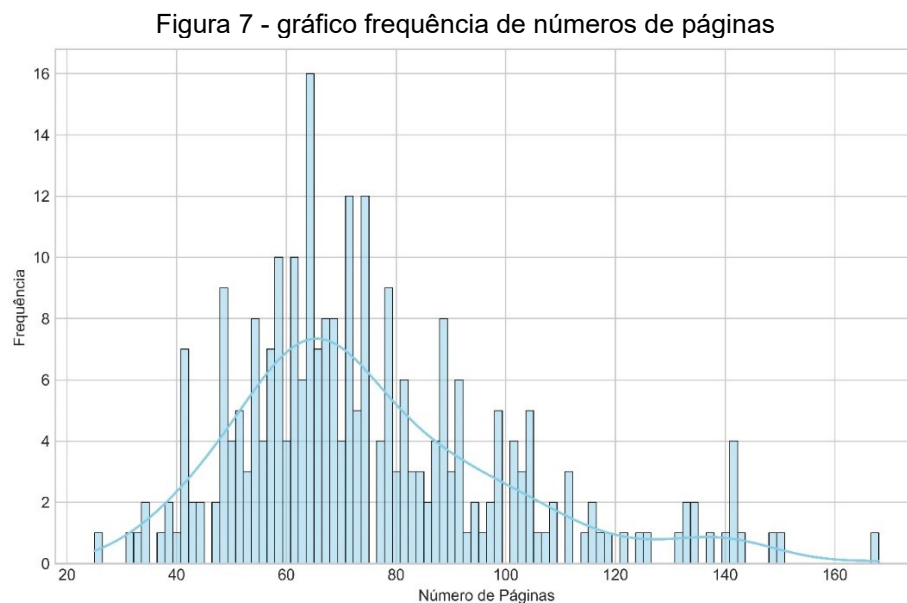
4.2 VISUALIZAÇÕES REALIZADAS

A seguir são apresentadas as visualizações geradas durante a análise dos dados, com o objetivo de sintetizar os resultados obtidos e facilitar a compreensão dos padrões e tendências identificados. Ressalta-se que, em respeito às questões éticas, todos os nomes de pessoas foram reduzidos a iniciais ou modificados.

4.2.1 Visualizações e gerais

4.2.1.1 *Quantidade de Trabalhos por Número de Páginas*

A Figura 7 revela a frequência de TCCs por intervalo de páginas, com destaque para uma forte tendência de concentração entre 50 e 80 páginas, faixa que segue as diretrizes padrão para monografias acadêmicas. A suavização entre os valores, representada por uma linha de tendência, reforça o padrão esperado.

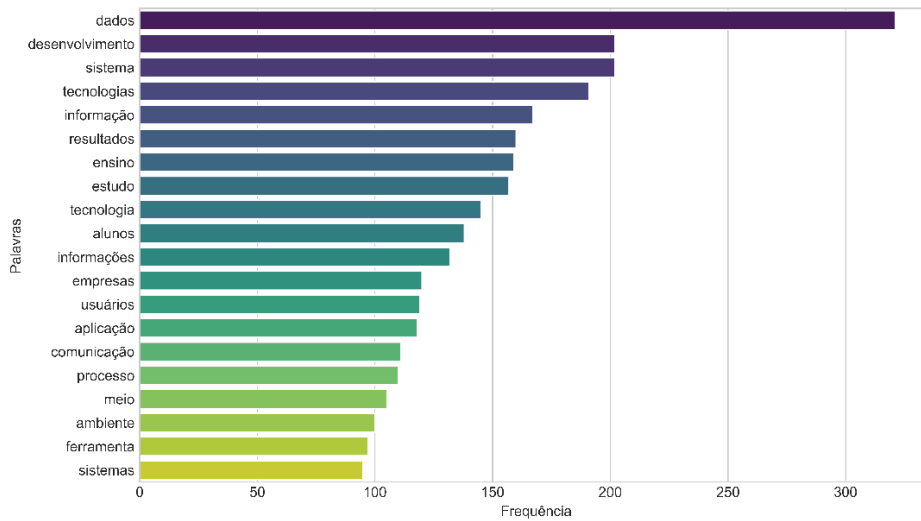


Fonte: elaborado pelo autor

4.2.1.2 *Nuvem de Palavras dos Resumos, Títulos e Palavras-Chave*

A nuvem de palavras da Figura 8 destaca os termos mais recorrentes no corpus textual, evidenciando as palavras "pesquisa", "sistema", "ferramenta",

Figura 9 - gráfico das palavras mais frequentes no resumo em português

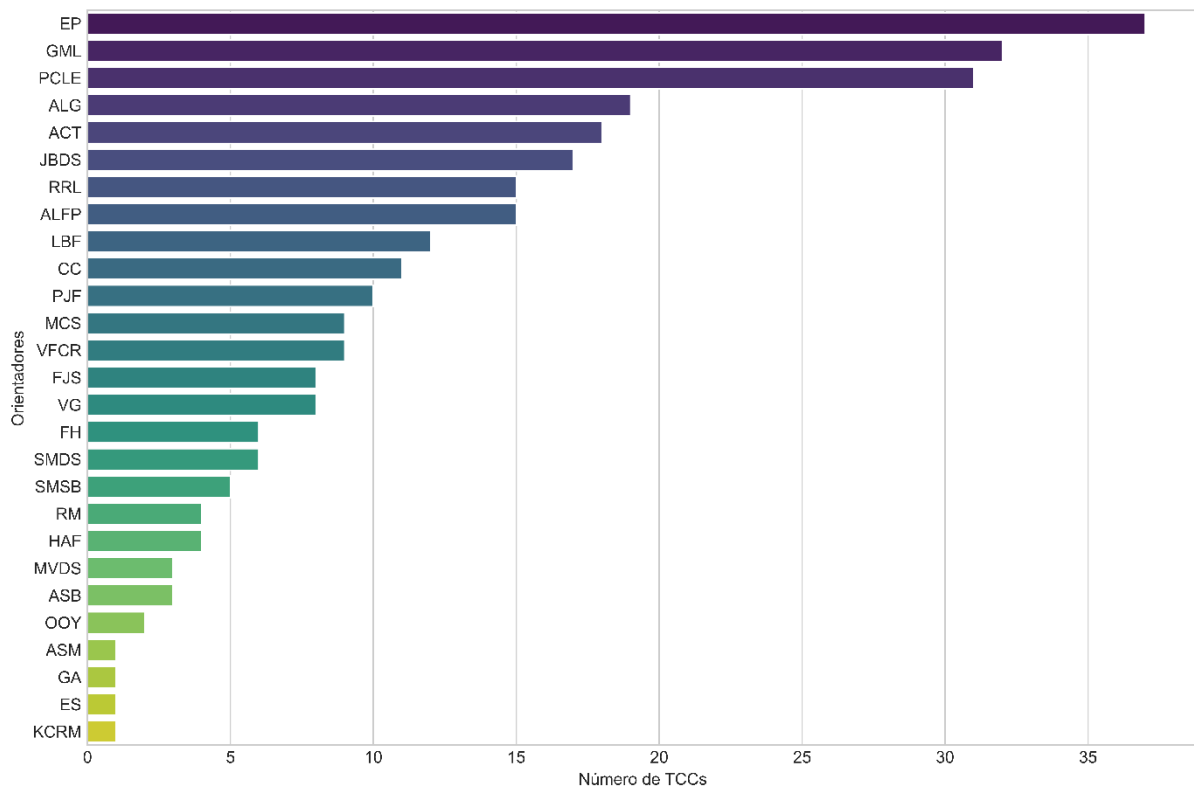


Fonte: elaborado pelo autor

4.2.1.4 Quantidade de trabalhos por orientador

O gráfico da Figura 10 apresenta o número de TCCs orientados por cada professor, evidenciando a distribuição de trabalhos acadêmicos entre os orientadores do curso. O orientador EP lidera com aproximadamente 35 TCCs, seguido de GML e PCLE, ambos com cerca de 30 orientações cada, indicando uma concentração significativa de trabalhos sob sua orientação. Orientadores como ALG, ACT e JBDS também possuem números expressivos, variando entre 20 e 25 orientações. Por outro lado, orientadores como KCRM, ES, GA e ASM aparecem com o menor número de TCCs, com menos de 5 trabalhos cada. Essa distribuição sugere diferenças na carga de orientação entre os docentes, possivelmente relacionadas às suas áreas de especialização, disponibilidade ou envolvimento com os temas mais populares do curso.

