



UNIVERSIDADE FEDERAL DE SANTA CATARINA
CAMPUS FLORIANÓPOLIS
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA E GESTÃO DO
CONHECIMENTO

Suzana Kilpp da Silva

**Uso de mineração de textos como ferramenta de avaliação da qualidade
informacional em Portal de Atendimento Institucional**

Florianópolis

2022

Suzana Kilpp da Silva

**Uso de mineração de textos como ferramenta de avaliação da qualidade informacional
em Portal de Atendimento Institucional**

Dissertação submetida ao Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento da Universidade Federal de Santa Catarina para a obtenção do título de Mestre em Engenharia e Gestão do Conhecimento. Área de concentração: Gestão do Conhecimento. Linha de pesquisa: Gestão do Conhecimento Organizacional.

Orientador: Prof. Rogério Cid Bastos, Dr.

Coorientador: Prof^a. Lia Caetano Bastos, Dra.

Florianópolis

2022

Ficha de identificação da obra elaborada pela autora,
atráves do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Silva, Suzana Kilpp da
Uso de mineração de textos como ferramenta de avaliação
da qualidade informacional em Portal de Atendimento
Institucional / Suzana Kilpp da Silva ; orientador,
Rogério Cid Bastos, coorientador, Lia Caetano Bastos, 2022.
141 p.

Dissertação (mestrado) - Universidade Federal de Santa
Catarina, Centro Tecnológico, Programa de Pós-Graduação em
Engenharia e Gestão do Conhecimento, Florianópolis, 2022.

Inclui referências.

1. Engenharia e Gestão do Conhecimento. 2. Mineração de
textos. 3. Tomada de decisão. 4. Descoberta de
conhecimento. 5. Portal de atendimento. I. Bastos, Rogério
Cid. II. Bastos, Lia Caetano. III. Universidade Federal de
Santa Catarina. Programa de Pós-Graduação em Engenharia e
Gestão do Conhecimento. IV. Título.

Suzana Kilpp da Silva

**Uso de mineração de textos como ferramenta de avaliação da qualidade informacional
em Portal de Atendimento Institucional**

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

Prof. Alexandre Marino Costa, Dr.
Universidade Federal de Santa Catarina

Prof. Fernando Alvaro Ostuni Gauthier, Dr.
Universidade Federal de Santa Catarina

Prof. João Bosco da Mota Alves, Dr.
Universidade Federal de Santa Catarina

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

Prof. Roberto Carlos dos Santos Pacheco, Dr.
Coordenador do programa

Prof. Rogério Cid Bastos, Dr.
Orientador

Florianópolis, 4 de março de 2022.

Este trabalho é dedicado ao meu esposo, minha filha, meu filho e às minhas queridas amigas, pelo apoio e carinho recebido.

RESUMO

A mineração de textos é uma ferramenta de Gestão do Conhecimento aplicada em bancos de dados não estruturados, que busca extrair informação útil e relevante para fornecer subsídios de tomada de decisão. Esta pesquisa tem por objetivo avaliar o uso de informações não estruturadas no processo de descoberta do conhecimento e de tomada de decisão, visando a melhoria da comunicação entre usuários e organização. Para tanto, realiza um estudo de caso utilizando o Portal de Atendimento Institucional (PAI) da Pró-Reitoria de Extensão (PROEX) da Universidade Federal de Santa Catarina (UFSC) e uma busca de informações na rede social *Twitter*. O PAI é o principal meio de comunicação entre usuários e a Universidade. As informações do PAI encontram-se na forma de texto em linguagem natural, ou seja, dados não estruturados, e, apesar de estarem armazenadas em um banco de dados, o sistema não disponibiliza nenhum tipo de relatório informacional. A PROEX utiliza o PAI desde agosto de 2019 e recebe centenas de chamados mensalmente. A mineração de textos proporciona a descoberta e a análise de informações armazenadas no PAI e de informações veiculadas em redes sociais, configurando-se em uma estratégia de Gestão do Conhecimento. A seleção da ferramenta de mineração de textos KH Coder permitiu a elaboração de mapas temáticos que se constituem na base da criação de ontologias. A ferramenta permitiu a análise de informações não estruturadas por meio de seis diferentes técnicas: (i) classificador Naive Bayes, (ii) rede de coocorrência de palavras, (iii) rede de coocorrência de palavras e variáveis, (iv) análise por cluster, (v) análise hierárquica de cluster e (vi) mapa auto-organizável. Como resultados obtidos, a identificação de informações relevantes armazenadas no PAI forneceu subsídios importantes aos gestores para a tomada de decisão e para a melhoria da qualidade da comunicação entre os usuários e a Universidade. Foram identificados 30 tópicos de perguntas e 29 tópicos de respostas, sendo os mais frequentes: (i) aprovação de programas e projetos, (ii) cursos de curta duração, (iii) prestação de serviço, serviço eventual e consultoria, (iv) publicações, artigos científicos, revisão de artigos, revistas, e (v) registro de atividades docentes. Além disso, foi descoberto que 1/5 das respostas fazem referência a artigos da Resolução Normativa nº 88/2016/Cun, que regulamenta as ações de extensão da Universidade. A análise de *tweets* sobre a UFSC na rede social *Twitter* apontou que a ferramenta é utilizada mais para replicar mensagens do que para gerar textos. A conclusão desta pesquisa aponta que a mineração de textos é uma ferramenta de Gestão do Conhecimento que se destaca quando o objeto de análise envolve bancos de informações textuais em linguagem natural, tornando possível extrair conhecimento potencialmente relevante para a tomada de decisão. E ainda, o sistema PAI/UFSC tem informações relevantes para a Gestão Universitária e que são bem exploradas por técnicas de mineração de textos.

Palavras-Chave: Mineração de texto. Tomada de decisão. Descoberta de conhecimento. Gestão do conhecimento.

ABSTRACT

Text mining is a Knowledge Management tool applied in unstructured databases, which seeks to extract useful and relevant information to support decision-making. This research aims to evaluate the use of unstructured information in the discovery of knowledge and in the decision-making process, aiming to improve communication between users and the organization. A case study is carried out using the Institutional Service Portal (ISP) of the Pro-Rectorate of Extension (PROEX) of the Federal University of Santa Catarina and a search for information on the social network *Twitter*. ISP is the main means of communication between users and the University. ISP information is in the form of text in natural language, that is, unstructured data, and, despite being stored in a database, the system does not provide any type of informational report. PROEX has been using ISP since August 2019 and receives hundreds of calls monthly. Text mining provides the discovery and analysis of information stored in the ISP and information transmitted on social networks, configuring itself in a Knowledge Management strategy. The selection of the KH Coder text mining tool allowed the elaboration of thematic maps that constitute the basis for the creation of ontologies. The tool allowed the analysis of unstructured information through six different techniques: (i) Naive Bayes classifier, (ii) word co-occurrence network, (iii) word and variable co-occurrence network, (iv) cluster analysis, (v) hierarchical cluster analysis and (vi) self-organizing map. The identification of relevant information stored in the ISP provided important subsidies to managers for decision-making and to improve the quality of communication between users and the University. Thirty questions and twenty-nine answer topics were identified, the most frequent being: (i) approval of programs and projects, (ii) short-term courses, (iii) service provision, occasional service and consultancy, (iv) publications, scientific articles, review of articles, journals, and (v) registration of teaching activities. In addition, it was found that 1/5 of the answers refer to articles from Normative Resolution nº 88/2016/Cun, which refers to the University's extension norms. The analysis of *tweets* about Federal University of Santa Catarina on the social network *Twitter* reveals that the tool is used more to replicate messages than to generate texts. The conclusion of this research points out that text mining is a Knowledge Management tool that stands out when the object of analysis involves textual information banks (natural language), making it possible to extract potentially relevant knowledge for decision-making. Furthermore, Federal University of Santa Catarina ISP system has relevant information for University Management that is well explored by text mining techniques.

Keywords: Text mining. Decision-making. Discovery of knowledge. Knowledge management.

LISTA DE FIGURAS

Figura 1 - Evolução da espiral do conhecimento	29
Figura 2 - A organização do conhecimento	30
Figura 3 - Processo de descoberta de conhecimento em base de dados	32
Figura 4 - Processo de descoberta do conhecimento textual	33
Figura 5 - Processo de descoberta do conhecimento textual	34
Figura 6 - Processo de mineração de textos	40
Figura 7 - Rede de Coocorrência (exemplo 1)	53
Figura 8 - Rede de Coocorrência (exemplo 2)	54
Figura 9 - Dendograma (exemplo)	55
Figura 10 - Mapa Auto-Organizável (exemplo)	55
Figura 11 – Aplicação Mineração de Textos	58
Figura 12 - Extração da informação	58
Figura 13 - Pré-processamento	61
Figura 14 - Mineração de Textos	68
Figura 15 - Rede de Coocorrência de Palavras - planilha PERGUNTAS	73
Figura 16 - Rede de Coocorrência entre Palavras e Variáveis - planilha PERGUNTAS	75
Figura 17 - Rede de Coocorrência de Palavras e Variável “projeto” - planilha PERGUNTAS	76
Figura 18 - Rede de coocorrência do Cluster 1 - planilha PERGUNTAS	79
Figura 19 - Análise Hierárquica de Cluster - planilha PERGUNTAS	80
Figura 20 - Mapa Auto-Organizável em 5 clusters - planilha PERGUNTAS	81
Figura 21 - Mapa Auto-Organizável em 10 clusters - planilha PERGUNTAS	81
Figura 22 - Pós-processamento	82
Figura 23 - Rede de coocorrência do cluster 2 - planilha PERGUNTAS	117
Figura 24 - Rede de coocorrência do cluster 3 - planilha PERGUNTAS	118
Figura 25 - Rede de coocorrência do cluster 4 - planilha PERGUNTAS	119
Figura 26 - Rede de coocorrência do cluster 5 - planilha PERGUNTAS	120
Figura 27 - Rede de coocorrência de palavras - planilha RESPOSTAS	121
Figura 28 - Rede de coocorrência entre palavras e variáveis - planilha RESPOSTAS	122
Figura 29 - Rede de coocorrência do cluster 1 - planilha RESPOSTAS	131
Figura 30 - Rede de coocorrência do Cluster 2 - planilha RESPOSTAS	132
Figura 31 - Rede de coocorrência do Cluster 3 - planilha RESPOSTAS	133
Figura 32 - Rede de coocorrência do Cluster 4 - planilha RESPOSTAS	134
Figura 33 - Rede de coocorrência do cluster 5 - planilha RESPOSTAS	135
Figura 34 - Análise Hierárquica de Cluster - planilha RESPOSTAS	136
Figura 35 - Mapa auto-organizável - planilha RESPOSTAS	137

LISTA DE QUADROS

Quadro 1 - Portal de Atendimento Institucional da Pró-Reitoria de Extensão.....	19
Quadro 2 - Dissertações e Teses do PPGEGC/UFSC aderentes ao contexto.....	23
Quadro 3 - Definições para "Conhecimento", segundo autores	26
Quadro 4 - Avaliação das ferramentas de mineração de textos.....	45
Quadro 5 - Frequência de Termo (exemplo)	48
Quadro 6 - Frequência de Documento (exemplo)	48
Quadro 7 - Gráfico TF-DF (exemplo)	49
Quadro 8 - Concordância de Palavras-Chave em Contexto (exemplo).....	49
Quadro 9 - Arquivo destino, destaque feito pelo KH Coder (exemplo).....	50
Quadro 10 - Estatísticas de Colocação (exemplo).....	51
Quadro 11 - Associação de Palavras (exemplo).....	52
Quadro 12 - Portal de Atendimento Institucional da Pró-Reitoria de Extensão.....	59
Quadro 13 - Relatório dos chamados de atendimento PAI-PROEX.....	59
Quadro 14 - Planilha Excel com dados extraídos do PAI	60
Quadro 15 - Etapas de preparação dos dados selecionados.....	61
Quadro 16 - WFL - planilha PERGUNTAS.....	62
Quadro 17 - Dados WFL da planilha PERGUNTAS em Excel.....	63
Quadro 18 - Seleção das palavras para análise – planilha PERGUNTAS	63
Quadro 19 - Concordância textual da palavra “participação” – planilha PERGUNTAS.....	64
Quadro 20 - Stemming do verbo “receber” – planilha PERGUNTAS.....	64
Quadro 21 - Conferência da classificação da palavra “docente” – planilha PERGUNTAS	65
Quadro 22 - Frequência de Termo (TF) – planilha PERGUNTAS.....	65
Quadro 23 - Frequência de Documento (DF) – planilha PERGUNTAS	65
Quadro 24 - Gráfico TF – DF – planilha PERGUNTAS	66
Quadro 25 - Seleção das palavras para análise – planilha RESPOSTAS.....	67
Quadro 26 - Frequência de Termo (TF) - planilha RESPOSTAS.....	67
Quadro 27 - Frequência de Documento (DF) - planilha RESPOSTAS	67
Quadro 28 - Gráfico TF-DF - planilha RESPOSTAS	68
Quadro 29 – Modelo de aprendizagem (exemplo)	70
Quadro 30 - Classificação de documento - Naive Bayes (exemplo).....	71
Quadro 31 - Classificação dos documentos - planilha PERGUNTAS.....	72
Quadro 32 – Matriz de tabulação - planilha PERGUNTAS.....	72
Quadro 33 - Concordância textual palavra “projeto” - planilha PERGUNTAS	76
Quadro 34 - Estatística de localização palavra “projeto” - planilha PERGUNTAS	76
Quadro 35 - Análise de Cluster do Documento - planilha PERGUNTAS.....	77
Quadro 36 - Documentos classificados no Cluster 1 - planilha PERGUNTAS.....	78
Quadro 37 - Lista de associação de palavras Cluster 1 - planilha PERGUNTAS.....	78
Quadro 38 - Síntese de tópicos das perguntas	82
Quadro 39 - Síntese de tópicos de respostas.....	84
Quadro 40 - Artigos da RN 88/2016/CUn mais referenciados nas orientações	85
Quadro 41 - Resultados obtidos	93
Quadro 42 - Resultados obtidos	95
Quadro 43 - Seleção das palavras para análise - planilha RESPOSTAS	107
Quadro 44 - Lista de frequência de palavras - planilha RESPOSTAS.....	108

Quadro 45 - Dados WFL da planilha RESPOSTAS em Excel	108
Quadro 46 - Concordância textual da palavra “registro” - planilha RESPOSTAS	109
Quadro 47 - Stemming do verbo “registrar” - planilha RESPOSTAS	109
Quadro 48 - Checagem da classificação da palavra “docente” - planilha RESPOSTAS.....	110
Quadro 49 - Concordância textual palavra “programa”- planilha PERGUNTAS	111
Quadro 50 - Estatística de localização palavra “programa” - planilha PERGUNTAS	111
Quadro 51 - Concordância textual palavra “publicação” - planilha PERGUNTAS	112
Quadro 52 - Estatística de localização palavra “publicação” - planilha PERGUNTAS	112
Quadro 53 - Concordância textual palavra “banca” - planilha PERGUNTAS	112
Quadro 54 - Estatística de localização palavra “banca” - planilha PERGUNTAS	112
Quadro 55 - Concordância textual palavra “curso” - planilha PERGUNTAS	113
Quadro 56 - Estatística de localização palavra “curso” - planilha PERGUNTAS.....	113
Quadro 57 - Concordância textual palavra “prestação”, condicional serviço em R2 - planilha PERGUNTAS.....	114
Quadro 58 - Estatística de localização palavra “prestação” - planilha PERGUNTAS	114
Quadro 59 - Concordância textual palavra “evento” - planilha PERGUNTAS	114
Quadro 60 - Estatística de localização palavra “evento” - planilha PERGUNTAS.....	114
Quadro 61 - Documentos classificados no cluster 2 - planilha PERGUNTAS.....	116
Quadro 62 - Lista de associação de palavras cluster 2 - planilha PERGUNTAS	116
Quadro 63 - Documentos classificados no cluster 3 - planilha PERGUNTAS.....	117
Quadro 64 - Lista de associação de palavras cluster 3 - planilha PERGUNTAS	117
Quadro 65 - Documentos classificados no cluster 4 - planilha PERGUNTAS.....	118
Quadro 66 - Lista de associação de palavras cluster 4 - planilha PERGUNTAS	118
Quadro 67 - Documentos classificados no cluster 5 - planilha PERGUNTAS.....	119
Quadro 68 - Lista de associação de palavras cluster 5 - planilha PERGUNTAS	119
Quadro 69 - Concordância textual palavra “projeto” - planilha RESPOSTAS.....	123
Quadro 70 - Estatística de localização palavra “projeto” - planilha RESPOSTAS.....	123
Quadro 71 - Concordância textual das palavras “projeto” e “encerrar” em posição R2 - planilha RESPOSTAS	123
Quadro 72 - Concordância textual das palavras “projeto” e “aprovar” em posição R3 - planilha RESPOSTAS	124
Quadro 73 - Concordância Textual das palavras “projeto” e “registrar” em posição L2 - planilha RESPOSTAS	124
Quadro 74 - Concordância textual palavra “programa” - planilha RESPOSTAS.....	125
Quadro 75 - Estatística de localização palavra “programa” - planilha RESPOSTAS.....	125
Quadro 76 - Concordância textual das palavras “programa” e “projeto” em posição R2 - planilha RESPOSTAS	125
Quadro 77 - Concordância textual das palavras “programa” e “vincular” em posição L3 - planilha RESPOSTAS	125
Quadro 78 - Concordância textual palavra “publicação” - planilha RESPOSTAS.....	126
Quadro 79 - Estatística de localização palavra “publicação” - planilha RESPOSTAS.....	126
Quadro 80 - Concordância textual das palavras “publicação” e “semestre” em posição L2 - planilha RESPOSTAS	126
Quadro 81 - Concordância textual palavra “banca” - planilha RESPOSTAS.....	127
Quadro 82 - Estatística de localização palavra “banca” - planilha RESPOSTAS.....	127
Quadro 83 - Concordância textual palavra “curso” - planilha RESPOSTAS	127
Quadro 84 - Estatística de localização palavra “curso” - planilha RESPOSTAS	127

Quadro 85 - Concordância textual palavra “prestação de serviço” - planilha RESPOSTAS.	128
Quadro 86 - Estatística de localização palavra “prestação de serviço” - planilha RESPOSTAS	128
Quadro 87 - Concordância textual palavra “evento” - planilha RESPOSTAS	129
Quadro 88 - Estatística de localização palavra “evento” - planilha RESPOSTAS	129
Quadro 89 - Concordância textual das palavras “evento” e “palestra” em posição R2 - planilha RESPOSTAS.....	129
Quadro 90 - Concordância textual das palavras “evento” e “curso” em posição L2 - planilha RESPOSTAS.....	129
Quadro 91 - Análise de Cluster do Documento - planilha RESPOSTAS	130
Quadro 92 - Documentos classificados no cluster 1 - planilha RESPOSTAS	131
Quadro 93 - Lista de associação de palavras cluster 1 - planilha RESPOSTAS.....	131
Quadro 94 - Documentos classificados no cluster 2 - planilha RESPOSTAS	132
Quadro 95 - Lista de associação de palavras cluster 2 - planilha RESPOSTAS.....	132
Quadro 96 - Documentos classificados no Cluster 3 - planilha RESPOSTAS	133
Quadro 97 - Lista de associação de palavras Cluster 3 - planilha RESPOSTAS.....	133
Quadro 98 - Documentos classificados no cluster 4 - planilha RESPOSTAS	134
Quadro 99 - Lista de associação de palavras cluster 4 - planilha RESPOSTAS.....	134
Quadro 100 - Documentos classificados no cluster 5 - planilha RESPOSTAS	135
Quadro 101 - Lista de associação de palavras cluster 5 - planilha RESPOSTAS.....	135
Quadro 102 - Classificação dos documentos - planilha RESPOSTAS	138
Quadro 103 - Matriz de tabulação - planilha RESPOSTAS.....	138
Quadro 104 - Concordância textual da palavra “RNE882016” - planilha RESPOSTAS	139

LISTA DE TABELAS

Tabela 1 - Número de alunos ensino superior por curso de graduação - ano 2018.....	38
Tabela 2 - Matriz de confusão do modelo de aprendizagem.....	70
Tabela 3 - Critérios de busca e ferramentas de visualização.....	93

LISTA DE ABREVIATURAS E SIGLAS

DCBD – Descoberta de Conhecimento em Banco de Dados

DCT – Descoberta de Conhecimento Textual

DF – Document Frequency (Frequência de Documento)

KWIC – Key Word in Context (Concordância de Palavras-Chave em Contexto)

L – left (esquerda)

PAI – Portal de Atendimento Institucional

PLN – Processamento de Linguagem Natural

PPGEGC – Programa de Pós-Graduação em Engenharia e gestão do Conhecimento

PROEX – Pró-Reitoria de Extensão

R – right (direita)

SECI – Socialização, Externalização, Combinação, Internalização

SeTIC – Superintendência de Governança Eletrônica e Tecnologia da Informação e Comunicação

SIGPEX – Sistema Integrado de Gerenciamento de Projetos de Pesquisa e de Extensão

TF – Term Frequency (Frequência do Termo)

UFSC – Universidade Federal de Santa Catarina

SUMÁRIO

1	INTRODUÇÃO	17
1.1	PROBLEMA DE PESQUISA.....	18
1.2	OBJETIVOS.....	19
1.2.1	Objetivo geral.....	20
1.2.2	Objetivos específicos.....	20
1.3	JUSTIFICATIVA.....	20
1.4	ESTRUTURA DO TRABALHO	22
1.5	LIMITAÇÕES DO TRABALHO	22
1.6	ADERÊNCIA AO PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA E GESTÃO DO CONHECIMENTO	23
2	CONCEITOS BÁSICOS	25
2.1	DADO	25
2.2	INFORMAÇÃO	25
2.3	CONHECIMENTO	26
2.4	TOMADA DE DECISÃO.....	30
2.5	DESCOBERTA DE CONHECIMENTO EM BANCOS DE DADOS.....	31
3	MINERAÇÃO DE TEXTOS: APLICAÇÃO, TÉCNICAS E PROCESSO	35
3.1	TWITTER E MINERAÇÃO DE TEXTOS.....	36
3.2	TÉCNICAS DE MINERAÇÃO DE TEXTOS	38
3.3	PROCESSO DE MINERAÇÃO DE TEXTOS	40
3.3.1	Seleção de documentos	40
3.3.2	Preparação dos dados.....	40
3.3.3	Indexação e normalização.....	41
3.3.4	Cálculo da relevância dos termos.....	41
3.3.5	Aplicação das técnicas de mineração de textos.....	42
3.3.6	Análise e interpretação de resultados	42
4	FERRAMENTAS DE MINERAÇÃO DE TEXTOS	43
4.1	SELEÇÃO DA FERRAMENTA DE MINERAÇÃO DE TEXTO	46
4.2	SOFTWARE KH CODER.....	46
4.2.1	Técnicas de Mineração de Textos do KH Coder	47
4.2.1.1	<i>WFL – Words Frequency List – Lista de Frequência das Palavras</i>	<i>47</i>
4.2.1.2	<i>KWIC Concordance – Concordância de Palavras-Chave em Contexto.....</i>	<i>49</i>
4.2.1.3	<i>Word Association – Associação de Palavras.....</i>	<i>52</i>