Figura 10 - quantidade de trabalhos orientados por orientador



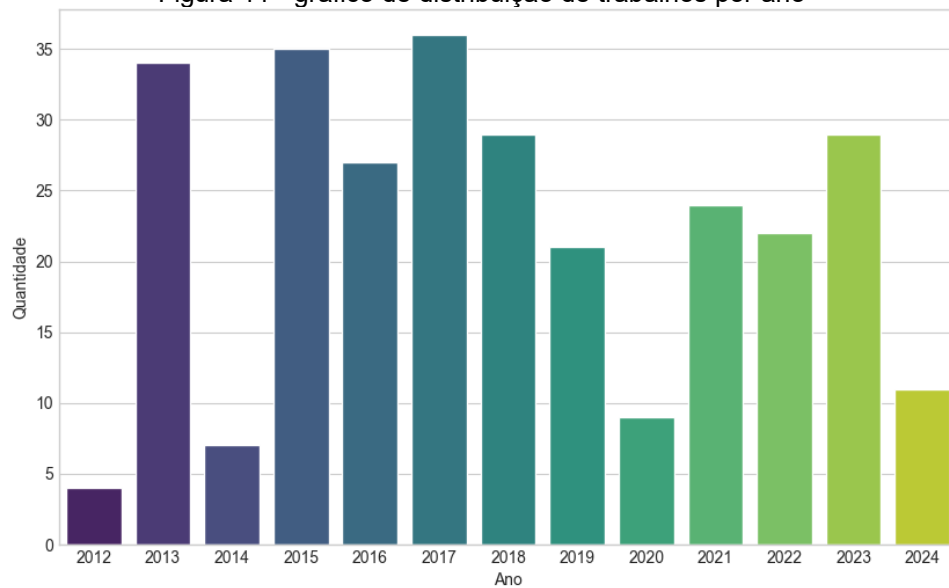
Fonte: elaborado pelo autor

4.2.2 Distribuição de TCCs pelo tempo

4.2.2.1 TCCs por ano

O gráfico de distribuição anual, apresentado na Figura 11, demonstra uma forte redução na quantidade de TCCs defendidos em 2020, coincidindo com o início da pandemia de COVID-19. Após esse período, a produção voltou a crescer, embora em ritmo moderado.

Figura 11 - gráfico de distribuição de trabalhos por ano

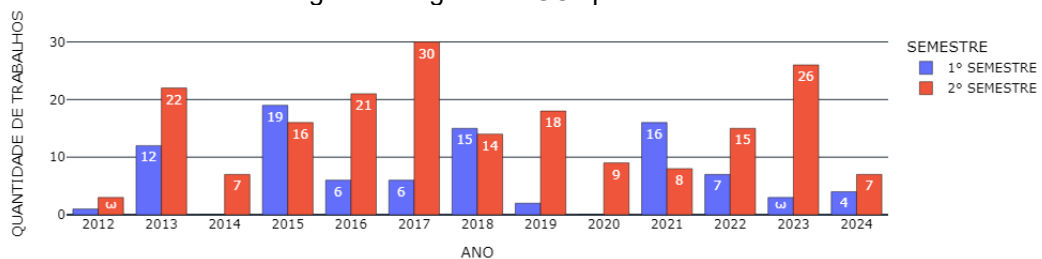


Fonte: elaborado pelo autor

4.2.2.2 TCCs por semestre

A análise por semestre revelou uma predominância de publicações no segundo semestre, o que pode estar relacionado ao cronograma acadêmico do curso, que concentra defesas ao final do ano letivo. A Figura 12 ilustra a distribuição por semestre.

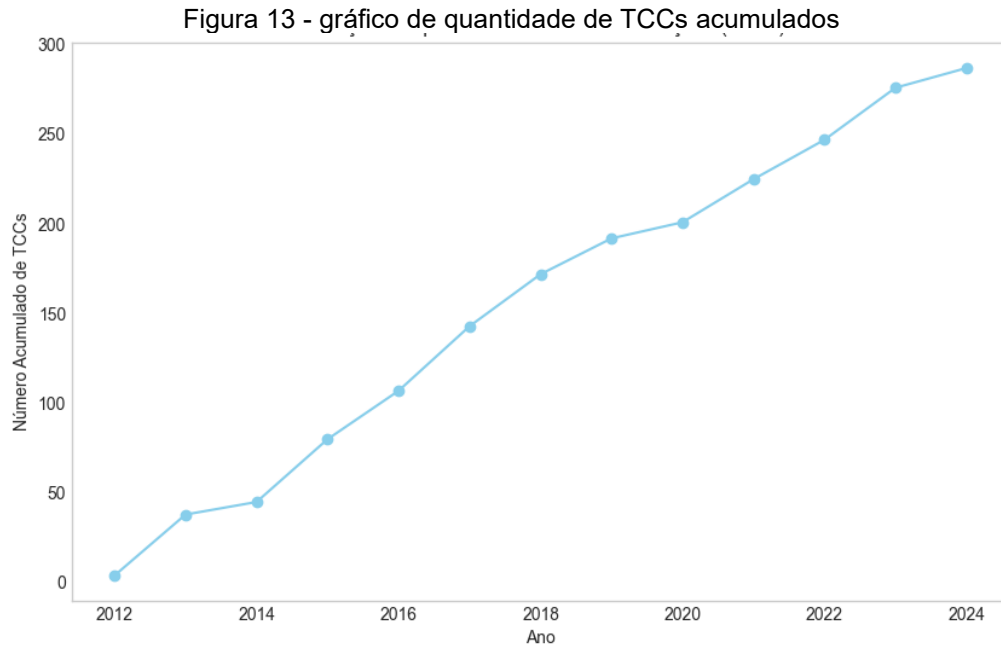
Figura 12 – gráfico TCCs por semestre



Fonte: elaborado pelo autor

4.2.2.3 Evolução temporal acumulada de TCCs por ano

O gráfico de evolução acumulada demonstra um crescimento consistente na quantidade de TCCs ao longo dos anos, indicando a continuidade e a estabilidade da produção acadêmica, conforme a Figura 13.

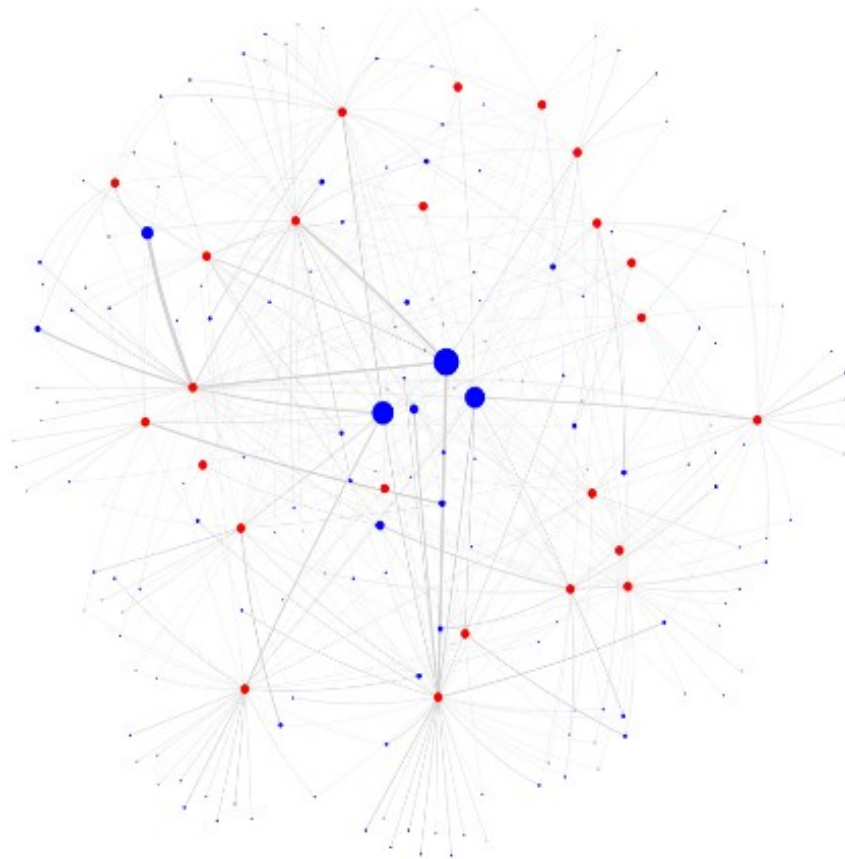


Fonte: elaborado pelo autor

4.2.3 Gráficos de rede

O gráfico da Figura 14 corresponde a todos os professores e as palavras chaves dos trabalhos que eles orientaram, as palavras chaves estão em azul e aumentam com a quantidade de citações, e os professores estão em vermelho, a ligação entre eles possuem peso demonstrado pela grossura da linha, quanto mais determinado orientador tem vínculo daquela palavra-chave. Para facilitar a visualização foi selecionado somente as palavras chaves que tivessem mais de 1 conexão com algum professor ou com vários. Isso filtra a quantidade de palavras chaves e aumenta a visibilidade e o valor entregue pelo gráfico.

Figura 14 - gráfico de rede de orientadores e temas

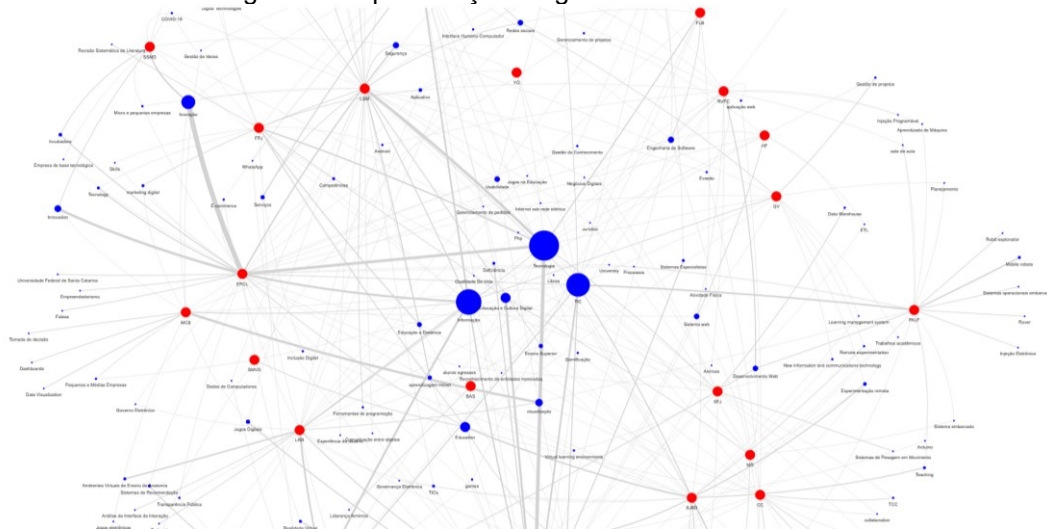


Fonte: elaborado pelo autor

Como pode ser observado existem poucos temas que estão vinculados a somente um orientador, geralmente são temas específicos de uma das três áreas do TIC.

Na aproximação do gráfico Figura 15 abaixo conseguimos notar isso com mais clareza, o orientador PALF (na parte mais à direita) trabalha numa área pouco abrangida pelo curso de TIC a de *hardware* como robôs e até mesmo injeção eletrônica, passando por sistemas operacionais embarcados. Podemos observar o mesmo padrão no orientador MCS (mais à esquerda) que é o único a trabalhar com tomada de decisão e possui forte predominância com visualização de dados.

Figura 15 - aproximação do gráfico de rede anterior



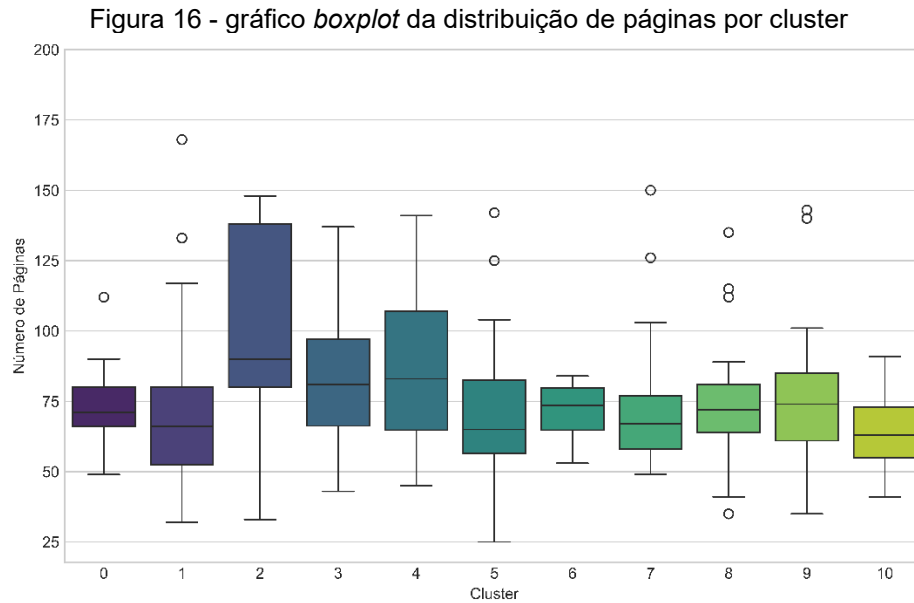
Fonte: elaborado pelo autor

Além disso expressa-se com clareza que os temas TIC, tecnologia e informação que são parte do nome do curso têm as mais altas relevâncias, o que caracteriza ampla integração do TCC com as partes fundamentais do curso.

4.2.4 Gráficos de agrupamento (*cluster*)

4.2.4.1 Páginas de TCCs por agrupamentos *Boxplot*

No Figura 16 abaixo nota-se que a mediana de todos os agrupamentos fica entre 50 e 100 páginas, que a maior parte dos grupos possui *outliers*, somente os grupos 2, 3, 4, 6 e 10 não os possuem. Além disso a quantidade mínima de páginas pertence ao agrupamento 5 com 25 páginas.