4.2.1.4	<i>Co-occurrence Network – Rede de Coocorrência</i>	53
4.2.1.5	<i>Cluster Analysis – Análise de Cluster</i>	54
4.2.1.6	<i>Hierarchical Cluster Analysis – Análise Hierárquica de Cluster</i>	54
4.2.1.7	<i>Self-Organizing Map – Mapa Auto-organizável</i>	55
4.2.1.8	<i>Naive Bayes – Classificador Naive Bayes</i>	56
5	APLICAÇÃO DE MINERAÇÃO DE TEXTOS NO PAI/PROEX	58
5.1	EXTRAÇÃO DA INFORMAÇÃO	58
5.1.1	Seleção de documentos	58
5.1.2	Coleta dos dados	60
5.2	PRÉ-PROCESSAMENTO.....	61
5.2.1	Preparação dos dados	61
5.2.2	Indexação e normalização	62
5.2.3	Cálculo da relevância dos termos para a planilha PERGUNTAS	62
5.2.4	Cálculo da relevância dos termos para a planilha RESPOSTAS	66
5.3	MINERAÇÃO DE TEXTOS	68
5.3.1	Mineração por meio das técnicas do KH Coder – planilha PERGUNTAS 69	
5.3.1.1	<i>Classificador Naive Bayes</i>	69
5.3.1.2	<i>Rede de coocorrência de palavras</i>	73
5.3.1.3	<i>Análise de rede de coocorrência de palavras e variáveis</i>	74
5.3.1.4	<i>Análise por cluster</i>	77
5.3.1.5	<i>Análise hierárquica de cluster</i>	79
5.3.1.6	<i>Mapa auto-organizável</i>	80
5.3.2	Mineração por meio das técnicas do KH Coder – planilha RESPOSTAS ..	82
5.4	PÓS-PROCESSAMENTO.....	82
6	Análise de Twitter e a Mineração de Textos aplicado a @UFSC	92
6.1	RESULTADOS UFSC X BITCOIN.....	92
7	CONCLUSÃO E TRABALHOS FUTUROS	97
	REFERÊNCIAS	99
	APÊNDICE 1	104
	APÊNDICE 2	105
	APÊNDICE 3	106
	APÊNDICE 4	107
	APÊNDICE 5	111
	APÊNDICE 6	116

APÊNDICE 7	121
------------------	-----

1 INTRODUÇÃO

Desde a década de 1950, com o desenvolvimento e o avanço das tecnologias, muitas transformações sociais e econômicas foram vivenciadas, entre elas, destaca-se a evolução da era digital para a era tecnológica e, mais recentemente, para a era da informação.

A *Internet* e a popularização de dispositivos portáteis ocasionaram um aumento na quantidade de informações disponíveis na *Web*, transformando-a numa fonte inesgotável de dados e impulsionando o processo de disseminação do conhecimento.

Entretanto, esta evolução tecnológica trouxe como consequência uma sobrecarga de informações produzidas em linguagem natural que, devido a sua complexidade, demandam coleta, tratamento e análise adequada para que se convertam em informações relevantes (SANTOS, R. E. S. *et al.*, 2014; CAVALCANTI, 2020).

A evolução tecnológica fez da *Internet* o principal agente de mudança econômica e uma fonte inesgotável de dados – que é a unidade básica da informação. Esses dados são coletados, armazenados, analisados e transformados em informações pelas plataformas online. O crescimento do volume de coleta e análise de dados veio acompanhado do crescente uso de algoritmos para possibilitar o tratamento de dados e a tomada de decisões pela plataforma online (DIAS, 2021, p. 3).

Nesse processo de desenvolvimento, a informação tornou-se um recurso fundamental para as organizações sendo considerada uma fonte de sucesso ou de fracasso, pois traz consigo um grande desafio: saber como lidar com o grande volume de informações produzidas e que circulam nas organizações e redes sociais (CALAZANS, 2008; FERREIRA, MOURA, BARROS, 2014; CHISTOL, 2020).

Grande parte das informações está armazenada em formato eletrônico e é acessível via *Web*, contudo, precisa ser tratada adequadamente para transformar-se em informação útil que gera conhecimento. Essa preocupação com a busca da informação de qualidade está relacionada ao fato da informação ser o insumo básico no processo de tomada de decisão das organizações (CHOO, 2003; DANTAS, 2013; ZHU *et al.*, 2014; SHUHAI *et al.*, 2019).

Assim, a descoberta do conhecimento a partir de bancos de dados tem mudado a forma como as organizações tomam suas decisões, tornando-se uma estratégia fundamental para o desenvolvimento e crescimento organizacional. Consequentemente, várias técnicas e ferramentas estão emergindo para ajudar gestores a analisar a informação armazenada com o objetivo de extrair informação de qualidade e construir conhecimento novo.

Dentre elas, destaca-se a mineração de textos, uma ferramenta que vem sendo aplicada em áreas científicas e comerciais, gerando conhecimento a partir de dados não estruturados.

Utiliza técnicas de análise e extração de dados a partir de textos, frases ou apenas palavras e, a partir de algoritmos computacionais, identifica informações úteis e muitas vezes implícitas para as organizações.

Trata-se de um processo de mineração do conhecimento a partir de grande volume de dados não estruturados, que aplica tecnologias de bancos de dados, reconhecimento de padrões, inteligência artificial, redes neurais, estatística e recuperação da informação. Podendo, inclusive, ser utilizada para análise de sentimentos em pesquisas.

A mineração de texto refere-se à extração de informações e padrões que estão implícitos, anteriormente desconhecido e potencialmente valioso de maneira automática ou semiautomática a partir de imensos dados textuais não estruturados, como textos em linguagem natural (HASSANI *et al.*, 2020).

1.1 PROBLEMA DE PESQUISA

O Portal de Atendimento Institucional (PAI) é um canal de atendimento ao usuário disponibilizado pela Superintendência de Governança Eletrônica e Tecnologia da Informação e Comunicação (SeTIC) da Universidade Federal de Santa Catarina (UFSC).

Através do PAI, qualquer pessoa pode abrir um chamado de atendimento e encaminhar sua dúvida ou solicitação para diferentes setores da Universidade.

O acesso é simples e rápido, necessitando apenas de um navegador *Web* (*Firefox*, *Internet Explorer*, *Chrome*, etc.) para a abertura de um chamado. A resposta é encaminhada para o *e-mail* do usuário.

Dentre as vantagens e facilidades da utilização do PAI destacam-se:

- Celeridade na resposta;
- Padronização do atendimento;
- Rastreabilidade das solicitações;
- Atendimento eletrônico via *e-mail*;
- Divisão de papéis – atribuição de responsabilidades nos setores.

Principal meio de comunicação entre os usuários e a Universidade, o PAI recebe centenas de chamados a cada mês. A Pró-Reitoria de Extensão (PROEX) utiliza o PAI desde de agosto de 2019 e em menos de 2 anos recebeu cerca de 2.500 chamados.

Ao registrar um chamado de atendimento, o usuário seleciona um dos serviços listados no menu (Quadro 1). O tipo de serviço selecionado permite ao atendente identificar o assunto antes de abrir o chamado.



Fonte: Autora (2021) - (PAI-PROEX)

A maioria dos chamados encaminhados para a PROEX está relacionada ao Sistema Integrado de Gerenciamento de Projetos de Pesquisa e de Extensão (SIGPEX). E muitas dúvidas são recorrentes entre os usuários.

As informações do PAI encontram-se na forma de texto, linguagem natural, ou seja, dados não estruturados, e apesar de estarem armazenadas em um banco de dados, **o sistema não disponibiliza nenhum tipo de relatório informacional.**

A identificação das dúvidas e respostas mais frequentes poderia conferir aos gestores subsídios para a gestão do conhecimento e para a adoção de estratégias de compartilhamento de conhecimento visando melhorar a comunicação entre usuários e organização.

Nesse contexto, dois questionamentos foram levantados: Quais são as principais dúvidas dos usuários do PAI-PROEX? Quais são as principais respostas dos atendentes do PAI-PROEX?

Além disso, levando em consideração que o público da Universidade é formado majoritariamente de jovens estudantes e que o uso de redes sociais como forma de comunicação também é uma fonte importante de informações, surgiu o seguinte questionamento: As informações veiculadas nas redes sociais podem ser utilizadas para melhorar a comunicação entre os usuários e a Universidade?

Assim, levantou-se o seguinte problema de pesquisa: Quais informações armazenadas no banco de dados do PAI e veiculadas em redes sociais podem ser descobertas e utilizadas na Gestão do Conhecimento, melhorando a comunicação entre os usuários e a PROEX?

1.2 OBJETIVOS

A presente pesquisa apresenta os seguintes objetivos geral e específicos.

1.2.1 Objetivo geral

Avaliar o uso de informações não estruturadas no processo de descoberta do conhecimento e de tomada de decisão, visando a melhoria da comunicação entre usuários e organização.

Para tanto, é realizado um estudo de caso utilizando a aplicação de diferentes técnicas de mineração de textos no Portal de Atendimento Institucional (PAI) da Pró-Reitoria de Extensão (PROEX) da Universidade Federal de Santa Catarina (UFSC) e a aplicação de mineração de textos na rede social *Twitter*.

1.2.2 Objetivos específicos

- a) Selecionar uma ferramenta para a aplicação de mineração de textos;
- b) Levantar e analisar diferentes técnicas de análise de informação não estruturada a serem utilizadas na mineração de textos;
- c) Aplicar técnicas de mineração de textos em informações coletadas do PAI-PROEX, visando a descoberta de informações úteis e relevantes para a tomada de decisão;
- d) Aplicar técnicas de mineração de textos em análise de rede social, visando a descoberta de informações úteis e relevantes para a tomada de decisão;
- e) Analisar a qualidade informacional descoberta por meio da mineração de textos.

1.3 JUSTIFICATIVA

O desenvolvimento das tecnologias, principalmente a criação da *Internet* e a popularização de dispositivos portáteis como celulares e *tablets*, tem proporcionado uma maior interação entre as pessoas e tem facilitado a criação de conteúdo digital de diversas fontes e assuntos.

Conseqüentemente, gerou-se um aumento significativo na quantidade de informações disponíveis na *Web*, transformando-a numa fonte inesgotável de dados em formato eletrônico, grande parte produzida em linguagem natural (CAVALCANTI, 2020; NASEEM *et al.*, 2020).

O aumento significativo e substancial da quantidade de dados textuais digitais disponíveis tem aberto muitas oportunidades de pesquisa, principalmente no campo de técnicas

analíticas de *big data*, onde a mineração de textos ganhou atenção significativa e uma ampla gama de aplicações.

Contudo, a existência de uma grande quantidade de informações armazenadas em bancos de dados não estruturados, em formato de textos e que utilizam a linguagem natural, demanda uma coleta, tratamento e análise adequada para que se converta em informação útil e relevante (SHUHAI *et al.*, 2019).

A informação é considerada uma estratégia fundamental na tomada de decisão (CHISTOL, 2020). Assim, devido à crescente importância da inteligência artificial e a implementação de plataformas digitais, diferentes modelos de negócios têm buscado soluções para seus problemas em tecnologias de informação.

A mineração de textos é um processo de extração de informações previamente desconhecidas e potencialmente úteis de documentos textuais escritos em linguagem natural e disponibilizados em formato eletrônico (HASSANI *et al.*, 2020).

Ao se trabalhar com dados provenientes da organização, utiliza-se um conjunto de artefatos que trabalham sobre informações não estruturadas e que podem ser compartilhadas com o propósito de agregar valor ao processo decisório envolvido. São estudadas técnicas de tratamento de informação não estruturada que potencializam a Gestão do Conhecimento Organizacional.

Trabalha-se na gestão de processos e no desenvolvimento de soluções que, embora altamente técnicas, permitem uma interação em linguagem natural. Também são estabelecidas condições para promoção da inovação dado que se trabalha em nível de cultura organizacional.

O PAI é uma importante fonte de informação, porém não produz nenhum relatório informacional, estando no momento atuando apenas como um banco de dados textuais. A aplicação de uma ferramenta de mineração permitirá explorar técnicas de processamento de linguagem natural, cujo objetivo é promover a compreensão das informações armazenadas por meio do uso de recursos computacionais para processamento de textos.

Da mesma forma, a análise de conteúdo dos textos de redes sociais, aplicando os conceitos de processamento de linguagem natural, permitirá identificar atitudes e posições expressas pelos seus usuários. Assim, grandes redes sociais podem ser utilizadas para definir, por exemplo, uma política de comunicação eficiente entre o usuário e a organização.

Nesta pesquisa, utiliza-se a aplicação de uma ferramenta de mineração de textos e diferentes técnicas como forma de avaliar a qualidade informacional do PAI-PROEX e da rede social *Twitter*, que podem promover a descoberta de informações relevantes e importantes que a Universidade já possui, mas que são desconhecidas para a mesma.

1.4 ESTRUTURA DO TRABALHO

Esta pesquisa está dividida em 7 capítulos. O capítulo 1 apresenta a introdução, o problema de pesquisa, os objetivos, a justificativa, as limitações e a aderência da pesquisa ao Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento (PPGEGC).

O capítulo 2 apresenta os conceitos básicos e relações teóricas que fornecem subsídios para uma melhor compreensão e desenvolvimento da pesquisa. Relaciona de forma concisa os conceitos dos seguintes temas: dado, informação, conhecimento, tomada de decisão e descoberta de conhecimento em bancos de dados.

O capítulo 3 apresenta a evolução da mineração de dados para a mineração de textos; a aplicação de mineração de textos na rede social *Twitter*; apresentação de diferentes técnicas que utilizam análise estatística para descobrir regras de associação e padrões sobre distribuição e associações de palavras-chave; e a apresentação das etapas que envolvem o processo de mineração de textos, desde a seleção dos documentos até a análise dos resultados.

O capítulo 4 apresenta um comparativo de ferramentas de mineração de textos que embasaram a seleção da ferramenta utilizada nesta pesquisa. Apresenta também as características principais da ferramenta escolhida e analisa e descreve as técnicas utilizadas no estudo de caso.

O capítulo 5 apresenta a metodologia para aplicação da ferramenta de mineração de texto. Apresenta os processos de (i) extração da informação, com a seleção dos documentos e coleta dos dados; (ii) pré-processamento, com a preparação dos dados, indexação, normatização e cálculo de relevância dos termos; e (iii) mineração de texto, com a aplicação das técnicas ou comandos da ferramenta escolhida; e (iv) pós-processamento, com a interpretação dos resultados obtidos por meio da mineração de textos.

O capítulo 6 apresenta a análise de *Twitter* e a mineração de textos aplicada à #UFSC. Por fim, o capítulo 7, traz a conclusão e contribuições desta pesquisa.

1.5 LIMITAÇÕES DO TRABALHO

Esta pesquisa concentra-se na área de Qualidade da Informação e Gestão do Conhecimento, não faz parte de seu escopo aprofundar-se nas áreas de Estatística e Ciências da Computação.

Dentre os serviços disponibilizados pela PROEX por meio do PAI, esta pesquisa se limita aos chamados de atendimento relacionados ao SIGPEX.

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

Esta dissertação aborda a descoberta de conhecimento por meio do uso de uma ferramenta de mineração de dados não estruturados e a sua utilização na tomada de decisão de organizações, para tanto, é realizado um estudo de caso.

Como tal, está inserida na área de Gestão do Conhecimento e na linha de pesquisa Gestão do Conhecimento Organizacional, que aborda os estudos teóricos e práticos sobre a utilização do conhecimento como fator de produção estratégico no gerenciamento de negócios relacionados à economia do conhecimento.

Quadro 2 - Dissertações e Teses do PPGEGC/UFSC aderentes ao contexto

Título	Autor/Orientador	Objetivo	Ano/Tipo
Método de identificação de padrões em discurso político a partir da descoberta de conhecimento	Autor: Márcio Welter Orientador: João Bosco da Mota Alves	Desenvolver um método de identificação e representação de padrões em discursos políticos por intermédio do processo de descoberta de conhecimento em texto.	2021 Dissertação
Modelo de avaliação de potenciais ideias alinhadas ao contexto organizacional	Autor: Marina Carradore Sérgio Orientador: Alexandre Leopoldo Gonçalves	Avaliar o potencial de implementação de ideias alinhadas ao contexto organizacional, propondo um modelo com base em ferramentas.	2020 Tese
Modelo de reconhecimento de padrões em ideias usando técnicas de descoberta de conhecimento em textos	Autor: Alessandro Costa Ribeiro Orientador: Gertrudes Aparecida Dandolini	Desenvolver um modelo de reconhecimento de padrões em ideias amparado por técnicas de descoberta de conhecimento em texto.	2018 Dissertação
Análise de Agrupamentos e Mineração de Opinião como Suporte à Gestão de Ideias	Autor: Guilherme Martins Alvarez Orientador: Alexandre Leopoldo Gonçalves	Propor um método baseado em Mineração de Opinião e Análise de Agrupamentos como suporte à Gestão de Ideias, objetivando auxiliar o processo de análise e seleção de ideias inovadoras	2018 Dissertação
Um modelo baseado em ontologia e extração de informação como suporte ao processo de design instrucional na geração de mídias do conhecimento	Autor: Israel de Alcântara Braglia Orientador: Alice T. Cybis Pereira	Desenvolver um modelo que tivesse como suporte ontologias de domínio para a EAD, pois todo material instrucional de um curso de EAD nasce de um texto bruto (corpus).	2014 Tese
Um modelo de descoberta de conhecimento inerente à evolução temporal dos	Autor: Alessandro Botelho Bovo	Desenvolver um modelo para descoberta de conhecimento a partir de informações não estruturadas analisando a	2011 Tese

relacionamentos entre elementos textuais	Orientador: Vinícius Medina Kern	evolução dos relacionamentos entre os elementos textuais ao longo do tempo.	
--	----------------------------------	---	--

Fonte: Autora (2021)

Neste trabalho, focou-se no uso de uma ferramenta de mineração de dados não estruturados para verificar como informações armazenadas em banco de dados de textos podem ser utilizadas para melhorar a comunicação entre usuários e organização, objetivando a descoberta do conhecimento e fornecendo insumos para a tomada de decisão.

No histórico do PPGE GC, foram encontradas pesquisas que guardam afinidade com o tema desta dissertação, destacadas no Quadro 2.

Observa-se que os trabalhos de Bovo (2011) e Ribeiro (2018) propõe modelos de descoberta do conhecimento em texto. Enquanto, Braglia (2014) propõe um modelo de suporte de ontologias.

Contudo, os trabalhos que mais se aproximam desta dissertação são Welter (2021), pois analisa métodos, técnicas e ferramentas utilizadas para tratamento de dados textuais em discursos políticos, e Sérgio (2020), pois identifica métodos e técnicas de mineração de dados e texto para a avaliação de ideias.

Observando o histórico de trabalhos do PPGE GC, verifica-se que a presente pesquisa traz uma contribuição diversa e específica, focada na avaliação do uso de informações não estruturadas no processo de tomada de decisão, por meio da aplicação de uma ferramenta de mineração de textos.

2 CONCEITOS BÁSICOS

Este capítulo apresenta de forma concisa os conceitos e relações teóricas que fornecem subsídios para uma melhor compreensão e desenvolvimento da pesquisa. São abordados os seguintes temas: dado, informação, conhecimento, tomada de decisão e descoberta de conhecimento em bancos de dados.

2.1 DADO

Dado é a representação ou registro de alguma coisa. Pode ser uma letra, um número, uma palavra, bem como, conjuntos de números e vocábulos desorganizados, os quais não transmitem nenhuma informação ou conhecimento. Um dado quando observado individualmente não apresenta significado ou sentido definido (URIARTE, 2008; WOLSKI; GOMOLIŃSKA, 2020).

Para trazer significado a um dado é necessário tratá-lo e transformá-lo em informação. O dado se transforma em informação quando se relaciona com outros dados. Nesta pesquisa, será adotada a abordagem que mais se sobressai, onde o dado propicia a informação que propicia o conhecimento (OLETO, 2006; COSTA *et al.*, 2019).

2.2 INFORMAÇÃO

Informação é um conjunto de dados estruturados, organizados, processados, contextualizados ou interpretados, que transmite uma mensagem dentro de um contexto real, provida de propósito, significado e relevância.

A informação é formada por um conjunto de dados que, quando fornecida na forma e tempo adequado, aprimora o conhecimento do indivíduo que a recebe, tornando-o mais habilitado a desenvolver determinada atividade ou a tomar determinada decisão (CHIAVENATO, 2003; COSTA *et al.*, 2019).

Assim, compreende-se que a informação proporciona um novo ponto de vista para a interpretação de conceitos e eventos, tornando visíveis os significados antes não percebidos e permitindo extrair e construir o conhecimento.

2.3 CONHECIMENTO

É fundamental distinguir informação e conhecimento. O conhecimento não é encontrado no conteúdo, estrutura, precisão ou utilidade da informação. O conhecimento é informação possuída na mente dos indivíduos, é informação personalizada (que pode ou não ser nova, única, útil ou precisa) relacionada a fatos, procedimentos, conceitos, interpretações, ideias, observações e julgamentos (ALAVI; LEIDNER, 2001; ABUBAKAR *et al.*, 2019).

Além disso, o conhecimento é criado e organizado por fluxos de informação, moldados por seu detentor. Portanto, o conhecimento não existe fora de um agente (um conhecedor), ele é indelevelmente moldado pelas necessidades de alguém, bem como pelo seu conhecimento prévio e experiências (SERRAT, 2008; FAHEY; PRUSAK, 1998; TUOMI, 1999).