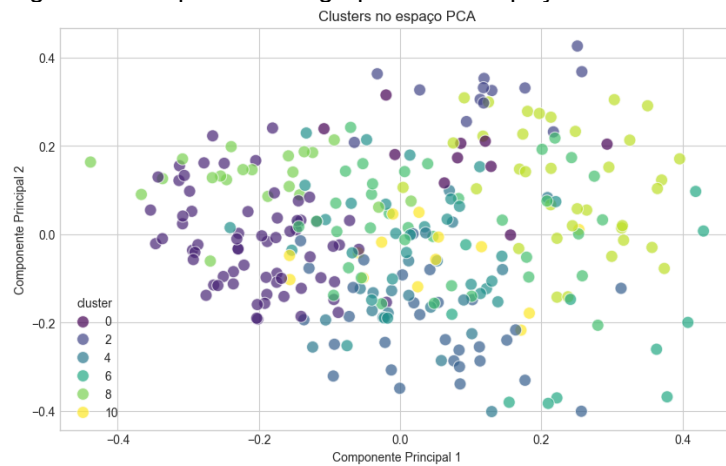


Fonte: elaborado pelo autor

4.2.4.2 Dispersões dos trabalhos agrupadas

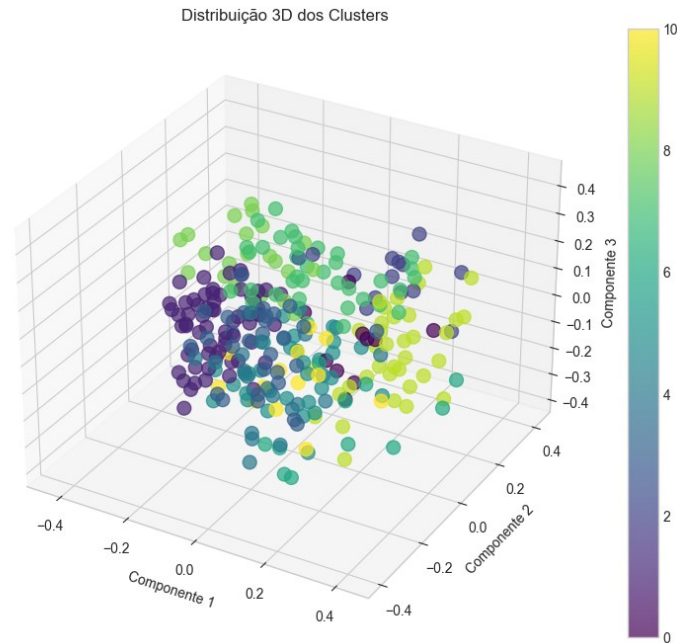
O gráfico da Figura 17 representa todos os TCCs do TIC apresenta no espaço reduzido por PCA, com duas componentes principais representadas nos eixos. Cada ponto no gráfico corresponde a um TCC, e a coloração indica a atribuição a diferentes *clusters*. A proximidade de alguns clusters sugere sobreposição ou semelhanças em características dos TCCs, enquanto *clusters* mais isolados indicam temas ou abordagens mais específicas. A dispersão global demonstra uma ampla diversidade temática entre os trabalhos, enquanto os tamanhos relativos dos clusters podem apontar para grupos com maior ou menor frequência de características compartilhadas nos dados analisados.

Figura 17 - dispersão de agrupamentos espaço reduzido PCA



Fonte: elaborado pelo autor

Figura 18 -dispersão 3D de agrupamentos espaço reduzido PCA



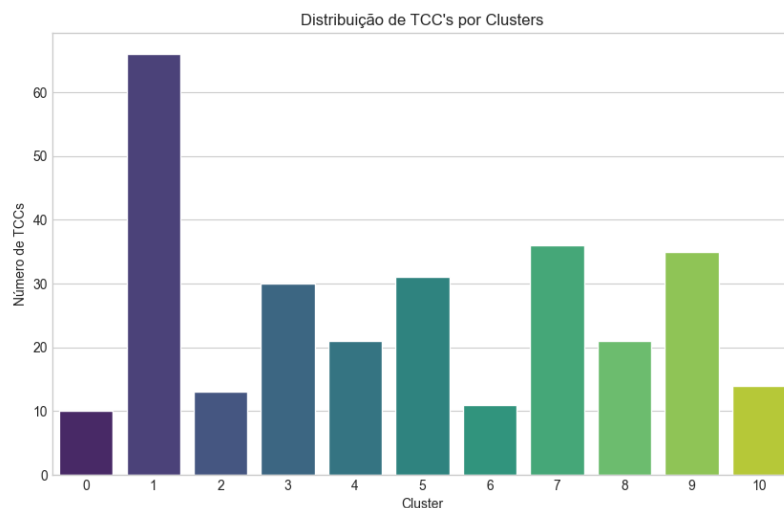
Fonte: elaborado pelo autor

Na Figura 18 os grupos ainda estão bastante próximos, com uma sobreposição significativa entre eles, indicando que os TCCs compartilham muitos temas ou características. A densidade e dispersão dos pontos permanecem semelhantes, mas a visualização tridimensional reforça que alguns trabalhos podem se destacar em aspectos adicionais que a segunda dimensão não capturava.

4.2.4.3 Distribuição de TCCs por agrupamentos

Como mostrado na figura 19 abaixo o agrupamento 1 é o mais populoso, contendo mais de 60 TCCs, enquanto o *Cluster* 0 é o menor, com menos de 10 trabalhos associados. Outros *clusters*, como o 7 e o 9, também apresentam uma quantidade considerável de TCCs, com valores próximos a 40. Essa distribuição indica que há áreas temáticas ou características que predominam entre os TCCs, representadas pelos *clusters* mais populosos, enquanto temas menos recorrentes são agrupados nos *clusters* com menor frequência. Essa variação sugere uma heterogeneidade temática entre os trabalhos, com concentração em áreas específicas, conforme evidenciado pela diferença no número de TCCs entre os *clusters*.

Figura 19 - quantidade de TCCs por agrupamento

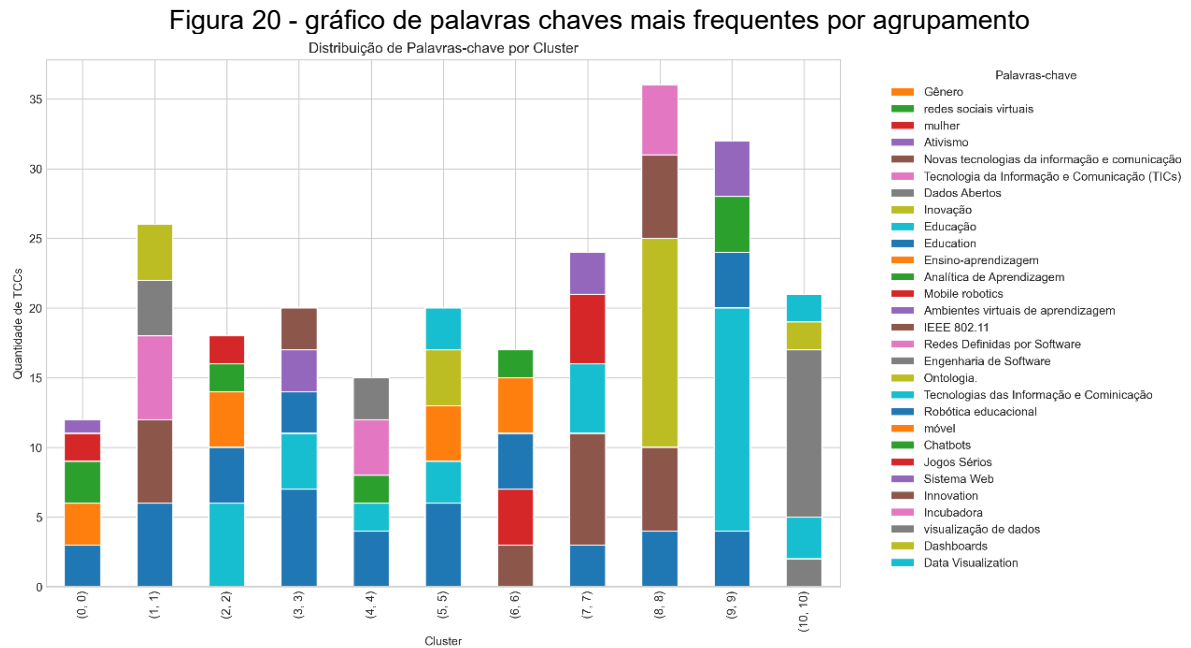


Fonte: elaborado pelo autor

4.2.4.4 *Cinco palavras chaves mais relevantes por agrupamento*

No gráfico abaixo, Figura 20, é possível observar a distribuição de palavras-chave entre os *clusters*, o que destaca as áreas temáticas predominantes nos TCCs. O cluster 7 apresenta forte concentração em temas relacionados a "Educação", "Analítica de Aprendizagem" e "Ambientes Virtuais de Aprendizagem", indicando que é um grupo focado em tecnologias educacionais. Já o *cluster* 1 se destaca por agrupar trabalhos com palavras-chave como "Inovação" e "Tecnologias da Informação e Comunicação", evidenciando uma abordagem mais geral ou

abrangente dentro do campo de TIC. Por outro lado, o *cluster* 8 possui um foco marcante em "Redes Definidas por Software" e "Engenharia de Software", o que o diferencia como uma área mais técnica. Esse padrão reflete a diversidade temática entre os TCCs e reforça a especialização de certos *clusters* em áreas específicas.



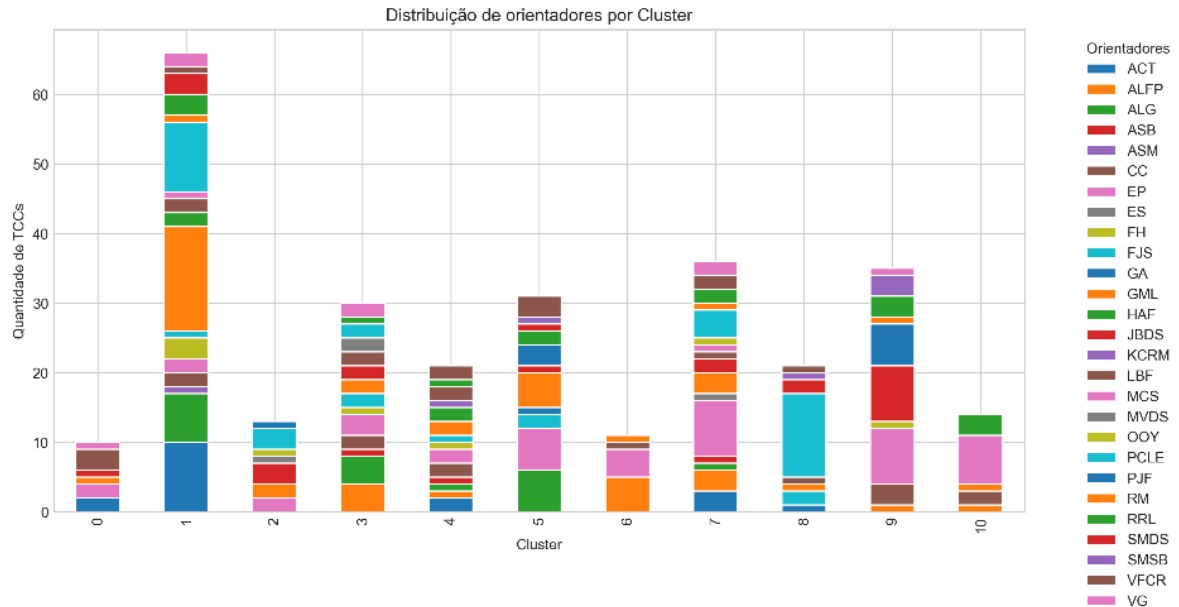
Fonte: elaborado pelo autor

4.2.4.5 Orientadores por agrupamento

No Figura 21 observamos a distribuição de orientadores por *cluster*, evidenciando a contribuição individual em diferentes grupos temáticos de TCCs. O *cluster* 1, que é o mais populoso, apresenta uma ampla diversidade de orientadores, com destaque para a predominância de ALG e ALFP, sugerindo uma influência significativa desses docentes em áreas mais abrangentes ou populares. Por outro lado, no *cluster* 8, percebe-se uma maior concentração de orientações por parte de MVDS e SMSB, sugerindo especialização em temas mais técnicos ou específicos. Já no *cluster* 7, orientadores como FJS e CC possuem uma presença notável, reforçando o foco em áreas relacionadas à educação e tecnologias educacionais. Essa distribuição reflete tanto a diversidade temática dos TCCs quanto a

especialização de certos orientadores em clusters específicos, permitindo uma análise mais direcionada de suas contribuições.

Figura 21 - gráfico orientadores por agrupamentos

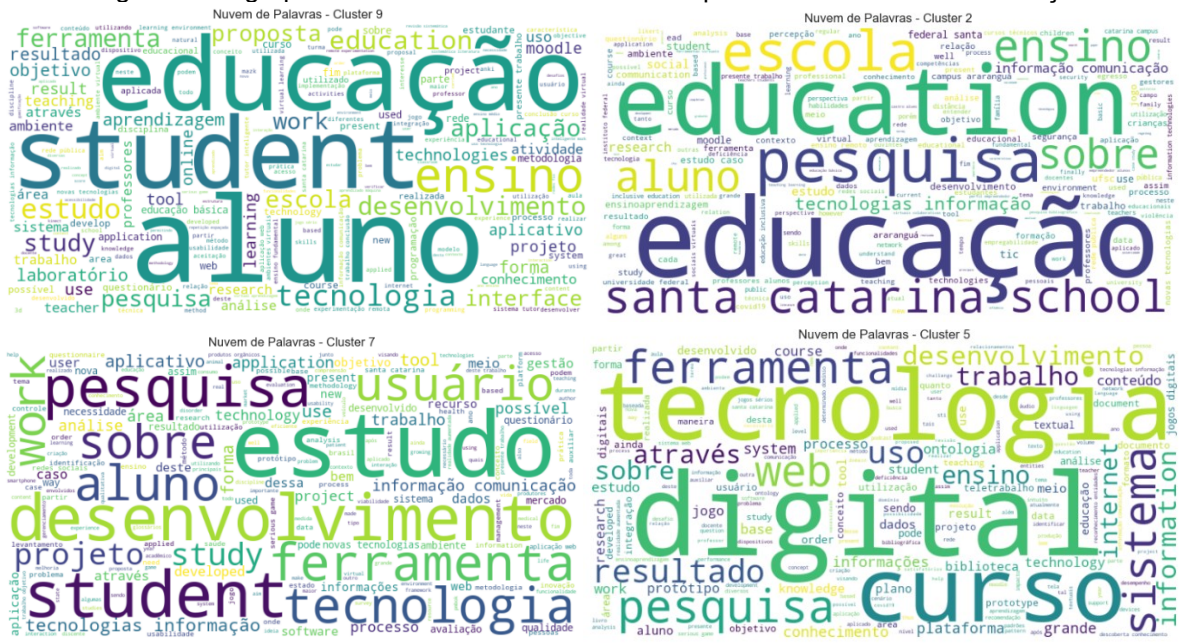


Fonte: elaborado pelo autor

4.2.4.6 Nuvens de Palavras dos agrupamentos

As nuvens de palavras nós dão uma visão rápida e concisa sobre de que se trata cada agrupamento de TCCs, podemos assim identificar padrões e correlações entre os próprios agrupamentos e entender melhor sobre o conteúdo deles.

Figura 22 - agrupamentos com maior incidência de palavras associadas a educação



Fonte: elaborado pelo autor

No conjunto de imagens Figura 22, as nuvens de palavras destacam os principais termos associados à educação em diferentes *clusters*. No *cluster 9*, os termos "educação", "aluno" e "student" predominam, sugerindo um foco em projetos educacionais centrados em estudantes e aprendizagem. O *cluster 2*, por sua vez, apresenta "educação", "escola" e "Santa Catarina", indicando um possível contexto regional e institucional em trabalhos educacionais. No *cluster 7*, observamos uma ênfase em palavras como "pesquisa", "usuário" e "tecnologia", o que aponta para uma abordagem mais técnica e voltada ao desenvolvimento de ferramentas educacionais. Já o *cluster 5* destaca "tecnologia", "digital" e "curso", sugerindo uma exploração de educação digital e plataformas online. Essa distribuição reflete a diversidade temática e as diferentes abordagens na aplicação de tecnologia na educação, desde contextos regionais até o desenvolvimento de soluções digitais.

No conjunto de nuvens da Figura 23, as nuvens de palavras ilustram os agrupamentos com maior incidência de termos associados ao desenvolvimento de sistemas e trabalho. O *cluster 1* evidencia uma forte associação com palavras como "sistema", "empresa" e "dados", indicando foco em soluções corporativas e análise de informações empresariais. Já no *cluster 3*, destacam-se termos como "sistema", "trabalho" e "ambiente", sugerindo projetos relacionados a infraestrutura tecnológica

e aplicação de sistemas em diferentes contextos. O cluster 4 apresenta ênfase em palavras como "software", "usuário" e "processo", refletindo a relevância de trabalhos voltados à usabilidade e ao desenvolvimento de aplicações práticas e interfaces. Esses agrupamentos destacam a diversidade temática nos projetos, com foco em abordagens técnicas, gerenciais e de implementação prática no campo de sistemas e trabalho.