Pode-se afirmar que a informação é convertida em conhecimento quando processada na mente dos indivíduos. Ao mesmo tempo que o conhecimento se torna informação, quando articulado e apresentado na forma de texto, gráficos, palavras ou outras formas simbólicas (ALAVI; LEIDNER, 2001; ABUBAKAR *et al.*, 2019).

Sabino (2019), em sua tese de doutorado, apresenta uma síntese sobre das definições de conhecimento sob o olhar de diversos autores, apresentada no Quadro 3.

Quadro 3 - Definições para "Conhecimento", segundo autores

Definições de “conhecimento”	Autor
Nós sabemos mais do que somos capazes de expressar.	Polanyi (1967)
O conhecimento consiste numa construção contínua e é resultante da interação entre o homem e o mundo.	Maturana e Varela (1995)
Conhecimento é essencialmente dado, já existe com a organização, ou pode ser apreendido ou adquirido de outras fontes.	Nonaka, Umemoto e Senoo (1996)
O conhecimento é uma informação cuja validade foi estabelecida através de testes para sua validação.	Liebeskind (1996)
O conhecimento refere-se tanto à experiência física e à tentativa e erro quanto à geração de modelos mentais e ao aprendizado com os outros.	Nonaka e Takeuchi (1997)
Conhecimento é o que compramos, vendemos e produzimos.	Stewart (1998)
Conhecimento é uma mistura fluída de experiência condensada, valores, informação contextual e insight experimentado – mistura que proporciona uma estrutura de avaliação e incorporação de novas experiências e informações.	Davenport e Pruzak (1998)
O conhecimento consiste numa construção contínua e é resultante da interação entre o homem e o mundo. A definição do conhecimento é algo amplo e não existe uma palavra que seja aceita de modo geral.	Sveiby (1998)

Conhecimento é um conjunto de declarações organizadas sobre fatos ou ideias. Apresenta um julgamento ponderado ou um resultado experimental que é transmitido a outros por intermédio de algum meio de comunicação, de alguma forma sistemática.	Castells (1999)
Conhecimento é um significado feito para a mente.	Marakas (1999)
Conhecimento é prática compartilhada, como a propriedade da comunidade de prática que necessita, cria, usa, debate, distribui, adapta e o transforma.	Despres e Chauvel (2000)
Conhecimento é o conjunto de insights, experiências, e procedimentos que são considerados corretos e verdadeiros; que guiam pensamentos, comportamentos e a comunicação entre pessoas; e que, além disso, aumentam a compreensão ou o desempenho numa área ou disciplina.	Queiroz (2001)
Conhecimento é o entendimento obtido por meio da inferência realizada no contato com dados e informações que traduzem a essência de qualquer elemento.	Cruz (2002)
O conhecimento é um conjunto total que inclui cognição e habilidades que os indivíduos utilizam para resolver problemas. O conhecimento se baseia em dados e informações, mas, ao contrário deles, está sempre ligado a pessoas.	Probst, Raub e Romhardt (2002)
O conhecimento é uma mistura fluída de experiência condensada, valores, informação contextual e insight experimentado, a qual proporciona uma estrutura para avaliação e incorporação de novas experiências e informações. Ele tem origem e é aplicado na mente dos conhecedores. Nas organizações, ele costuma estar embutido não só em documentos ou repositórios, mas também em rotinas, processos, práticas e normas organizacionais.	Prusak e Davenport (2003)
O conhecimento é uma construção social, historicamente datada, não neutra, que atende diferentes fins em cada sociedade, reproduzindo e produzindo relações sociais, inclusive as que se referem à vinculação de saber e poder.	Loureiro (2006)
O conhecimento é uma construção social que só ganha sentido quando circula publicamente e se coloca a serviço das comunidades.	Grusmann e Siqueira (2007)
Conhecimento inclui tudo aquilo que sabemos sobre o mundo.	Molaei (2010)
Conhecimento é a compreensão humana de um campo especializado de interesse, adquirida por meio de estudo e experiência.	Koskinen (2013)

Fonte: Autora (2021) - Adaptado de Sabino (2019).

Para a área de concentração de Gestão do Conhecimento do Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento da UFSC, conforme definido em seu Planejamento Estratégico de 2016, “o conhecimento é processo e produto efetivado na relação entre pessoas e agentes não humanos para a geração de valor” (SANTOS, N.; RADOS, 2020).

Pode-se identificar dois tipos de conhecimento, o tácito e o explícito. De acordo com Serrat (2008, p.1):

O **conhecimento tácito** é o conhecimento não verbalizado, intuitivo e não articulado que as pessoas carregam em suas cabeças. É difícil formalizar e comunicar porque está enraizado em habilidades, experiências, percepções, intuição e julgamento, mas pode ser compartilhado em discussões, narrativas e interações pessoais. Possui uma

dimensão técnica, que engloba competências e capacidades denominadas know-how. Tem uma dimensão cognitiva, que consiste em crenças, ideais, valores, esquemas ou modelos mentais. O **conhecimento explícito** é o conhecimento codificado que pode ser expresso por escrito, desenhos ou programas de computador, por exemplo, e transmitido de várias formas. O conhecimento tácito e o conhecimento explícito são formas de significado mutuamente complementares.

Apesar de serem muitas vezes retratados como extremos polares, Takeuchi e Nonaka (2008) afirmam que o conhecimento não é explícito ou tácito, mas é tanto explícito quanto tácito, e mais do que complementares, os conhecimentos são também interpenetrantes.

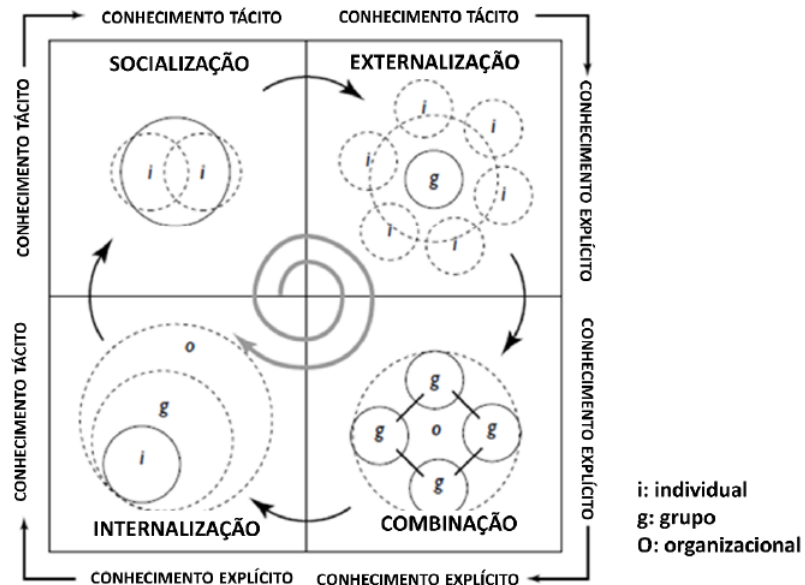
Para os autores, conseguimos entender o conhecimento tácito no momento que entendemos o conhecimento explícito, pois “existe algum conhecimento explícito em cada conhecimento tácito e algum conhecimento tácito em todo conhecimento explícito” (TAKEUCHI; NONAKA, 2008, p. 22), tornando-os contínuos, porém separáveis.

Compreende-se a conversão do conhecimento tácito em explícito, e vice-versa, como um processo de criação do conhecimento. Nonaka e Takeuchi propõe um modelo, conhecido como SECI (Socialização, Externalização, Combinação, Internalização), que descreve como os conhecimentos tácito e explícito são amplificados em termos de qualidade e quantidade, e sua transição de indivíduo para grupo e para organização (OLIVEIRA *et al.*, 2020).

Somando-se ao modelo SECI, Nonaka e Konno (1998) apresentam o conceito de *Ba* e representam a evolução da espiral do conhecimento e o processo de autotranscendência por meio da Figura 1, que enfatiza as trocas de conhecimento e os níveis em que elas ocorrem (VICENTE; DA CUNHA, 2021).

Para aqueles não familiarizados com o conceito, *Ba* pode ser pensado como um espaço compartilhado para relacionamentos emergentes. Este espaço pode ser físico (por exemplo, escritório, espaço comercial disperso), virtual (por exemplo, e-mail, teleconferência), mental (por exemplo, experiências compartilhadas, ideias, ideais) ou qualquer combinação deles. O que diferencia *Ba* da interação humana comum é o conceito de criação de conhecimento. *Ba* fornece uma plataforma para o avanço do conhecimento individual e/ou coletivo. É de tal plataforma que uma perspectiva transcendental integra toda a informação necessária. *Ba* também pode ser pensado como o reconhecimento do eu em tudo. De acordo com a teoria do existencialismo, *Ba* é um contexto que abriga significado. Assim, consideramos que o *Ba* é um espaço compartilhado que serve de base para a criação de conhecimento (NONAKA; KONNO, 1998, p. 40).

Figura 1 - Evolução da espiral do conhecimento



Fonte: Nonaka e Konno (1998)

Assim, na espiral do conhecimento, o conhecimento pode ser convertido e transmitido, por meio de quatro processos:

- Socialização** – *Ba* origem (Tácito em Tácito). Compartilhar e criar conhecimento tácito através de experiências diretas e modelos mentais no nível indivíduo para indivíduo.
- Externalização** – *Ba* interação (Tácito em Explícito). Articular conhecimento tácito através do diálogo e da reflexão no nível indivíduo para grupo.
- Combinação** – *Cyber Ba* (Explícito em Explícito). Sistematizar e aplicar o conhecimento explícito e a informação no nível grupo para organização.
- Internalização** – *Ba* exercício (Explícito em Tácito). Aprender e adquirir novo conhecimento tácito na prática no nível organização para indivíduo.

Os quatro processos se retroalimentam, numa espiral contínua de construção do conhecimento. Do ponto de vista de organizações, resumidamente, a construção do conhecimento se inicia com os indivíduos, por meio de *insights* ou intuições. O *know-how* dos indivíduos é compartilhado por meio da socialização. Sua exploração pode ser combinada e reconfigurada em novas formas de conhecimento explícito. O novo conhecimento explícito gerado é revivenciado e reinternalizado na forma de novo conhecimento tácito.

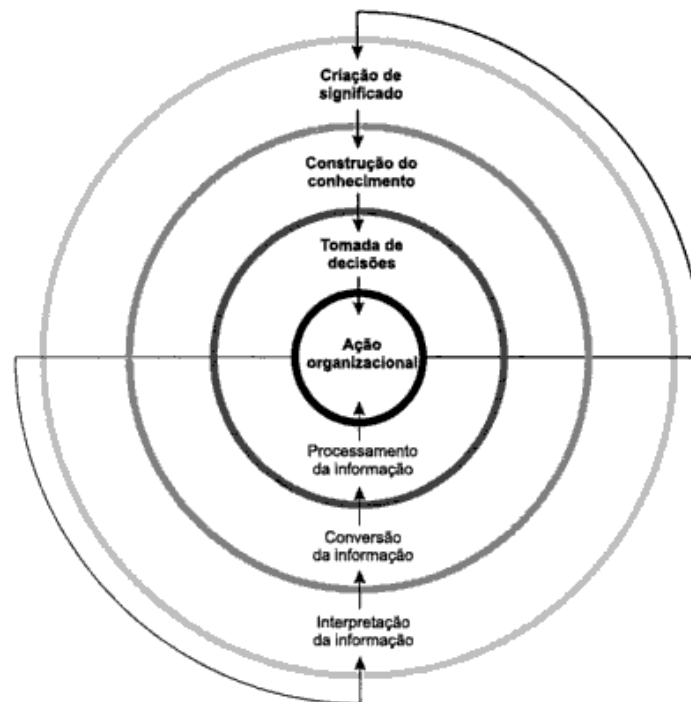
2.4 TOMADA DE DECISÃO

Para Choo (2003, p. 30), “durante a construção do conhecimento, o principal processo de informação é a conversão do conhecimento”. E, “durante a tomada de decisão, a principal atividade é o processamento e a análise da informação a partir de alternativas disponíveis, cujas vantagens e desvantagens são pesadas”.

Embora sejam quase sempre tratadas como processos independentes de informação organizacional, as três arenas de uso da informação – criar significado, construir conhecimento e tomar decisões – são de fato processos interligados, de modo que, analisando como essas três atividades se alimentam mutuamente, teremos uma visão holística do uso da informação (CHOO, 2003, p. 29).

Para o autor, a interpretação, a conversão e o processamento de informações, são processos sociais dinâmicos que reconstituem significados e conhecimento, bem como, embasam a tomada de decisão (Figura 2).

Figura 2 - A organização do conhecimento



Fonte: Choo (2003)

A informação é o insumo básico do processo de tomada de decisão. Assim, torna-se primordial e indispensável às organizações dispor de informação confiável, adequada e em tempo certo, para que se tomem decisões eficazes e eficientes (WANG, STRONG, 1996; LOH, OLIVEIRA, GAMEIRO, 2003; CAMPELO *et al.*, 2020).

Segundo Cassaro (1995, p.45), “uma decisão nada mais é do que uma escolha entre alternativas, obedecendo a critérios previamente estabelecidos. Estas alternativas poderão ser

os objetivos, os programas ou políticas – em uma atividade de planejamento – ou os recursos, estrutura e procedimentos – em uma atividade organizacional”.

Contudo, a forma como as pessoas tomam as decisões e a qualidade de suas escolhas finais depende muito da percepção de quem toma a decisão (ROBBINS, 2005). Assim, o embasamento para a tomada de decisão nas organizações deve partir da busca por informações de qualidade, que agreguem valor e respaldem as escolhas dos indivíduos que tomam as decisões.

[...] os indivíduos que tomam decisões numa organização também são influenciados por sua tendência a buscar e usar seletivamente as informações que confirmem suas crenças e facilitem os resultados desejados. Esse processamento seletivo não implica que os indivíduos abreviem a busca da informação. Ao contrário, eles buscam mais informações do que seriam necessárias e as utilizam para aumentar sua confiança em suas escolhas. Nas situações cercadas por alto nível de incerteza, as preferências por certos resultados podem ser o componente menos ambíguo do processo decisório, mais certo que a definição do problema, o número de alternativas plausíveis ou as probabilidades associadas às várias alternativas. Portanto, os que tomam as decisões podem reduzir a incerteza concentrando-se nas informações que os ajudem a alcançar os resultados desejados (CHOO, 2003, p. 319).

Pode-se afirmar que a tomada de decisão requer interpretação e avaliação de dados e informações de diversas origens, que precisam ser selecionados, processados e interpretados. Para (LOH; WIVES; OLIVEIRA, 2000b, p. 2):

A prática da inteligência possibilita aos empresários estarem sempre bem informados e preparados para minimizar riscos, antecipar crises e tornar seus produtos mais competitivos. [...] Neste sentido, a área de inteligência busca suprir suas deficiências adotando técnicas provenientes da área de recuperação de informações, extração de informações e descoberta de conhecimento em textos.

2.5 DESCOBERTA DE CONHECIMENTO EM BANCOS DE DADOS

Os constantes avanços tecnológicos e a proeminente quantidade de dados gerados, bem como, a necessidade de selecionar informações de qualidade que gerem conhecimento têm levado as organizações a procurarem ferramentas de gestão que auxiliem sua tomada de decisão.

Atualmente, uma quantidade significativa de informações importantes se encontra armazenada em formato de textos, disponíveis de forma eletrônica e acessíveis via ferramentas digitais, e que requerem tratamento e análise adequada.

Desde o final dos anos 80, pesquisadores em Descoberta de Conhecimento em Banco de Dados (DCBD) vêm se dedicando para disponibilizar ferramentas para a extração de padrões

desconhecidos a partir de bancos de dados estruturados e não estruturados, procurando tornar essa tarefa a mais automatizada possível (SILVA, E. M., 2002; COSTA *et al.*, 2019).

Assim, a DCBD está preocupada com o desenvolvimento de métodos e técnicas para dar sentido aos dados.

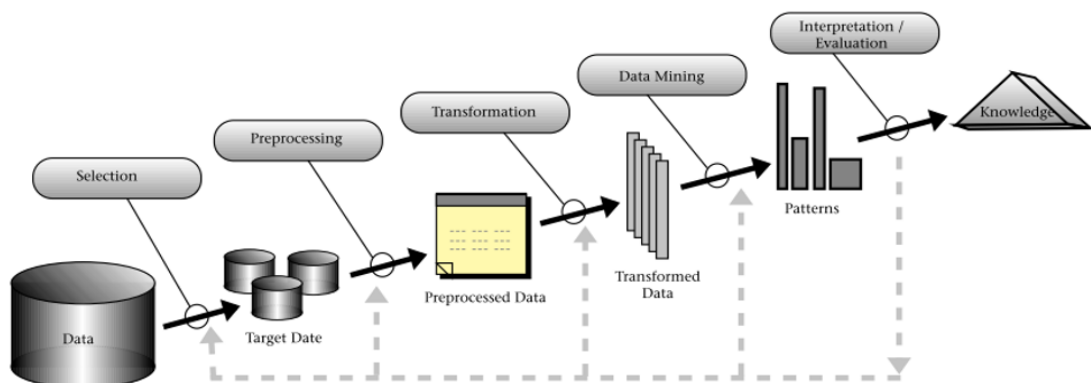
O problema básico abordado pelo processo de descoberta de conhecimento em banco de dados é o de mapear dados de baixo nível (que normalmente são muito volumosos para entender e digerir facilmente) em outras formas que podem ser mais compactas (por exemplo, um breve relatório), mais abstrato (por exemplo, uma aproximação descritiva ou modelo do processo que gerou os dados), ou mais útil (por exemplo, um modelo preditivo para estimar o valor de casos futuros). No cerne do processo está a aplicação de métodos específicos de mineração de dados para a descoberta e extração de padrões (Fayyad *et al.*, 1996, p. 37).

As primeiras aplicações da DCBD surgiram em bancos de dados estruturados, cujo principal objetivo está ligado a descoberta de relacionamentos de dados em registro de bancos de dados e a produção de relatórios para análise (CAMILO, 2010). Os métodos e ferramentas utilizadas na DCBD estão baseados em três áreas: estatística, inteligência artificial e recuperação de informações (WIVES, 1999).

Dependendo do seu objetivo, a DCBD pode ocorrer de forma reativa ou proativa (LOH; WIVES; OLIVEIRA, 2000a). Na forma reativa, o usuário define claramente sua necessidade e procura a solução para um problema específico. Já na forma proativa, o usuário procura informações relevantes que possam levar à descoberta de padrões que lhe auxiliem na tomada de decisões.

O processo da DCBD, conforme Fayyad *et al.* (1996), consiste das seguintes fases: compreensão do negócio, seleção dos dados, limpeza e pré-processamento dos dados, codificação ou transformação dos dados, mineração dos dados, interpretação e avaliação dos dados (Figura 3) (SILVA; VIERA, 2021).

Figura 3 - Processo de descoberta de conhecimento em base de dados



Fonte: Fayyad *et al.* (1996)

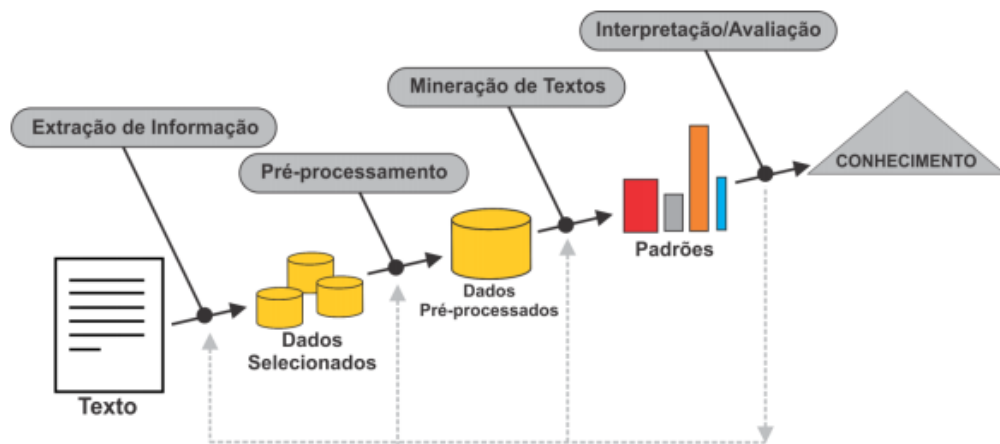
Contudo, com o advento da *internet* e a popularização de serviços *online*, surgiu uma nova área de descoberta do conhecimento, cuja principal fonte está relacionada a dados não estruturados, armazenados na forma textual, sem organização e padronização prévia. Surge então, a Descoberta de Conhecimento Textual (DCT).

A descoberta do conhecimento em bases de dados não estruturados é uma atividade complexa e que, por consequência, exige técnicas e ferramentas específicas que auxiliem na análise de grandes volumes de dados e na busca por informação útil que gere conhecimento.

Semelhante ao processo DCBD, para Silva (2012) o processo de DCT consiste nas seguintes etapas: extração da informação, pré-processamento, mineração de textos, interpretação e avaliação (Figura 4). Para o autor as principais diferenças entre os processos DBCD e DCT são:

- a) Extração de Informação. Nesta etapa são selecionados os textos em geral de acordo com o domínio do problema, considerando os objetivos que se deseja alcançar.
- b) Pré-processamento. O objetivo desta fase é eliminação de termos não relevantes (*stopwords*), redução das palavras aos seus radicais (*stemming*), correções ortográficas e outros aspectos morfológicos e também sintáticos que as expressões textuais possuem.

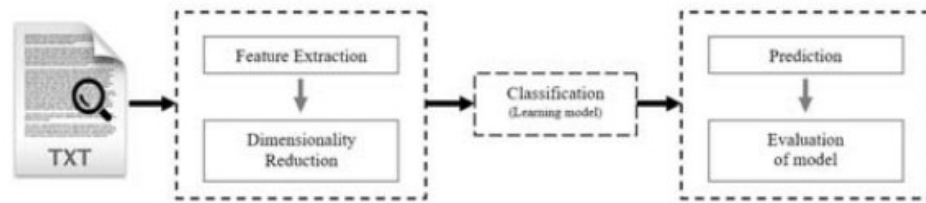
Figura 4 - Processo de descoberta do conhecimento textual



Fonte: Silva (2012)

Assim também para Chistol (2020), o processo de DCT consistem em: entendimento do negócio e compreensão de dados, coleta de dados, pré-processamento, mineração de textos e avaliação e interpretação de modelos. Tal processo é realizado em três etapas: extração de recursos, redução de dimensionalidade e classificação (Figura 5).