Figura 23 - agrupamentos com maior incidência com desenvolvimento de sistemas e trabalho



Fonte: elaborado pelo autor

Na figura 24, observa-se uma nuvem de palavras do *cluster 6*, destacando um agrupamento com predominância de termos relacionados à robótica. Palavras como "robô", "robot", "robótica" e "sistema" aparecem em destaque, indicando que os trabalhos nesse *cluster* estão focados no desenvolvimento e aplicação de sistemas robóticos. Termos como "tecnologia", "tarefas" e "educacional" sugerem a presença de projetos voltados para a robótica aplicada em contextos de ensino ou tarefas específicas, enquanto "desenvolvimento" e "processo" apontam para uma abordagem prática e técnica. A presença de termos como "chatbots" e "embarcado" reforça a diversidade de aplicações tecnológicas exploradas no *cluster*, abrangendo desde robôs físicos até sistemas interativos. Esse agrupamento reflete um nicho temático de grande relevância no campo da tecnologia da informação e comunicação.

projetos, sugerindo que o planejamento das disciplinas e recursos institucionais está fortemente alinhado ao ciclo anual.

Os gráficos de quantidade de páginas por trabalho e os *boxplots* de distribuição dentro dos *clusters* mostram uma padronização clara em termos de extensão dos TCCs, com a maioria se concentrando entre 50 e 100 páginas. No entanto, a presença de outliers em vários agrupamentos reflete a heterogeneidade na complexidade e no alcance dos projetos desenvolvidos. Essa variação também é evidenciada pela análise de palavras-chave, onde temas como "educação digital" e "engenharia de *software*" aparecem com diferentes níveis de relevância em clusters específicos, apontando para prioridades distintas dentro das mesmas áreas de pesquisa.

A nuvem de palavras gerais, comparada com as nuvens segmentadas por *cluster*, fornece um panorama das palavras mais utilizadas no corpus completo e suas relações contextuais. Palavras como "sistema" e "dados" aparecem consistentemente, mas os termos variam amplamente entre agrupamentos, evidenciando a diversidade temática. Essa análise combinada com os gráficos de rede e as distribuições temporais oferece uma visão mais detalhada da dinâmica de produção acadêmica, revelando interseções e particularidades que definem a identidade científica do curso.

4.4 LIMITAÇÕES E DESAFIOS

Por tratar-se de um escopo bem definido e diminuto a capacidade de utilizar a *LLM* para agrupamento eficiente tornou-se uma das grandes limitações desse trabalho, com corpora textuais maiores os resultados seriam ainda mais impressionantes e valorosos.

Além disso a falta de padrões de inserções dos dados na fonte tornou-se um desafio que precisou de técnicas adicionais como detecção de idiomas para contornar, além disso tínhamos muitos dados que não estavam corretamente organizados apesar das instruções claras da ferramenta, o que demandou vários tipos de organização dos dados, adição de filtros e redução do escopo dos dados.

5 CONSIDERAÇÕES FINAIS

Este trabalho explorou a produção acadêmica do curso de Tecnologias da Informação e Comunicação (TIC) da Universidade Federal de Santa Catarina por meio de uma análise detalhada dos metadados dos Trabalhos de Conclusão de Curso (TCCs). A aplicação de técnicas avançadas de visualização de dados, como gráficos de rede, *clusters* temáticos e nuvens de palavras, permitiu identificar padrões, tendências e interseções relevantes, contribuindo para uma compreensão mais ampla das linhas de pesquisa, dos temas predominantes e das interações entre orientadores e estudantes.

Os resultados demonstraram que a produção acadêmica do curso é caracterizada por uma significativa interdisciplinaridade, com forte presença de temas ligados a educação, análise de dados e desenvolvimento de sistemas. A predominância de termos como "dados", "sistema" e "educação" em diversas visualizações reforça a conexão entre os trabalhos acadêmicos e as demandas do mercado e da sociedade. A análise também destacou a especialização de certos orientadores em áreas específicas, como visualização de dados e tecnologias aplicadas à educação, o que sugere a importância do papel dos docentes na definição de direções de pesquisa.

A análise temporal evidenciou os efeitos da pandemia de COVID-19 sobre a produção acadêmica, marcando uma queda em 2020, seguida por uma recuperação gradual. Essa dinâmica reflete a resiliência do curso e sua capacidade de adaptação a contextos desafiadores. Além disso, a distribuição dos trabalhos por semestre aponta para uma organização acadêmica que maximiza a utilização de recursos e incentiva a conclusão das pesquisas ao final do ano letivo.

Embora as visualizações e os *insights* obtidos tenham oferecido uma visão abrangente, o estudo enfrentou limitações relacionadas à qualidade e à consistência dos dados. Informações inseridas manualmente, como nomes de orientadores e palavras-chave, apresentaram variações que exigiram técnicas avançadas de padronização e processamento. Ainda assim, as técnicas empregadas, incluindo clusterização e análise semântica, demonstraram sua eficácia em lidar com dados complexos e extrair informações úteis.

Por fim, este trabalho não apenas mapeou a produção científica do curso de TIC, mas também forneceu ferramentas e abordagens que podem ser replicadas em outros contextos acadêmicos. Os *insights* obtidos podem orientar futuras pesquisas, aprimorar as políticas institucionais e estimular novos projetos, contribuindo para o fortalecimento das áreas de pesquisa e inovação no campo das Tecnologias da Informação e Comunicação.

REFERÊNCIAS

- ALVES, M. DAS D. R.; SOUZA, M. I. F. Estudo de correspondência de elementos metadados: DUBLIN CORE e MARC 21. **RDBCI: Revista Digital de Biblioteconomia e Ciência da Informação**, v. 5, n. 1, p. 20–38, 31 jan. 2007.
- BARABÁSI, A.-L.; PÓSFAL, M. **Network science**. Cambridge: Cambridge university press, 2016.
- Beautiful Soup Documentation — Beautiful Soup 4.12.0 documentation**. Disponível em: <<https://www.crummy.com/software/BeautifulSoup/bs4/doc/>>. Acesso em: 10 dez. 2024.
- BITENCOURT, L. D. DE. **Um sistema voltado à indexação e recuperação de informação integrado à ontologia**. , 2013.
- BOHNERT, M. K.; PAVÃO, C. M. G.; SILVA, F. C. C. DA. Visualização de Dados de Saúde Pública: um estudo de caso sobre a Covid-19. **InCID: Revista de Ciência da Informação e Documentação**, v. 13, n. 2, p. 283–304, 20 dez. 2022.
- BRESSOUD, T.; WHITE, D. **Introduction to Data Systems: Building from Python**. Cham: Springer International Publishing, 2020.
- BROWN, T. B. et al. **Language Models are Few-Shot Learners**. arXiv, , 22 jul. 2020. Disponível em: <<http://arxiv.org/abs/2005.14165>>. Acesso em: 8 dez. 2024
- CAIRO, A. **The Truthful Art: Data, Charts, and Maps for Communication**. [s.l.: s.n.].
- CARD, S. K.; MACKINLAY, J.; SHNEIDERMAN, B. **Readings in Information Visualization: Using Vision to Think**. [s.l.] Morgan Kaufmann, 1999.
- FEW, S. **Show me the numbers: designing tables and graphs to enlighten**. second edition ed. Burlingame, Calif: Analytics Press, 2012.
- FuzzyWuzzy: Fuzzy String Matching in Python - ChairNerd**. Disponível em: <<https://chairnerd.seatgeek.com/fuzzywuzzy-fuzzy-string-matching-in-python/>>. Acesso em: 8 dez. 2024.
- GEERTS, G. L. A design science research methodology and its application to accounting information systems research. **International Journal of Accounting Information Systems**, Special Issue on Methodologies in AIS Research. v. 12, n. 2, p. 142–151, 1 jun. 2011.
- HEER, J.; BOSTOCK, M.; OGIEVETSKY, V. A tour through the visualization zoo. **Commun. ACM**, v. 53, n. 6, p. 59–67, 1 jun. 2010.
- HOLANDA, P. M. C.; GONTIJO, M. C. A. Análise das redes bibliométricas da produção científica sobre visualização da informação: uso do software VOSviewer. **Múltiplos Olhares em Ciência da Informação**, v. 14, p. e051020–e051020, 3 maio 2024.

JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition**. [s.l.] Prentice Hall, 2009.

langdetect: Language detection library ported from Google's language-detection. , [s.d.]. Disponível em: <<https://github.com/Mimino666/langdetect>>. Acesso em: 10 dez. 2024

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **Nature**, v. 521, n. 7553, p. 436–444, maio 2015.

LINHARES, R. Técnicas De Visualização e Softwares De Análise De Dados Qualitativos Na Educação. 1 jan. 2020.

LOHMANN, S. et al. **Concentri Cloud: Word Cloud Visualization for Multiple Text Documents**. 2015 19th International Conference on Information Visualisation. **Anais...** Em: 2015 19TH INTERNATIONAL CONFERENCE ON INFORMATION VISUALISATION. jul. 2015. Disponível em: <https://ieeexplore.ieee.org/abstract/document/7272588?casa_token=IYYgMsphEDgAAAAA:Yn5ooB4_cU34Qyc91e5PPAqpP_VFNiY1yggVmUZXMdjXxe5qwNhlXC2jwrO9wxbqbTtZwwVg>. Acesso em: 6 dez. 2024

LUDERMIR, T. B. Inteligência Artificial e Aprendizado de Máquina: estado atual e tendências. **Estudos Avançados**, v. 35, p. 85–94, 19 abr. 2021.

MORAES, L. DE C. et al. **Análise de ambiguidade linguística em modelos de linguagem de grande escala (LLMs)**. arXiv, , 25 abr. 2024. Disponível em: <<http://arxiv.org/abs/2404.16653>>. Acesso em: 10 dez. 2024

multiprocessing — Process-based parallelism. Disponível em: <<https://docs.python.org/3/library/multiprocessing.html>>. Acesso em: 10 dez. 2024.

MUNZNER, T. What's Vis, and Why Do It? Em: **Visualization Analysis and Design**. [s.l.] A K Peters/CRC Press, 2014.

MURPHY, K. **Machine Learning - A Probabilistic Perspective**. Cambridge: MIT Press, 2014.

NETO, J. V. Boxplot: um recurso gráfico para a análise e interpretação de dados quantitativos. 2017.

NEWMAN, M. E. J. The Structure and Function of Complex Networks. **SIAM Review**, v. 45, n. 2, p. 167–256, 1 jan. 2003.

NLTK :: Natural Language Toolkit. Disponível em: <<https://www.nltk.org/>>. Acesso em: 10 dez. 2024.

NumPy documentation — NumPy v2.1 Manual. Disponível em: <<https://numpy.org/doc/stable/>>. Acesso em: 10 dez. 2024.

pandas documentation — pandas 2.2.3 documentation. Disponível em: <<https://pandas.pydata.org/docs/>>. Acesso em: 10 dez. 2024.

PEDRINA, G. **Coleta de dados na internet precisa ser mais debatida**. Disponível em: <<https://www.conjur.com.br/2014-set-23/gustavo-pedrina-coleta-dados-internet-debatida/>>. Acesso em: 10 nov. 2024.

Plotly. Disponível em: <<https://plotly.com/python/>>. Acesso em: 10 dez. 2024.

PYTHON SOFTWARE FOUNDATION. **General Python FAQ**. Disponível em: <<https://docs.python.org/3/faq/general.html>>. Acesso em: 1 dez. 2024.

REGLY, T.; SOUZA, R. F. DE. Aportes para análise de plataformas de dados abertos dotadas de recursos visuais: requisitos fundamentados nos campos da Visualização, Arquitetura da Informação e Usabilidade. **Em Questão**, v. 29, p. e, 16 jun. 2023.

Requests: HTTP for Humans™ — Requests 2.32.3 documentation. Disponível em: <<https://docs.python-requests.org/en/latest/>>. Acesso em: 8 dez. 2024.

RODRIGUES, A.; DIAS, G. ESTUDOS SOBRE VISUALIZAÇÃO DE DADOS CIENTÍFICOS NO CONTEXTO DA DATA SCIENCE E DO BIG DATA. 2017.

ROSEN, P.; QUADRI, G. J. **LineSmooth: An Analytical Framework for Evaluating the Effectiveness of Smoothing Techniques on Line Charts**. arXiv, , 19 set. 2020. Disponível em: <<http://arxiv.org/abs/2007.13882>>. Acesso em: 7 dez. 2024

SANTOS, T. C. B. DOS; PERRY, G. T. Revisão sistemática sobre painéis de visualização de dados educacionais. **RENOTE**, v. 21, n. 1, p. 87–96, 31 jul. 2023.

SCOTT, J. **Social Network Analysis**. [s.l.] SAGE Publications Ltd, 2017.

SentenceTransformers Documentation — Sentence Transformers documentation. Disponível em: <<https://www.sbert.net/>>. Acesso em: 10 dez. 2024.

SÉRGIO, M. C. **Uma arquitetura de descoberta de conhecimento baseada na correlação e associação temporal de padrões textuais**. , 2013.

SILVA, C. G. DA. Considerações sobre o uso de Visualização de Informação no auxílio à gestão de informação. **XXXIV Seminário Integrado de Software e Hardware**, 2007.

SILVA, F. C. C. DA. Visualização de dados : passado, presente e futuro. 2019.

SILVA, T. DO N. DA. **Uma arquitetura para descoberta de conhecimento a partir de bases textuais**. , 2012.

SMITH, M. et al. DSpace: An Open Source Dynamic Digital Repository. **DSpace: An Open Source Dynamic Digital Repository**, v. 9, n. 1, jan. 2003.

STOLL, B. B. et al. Análise de dados acadêmicos baseado em previsão, recomendação e visualização. **RENOTE**, v. 17, n. 1, p. 286–295, 28 jul. 2019.

TIM BERNERS-LEE et al. **Hypertext Transfer Protocol – HTTP/1.1**. [s.l.] Internet Engineering Task Force, jun. 1999. Disponível em: <<https://datatracker.ietf.org/doc/rfc2616>>. Acesso em: 7 dez. 2024.

Using Matplotlib — Matplotlib 3.9.3 documentation. Disponível em: <<https://matplotlib.org/stable/users/index.html>>. Acesso em: 10 dez. 2024.

VASWANI, A. et al. **Attention Is All You Need**. arXiv, , 2 ago. 2023. Disponível em: <<http://arxiv.org/abs/1706.03762>>. Acesso em: 8 dez. 2024

WASKOM, M. seaborn: statistical data visualization. **Journal of Open Source Software**, v. 6, n. 60, p. 3021, 6 abr. 2021.

YAU, N. **Data Points: Visualization That Means Something | Wiley**. [s.l.: s.n.].

APENDICE A – BIBLIOTECAS PYTHON UTILIZADAS

Requests

Requests é uma biblioteca desenvolvida para simplificar a realização de requisições HTTP em Python. Ela permite enviar e receber dados de servidores de maneira eficiente, sendo amplamente utilizada em projetos de integração com APIs e web scraping. Com funcionalidades como autenticação, sessões persistentes e suporte a cookies, oferece uma interface amigável para operações complexas, como upload de arquivos e manipulação de cabeçalhos HTTP. Sua simplicidade e robustez tornam-na indispensável para desenvolvedores que precisam interagir com serviços online. Segundo a documentação, é considerada uma das bibliotecas mais fáceis e poderosas para trabalhar com protocolos HTTP (“Requests: HTTP for Humans™ — Requests 2.32.3 documentation”, [s.d.]).

Multiprocessing

O módulo *Multiprocessing* é uma solução robusta para execução de processos paralelos em Python. Ele permite que tarefas sejam distribuídas entre diferentes núcleos de CPU, maximizando o uso de hardware multi-core. Esse módulo é ideal para operações que envolvem computação intensiva ou grandes volumes de dados, reduzindo significativamente o tempo de processamento. Além disso, oferece suporte a comunicação entre processos e compartilhamento de dados por meio de filas e pipes. A documentação oficial destaca sua capacidade de criar sistemas escaláveis e eficientes em termos de desempenho, tornando-o uma escolha comum em aplicações de ciência de dados e aprendizado de máquina (“multiprocessing — Process-based parallelism”, [s.d.]).