Figura 5 - Processo de descoberta do conhecimento textual



Fonte: Chistol (2020)

Segundo Chistol (2020), na etapa de extração de recursos o texto não estruturado deve ser pré-processado e transformado, de forma a possibilitar a aplicação de classificadores de modelagem matemática. A etapa da redução da dimensionalidade envolve a aplicação de algoritmos que procuram reduzir o tempo de processamento e a memória utilizada. E, finalmente, a etapa de classificação, considerada a mais importante, requer um bom entendimento de cada ferramenta escolhida, porque a eficiência do modelo depende do método escolhido.

Na prática, a DCT é centrada na mineração de textos e envolve técnicas como a recuperação de informação, análise textual, extração da informação, clusterização, categorização, visualização e tecnologias de mineração de base de dados (MORAIS; AMBRÓSIO, 2007). A mineração de textos é uma nova abordagem de descoberta de conhecimento por meio da busca por informação em dados não estruturados.

3 MINERAÇÃO DE TEXTOS: APLICAÇÃO, TÉCNICAS E PROCESSO

Este capítulo apresenta os conceitos e relações teóricas que fornecem subsídios para a compreensão da mineração de textos. São abordados os seguintes temas: a evolução da mineração de dados para a mineração de textos; a aplicação de mineração de textos na rede social *Twitter*; apresentação de diferentes técnicas que utilizam análise estatística para descobrir regras de associação e padrões sobre distribuição e associações de palavras-chave; e a apresentação das etapas que envolvem o processo de mineração de textos, desde a seleção dos documentos até a análise dos resultados.

Shrihari e Desai (2015) descrevem a mineração de dados como um processo de descoberta de informação potencial, útil e de padrão anteriormente desconhecido a partir de uma grande quantidade de dados. Para os autores, com o uso de um algoritmo apropriado, pode-se encontrar informações relevantes que levarão ao “descobrimento de conhecimento a partir de dados”.

A mineração de dados é uma ferramenta usada na análise de pesquisa de dados estruturados e é capaz de extrair conhecimento valioso de bancos de informações. Sua aplicação tem aumentado nas últimas décadas, impulsionada pelo crescimento da geração de dados capturados pelas Tecnologias de Informação (SHRIHARI, DESAI, 2015; OJO, AKINNULI, FARAYIBI, 2019; HASSANI *et al.*, 2020).

Consequentemente, surgiu a mineração de textos, uma intersecção de várias áreas de pesquisa, como a Recuperação de Informação, o Processamento de Linguagem Natural e a Mineração de Dados; e deu origem a uma nova disciplina científica e de engenharia, que procura a solução de problemas por meio da análise de dados não estruturados e não numéricos existentes em bancos de informações.

A mineração de textos é uma extensão da mineração de dados, e pode ser definida como um processo de extração de informações desconhecidas e úteis de documentos textuais escritos em linguagem natural. Como a maioria das informações são armazenadas em forma de texto, a mineração de textos possui alto valor comercial, e pode ser aplicada em áreas como medicina e atendimento ao cliente (PEZZINI, 2016, p.58).

O que diferencia a mineração de dados da mineração de textos é tipo de dados manipulados. Enquanto a mineração de dados lida com dados estruturados provenientes de sistemas, como bancos de dados e planilhas, a mineração de textos lida com dados não estruturados encontrados em documentos, *e-mails*, mídia social e na *Web*. Assim, a principal diferença entre a mineração de dados e texto é que na mineração de textos os padrões são

extraídos de texto em linguagem natural, em vez de bancos de dados estruturados (HASSANI *et al.*, 2020).

Para Morais e Ambrósio (2007,p.1):

A Mineração de textos (*text mining*) é um Processo de Descoberta de Conhecimento, que utiliza técnicas de análise e extração de dados a partir de textos, frases ou apenas palavras. Envolve a aplicação de algoritmos computacionais que processam textos e identificam informações úteis e implícitas, que normalmente não poderiam ser recuperadas utilizando métodos tradicionais de consulta, pois a informação contida nestes textos não pode ser obtida de forma direta, uma vez que, em geral, estão armazenadas em formato não estruturado.

Assim, também para Serapião *et al.* (2010, p.102):

A mineração de textos (*text mining*) é uma técnica adequada para manipulação automática de grandes volumes de dados, pertencente ao campo da ciência da computação, cientificamente ligada ao desenvolvimento de ferramentas de recuperação automática da informação. O método básico consiste em explorar e identificar termos relevantes em um grupo textual ou documental, bem como estabelecer padrões textuais e desenvolver grupos temáticos de assuntos pela frequência de aparecimento de termos no domínio a ser analisado. Com base no resultado da mineração de texto, é possível identificar com segurança os termos que fazem parte de um determinado conjunto de relatórios.

Contudo, a mineração de textos não deve ser confundida com um mecanismo de busca, onde o usuário já sabe o que pretende procurar. Trata-se de uma ferramenta que auxilia o usuário a descobrir informações relevantes e importantes que já possui, mas que são desconhecidas para o mesmo. É uma forma proativa de DCBD.

Ao utilizar os recursos de mineração de textos, um usuário não solicita exatamente uma busca, mas sim uma análise de um documento. Entretanto, este não recupera o conhecimento em si. É importante que o resultado da consulta seja analisado e contextualizado para posterior descoberta de conhecimento. [...] Na prática, a mineração de textos define um processo que auxilia na descoberta de conhecimento inovador a partir de documentos textuais, que pode ser utilizado em diversas áreas do conhecimento. (MORAIS; AMBROSIO, 2007, p. 6).

Os sistemas de mineração de textos utilizam técnicas de processamento de linguagem natural (PLN), cujo objetivo é promover a compreensão da linguagem natural através do uso de recursos computacionais para processamento de texto. Trata-se de um mecanismo criado não somente para extrair as informações de textos, mas também para facilitar a entrada de dados nos sistemas e a estruturação desses dados (SANTOS, R. E. S. *et al.*, 2014).

3.1 TWITTER E MINERAÇÃO DE TEXTOS

Atualmente, considerável atenção tem sido dada ao conteúdo gerado pelos usuários de *Internet*, pois uma grande quantidade de opiniões e emoções sobre produtos e serviços é publicada a cada dia, em todo o mundo. Consequentemente, empresas e organizações vêm se esforçando para “ouvir” seus usuários e procuram tomar decisões levando em conta esse conteúdo opinativo cada vez maior (ADWAN *et al.*, 2020; CARVALHO; PLASTINO, 2021).

Os dispositivos portáteis democratizaram a criação de conteúdo devido ao uso extensivo das mídias sociais e proporcionaram a geração de uma enorme quantidade de pequenos textos informais. A análise de conteúdo dos textos, aplicando os conceitos de PLN, permite identificar atitudes e posições expressas pelos internautas. Dessa forma, grandes redes sociais podem ser utilizadas para definir, por exemplo, uma política de comunicação eficiente entre o usuário e a organização.

No Brasil, as dez redes sociais mais utilizadas em 2021 (Resultados Digitais, in: <https://resultadosdigitais.com.br/blog/redes-sociais-mais-usadas-no-brasil/> acessada em 06.01.2022) foram:

- 1º. Facebook (130 milhões de usuários)
- 2º. YouTube (127 milhões de usuários)
- 3º. WhatsApp (120 milhões de usuários)
- 4º. Instagram (110 milhões de usuários)
- 5º. Facebook Messenger (77 milhões de usuários)
- 6º. LinkedIn (51 milhões de usuários)
- 7º. Pinterest (46 milhões de usuários)
- 8º. Twitter (17 milhões de usuários)
- 9º. TikTok (16 milhões de usuários)
- 10º. Snapchat (8,8 milhões de usuários)

Fazendo uma comparação, observa-se que, de acordo com dados do INEP, ano 2018, o número total de alunos em cursos superiores foi de 8,45 milhões. Desses, cerca de 68% estavam distribuídos em doze cursos (Tabela 1 - obtida em <https://abres.org.br/estatisticas/> acesso 06.01.22). Assim, o número de adeptos do *Twitter* é quase o dobro da população brasileira matriculada no ensino superior.

Tabela 1 - Número de alunos ensino superior por curso de graduação - ano 2018

Curso	Matriculados			Porcentagem Total
	Presencial	EAD	Total	
Engenharia	1.073.782	96.878	1.170.660	13,85%
Administração	600.037	525.247	1.125.284	13,32%
Direito	862.972	129	863.101	10,21%
Pedagogia	269.787	478.103	747.890	8,85%
Ciências Contábeis	227.439	132.401	359.840	4,26%
Enfermagem	291.602	21.635	313.237	3,71%
Psicologia	260.725	0	260.725	3,09%
Comunicação Social	60.440	1.213	61.653	0,73%
Letras	119.763	70.904	190.667	2,26%
Educação Física	131.787	58.361	190.148	2,25%
Ciências Biológicas	119.211	610	119.821	1,42%
Tecnologias da Informação	252.976	81.020	333.996	3,95%
Total dos doze	4.270.521	1.466.501	5.737.022	67,89%
Total Brasil	6.394.244	2.056.511	8.450.755	100,00%

Fonte: INEP (2018)

O *Twitter* trabalha com pequenos textos informais, os *tweets*, que são postados pelos usuários e que expõem pensamentos, interesses e opiniões em uma variedade de contextos e domínios (ADWAN *et al.*, 2020).

Em termos de Gestão do Conhecimento, a análise de *tweets* permite estudar a difusão de notícias em uma rede social levando à identificação de possíveis características de interesse (GUPTA, 2021). A análise desses textos é desafiadora, dado que os mesmos costumam ser curtos, informais, com ruídos (presença de erros de ortografia, descuido com a gramática, linguagem popular, abreviações, entre outros) e ricos em ambiguidades de linguagem como costuma ser a linguagem dos jovens universitários (NASEEM *et al.*, 2020).

Entre as redes sociais, o *Twitter* possui o maior número de ferramentas para análise de dados sociais. Seus dados são ideais para estudos, pois contêm informações espaciais e temporais e podem ser facilmente usadas por um usuário em movimento em tempo real, além de serem de fácil manipulação e acesso rápido. Contudo, os dados do *Twitter* exigem uma grande capacidade de armazenamento, preferencialmente usando uma ferramenta de *big data* (MARTÍN; JULIÁN; COS-GAYÓN, 2019).

3.2 TÉCNICAS DE MINERAÇÃO DE TEXTOS

Segundo Fayyad *et al.* (1996), a mineração tem dois objetivos primários: a previsão e a descrição, que podem ser alcançados usando uma variedade de técnicas específicas de mineração. Embora o limite entre eles não seja nítido, enquanto a previsão envolve o uso de

variáveis ou campos no banco de dados para prever valores desconhecidos ou futuros de outras variáveis de interesse; a descrição se concentra em encontrar padrões interpretáveis pelos usuários que descrevem os dados.

As técnicas de mineração usam análise estatística para descobrir regras de associação e padrões interessantes sobre distribuições e associações de palavras-chave.

Algumas das principais técnicas de mineração de textos são:

- a) **Categorização ou classificação.** A categorização é semelhante à classificação de texto. O classificador de texto é usado para categorizar palavras e predefinir suas classes. Um processo típico de categorização de texto consiste em pré-processamento, indexação, reduções de dimensões e classificação (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996; LAROSE, 2005; SHRIHARI R; DESAI, 2015; FERREIRA, M. H. W.; CORREA, 2021).
- b) **Clustering.** Técnica de agrupamento usada para agrupar documentos semelhantes. Este método é baseado no conceito de divisão de texto semelhante no mesmo cluster. Cada cluster contém vários documentos semelhantes (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996; LAROSE, 2005; SHRIHARI R; DESAI, 2015; FERREIRA, M. H. W.; CORREA, 2021).
- c) **Extração de informações.** Etapa principal de análise do texto não estruturado e seu relacionamento. O processo é feito por correspondência de padrões, usada para procurar uma sequência predefinida de texto. Esta técnica é muito útil para documentos de texto grandes (MACHADO *et al.*, 2010; SHRIHARI R; DESAI, 2015; FERREIRA, M. H. W.; CORREA, 2021).
- d) **Recuperação da informação.** Utiliza métodos e medidas estatísticas ou semânticas para automaticamente processar o texto de documentos para encontrar quais documentos possuem a resposta para a questão (mas não a resposta em si) (LOH; WIVES; OLIVEIRA, 2000b; MACHADO *et al.*, 2010).
- e) **Regressão.** A regressão consiste na realização de uma análise estatística com o objetivo de verificar a existência de uma variável dependente com uma ou mais variáveis independentes, ou seja, estimar o valor de uma variável analisando os valores das demais (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996; LAROSE, 2005; FERREIRA, M. H. W.; CORREA, 2021).
- f) **Sumarização.** Resume os dados sem alterar o significado do conteúdo e o comprimento dos dados. Portanto, todo o conjunto de documentos é substituído pelo de resumo. O resumo é útil para o usuário ler documentos curtos em vez de longos

(FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996; SHRIHARI R; DESAI, 2015; FERREIRA, M. H. W.; CORREA, 2021).

- g) **Visualização.** Grupo de documentos ou um documento único destacado, usado para mostrar o documento e a cor usada. Este método fornece informações mais compreensíveis, que ajudam a descobrir ou explorar o padrão da coleção de documentos. Ele usa cores diferentes e distância de relacionamento (SHRIHARI R; DESAI, 2015).

3.3 PROCESSO DE MINERAÇÃO DE TEXTOS

Conforme Moraes e Ambrósio (2007), o processo de mineração de textos passa pelas seguintes etapas: seleção de documentos, preparação dos dados, indexação e normalização, cálculo da relevância dos termos, mineração e análise de resultados (Figura 6).

Figura 6 - Processo de mineração de textos



Fonte: Autora (2021) - Adaptado de Moraes e Ambrósio (2007)

3.3.1 Seleção de documentos

Definição do conjunto de textos ou documentos a ser analisado e coleta dos dados que serão analisados.

3.3.2 Preparação dos dados

Preparação dos dados que constituem a base de textos de interesse e o trabalho inicial para tentar selecionar o núcleo que melhor expressa o conteúdo desses textos, eliminação de ruídos.

3.3.3 Indexação e normalização

Identificação de características de um documento e de similaridade de significado entre suas palavras utilizando uma estrutura denominada índice. Envolve as fases:

- a) Identificação de termos. Objetivo principal é a identificação dos termos contidos no texto, por meio da utilização de um dicionário ou aplicação de um analisador léxico.
- b) Remoção de palavras de parada (*stopwords*). As palavras de parada são palavras utilizadas para dar sentido ao texto, mas que podem ser consideradas irrelevantes para o conjunto de resultados a ser exibido em uma busca realizada em ferramentas de processamento de linguagem natural. Portanto, remover palavras de parada reduz o ruído dos dados analisados e contribui para aumentar a rapidez de operação (WIVES, 1999).
- c) Normalização morfológica (*stemming*). Consiste na redução das palavras flexionadas, por meio da retirada de afixos, permanecendo apenas a sua raiz ou radical. Seu objetivo é dar uma forma neutra às palavras e facilitar a classificação das mesmas. Segundo Chaves (2004, p. 2), um *stem* “captura uma palavra com generalidade suficiente para permitir um sucesso na combinação de caracteres, mas sem perder muito detalhe e precisão”. Duas situações possíveis no processo de *stemming* são *overstemming*, remoção de caracteres que fazem parte da palavra raiz, e *understemming*, não remoção de caracteres que fazem parte do afixo da palavra.

3.3.4 Cálculo da relevância dos termos

A frequência do termo (FT) ou *term frequency* (TF) representa a medida da quantidade de vezes que um termo aparece em um documento. O cálculo de relevância de um termo em relação ao texto está geralmente baseado na FT, na análise estrutural do documento ou na sua posição sintática. A FT pode ser classificada como:

- a) Absoluta. Quantidade de vezes que o termo aparece em um documento. Não leva em consideração a quantidade de termos existentes em um documento.

- b) Relativa. Leva em consideração o tamanho do documento e normaliza os pesos de acordo com essa informação.
- c) Frequência inversa de documentos. Com base na informação da frequência absoluta (quantidade de vezes que o termo aparece em um documento) e da frequência de documento (quantidade de documentos em que o termo aparece) temos a frequência inversa de documentos, utilizada para aumentar a importância de termos que aparecem em poucos documentos e diminuir a importância de termos que aparecem em muitos documentos.

3.3.5 Aplicação das técnicas de mineração de textos

Aplicação das técnicas de mineração de textos, análise de um conjunto de documentos a fim de descobrir coocorrências estatísticas de palavras que aparecem juntas, as quais fornecem *insights* sobre os tópicos dessas palavras e documentos.

3.3.6 Análise e interpretação de resultados

Análise dos resultados do processo de mineração de textos. Esta análise pode ser realizada com base em bibliometria, uma subárea da biblioteconomia encarregada de estudar e aplicar métodos matemáticos e estatísticos em documentos.

Portanto, por meio do processo de mineração de textos percebe-se que, devido a sua complexidade, a linguagem natural demanda coleta, tratamento e análise adequada para que se convertam textos em informações úteis e relevantes. Assim, diversas técnicas foram desenvolvidas utilizando análise estatística com o objetivo de descobrir regras de associação e padrões interessantes sobre distribuições e associações de palavras-chave e que possam revelar conhecimento existente, mas desconhecido.

4 FERRAMENTAS DE MINERAÇÃO DE TEXTOS

A mineração de textos envolve diferentes técnicas que exploram e identificam termos relevantes em um grupo textual ou documental. Por meio de padrões textuais, desenvolvem grupos temáticos de assuntos pela frequência de aparecimento de termos no domínio a ser analisado.

As ferramentas de mineração de textos utilizam recuperação de informação, análise textual, extração da informação, clusterização, categorização, visualização e tecnologias de mineração de base de dados. Atualmente, várias ferramentas encontram-se disponíveis no mercado.

Em sua pesquisa, Monteiro (2017) apresenta um comparativo entre diferentes ferramentas de mineração de textos. Para tanto, a autora pesquisou as ferramentas utilizadas em trabalhos recentes e aplicou critérios de avaliação baseados no método Reeves. O método de Thomas Reeves tem por objetivo avaliar a qualidade de software educacional por meio de critérios pedagógicos e de uso de interface.

Como primeira etapa de seleção, Monteiro (2017) utilizou os seguintes critérios de seleção:

- a) Ferramenta gratuita;
- b) Para mineração de textos no idioma português;
- c) Disponível para download ou execução online;
- d) Não oriunda de empresa comercial.

Assim, sete ferramentas foram selecionadas para a pesquisa:

- KH Coder,
- NLTK – Natural Language Toolkit,
- PyPLN,
- Sobek Mining,
- TagCrowd,
- TextAlyser e
- Wordcounter.

Contudo, a avaliação não pode ser aplicada nos softwares NLTK – Natural Language Toolkit e PyPLN, pois a utilização das ferramentas exigia conhecimentos prévios de programação e por este motivo não foram testadas.

Com relação às ferramentas selecionadas, apresentam-se as seguintes descrições de forma concisa:

- a) **KH Coder**. Ferramenta livre para análise quantitativa de conteúdo ou mineração de textos, utilizada para a linguística computacional. É possível analisar diversos idiomas, inclusive o Português. Oferece várias técnicas de pesquisa e análise estatística, como: Lista de Frequência, KWIC Concordância, Collocation Stats, Análise de Correspondência, Escalonamento Multi-Dimensional, Rede de Coocorrência, Análise de Agrupamento Hierárquico, Clustering e Classificador Naive Bayes.
- b) **Sobek Mining**. Ferramenta desenvolvida no ano de 2007, pela Universidade Federal do Rio Grande do Sul (UFRGS) no Grupo de Pesquisa GTech.Edu. Identifica os termos relevantes em um texto a partir da análise de frequência destes termos no material textual. Foi criada inicialmente como uma ferramenta de mineração de textos para auxiliar os professores do ensino a distância a avaliarem o trabalho dos alunos.
- c) **TagCrowd**. Ferramenta para aplicação web que permite a visualização de frequência de palavras em qualquer texto criando uma nuvem de palavra, nuvem texto ou nuvem de *tags*. As nuvens de *tags* reúnem um conjunto de palavras mais frequentes num determinado texto, dispostas em ordem alfabética e o tamanho da fonte da palavra exibida na nuvem é proporcional à frequência da palavra no texto. A ferramenta não busca encontrar relações entre as palavras. TagCrowd é especializada na criação de nuvens de palavras que sejam fáceis de ler, analisar e comparar. Foi criada em julho de 2006, por um estudante de doutorado em Design e Educação na Universidade de Stanford.
- d) **TextAlyser**. Ferramenta de análise de texto online, exibe estatísticas detalhadas e serve para descobrir o assunto de um texto.

A ferramenta faz uma contagem dos termos utilizados no texto apontando o número total de palavras e apresentando uma série de estatísticas sobre palavras e termos mais frequentes. Analisa a complexidade e capacidade de leitura de qualquer texto ou website. O programa aponta também a frequência com que as palavras mais utilizadas ocorrem no texto, bem como número de palavras, número de sílabas, dentre outros. Além destes fatores, a ferramenta ainda apresenta um índice relativo à “facilidade de leitura” (readability), critério obtido a partir do tamanho das frases e estatísticas encontradas. O programa não apresenta nenhuma ferramenta gráfica para visualização das principais informações contidas no texto. (KLEMANN et al., 2011, p. 1101 apud Monteiro, 2017).

- e) **WordCounter**. Ferramenta online gratuita que apresenta a relação das palavras mais utilizadas em um texto em uma lista. Não disponibiliza ferramentas gráficas mais complexas para visualização dos resultados.

Na segunda etapa, as cinco ferramentas selecionadas foram avaliadas segundo critérios pré-definidos de uso de interface. Os resultados são apresentados no Quadro 4.

Quadro 4 - Avaliação das ferramentas de mineração de textos

CRITÉRIOS DE AVALIAÇÃO	KH Coder	SOBEK MINING	TAGCROWD	TEXTALYZER	WORDCOUNTER	Condições desejáveis para esta pesquisa
Instalação (Local / Online)	L	L	O	O	O	L/O
Possui documentação sobre a ferramenta (Sim / Não)	S	S	N	N	N	S
Facilidade de operação (Fácil / Médio / Difícil)	M	F	F	F	F	F/M
Processamento do teste executou satisfatoriamente (Sim / Parcialmente / Não)	P	S	S	S	S	S
Linguagem da Interface (Inglês / Espanhol / Português)	I/E	P	I	I	I	I/E/P
Entrada de dados (Manual / Importação Arquivo)	IA	IA	M/IA	M/IA	M	IA
Saída de dados (Tela / Exportação Arquivo)	EA	T / EA	T / EA	T	T	EA
Análise do texto no idioma Português (Sim / Não)	S	S	S	S	S	S
Permite lista de <i>stopwords</i> (Sim / Não)	S	S	S	S	N	S
Permite lista de <i>stopwords</i> em Português (Sim / Não)	S	S	S	N	N	S
Entrada da lista de <i>stopwords</i> (Manual / Importação de Arquivo / Não Permite Lista)	M	IA	M	M	NPL	M/IA
Usa <i>stemming</i> ou lematização (Sim / Não)	S	N	N	N	N	S
Considera palavras acentuadas (Sim / Não)	S	S	S	S	S	S
Exibe número de frequência das palavras (Sim / Não)	S	S	S	S	S	S
Permite visualização gráfica do resultado (Sim / Não)	S	S	S	N	N	S
Permite visualização gráfica dos relacionamentos entre palavras (Sim / Não)	S	S	N	N	N	S
Existe integração com outras ferramentas (Sim / Não)	N	N	N	N	N	S/N
Exige conhecimento prévio de programação computacional (Sim / Não)	N	N	N	N	N	N

Fonte: Autora (2021) - Adaptado de Monteiro (2017)

4.1 SELEÇÃO DA FERRAMENTA DE MINERAÇÃO DE TEXTO

Para a seleção da ferramenta de mineração de texto, considerou-se a análise e avaliação de Monteiro (2017), pois os critérios aplicados pela autora foram ponderados como suficientes para a escolha de uma ferramenta de mineração e atendem aos objetivos desta pesquisa.

Assim, uma coluna foi adicionada ao Quadro 4, para comparação entre os dados das ferramentas apresentadas e as condições desejáveis para a ferramenta desta pesquisa. Buscando a ferramenta que mais atende aos critérios desejáveis, destacaram-se duas ferramentas: KH Coder e Sobek Mining. Então, foi realizada uma comparação considerando as técnicas de cada ferramenta.

Enquanto a principal técnica do Sobek Mining é a construção de gráficos dos termos frequentes de um texto, o KH Coder apresenta uma ampla variedade de técnicas: lista de frequência de palavras, concordância de palavras-chave em contexto, associação de palavras, rede de coocorrência de palavras, análise de cluster, análise hierárquica de cluster, mapa de auto-organização e classificador Naive Bayes.

Uma vez que todas as informações escritas ou faladas podem ser representadas na forma textual, a mineração de dados requer todos os tipos de ferramentas de mineração de texto quando se trata de interpretação e análise de sentenças, palavras, frases, discursos, declarações, anúncios e declarações (HASSANI *et al.*, 2020, p. 2).

Levando em consideração a importância de diferentes representações para a análise e interpretação dos dados, a ferramenta de mineração de textos selecionada para esta pesquisa foi o **KH Coder**. A ferramenta é gratuita, possibilita a mineração de textos no idioma Português, está disponível para download, não é oriunda de empresa comercial, não exige conhecimento prévio de programação e proporciona o maior número de técnicas para análise de mineração textual.

4.2 SOFTWARE KH CODER

O KH Coder é um software de código aberto utilizado para mineração de textos, que pode ser modificado e atualizado por usuários. Utiliza armazenamento em bancos de dados (MySQL) que podem ser recuperados diretamente, permitindo pesquisas de dados flexíveis. Suporta *Windows, Linux e Macintosh*. Uma vez instalado, as especificações e operações dos métodos de processamento são os mesmos, independentemente do sistema operacional usado.

Todo *software* necessário para trabalhar com KH Coder é fornecido na instalação do pacote para *Windows*, como dicionário, MySQL, Perl e R. O pacote realiza todas as configurações relevantes automaticamente.

Possui dicionário em chinês (simplificado), francês, alemão, italiano, coreano, português, russo e espanhol. Para idiomas de destino em inglês ou qualquer outro idioma da Europa Ocidental, é necessário introduzir as *stopwords*.

O KH Coder mostra que palavras estão fortemente conectadas encontrando as palavras que frequentemente aparecem juntas em um determinado caso.

Ao extrair as palavras flexionadas, como verbos ou adjetivos, o KH Coder converte essas palavras as suas formas base ou raiz (*stemming*) e, em seguida, realiza sua pesquisa. Esse processo tem por objetivo reduzir a carga necessária para contar a frequência de palavras, determinando a relação entre palavras e criando regras de codificação. O KH Coder usa o *Tagger* POS padrão para lematização e a *Snowball Stemmer* para *stemming*.

O KH Coder delimita um parágrafo com uma nova linha e uma frase com um ponto por padrão. Um valor numérico entre parênteses indica o número de palavras que o KH Coder reconhece como alvos de análise. Sem alterar nenhuma configuração, palavras comuns, incluindo artigos definidos ou indefinidos e verbos auxiliares, que existem em quaisquer documentos, são excluídos da análise.

Também é possível excluir palavras da análise através da técnica FORCE TO IGNORE, ou forçar sua inclusão através da técnica FORCE TO PICK UP. Além disso, classes gramaticais não relevantes podem ser excluídas da análise pelo usuário.

4.2.1 Técnicas de Mineração de Textos do KH Coder

Esta seção tem por objetivo levantar as técnicas de análise de informação não estruturada da ferramenta de mineração selecionada e descrever suas aplicabilidades. O KH Coder possui várias técnicas para análise dos textos, descritas a seguir.

4.2.1.1 WFL – Words Frequency List – Lista de Frequência das Palavras

Através da técnica *Words Frequency List*, as palavras encontradas são contabilizadas e relacionadas em uma lista. O número de vezes que cada palavra aparece é chamado de frequência de ocorrência. Conforme Brysbaert et al. (2018, p. 45), “quando o reconhecimento

de palavras é analisado, a frequência de ocorrência é um dos mais fortes preditores de eficiência de processamento”.

Esta técnica permite visualizar a lista de palavras extraídas pelo KH Coder, bem como, a classificação gramatical e a frequência encontrada. Além disso, permite extração para uma planilha Excel com lista das palavras por classificação em ordem decrescente de suas frequências.

Por meio desta técnica, encontra-se duas estatísticas descritivas possíveis:

- TF – *Term Frequency* – Frequência de Termo. Número de ocorrências de cada palavra na totalidade dos dados (Quadro 5).
- DF – *Document Frequency* – Frequência de Documento. Número de documentos em que cada palavra é encontrada (Quadro 6).

Quadro 5 - Frequência de Termo (exemplo)

The screenshot shows the 'Term Frequency Distribution' window. Under 'Descriptives', it lists: Types of Words (n) 1675, Mean of TF 8.25, and Std. Deviation of TF 27.02. Below is a 'Frequency Table' with columns: TF, Frequency, Percent, Cumulative Frequency, and Cumulative Percent. The table contains 10 rows of data.

TF	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	711	42.45	711	42.45
2	261	15.58	972	58.03
3	139	8.30	1111	66.33
4	97	5.79	1208	72.12
5	79	4.72	1287	76.84
6	58	3.46	1345	80.30
7	40	2.39	1385	82.69
8	25	1.49	1410	84.18
9	19	1.13	1429	85.31
10	13	0.78	1442	86.09

Fonte: Autora (2021) - KH Coder

Quadro 6 - Frequência de Documento (exemplo)

The screenshot shows the 'Document Frequency Distribution' window. Under 'Descriptives', it lists: Types of Words (n) 1675, Mean of DF 6.56, and Std. Deviation of DF 18.20. Below is a 'Frequency Table' with columns: DF, Frequency, Percent, Cumulative Frequency, and Cumulative Percent. The table contains 10 rows of data.

DF	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	751	44.84	751	44.84
2	253	15.10	1004	59.94
3	144	8.60	1148	68.54
4	95	5.67	1243	74.21
5	72	4.30	1315	78.51
6	58	3.46	1373	81.97
7	32	1.91	1405	83.88
8	25	1.49	1430	85.37
9	22	1.31	1452	86.69
10	15	0.90	1467	87.58

Fonte: Autora (2021) - KH Coder

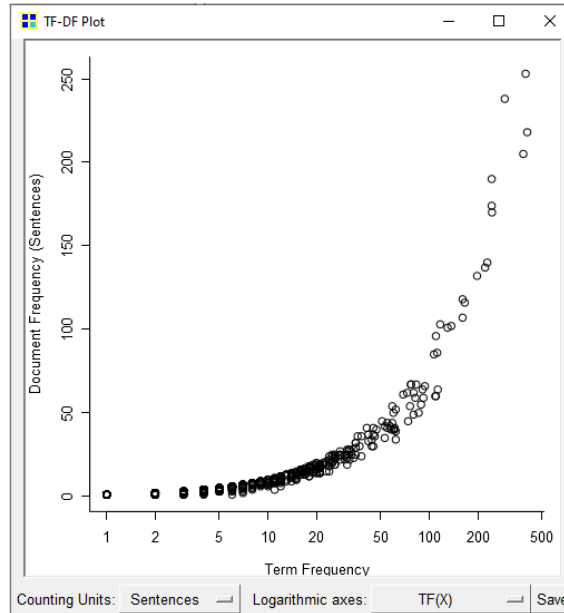
A TF tabula apenas palavras-alvo para análise, ou seja, não considera *stopwords*, palavras relacionadas no FORCE TO IGNORE ou classes de palavras excluídas pelo usuário.

Portanto, esta técnica é útil para identificar:

- Quantos tipos de palavras serão analisados;
- TF médio dessas palavras; e
- TF mínimo de palavras que deve ser incluído no escopo de análise.

Com as informações de TF e DF, é possível gerar um Gráfico TF-DF (Quadro 7), com o objetivo de avaliar a correlação entre as frequências. Normalmente, há forte correlação entre esses valores, ou seja, quanto maior o DF, maior o FT.

Quadro 7 - Gráfico TF-DF (exemplo)



Fonte: Autora (2021) - KH Coder

4.2.1.2 KWIC Concordance – Concordância de Palavras-Chave em Contexto

Por meio da técnica *KWIC Concordance* é possível visualizar as sentenças onde determinada palavra é encontrada, permitindo analisar como ela é usada no documento de destino.

Quadro 8 - Concordância de Palavras-Chave em Contexto (exemplo)

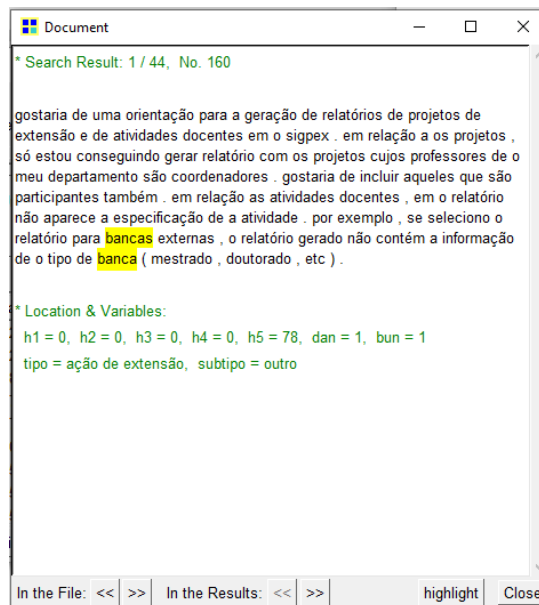
Fonte: Autora (2021) - KH Coder

A técnica permite restringir os resultados da pesquisa de concordância adicionando opções para indicar quando certas palavras aparecem imediatamente antes ou depois da palavra pesquisada.

No exemplo (Quadro 8), a informação HITS traz o número de vezes que a palavra pesquisada foi encontrada, neste caso, a palavra BANCA foi encontrada 44 vezes.

Clicando na linha 1, é possível ler o documento de destino (Quadro 9), onde a palavra pesquisada é destacada automaticamente.

Quadro 9 - Arquivo destino, destaque feito pelo KH Coder (exemplo)



Fonte: Autora (2021) - KH Coder

Estatística descritiva possível:

- *Collocation Stats* – Estatísticas de Colocação. Identifica quais palavras aparecem frequentemente antes e depois da palavra alvo (ou palavra do nó), relacionando palavras que tenham uma forte relação com a palavra pesquisada.

Utilizando o mesmo exemplo, clicando em STATS (Quadro 8), é possível gerar as estatísticas de colocação (Quadro 10), onde encontra-se uma lista palavras relacionadas a palavra pesquisada, também chamada PALAVRA NÓ.

Quadro 10 - Estatísticas de Colocação (exemplo)

The screenshot shows a window titled 'Collocation Stats'. At the top, there are input fields for 'Node Word' (set to 'banca'), 'POS' (set to 'N'), and 'Conj.' (empty). It also shows 'Hits: 44'. Below this is a 'Result' section containing a table with 17 columns: N, Word, POS, Total, LT, RT, L5, L4, L3, L2, L1, R1, R2, R3, R4, R5, and The Score. The table lists 10 results, with the first result 'externo' having a score of 17.333. The table is sorted by 'The Score' in descending order. At the bottom, there are controls for 'Copy', 'Filter', 'Sort' (set to 'The Score'), and 'Window span' (set to 'L5 - R5').

N	Word	POS	Total	LT	RT	L5	L4	L3	L2	L1	R1	R2	R3	R4	R5	The Score
1	externo	AQ	18	0	18	0	0	0	0	0	17	0	1	0	0	17.333
2	participação	N	13	12	1	2	1	1	8	0	0	0	0	1	0	5.233
3	avaliação	N	5	1	4	1	0	0	0	0	0	3	1	0	0	2.033
4	atividade	N	6	5	1	3	0	1	1	0	0	0	0	0	1	1.633
5	exemplo	N	4	3	1	1	0	0	2	0	0	0	1	0	0	1.533
6	tcc	N	3	0	3	0	0	0	0	0	0	3	0	0	0	1.500
7	mestrado	N	4	1	3	1	0	0	0	0	0	1	0	2	0	1.200
8	tese	N	3	0	3	0	0	0	0	0	0	2	0	0	1	1.200
9	palestra	N	3	1	2	0	0	0	1	0	0	0	2	0	0	1.167
10	concurso	N	3	0	3	0	0	0	0	0	0	1	1	0	1	1.033

Fonte: Autora (2021) - KH Coder

A lista é apresentada em ordem decrescente de frequência das palavras e contém:

- Palavra (WORD)
- Classificação (POS – Part of Speech)
- Frequência total (TOTAL)
- Frequência nas posições a esquerda (L – LEFT) da palavra nó (LT = L1 a L5)
- Frequência nas posições a direita (R – RIGHT) da palavra nó (RT = R1 a R5)
- Maiores frequências em destaque na cor vermelha
- *Score* – pontuação

A pontuação (*score*) é calculada pela fórmula:

$$f(w) = \sum_{i=1}^5 \frac{(l_i + r_i)}{i} \quad (1)$$

Em geral, quanto maior a frequência com que uma determinada palavra w aparece antes ou depois da palavra do nó ($l_i + r_i$), maior será o valor $f(w)$.

No cálculo do valor $f(w)$, as frequências ($l_i + r_i$) são divididas por “ i ”, que pesa as frequências de acordo com sua distância da palavra do nó. Assim, as palavras que aparecem mais próximas da palavra do nó (ou seja, com um “ i ” menor) têm peso maior do que aquelas que ocorrem cinco palavras antes ou depois da palavra do nó (HIGUCHI, 2017).

4.2.1.3 Word Association – Associação de Palavras

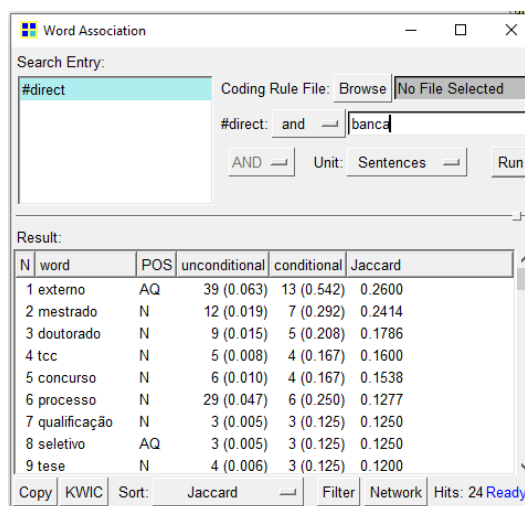
Esta técnica identifica as palavras que estão intimamente associadas a uma determinada palavra nó e indica o número de sentenças que atendem a condições especificadas (HITS).

A coluna “incondicional” indica o número de sentenças em que cada palavra aparece no arquivo de destino e a probabilidade de ocorrência em todas as sentenças (probabilidade incondicional).

A coluna “condicional” indica o número de sentenças em que cada palavra aparece e que satisfazem as condições especificadas, e a probabilidade de ocorrência em sentenças que satisfaçam as condições especificadas (probabilidade condicional).

A coluna rotulada “Jaccard” determina a ordem de exibição selecionada e mostra os valores usados para classificar os resultados. O coeficiente de Jaccard enfatiza se palavras específicas se relacionam ou não, independentemente de a palavra aparecer uma ou dez vezes em um documento. Assim, a associação de palavras é calculada independentemente da frequência de aparecimento. Mesmo quando muitos documentos não têm a palavra A nem a palavra B, o coeficiente de Jaccard entre A e B não aumenta (HIGUCHI, 2017).

Quadro 11 - Associação de Palavras (exemplo)



The screenshot shows the 'Word Association' application window. The search entry is '#direct'. The coding rule file is 'No File Selected'. The search criteria are '#direct: and banca'. The results table is as follows:

N	word	POS	unconditional	conditional	Jaccard
1	externo	AQ	39 (0.063)	13 (0.542)	0.2600
2	mestrado	N	12 (0.019)	7 (0.292)	0.2414
3	doutorado	N	9 (0.015)	5 (0.208)	0.1786
4	tcc	N	5 (0.008)	4 (0.167)	0.1600
5	concurso	N	6 (0.010)	4 (0.167)	0.1538
6	processo	N	29 (0.047)	6 (0.250)	0.1277
7	qualificação	N	3 (0.005)	3 (0.125)	0.1250
8	seletivo	AQ	3 (0.005)	3 (0.125)	0.1250
9	tese	N	4 (0.006)	3 (0.125)	0.1200

At the bottom of the window, there are buttons for 'Copy', 'KWIC', 'Sort: Jaccard', 'Filter', 'Network', and 'Hits: 24 Ready'.

Fonte: Autora (2021) - KH Coder

No exemplo, Quadro 11, a palavra BANCA é a palavra nó. A associação de palavras mostra que a palavra EXTERNO é encontrada em 39 sentenças, com uma probabilidade de 6,3% de ser encontrada em qualquer sentença (probabilidade incondicional). Porém, a probabilidade da palavra EXTERNO aparecer na mesma sentença que a palavra BANCA é de 54,2%, 13 sentenças (probabilidade condicional).

Esta técnica permite gerar uma rede de coocorrência de palavras selecionando a técnica NETWORK do quadro associação de palavras (Figura 7).

Figura 7 - Rede de Coocorrência (exemplo 1)



Fonte: Autora (2021) - KH Coder

4.2.1.4 *Co-occurrence Network – Rede de Coocorrência*

Técnica que exibe um gráfico da rede de coocorrência das palavras. Neste gráfico, palavras intimamente associadas umas às outras são conectadas por linhas. Permite agrupamento por palavras e por variáveis pré-definidas.

Conforme Robredo e Cunha (1998), a rede de coocorrência é uma forma de representação que permite identificar a maior ou a menor frequência dos elementos componentes em uma rede, representados por meio de círculos ou quadrados de tamanho proporcional aos valores das respectivas ocorrências (Figura 8). Além disso, linhas de enlace mais ou menos destacadas realçam a “força” de associação entre pares de termos.

Mediante a análise das coocorrências entre pares de palavras, é possível estabelecer índices estatísticos que representam a “força” de associação entre esses pares e, a partir dos valores encontrados, elaborar diversos tipos de representações gráficas (árvores, redes, agrupamentos diversos) e, assim, visualizar (ou, utilizando um anglicismo bem em voga, ‘mapear’) o estado de um campo do conhecimento, em um determinado momento (ROBREDO; CUNHA, 1998, p. 11).

Contudo, ao contrário dos resultados de escalonamento multidimensional, o fato das palavras estarem próximas umas das outras nem sempre significa que elas têm uma forte coocorrência. Embora as palavras estejam próximas, se não estiverem conectados por linhas (arestas), não há coocorrência forte identificada (HIGUCHI, 2017).

Higuchi (2017) explica que, para determinar a localização das palavras, este comando usa o método desenvolvido por Fruchterman e Reingold (1991) para desenhar uma rede de palavras e o método desenvolvido por Kamada e Kawai (1988) para desenhar uma rede de

palavras e variáveis. Esses métodos organizam as palavras de forma que a rede resultante é fácil de ser lida e interpretada (Figura 8).

Figura 8 - Rede de Coocorrência (exemplo 2)



Fonte: Autora (2021) - KH Coder

4.2.1.5 Cluster Analysis – Análise de Cluster

A análise de Cluster é um método exploratório multivariado amplamente utilizado em várias áreas, que tem por objetivo ajudar na identificação de padrões de decisão dinâmicos em organizações e empresas (LOPES; GOSLING, 2021).

Esta técnica classifica palavras, alocando-as em grupos internamente homogêneos, mas também em grupos heterogêneos. Portanto, sua lógica é reunir o que é semelhante separando o que é diferente. Para definir a **semelhança** ou a **diferença** entre as palavras, é utilizada uma técnica estatística que determina a **distância entre as palavras** de um cluster.

A análise de cluster é executada em nível de documento e os resultados são salvos como variáveis. Ou seja, identifica quais documentos contêm palavras semelhantes e agrupa-os.

Sua visualização pode ser feita por meio de dendrograma, gráfico da rede de coocorrência de documentos ou palavras e associação de documentos ou palavras.