Re (regex)

O módulo Re é a principal ferramenta do Python para manipulação de expressões regulares, possibilitando buscas, substituições e extrações em strings de forma eficiente. Ele suporta padrões complexos, permitindo identificar e trabalhar com sequências específicas de caracteres. Essa funcionalidade é essencial para validação de formatos, como e-mails, números de telefone ou padrões de texto em arquivos. Além disso, o Re oferece opções avançadas, como agrupamento de

capturas e uso de flags para personalizar a busca. Conforme a documentação oficial, é amplamente utilizado em aplicações de processamento de texto e análise de dados (“Requests: HTTP for Humans™ — Requests 2.32.3 documentation”, [s.d.]).

Pandas

Pandas é uma biblioteca poderosa para manipulação e análise de dados estruturados, amplamente utilizada em ciência de dados. Ela fornece estruturas de dados flexíveis, como DataFrame e Series, permitindo realizar operações complexas de forma eficiente e intuitiva. Com funcionalidades para leitura e escrita de arquivos em diversos formatos (CSV, Excel, SQL, etc.), Pandas facilita a integração de dados em fluxos de trabalho variados. Além disso, sua capacidade de lidar com dados ausentes e executar operações vetorizadas a torna uma escolha popular entre analistas. Segundo a sua equipe de desenvolvimento, é uma ferramenta essencial para qualquer projeto que envolva análise e manipulação de grandes volumes de informações (“pandas documentation — pandas 2.2.3 documentation”, [s.d.]).

Numpy

Numpy é a biblioteca principal para computação numérica em Python, oferecendo suporte para arrays multidimensionais e funções matemáticas de alto desempenho. É amplamente empregada em aplicações científicas, como modelagem matemática, simulações e aprendizado de máquina. Sua eficiência se deve ao uso de implementações em C, que garantem operações rápidas mesmo com grandes conjuntos de dados. Além disso, Numpy fornece funções para álgebra linear, transformadas de Fourier e geração de números aleatórios. Conforme a documentação, ela serve como base para muitas outras bibliotecas científicas, como Pandas e Scikit-learn, sendo indispensável para o ecossistema Python (“NumPy documentation — NumPy v2.1 Manual”, [s.d.])

BeautifulSoup

BeautifulSoup é uma biblioteca projetada para extrair e manipular dados de arquivos HTML e XML, facilitando o web scraping. Ela transforma o conteúdo de páginas web em uma estrutura navegável, permitindo localizar elementos específicos, como tabelas, links ou parágrafos. Sua integração com bibliotecas como

Requests aumenta a eficiência na coleta de informações da web. BeautifulSoup também oferece suporte a diferentes parsers, como lxml e html.parser, otimizando o desempenho. Segundo seus desenvolvedores, a ferramenta é ideal para projetos que envolvam automação e coleta de dados de sites (“Beautiful Soup Documentation — BeautifulSoup 4.12.0 documentation”, [s.d.]).

Fuzzywuzzy

Fuzzywuzzy é uma biblioteca especializada em comparação de strings e análise de similaridade textual. Ela utiliza o algoritmo Levenshtein para calcular a distância entre duas sequências, permitindo identificar padrões semelhantes mesmo quando existem diferenças sutis, como erros de digitação. Essa funcionalidade é útil em tarefas como limpeza de dados, correspondência de registros e detecção de duplicatas. A biblioteca também fornece métricas avançadas, como Token Sort Ratio e Token Set Ratio, para aumentar a precisão em diferentes contextos. De acordo com sua documentação, Fuzzywuzzy é amplamente utilizada em projetos que exigem processamento de linguagem natural e análise textual (“FuzzyWuzzy: Fuzzy String Matching in Python - ChairNerd”, 2011)

Langdetect

Langdetect é uma biblioteca de detecção automática de idiomas, baseada no algoritmo de aprendizado de máquina desenvolvido para o Google Translate. Ela analisa padrões em texto e atribui probabilidades aos idiomas mais prováveis. Sua implementação é simples e eficiente, com suporte para mais de 50 idiomas, tornando-se ideal para aplicações de processamento de linguagem natural. Além disso, Langdetect permite personalização de modelos e integração com fluxos de trabalho de análise textual. De acordo com a documentação, é amplamente usada em sistemas de categorização de conteúdo multilíngue (“langdetect”, [s.d.]).

Nltk

NLTK (Natural Language Toolkit) é uma das bibliotecas mais populares para processamento de linguagem natural (PLN) em Python. Ela fornece ferramentas para tokenização, análise sintática, classificação de texto, entre outros. NLTK também inclui um extenso corpus de dados linguísticos e métodos para treinamento de modelos de aprendizado de máquina. Com uma ampla gama de funcionalidades,

é amplamente utilizada em projetos acadêmicos e comerciais que envolvem análise de texto e interpretação semântica. Segundo seus criadores, NLTK é ideal tanto para iniciantes quanto para especialistas em PLN (“NLTK:: Natural Language Toolkit”, [s.d.]).

Matplotlib

Matplotlib é uma biblioteca de visualização de dados em Python, amplamente reconhecida por sua flexibilidade na criação de gráficos. Ela suporta uma grande variedade de estilos, como gráficos de linha, dispersão, barras e histogramas. Uma de suas maiores vantagens é a possibilidade de personalização detalhada, permitindo ajustes em elementos como títulos, rótulos e cores. Além disso, Matplotlib é frequentemente usada em conjunto com outras bibliotecas, como Pandas e Numpy, para criar análises visuais ricas. A documentação destaca sua importância para comunicações científicas e relatórios técnicos (“Using Matplotlib — Matplotlib 3.9.3 documentation”, [s.d.]

Seaborn

Seaborn é uma biblioteca voltada para gráficos estatísticos, construída sobre o Matplotlib. Ela simplifica a criação de visualizações complexas, como mapas de calor, gráficos de regressão e distribuições de dados. Seu design focado na estética garante visualizações informativas e visualmente agradáveis. Seaborn é amplamente utilizada em projetos de análise exploratória de dados, pois integra facilmente DataFrames do Pandas para geração de gráficos. Segundo a documentação, sua funcionalidade economiza tempo ao criar gráficos sofisticados com poucas linhas de código (Waskom, 2021).

SentenceTransformers

SentenceTransformers é uma biblioteca poderosa para criação de embeddings de texto, permitindo medir similaridade semântica e realizar busca contextual. Ela suporta modelos pré-treinados baseados em transformers, como BERT e RoBERT-a, otimizados para comparar sentenças ou parágrafos. Essa biblioteca é amplamente utilizada em aplicações de busca semântica, clusterização de texto e criação de sistemas de recomendação baseados em linguagem natural. Além disso, SentenceTransformers oferece suporte para personalização por meio de

fine-tuning, conforme descrito em sua documentação oficial (“SentenceTransformers Documentation — Sentence Transformers documentation”, [s.d.]).

Networkx

Networkx é uma biblioteca utilizada para a criação, manipulação e análise de grafos e redes complexas. Ela suporta grafos direcionados, não direcionados e multigrafos, permitindo modelar uma ampla gama de problemas relacionados a redes. Suas funcionalidades incluem cálculo de métricas como centralidade, conectividade e detecção de comunidades. Networkx é amplamente usada em áreas como ciência de redes sociais, biologia computacional e logística. Conforme a documentação, sua integração com bibliotecas de visualização, como Matplotlib, facilita a representação gráfica de redes (Newman, 2003).

Plotly

Plotly é uma biblioteca interativa para visualização de dados, conhecida por sua capacidade de gerar gráficos dinâmicos e responsivos. Ela suporta uma ampla gama de gráficos, incluindo linhas, dispersão, barras, mapas de calor e gráficos tridimensionais. Plotly também se integra facilmente com frameworks como Dash para criar dashboards interativos e visualizações em tempo real. Sua funcionalidade para exportação de gráficos em formatos como HTML a torna ideal para apresentações e relatórios. De acordo com sua documentação, Plotly é amplamente utilizada em setores como ciência de dados, engenharia e negócios (“Plotly”, [s.d.]).