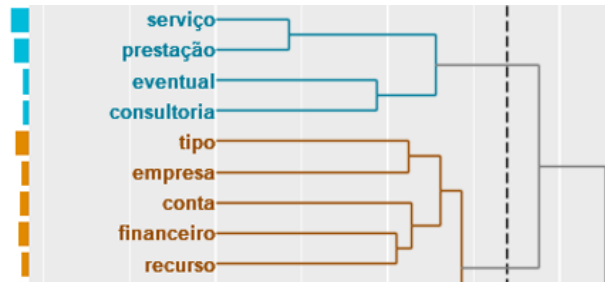
4.2.1.6 Hierarchical Cluster Analysis – Análise Hierárquica de Cluster

A análise hierárquica de cluster utiliza um algoritmo capaz de produzir uma representação hierárquica dos agrupamentos. Sua representação por meio de um dendrograma facilita a visualização da formação dos grupamentos em cada estágio e o grau de semelhança entre eles.

Esta técnica permite encontrar e analisar quais combinações ou grupos de palavras têm padrões de aparência semelhantes, para tanto, utiliza uma matriz de similaridade contendo as métricas de distância entre os agrupamentos em cada estágio do algoritmo (VALE, 2005).

No dendrograma (Figura 9), cada cluster é apresentado em uma cor diferente para indicar os agrupamentos. As barras do lado esquerdo do dendrograma indicam o TF de cada palavra. As linhas horizontais indicam a distância entre as palavras e as linhas verticais indicam as conexões entre as palavras e clusters.

Figura 9 - Dendrograma (exemplo)

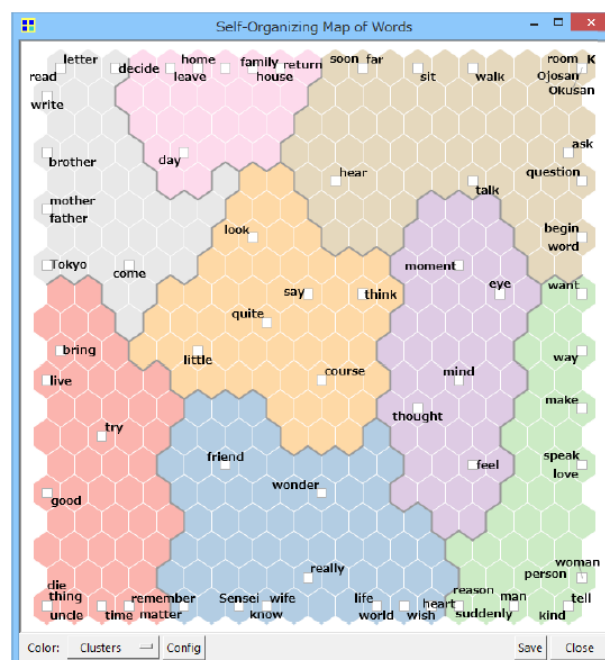


Fonte: Autora (2021) - KH Coder

4.2.1.7 Self-Organizing Map – Mapa Auto-organizável

O mapa auto-organizável é uma técnica que faz o agrupamento de dados de forma que palavras semelhantes pertençam ao mesmo grupo e palavras pouco semelhantes pertençam a grupos diferentes (Figura 10). É um algoritmo de aprendizado não supervisionado, que mapeia dados multidimensionais em subespaços dimensionais inferiores, onde as relações geométricas entre os pontos indicam sua similaridade. A redução na dimensionalidade permite a visualização e facilita a interpretação de dados complexos.

Figura 10 - Mapa Auto-Organizável (exemplo)



Fonte - Autora (2021) - KH Coder

Esta técnica permite uma representação multidimensional do agrupamento das palavras, ou seja, o mapa auto-organizável é capaz de aliar a análise de agrupamento com uma representação gráfica para visualização das distâncias.

Estes mapas são baseados na aprendizagem competitiva, onde os neurônios de saída da grade competem entre si para serem ativados. O neurônio que vence a competição é denominado de neurônio vencedor. [...] No modelo de Kohonen, os neurônios estão localizados em nós de uma grade, que geralmente possui uma ou duas dimensões. A grade procura estabelecer e preservar noções de vizinhança (preservação topológica). Durante o processo de aprendizado, é formado um mapa topográfico dos padrões de entradas. Neste mapa, as localizações espaciais dos neurônios nas grades são indicativas das características contidas nos padrões de entradas. Outra característica importante deste tipo de rede, é que elas utilizam treinamento não supervisionado. Neste tipo de treinamento, a rede busca encontrar similaridades baseando-se apenas nos padrões de entrada (ALVES; LEAL, 2021, p. 1960).

4.2.1.8 Naive Bayes – Classificador Naive Bayes

O classificador Naive Bayes é um algoritmo probabilístico de classificação que produz estimativas de probabilidade contando a frequência e combinações de valores nos dados históricos (BELLENZIER, 2013). O objetivo é calcular a probabilidade de um evento ocorrer, dada a probabilidade de outro evento que já ocorreu.

É chamado de classificador ingênuo, pois assume que os atributos são condicionalmente independentes uns dos outros, ou seja, a informação de um evento não é informativa sobre nenhum outro (ANDRADE, 2015).

De acordo com o teorema de Bayes, uma probabilidade condicional pode ser representada como:

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)} \quad (2)$$

Onde $p(x|y)$ é a probabilidade de ocorrência de x sob a condição y .

Em aplicações sobre textos, x é substituído por C e representa as categorias; y é substituído por W e representa as palavras.

Não é possível calcular diretamente $p(C|W)$, ou seja, a probabilidade de que um documento que contém uma série de **palavras W** pertencer à **categoria C**. No entanto, usando o teorema Naive Bayes, $p(C|W)$ pode ser representado como:

$$p(C|W) = \frac{p(W|C)p(C)}{p(W)} \quad (3)$$

Transformar a fórmula dessa maneira torna mais fácil estimar o valor de $p(C|W)$. O modelo Naive Bayes usado pelo KH Coder calcula esse valor para todas as categorias e classifica os documentos na categoria com o maior valor. Quando n tipos de palavras estão contidas, utiliza-se a forma simplificada:

$$p(W|C) = p(w_1|C)p(w_2|C) \cdots p(w_n|C) \quad (4)$$

Como o cálculo de um grande número de probabilidades resulta em um valor próximo à zero, aplica-se a seguinte propriedade logarítmica:

$$\log p(W|C)p(C) = \log p(w_1|C) + \log p(w_2|C) + \cdots + \log p(w_n|C) + \log p(C) \quad (5)$$

Os valores logarítmicos $\log p(w_i|C)$ podem ser determinados na fase de aprendizagem. Sua adição promove um cálculo mais rápido e preciso do que a multiplicação na fase de classificação.

Portanto, a classificação usando o Modelo *Naive Bayes* envolve as seguintes etapas:

- I. Na fase de aprendizagem, o valor de $\log p(w_i|C)$ é calculado a partir dos modelos de classificação.
- II. O $\log p(w_i|C)$ reflete a pontuação que deve ser adicionada à categoria C , quando a palavra i aparece uma vez no documento durante a classificação automática.
- III. Quanto mais frequentemente a palavra i aparece na categoria C nos exemplos de classificação manual, maior o valor de $\log p(w_i|C)$.

Este modelo é baseado em um conceito simples: **se uma palavra aparece muitas vezes na categoria C** nos exemplos de classificação do manual, então os documentos que contêm mais ocorrências dessa palavra são mais propensos a serem **classificados na categoria C também**. KH Coder usa um módulo Perl chamado “Algorithm Naive Bayes” criado por Ken Williams (HIGUCHI, 2017).

Neste capítulo, foram apresentadas oito técnicas de mineração de textos disponíveis por meio da ferramenta KH Coder. A aplicação e a combinação de diferentes técnicas permite a exploração de informação não estruturada e a descoberta de conhecimento.

5 APLICAÇÃO DE MINERAÇÃO DE TEXTOS NO PAI/PROEX

Para o PAI/PROEX, esta pesquisa adota a seguinte metodologia de aplicação de mineração de textos, adaptada do processo de Descoberta de Conhecimento Textual (SILVA, 2012) e do Processo de Mineração de Textos (MORAIS; AMBRÓSIO, 2007):

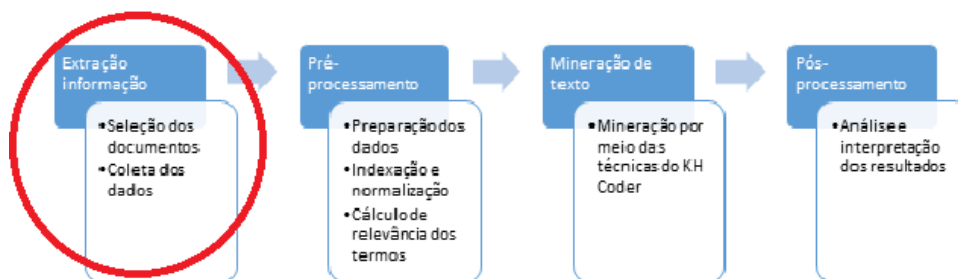
Figura 11 – Aplicação Mineração de Textos



Fonte: Autora (2021)

5.1 EXTRAÇÃO DA INFORMAÇÃO

Figura 12 - Extração da informação



Fonte: Autora (2021)

A etapa da extração da informação está dividida em seleção dos documentos e coleta de dados.

5.1.1 Seleção de documentos

Esta pesquisa utilizou as informações do Portal de Atendimento Institucional (PAI) da Pró-Reitoria de Extensão (PROEX) da Universidade Federal de Santa Catarina (UFSC) (Quadro 12).

Quadro 12 - Portal de Atendimento Institucional da Pró-Reitoria de Extensão



Fonte: Autora (2021) – (PAI - PROEX)

O PAI começou a ser utilizado pela PROEX em agosto de 2019. Portanto, foram analisados todos os chamados de atendimento registrados no período de 25/08/2019 a 31/07/2021. O acesso aos chamados foi liberado pela SETIC por meio de um relatório digital (Quadro 13).

Quadro 13 - Relatório dos chamados de atendimento PAI-PROEX

Fila	Cliente	link	name	Título	Criado	Modificado
PROEX::Professores	samira.mansur	http://ufsc.br/pai/2021092278000187	closed successful	SIGPEX	22/9/2021, 09:29	22/9/2021, 09:49
PROEX::Comunidade	paginas@systemas.ufsc.br	http://ufsc.br/pai/2021092178001179	closed successful	Escola de extensão	21/9/2021, 20:46	22/9/2021, 08:30
PROEX	paginas@systemas.ufsc.br	http://ufsc.br/pai/2021092178001188	duplicidade	Editais	21/9/2021, 20:47	22/9/2021, 08:26
PROEX::Professores	dfiogo.siebert	http://ufsc.br/pai/2021092178000992	closed successful	Editais	21/9/2021, 17:32	21/9/2021, 20:11
PROEX::TAE	andre.tiago	http://ufsc.br/pai/2021092178000965	closed successful	SIGPEX	21/9/2021, 17:16	21/9/2021, 17:52
PROEX::Professores	r.roesler	http://ufsc.br/pai/2021092178000652	closed successful	SIGPEX	21/9/2021, 13:15	21/9/2021, 16:26
PROEX::TAE	giullia.pimentel	http://ufsc.br/pai/2021092178000732	closed successful	Editais	21/9/2021, 14:19	21/9/2021, 15:18
PROEX::Professores	anelise.regiani	http://ufsc.br/pai/2021092178000803	closed successful	Outros	21/9/2021, 14:52	21/9/2021, 15:11
PROEX::Professores	poliana.bezerra	http://ufsc.br/pai/2021092178000492	closed successful	Outros	21/9/2021, 11:27	21/9/2021, 14:54
PROEX::Alunos	karine.jouille	http://ufsc.br/pai/2021092078000332	closed successful	SIGPEX	20/9/2021, 09:53	21/9/2021, 12:42
PROEX::Professores	poliana.bezerra	http://ufsc.br/pai/2021092178000509	duplicidade	Outros	21/9/2021, 11:27	21/9/2021, 12:41

Fonte: Autora (2021)

Ao registrar um atendimento, o usuário seleciona um tipo de serviço. Atendendo ao objetivo desta pesquisa, foram analisados somente os chamados com o **serviço SIGPEX**. Dentre os chamados de serviço SIGPEX, foram excluídos da coleta de dados os chamados com os seguintes assuntos:

- 1) reclamações de erro de sistema;
- 2) pedidos de troca de coordenação;
- 3) indicação de coordenadores de extensão (portarias);
- 4) pedidos de isenção de ressarcimentos;
- 5) pedidos de troca de situação;

- 6) dúvidas sobre emissão de certificados;
- 7) dúvidas sobre bolsas de extensão;
- 8) dúvidas sobre editais;
- 9) recebimento de propostas editais;
- 10) dúvidas sobre a escola de extensão;
- 11) solicitações de declarações;
- 12) dúvidas sobre o Curso SIGPEX de Coordenadores de Extensão.

Assim, foram selecionados 600 chamados, de um total de 2.417.

5.1.2 Coleta dos dados

Os dados dos chamados selecionados foram coletados e transferidos para uma planilha Excel (Quadro 14), contendo:

- 1) Tipo: ação de extensão ou atividade docente;
- 2) Subtipo: programa, projeto, curso, evento, prestação de serviço, curso de curta duração, publicação, banca externa, evento e palestra;
- 3) Ticket: número do chamado de atendimento;
- 4) Data do chamado de atendimento;
- 5) Pergunta;
- 6) Resposta.

Quadro 14 - Planilha Excel com dados extraídos do PAI

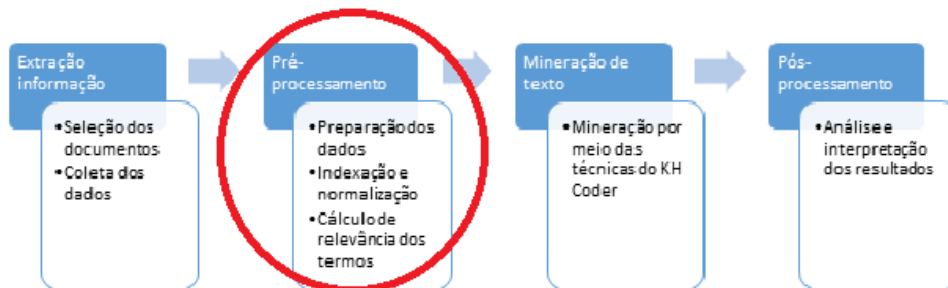
	A	B	D	E	F	G
11	tipo	subtipo	ticket	data	pergunta	resposta
12	ação de extensão	curso	2019120978000325	09/12/2019	A atividade de extensão número 2019122 Não é necessário solicitar a reabertura de	
13	ação de extensão	curso	2020012178000132	21/01/2020	O projeto de extensão cadastrado na plat Após devolvido para a conta única não há	
14	ação de extensão	curso	2020021278000111	12/02/2020	Estou em vias de sair para um pós-doutor As ações de extensão devem ser aprovad	
15	ação de extensão	curso	2020032778000092	27/03/2020	Ano passado fui coordenadora de uma aç Você pode sim fazer um curso e cobrar ta	
16	ação de extensão	curso	2020050778000689	07/05/2020	Ações de Extensão relacionadas a cursos Os campos indicados abaixo pela profess	
17	ação de extensão	curso	2020061278000248	12/06/2020	Prezados, possuo 3 cursos de extensão (n Como as ações não aconteceram, você de	
18	ação de extensão	curso	2020071778000282	17/07/2020	gostaria de saber porque não é possível c Conforme normativa do FORPROEX (Fóru	
19	ação de extensão	curso	2020072378000421	23/07/2020	No cadastramento de cursos de extensão A orientação é que o coordenador registr	
20	ação de extensão	curso	2020081078000631	10/08/2020	A dúvida é com relação a relatório final di Deve ser sempre comprovante de pagam	
21	ação de extensão	curso	2020090178001441	01/09/2020	Estou organizando com colegas argentina para registrar em ações de extensão/curs	
22	ação de extensão	curso	2020111078000661	10/11/2020	Venho por meio desse contato tirar uma O PAAD é de responsabilidade da PROGR	
23	ação de extensão	curso	2021020178002343	01/02/2021	Estou escrevendo uma proposta de curso Se a anuência desta carta deve partir de l	
24	ação de extensão	curso	2021020578000123	05/02/2021	A docente registrou um curso de extensã No SIGPEX a aba descrição apresenta os c	
25	ação de extensão	curso	2021032278001821	22/03/2021	Pretendemos oferecer um curso com dire Os certificados podem ser emitidos atrav	
26	ação de extensão	curso	2021042378000452	23/04/2021	Estou analisando um curso de extensão q Acho que vai depender do que está defir	
27	ação de extensão	curso	2021051078001312	10/05/2021	gostaria de saber se é possível fazer um c Conforme artigo 22 da RN 88/2016/CUn: §	
28	ação de extensão	curso	2021051178000428	11/05/2021	Sou docente da Pós em Linguística e, por Legislação da UFSC: A Resolução 88/CUn/	

Fonte: Autora (2021)

A classificação dos dados em tipo e subtipo seguiu o que determina a Resolução Normativa nº 88/2016/CUn, que dispõe sobre as normas que regulamentam as ações de extensão na Universidade Federal de Santa Catarina.

5.2 PRÉ-PROCESSAMENTO

Figura 13 - Pré-processamento



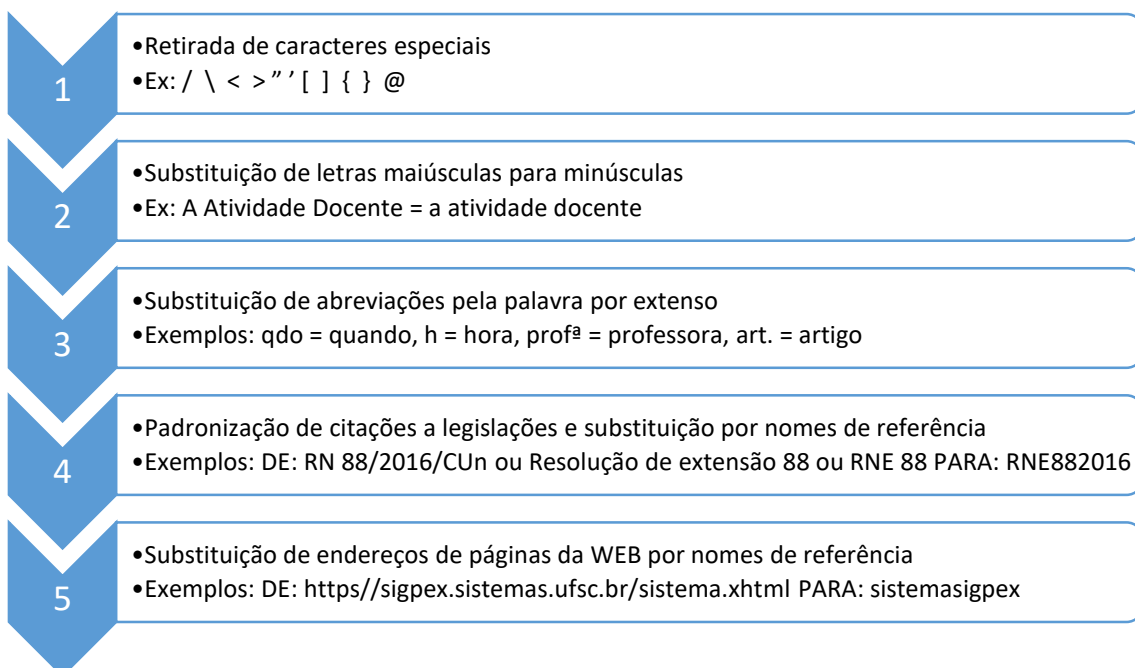
Fonte: Autora (2021)

A etapa de pré-processamento está dividida em: preparação dos dados, indexação e normalização e cálculo de relevância dos termos.

5.2.1 Preparação dos dados

A preparação dos dados da planilha Excel envolveu as seguintes etapas (Quadro 15):

Quadro 15 - Etapas de preparação dos dados selecionados



Fonte: Autora (2021)

A lista de nomes de referência à legislação padronizados na preparação dos dados encontra-se no Apêndice 1.

A lista de nomes de referência para os endereços de páginas da *Web* padronizados na preparação dos dados encontra-se no Apêndice 2.

Os dados tratados foram salvos em duas planilhas:

- 1) Planilha de PERGUNTAS, contendo os campos: pergunta, tipo e subtipo;
- 2) Planilha de RESPOSTAS, contendo os campos: resposta, tipo e subtipo.

5.2.2 Indexação e normalização

Como não existe uma lista universal de palavras de parada (*stopwords*), para esta pesquisa foi preparada uma lista com 268 palavras, que se encontra no Apêndice 3.

O KH Coder realiza o processo de *stemming* automaticamente. Na etapa de cálculo da relevância dos termos o processo de *stemming* foi testado.

5.2.3 Cálculo da relevância dos termos para a planilha PERGUNTAS

Para o cálculo da relevância dos termos foi utilizada a técnica WORD FREQUENCY LIST (WFL) do KH Coder, utilizando a planilha das PERGUNTAS.

A técnica gerou uma lista em ordem decrescente de frequência de palavras contendo os seguintes dados: palavra, classificação gramatical e número de vezes que a palavra aparece nos textos (frequência) (Quadro 16).

Quadro 16 - WFL - planilha PERGUNTAS

#	Word	POS / Conj.	Frequency
1	ser	V	1277
2	projeto	N	704
3	atividade	N	549
4	estar	V	480
5	sigpex	N	363
6	ir	V	359
7	professor	N	333
8	ação	N	310
9	registrar	V	287
10	relatório	N	275
11	registro	N	271
12	gostar	V	255
13	coordenador	N	248
14	curso	N	223
15	docente	N	213
16	aprovar	V	208
17	haver	V	195

Fonte: Autora (2021) - KH Coder

O software utiliza a seguinte classificação de palavras: adjetivos (AQ), numerais ordinais (AO) advérbios, preposições e conjunções (R), substantivos (N), verbos (V) e interjeição (I).

O resultado da WFL foi extraído para análise em planilha Excel (Quadro 17).

Quadro 17 - Dados WFL da planilha PERGUNTAS em Excel

	A	B	C	D	E	F	G	H	I	J	K
1	AQ		AO		R		N		V		I
2	final	187	segundo	8	mais	83	projeto	704	ser	1277	obrigada
3	docente	130	terceiro	3	assim	60	atividade	549	estar	480	uff
4	possível	126	terço	2	então	38	sigpex	363	ir	359	
5	horário	108	duplo	1	entanto	34	professor	333	registrar	287	
6	novo	86			somente	28	ação	310	gostar	255	
7	necessário	74			conforme	26	relatório	275	aprovar	208	
8	coordenador	65			aqui	25	registro	271	haver	195	
9	preciso	52			entretanto	25	coordenador	248	saber	175	
10	financeiro	51			novamente	25	curso	223	realizar	131	
11	externo	48			online	24	docente	213	receber	128	
12	seguinte	47			acima	22	aprovação	182	cadastrar	126	
13	semanal	44			além	22	departamento	160	solicitar	124	
14	científico	40			hoje	22	caso	157	aparecer	102	
15	ad	29			onde	22	sistema	155	proceder	102	
16	referente	28			bem	21	hora	153	conseguir	86	
17	total	28			dentro	18	forma	132	enviar	83	
18	anterior	27			junto	18	artigo	122	informar	67	
19	anexo	26			abaixo	15	carga	117	dar	65	
20	diferente	26			atual	15	data	111	inserir	65	

Fonte: Autora (2021)

Uma leitura dos resultados identificou palavras de menor importância para o objeto deste estudo. As palavras identificadas foram relacionadas no software na técnica FORCE IGNORE (Quadro 18).

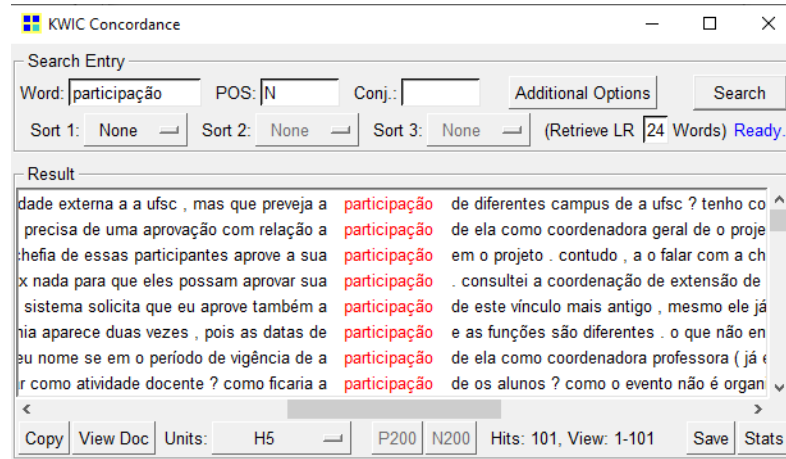
Além disso, foram subtraídas da análise palavras das classes gramaticais advérbios, preposições e conjunções (R) e interjeição (I), por serem consideradas de menor relevância para esta pesquisa.

Quadro 18 - Seleção das palavras para análise – planilha PERGUNTAS

Fonte: Autora (2021) - KH Coder

Através da técnica KWIC - CONCORDANCE, as palavras extraídas na WFL puderam ser analisadas dentro do seu contexto e sua concordância textual, esse recurso possibilitou uma conferência da classificação das palavras (quadro 19).

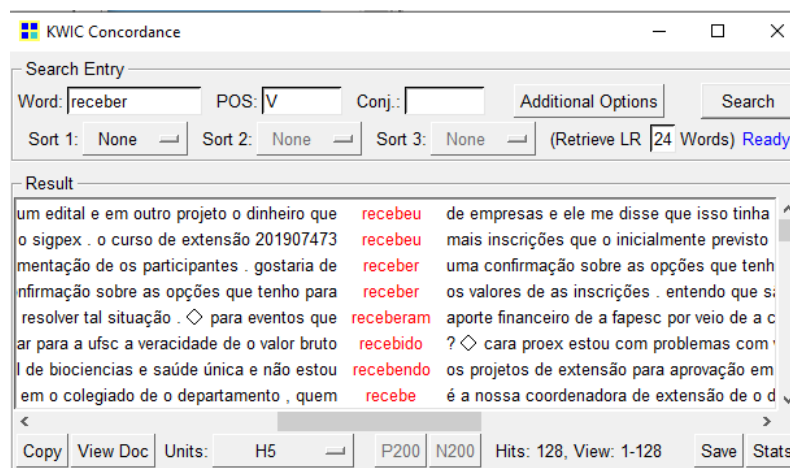
Quadro 19 - Concordância textual da palavra “participação” – planilha PERGUNTAS



Fonte: Autora (2021) - KH Coder

A mesma técnica permitiu conferir o processo automático de *Stemming* do software (Quadro 20).

Quadro 20 - Stemming do verbo “receber” – planilha PERGUNTAS

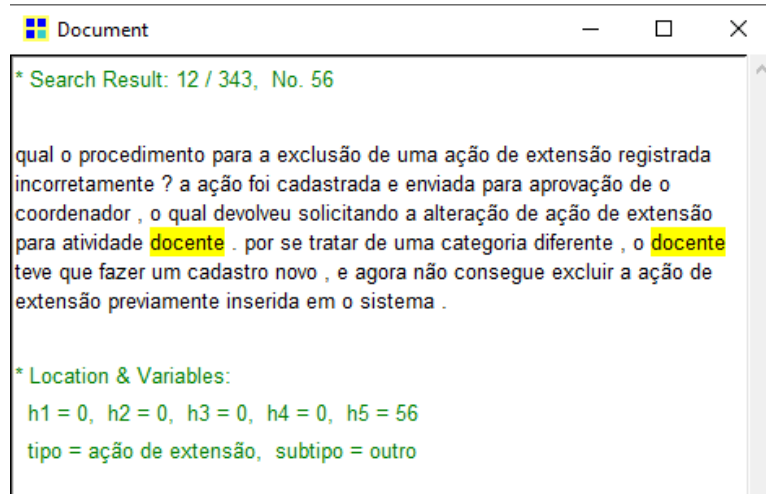


Fonte: Autora (2021) - KH Coder

Na conferência da extração e classificação das palavras, foram analisadas 50 palavras de forma aleatória utilizando a técnica KWIC - CONCORDANCE. Verificou-se que a extração e classificação das palavras foi satisfatória na amostra utilizada, bem como o processo de *stemming* automático.

No processo de conferência, confirmou-se a identificação de palavras classificadas em duas classes gramaticais na WFL. Como no caso da palavra DOCENTE, que aparece como substantivo e adjetivo (Quadro 21).

Quadro 21 - Conferência da classificação da palavra “docente” – planilha PERGUNTAS



Fonte: Autora (2021) - KH Coder

Utilizando a técnica TERM FREQUENCY DISTRIBUTION observou-se que foram encontradas 2.662 palavras alvo de análise, com uma média de 7,47 de Frequência de Termo (TF). Observou-se ainda, 1.264 palavras que ocorrem apenas uma vez (TF=1).

Quadro 22 - Frequência de Termo (TF) – planilha PERGUNTAS

Term Frequency Distribution

Descriptives

Types of Words (n) 2662
Mean of TF 7.47
Std. Deviation of TF 27.88

Frequency Table

TF	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	1264	47.48	1264	47.48
2	396	14.88	1660	62.36
3	199	7.48	1859	69.83
4	139	5.22	1998	75.06
5	82	3.08	2080	78.14
6	68	2.55	2148	80.69
7	76	2.85	2224	83.55
8	46	1.73	2270	85.27
9	41	1.54	2311	86.81
10	29	1.09	2340	87.90

Copy Plot Refresh Close

Fonte: Autora (2021) - KH Coder

Quadro 23 - Frequência de Documento (DF) – planilha PERGUNTAS

Document Frequency Dist...

Descriptives

Types of Words (n) 2662
Mean of DF 5.54
Std. Deviation of TF 15.86

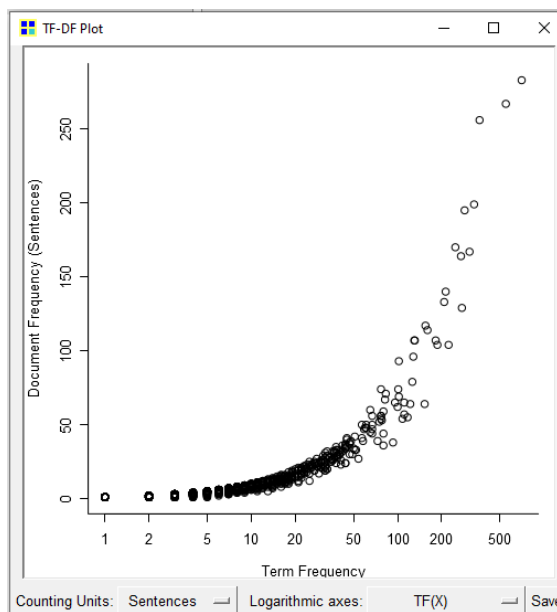
Frequency Table

DF	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	1360	51.09	1360	51.09
2	392	14.73	1752	65.82
3	188	7.06	1940	72.88
4	119	4.47	2059	77.35
5	97	3.64	2156	80.99
6	65	2.44	2221	83.43
7	60	2.25	2281	85.69
8	54	2.03	2335	87.72
9	28	1.05	2363	88.77
10	25	0.94	2388	89.71

Copy Plot Counting Units: H5 Close

Fonte: Autora (2021) - KH Coder

Quadro 24 - Gráfico TF – DF – planilha PERGUNTAS



Fonte: Autora (2021) - KH Coder

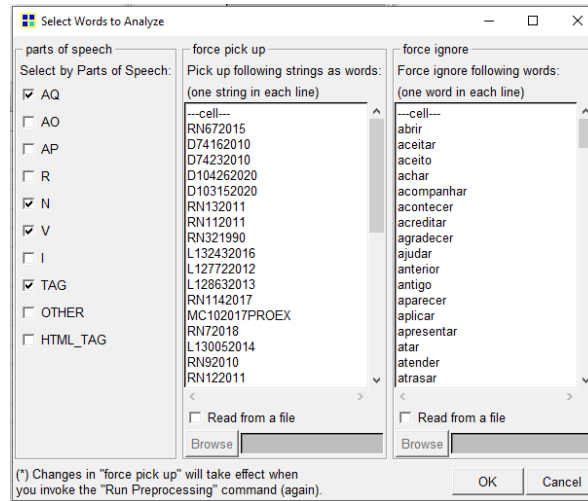
Assim, utilizando a TF média e excluindo as palavras com $TF \leq 7$, o número alvo de análise foi reduzido em 2.224 palavras (83,55%). **Restando 438 palavras para análise** (Quadro 22).

Utilizando a técnica DOCUMENT FREQUENCY DISTRIBUTION observou-se a distribuição de 2.662 palavras com uma frequência média de 5,54 de Frequência de Documento (DF). Observou-se que 1.360 palavras ocorrem em apenas um documento ($DF=1$) (Quadro 23). O Gráfico TF-DF (Quadro 24) permite avaliar a correlação entre TF e DF.

5.2.4 Cálculo da relevância dos termos para a planilha RESPOSTAS

O processo de cálculo de relevância dos termos utilizado para a planilha PERGUNTAS, foi repetido para a planilha RESPOSTAS, encontra-se no Apêndice 4. Para a planilha RESPOSTAS, foi identificada a necessidade de utilizar a técnica FORCE TO PICK UP (Quadro 25) para os nomes de referência utilizados nos passos 4 (legislação – Apêndice 1) e 5 (endereços páginas WEB – Apêndice 2) da preparação de dados.

Quadro 25 - Seleção das palavras para análise – planilha RESPOSTAS



Fonte: Autora (2021) - KH Coder

Utilizando a técnica TERM FREQUENCY DISTRIBUTION, observou-se que foram encontradas 1.675 palavras alvo de análise, com uma média de Frequência de Termo (TF) de 8,25. Observou-se ainda, 711 palavras que ocorrem apenas uma vez (TF=1) (Quadro 26).

Assim, utilizando a TF média e excluindo as palavras com $TF \leq 8$, o número alvo de análise foi reduzido em 1.410 palavras (84,18%). **Restando 265 palavras para análise.**

Quadro 26 - Frequência de Termo (TF) - planilha RESPOSTAS

TF	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	711	42.45	711	42.45
2	261	15.58	972	58.03
3	139	8.30	1111	66.33
4	97	5.79	1208	72.12
5	79	4.72	1287	76.84
6	58	3.46	1345	80.30
7	40	2.39	1385	82.69
8	25	1.49	1410	84.18
9	19	1.13	1429	85.31
10	13	0.78	1442	86.09

Fonte: Autora (2021) - KH Coder

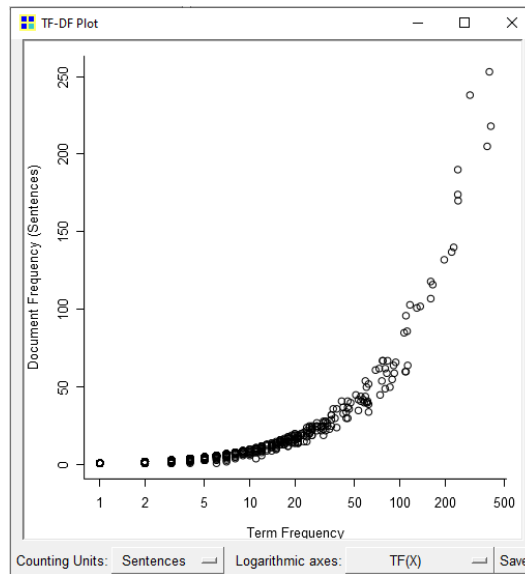
Quadro 27 - Frequência de Documento (DF) - planilha RESPOSTAS

DF	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	751	44.84	751	44.84
2	253	15.10	1004	59.94
3	144	8.60	1148	68.54
4	95	5.67	1243	74.21
5	72	4.30	1315	78.51
6	58	3.46	1373	81.97
7	32	1.91	1405	83.88
8	25	1.49	1430	85.37
9	22	1.31	1452	86.69
10	15	0.90	1467	87.58

Fonte: Autora (2021) - KH Coder

Utilizando a técnica DOCUMENT FREQUENCY DISTRIBUTION observou-se a distribuição de 1.675 palavras com uma frequência média de 6,56 de Frequência de Documento (DF). Observou-se que 751 palavras ocorrem em apenas um documento (DF=1) (Quadro 27). O Gráfico TF-DF (Quadro 28) permite avaliar a correlação entre TF e DF.

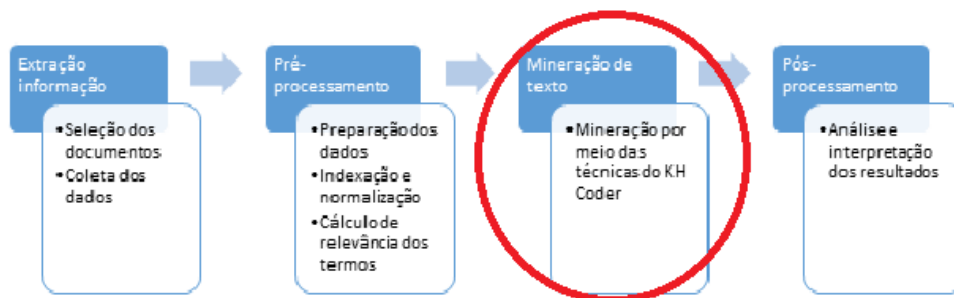
Quadro 28 - Gráfico TF-DF - planilha RESPOSTAS



Fonte: Autora (2021) - KH Coder

5.3 MINERAÇÃO DE TEXTOS

Figura 14 - Mineração de Textos



Fonte: Autora (2021)

A mineração de textos por meio das técnicas do KH Coder foi realizada em duas etapas, uma para a planilha PERGUNTAS e outra para a planilha RESPOSTAS.

5.3.1 Mineração por meio das técnicas do KH Coder – planilha PERGUNTAS

As informações da planilha PERGUNTAS foram analisadas por meio das seguintes técnicas: rede de coocorrência de palavras, rede de coocorrência de palavras e variáveis, análise por cluster, análise hierárquica de cluster, mapa auto-organizável e classificador Naive Bayes.

Importante ressaltar que a **semelhança** ou a **diferença** entre as palavras está relacionada à **distância entre as palavras** e é calculada estatisticamente e de forma automática pelas técnicas do software.

5.3.1.1 *Classificador Naive Bayes*

O Classificador Naive Bayes permitiu classificar automaticamente as perguntas em assuntos, por meio da identificação dos termos dos documentos e de cálculo probabilístico de ocorrência baseado em um modelo pré-definido de aprendizagem.

Para criar o modelo de aprendizagem, foram utilizados 92 documentos classificados por assunto de forma manual (Quadro 29).

A precisão do modelo de aprendizagem foi conferida por meio da matriz de confusão por validação cruzada. Este método de teste retém alguns documentos do exercício de treinamento para testar a classificação automática nesses documentos, ou seja, os documentos não utilizados para o treinamento tornam-se os primeiros sujeitos de teste para a classificação automática.

Com a validação cruzada, foi possível saber quão precisa será a classificação automática dos documentos. O procedimento envolveu a divisão dos documentos em 10 grupos, onde todos, exceto um grupo, foram alimentados para o treinamento. O grupo restante foi testado para estabelecer se a técnica classificou os documentos com precisão. Como resultado, foram classificadas corretamente 75% dos documentos, com coeficiente Kappa de 0,714 (Tabela 2).

Quadro 29 – Modelo de aprendizagem (exemplo)

Documento	Assunto
Estou avaliando um pedido de evento de extensão, o período de realização é de apenas 1 dia, e a carga horária semanal é de 6 horas. Esta orientação está correta ou o professor deve registrar as horas semanais do período preparatório do mesmo?	evento
Os programas de extensão registrados no sigpex permitem a emissão de certificados. Como outras ações estão vinculadas ao mesmo, não haverá duplicidade?	programa
Há alguma instrução regulamentação para carga horária em ações de extensão? Para aprovação de carga horária que vai para o paad é necessário contabilizar as horas dos projetos ativos e não permitir que ultrapasse 20h? se sim, estas 20 horas correspondem a somente projetos de extensão, ou devem ser contados também os projetos de pesquisa?	projeto
Realizei cursos de curta duração acima de 30 horas e o sistema apenas aceita até 30h. Por que cursos com maior carga horária não são contabilizados? Como faço para cadastrar estes cursos?	curso curta duração

Fonte: Autora (2021)

O Coeficiente Kappa pode ser definido como uma medida de associação usada para descrever e testar o grau de concordância (confiabilidade e precisão) na classificação. Assim, valores entre 0,41 e 0,60 são considerados regulares; valores entre 0,61 e 0,80 são considerados bons; valores entre 0,81 e 0,99 são considerados ótimos; 1,00 é considerado excelente (PERROCA; GAIDZINSKI, 2003).

Tabela 2 - Matriz de confusão do modelo de aprendizagem

Classificação	Classificado como:									
	prestação de serviço	evento e palestra	curso curta duração	projeto	programa	banca externa	publicação	evento	curso	PRECISÃO
prestação de serviço	10	0	1	0	0	0	0	0	0	90,91
evento e palestra	0	6	1	0	0	0	0	1	0	75,00
curso curta duração	0	1	11	0	0	0	0	0	0	91,67
projeto	0	0	0	15	1	0	0	1	0	88,24
programa	1	0	1	2	4	0	0	0	1	44,44
banca externa	0	0	1	0	0	8	0	0	0	88,89
publicação	0	0	2	0	0	0	3	0	0	60,00
evento	0	0	0	2	0	0	0	5	1	62,50
curso	1	0	2	1	1	0	0	1	7	53,85
TOTAL	12	7	19	20	6	8	3	8	9	

Documentos classificados corretamente: 69 / 92 (75.0%)

Kappa statistic: 0.714

Fonte: Autora (2021)

A partir do modelo de aprendizagem, as perguntas foram classificadas automaticamente de acordo com um dos seguintes assuntos: **programa, projeto, curso, evento, prestação de serviço, curso de curta duração, publicação, banca externa, evento e palestra.**

Exemplo de classificação automática, levando em consideração a PERGUNTA n° 341: *Para cadastrar a atividade docente curso de extensão de curta duração (participante), é obrigatório o comprovante de inscrição ou o docente pode criar a atividade e no relatório final anexar os documentos?*

No Quadro 30, observa-se uma linha com os *scores*, que indica 37,53% de probabilidade da sentença n° 341 ser uma pergunta relacionada à **curso de curta duração**. Esse *score* é calculado estatisticamente de acordo com as palavras encontradas no documento, por meio do Algoritmo Naive Bayes.

Quadro 30 - Classificação de documento - Naive Bayes (exemplo)

The screenshot shows a software window titled "Classification details file: log perguntas.nbi". It displays classification results for a document. The "Scores" section shows the following values: curso curta duração: 37.53, curso: 29.96, evento e palestra: 28.78, prestação de serviço: 27.56, projeto: 26.03, programa: 25.55, banca externa: 25.31, evento: 22.36, publicação: 21.50. The "Words" section is a table with columns for frequency and various categories. The categories include projeto, curso curta duração, evento e palestra, prestação de serviço, programa, banca externa, publicação, evento, curso, variance, projeto (%), curso curta duração (%), evento e palestra (%), and prior probability. The table lists various words and their associated scores for each category.

Words	frequency	projeto	curso curta duração	evento e palestra	prestação de serviço	programa	banca externa	publicação	evento	curso	variance	projeto (%)	curso curta duração (%)	evento e palestra (%)	pre:
curto-AQ	1	0.03	2.46	0.32	0.00	0.25	0.28	0.44	0.26	0.23	0.51	0.74	57.61	7.50	
duração-N	1	0.03	2.57	0.32	0.00	0.25	0.28	0.44	0.26	0.93	0.57	0.62	50.62	6.32	
curso-N	1	0.03	3.44	1.01	1.61	1.34	0.28	0.44	0.26	3.45	1.54	0.27	29.00	8.54	
cadastrear-V	1	0.03	2.06	1.01	0.69	2.19	0.97	0.44	0.95	0.93	0.43	0.34	22.17	10.92	
docente-AQ	1	0.72	2.21	1.93	2.40	0.94	0.97	1.13	0.26	0.93	0.47	6.30	19.24	16.79	
participante-N	1	2.11	1.65	0.32	0.00	0.94	0.28	0.44	1.87	1.33	0.54	23.60	18.47	3.59	
docente-N	1	1.82	2.35	1.42	1.10	1.34	2.07	1.83	0.95	1.33	0.19	12.83	16.50	9.99	
atividade-N	2	3.28	6.08	4.80	6.27	3.71	5.53	4.47	2.71	2.67	1.70	8.31	15.38	12.15	
inscrição-N	1	0.03	0.96	0.32	0.00	0.25	0.28	0.44	1.64	2.31	0.57	0.51	15.38	5.14	
anexar-V	1	1.42	1.65	2.72	1.61	0.25	1.38	0.44	0.26	1.33	0.57	12.83	14.94	24.59	
comprovante-N	1	0.72	0.96	1.93	1.39	0.25	0.28	0.44	0.26	0.93	0.30	10.13	13.41	26.98	
[prior probability]	1	5.69	5.16	4.88	5.48	4.88	4.79	4.34	4.88	4.96	0.14	12.63	11.46	10.83	
ser-V	1	3.50	3.26	3.41	3.61	2.95	3.33	3.15	3.30	3.37	0.03	11.70	10.92	11.42	
obrigatório-AQ	1	0.03	0.27	0.32	0.00	0.25	0.28	0.44	0.26	0.93	0.06	1.14	9.60	11.57	
criar-V	1	0.03	0.27	0.32	0.00	1.34	0.28	0.44	0.26	0.23	0.14	1.00	8.37	10.10	
final-AQ	1	2.43	0.96	1.71	1.79	1.63	1.38	0.44	1.87	1.84	0.30	17.29	6.83	12.15	
relatório-N	1	2.98	0.96	1.71	1.61	1.86	1.67	1.13	1.87	2.03	0.29	18.83	6.07	10.80	
documento-N	1	1.13	0.27	0.32	0.00	0.94	0.97	1.13	0.26	0.23	0.18	21.50	5.06	6.10	

Fonte: Autora (2021) - KH Coder

Após a introdução do modelo de aprendizagem, as 600 perguntas foram classificadas automaticamente de acordo com os assuntos, conforme o Quadro 31.

Quadro 31 - Classificação dos documentos - planilha PERGUNTAS

unit	variable	value	label	frequency
h5	Heading5	banca externa		55
h5	tipo	curso		38
h5	subtipo	curso curta duração		111
h5	model_10_fold-class	evento		35
h5	model_10_fold-is_correct	evento e palestra		13
h5	perguntas	prestação de serviço		62
		programa		32
		projeto		229
		publicação		25

Fonte: Autora (2021) - KH Coder

Também foi possível exportar uma matriz de tabulação (Quadro 32), mostrando a frequência de palavras específicas em cada documento classificado por assunto. Esta matriz possibilita o uso de informações em diferentes softwares estatísticos, permitindo análises mais especializadas das informações.

Quadro 32 – Matriz de tabulação - planilha PERGUNTAS

banca externa		curso		curso curta duração	
banca	0.3167	taxa	0.2034	curso	0.3375
docente	0.1925	inscrição	0.1837	atividade	0.3053
externo	0.1605	curso	0.1750	docente	0.2273
progressão	0.1579	cobrar	0.1463	registrar	0.2258
participação	0.1416	pessoa	0.1373	sigpex	0.2181
atividade	0.1408	necessário	0.1250	registro	0.2143
informação	0.1300	ministrar	0.1111	sistema	0.2011
professor	0.1222	ação	0.0978	certificado	0.1985
registrar	0.1222	evento	0.0978	docente	0.1981
mestrado	0.1167	fundação	0.0968	ser	0.1880

evento e palestra		prestação de serviço		programa	
palestra	0.2143	serviço	0.4762	programa	0.1930
link	0.2143	prestação	0.3810	vincular	0.1538
divulgação	0.1875	eventual	0.2424	criar	0.1087
foto	0.1429	pagamento	0.2113	indicar	0.1053
palestrante	0.1429	docente	0.2014	possibilidade	0.1034
evento	0.1343	valor	0.1778	ação	0.0950
tela	0.1111	consultoria	0.1739	mais	0.0947
falar	0.1053	ressarcimento	0.1467	coordenar	0.0930
resumo	0.1053	atividade	0.1454	cancelar	0.0889
online	0.1000	registro	0.1320	questionar	0.0870

A direita da Figura 12, é possível identificar círculos de referência para a frequência de palavras, quanto maior o círculo, maior a frequência da palavra destacada. Também é possível identificar a separação das palavras em vinte subgráficos. A conexão das palavras por meio de linhas, facilitou a visualização e o entendimento das estruturas de coocorrência, mostrando a associação entre grupos de palavras.

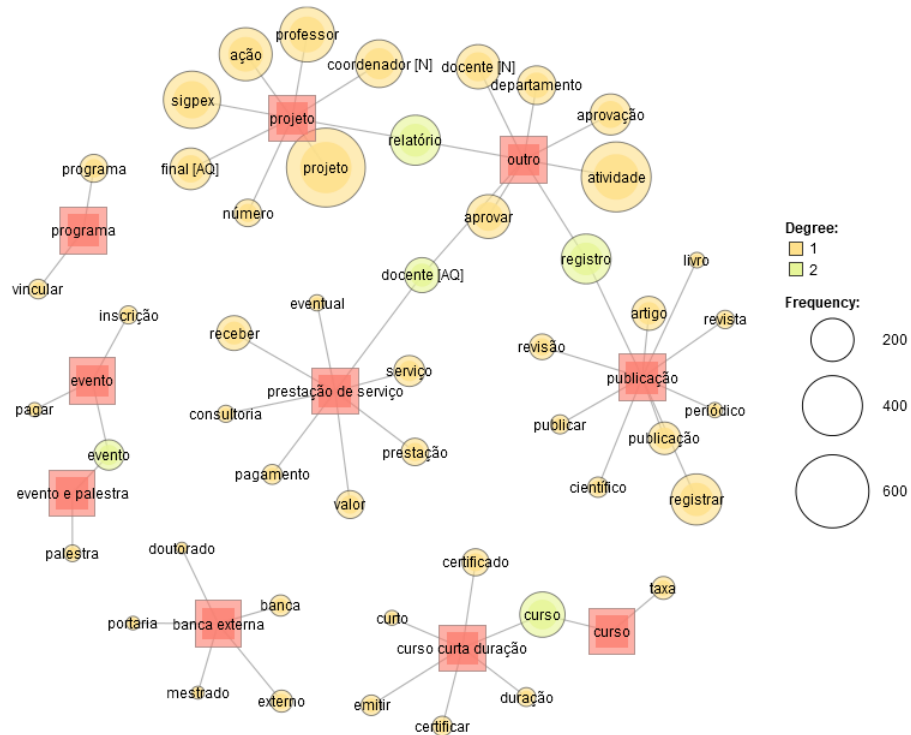
Observando a Figura 12, a análise por meio da rede de coocorrência permitiu a identificação dos seguintes **tópicos de dúvidas**:

- Análise ou aplicação de declaração de compatibilidade técnica e custos
- Aporte financeiro
- Aprovação
- Aprovação por reunião de colegiado ou ad referendum
- Banca externa
- Bancas de mestrado e doutorado
- Cobrança de inscrição
- Cursos de curta duração
- Emissão de certificados
- Envio do relatório final
- Fundação de apoio
- Grupo de estudos
- Membro de comissão
- Prestação de serviço, serviço eventual e consultoria
- Progressão funcional
- Prorrogação
- Publicações, artigos científicos, revisão de artigos, revistas
- Recolhimento de taxas
- Registro de atividades docentes
- Registro de carga horária semanal
- Registro de projetos, coordenação e departamento

5.3.1.3 *Análise de rede de coocorrência de palavras e variáveis*

Aplicando a rede de coocorrência de **palavras e variáveis**, pode-se identificar de forma mais detalhada as relações das palavras vinculadas a variáveis. Neste caso, as variáveis selecionadas são os tipos de ações de extensão (programa, projeto, curso e evento) e os tipos de atividades docentes (curso de curta duração, publicação, banca externa, evento e palestra, prestação de serviço) (Figura 13).

Figura 16 - Rede de Coocorrência entre Palavras e Variáveis - planilha PERGUNTAS



Fonte: Autora (2021) - KH Coder

Para a análise da rede de coocorrência entre palavras e variáveis, cada variável foi analisada individualmente por meio da concordância textual e da estatística de localização.

Enquanto a estatística de localização identificou quais palavras apareceram mais frequentemente antes e depois da palavra alvo (neste caso a variável), indicando uma forte relação, a concordância textual permitiu visualizar todos os documentos relacionados à variável selecionada.

a) Variável PROJETO

Analisando a rede de coocorrência entre palavras e a variável “projeto”, isolou-se a Figura 14.

Figura 17 - Rede de Coocorrência de Palavras e Variável “projeto” - planilha PERGUNTAS



Fonte: Autora (2021) - KH Coder

Por meio da concordância textual da variável PROJETO (Quadro 33) e a estatística de localização (Quadro 34). Apresenta-se a seguinte análise de resultados:

Quadro 33 - Concordância textual palavra “projeto” - planilha PERGUNTAS

Interface do software KWIC Concordance. O campo de busca contém a palavra 'projeto'. O resultado mostra trechos de texto com a palavra 'projeto' destacada em vermelho. Exemplo de trecho: '... projeto de outro departamento, gostaria de confirmar...'

Fonte: Autora (2021) - KH Coder

Quadro 34 - Estatística de localização palavra “projeto” - planilha PERGUNTAS

Interface do software Collocation Stats. O campo de busca contém a palavra 'projeto'. O resultado é uma tabela com as seguintes colunas: N, Word, POS, Total, LT, RT, L5, L4, L3, L2, L1, R1, R2, R3, R4, R5, The Score.

N	Word	POS	Total	LT	RT	L5	L4	L3	L2	L1	R1	R2	R3	R4	R5	The Score
1	sigpex	N	52	11	41	3	3	5	0	0	11	1	8	4	17	21.583
2	registrar	V	34	20	14	3	3	2	12	0	2	3	3	4	2	13.917
3	coordenador	N	46	34	12	1	4	27	2	0	0	0	0	6	6	13.900
4	cadastrar	V	25	16	9	0	0	1	14	1	2	0	6	0	1	12.533
5	aprovar	V	23	10	13	2	1	2	4	1	2	6	4	0	1	10.850
6	referir	V	15	14	1	2	1	3	0	8	1	0	0	0	0	10.650
7	número	N	22	3	19	1	0	2	0	0	4	4	7	2	2	10.100
8	financiar	V	11	0	11	0	0	0	0	0	8	1	1	0	1	9.033
9	professor	N	32	22	10	7	12	3	0	0	0	1	5	3	1	8.517
10	ação	N	24	11	13	3	4	1	2	1	0	3	3	1	6	7.883
11	aparecer	V	14	2	12	0	1	1	0	1	3	5	0	2	2	7.233
12	relatório	N	21	12	9	0	7	4	1	0	0	5	1	2	1	7.117
13	encerrar	V	12	4	8	0	0	0	4	0	3	3	1	0	1	7.033

Fonte: Autora (2021) - KH Coder

O quadro de estatística de localização (Quadro 34) mostra que o verbo “registrar” aparece 34 vezes, sendo 20 vezes em posição à esquerda (LT) da palavra PROJETO e 14 vezes em posição à direita (RT). O verbo “cadastrar” aparece 25 vezes, sendo 16 vezes em posição LT e 9 vezes em posição RT. O verbo “aprovar” aparece 23 vezes, sendo 10 vezes em posição LT e 13 vezes em posição RT.

A palavra “SIGPEX” é encontrada 52 vezes, em diferentes posições, trata-se do Sistema Integrado de Gerenciamento de Projetos de Pesquisa e de Extensão, onde todas as ações de extensão e atividades docentes são gerenciadas.

Esta análise revela que as dúvidas mais frequentes são relacionadas ao registro ou cadastro do projeto no SIGPEX e à aprovação. A mesma análise foi repetida para as variáveis

PROGRAMA, PUBLICAÇÃO, BANCA, CURSO, PRESTAÇÃO DE SERVIÇO, e EVENTO.

Os resultados encontram-se detalhados no Apêndice 5.

A análise por meio da rede de coocorrência e da concordância textual das palavras e variáveis permitiu a identificação dos seguintes **tópicos de dúvidas**:

- Participação em bancas externas
- Participação em eventos e palestras
- Registro da atividade docente prestação de serviço, serviço eventual
- Registro de publicação de artigo e data de publicação
- Registro ou cadastro do projeto no SIGPEX e à aprovação
- Registro/cadastro de cursos de curta duração
- Vínculo de ações ou projetos em programas e a criação de programas

5.3.1.4 Análise por cluster

Para esta técnica, utilizou-se o agrupamento por variáveis em cinco clusters e adotou-se o coeficiente de similaridade Jaccard, para adicionar peso às palavras mais relevantes que aparecem em apenas alguns documentos (Quadro 35).

Com a divisão dos documentos em clusters, foi possível identificar o número de documentos (sentenças) que fazem parte de cada cluster e os estágios de aglomeração.

Quadro 35 - Análise de Cluster do Documento - planilha PERGUNTAS

cluster	documents
Cluster1	95
Cluster2	121
Cluster3	117
Cluster4	38
Cluster5	229

stage	cluster 1	cluster 2	coefficients
1	-36	-160	0.111
2	-462	-538	0.250
3	-369	-479	0.286
4	-153	-171	0.333
5	-384	-419	0.333
6	-7	-353	0.385
7	-494	-506	0.385
8	-370	-371	0.400
9	-184	-242	0.429
10	-393	-394	0.429
11	-592	-593	0.455
12	-113	-270	0.471
13	-137	-449	0.471
14	-47	-77	0.500
15	-92	-454	0.500

Fonte: Autora (2021) - KH Coder

a) Análise Cluster 1

Verificando o conteúdo do Cluster 1, pode-se listar o número de documentos classificados (Quadro 36) e a lista de associação de palavras (Quadro 37), que mostra as palavras que aparecem com maiores probabilidades nos documentos do Cluster 1.

Quadro 36 - Documentos classificados no Cluster 1 - planilha PERGUNTAS

Fonte: Autora (2021) - KH Coder

Quadro 37 - Lista de associação de palavras Cluster 1 - planilha PERGUNTAS

N	word	POS	unconditional	conditional	Jaccard
1	serviço	N	62 (0.103)	39 (0.411)	0.3305
2	prestação	N	54 (0.090)	31 (0.326)	0.2627
3	registro	N	161 (0.268)	50 (0.526)	0.2427
4	professor	N	193 (0.322)	56 (0.589)	0.2414
5	registrar	V	193 (0.322)	50 (0.526)	0.2101
6	eventual	AQ	20 (0.033)	18 (0.189)	0.1856
7	curso	N	103 (0.172)	28 (0.295)	0.1647
8	atividade	N	261 (0.435)	44 (0.463)	0.1410
9	externo	AQ	39 (0.065)	16 (0.168)	0.1356
10	evento	N	63 (0.105)	18 (0.189)	0.1286

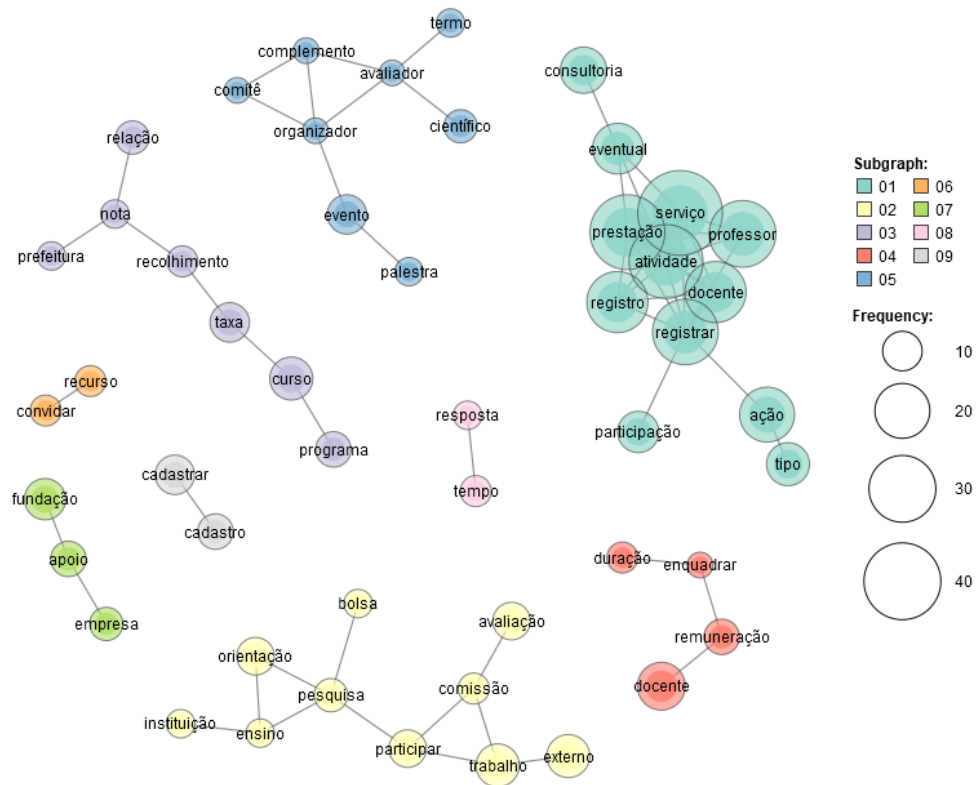
Fonte: Autora (2021) - KH Coder

Foram classificados 95 documentos no cluster 1. A palavra “professor” aparece em 58,9 % dos documentos. As palavras “registro” e “registrar” aparecerem em 52,6% dos documentos. A palavra “serviço” aparece em 41,1% dos documentos e “prestação” aparece em 32,6%. A rede de coocorrência do Cluster 1 pode ser observada na Figura 15, com destaque para a rede na cor verde.

Para o Cluster 1, pode-se identificar que as principais dúvidas estão relacionadas ao registro de prestação de serviço pelo professor, destacando-se serviço eventual e consultoria.

A mesma análise foi repetida para os demais clusters e encontra-se detalhada no Apêndice 6.

Figura 18 - Rede de coocorrência do Cluster 1 - planilha PERGUNTAS



Fonte: Autora (2021) - KH Coder

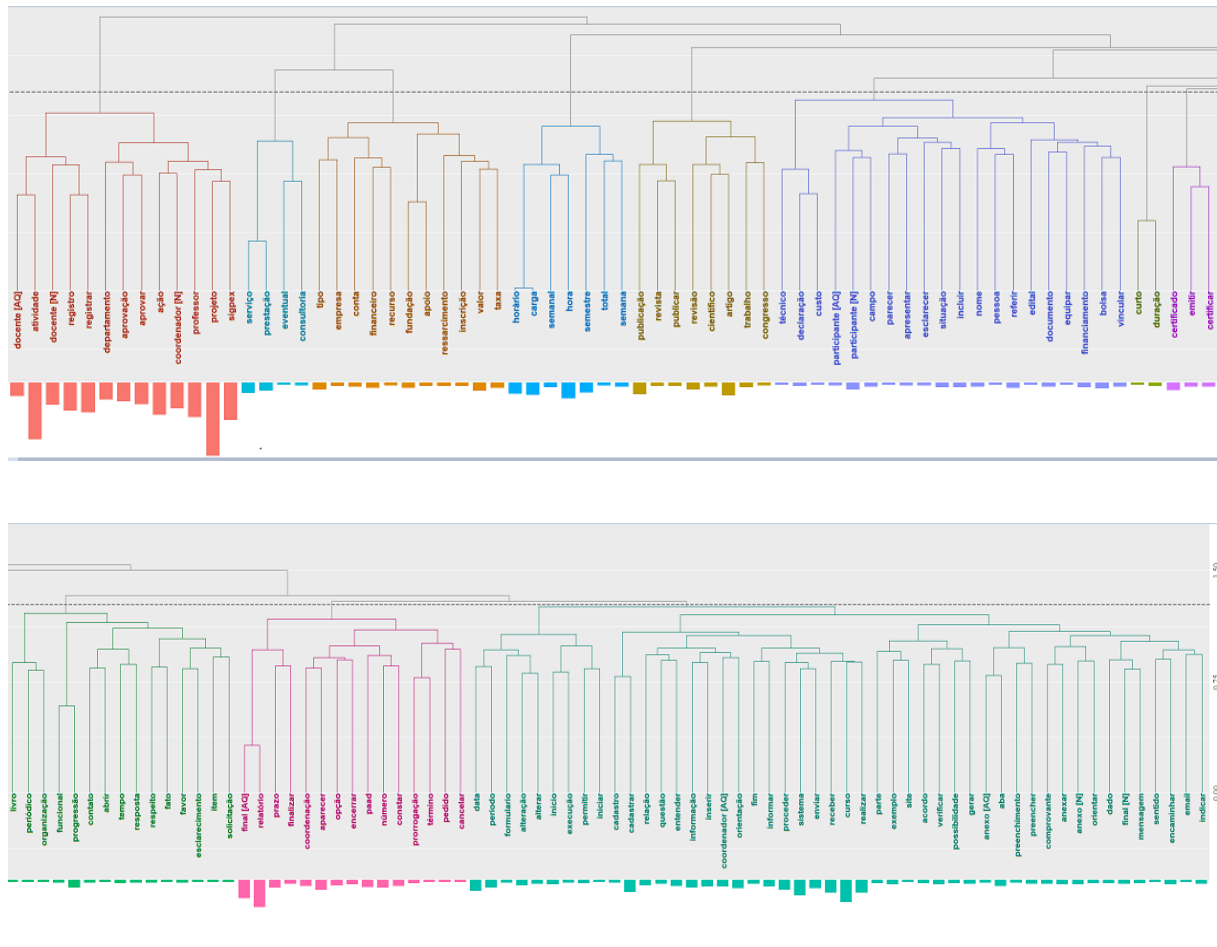
5.3.1.5 Análise hierárquica de cluster

A análise hierárquica de cluster (Figura 16) permitiu visualizar as combinações ou grupos de palavras que têm padrões de aparência semelhante usando análise de agrupamento hierárquico.

O dendograma gerado mostra as palavras que aparecem no documento e suas combinações, relacionadas por meio de linhas que indicam as posições e comprimentos de relacionamentos entre documentos.

As diferentes cores do dendograma permitiu distinguir os agrupamentos de palavras, os clusters. As linhas intra e extra cluster permitiram a visualização da organização hierárquica das palavras, criando subgrupos de palavras dentro do próprio cluster, o que facilita a análise do mesmo. As barras ligadas às palavras estão relacionadas à frequência das mesmas, quanto maior a barra, mais vezes a palavra foi encontrada.

Figura 19 - Análise Hierárquica de Cluster - planilha PERGUNTAS

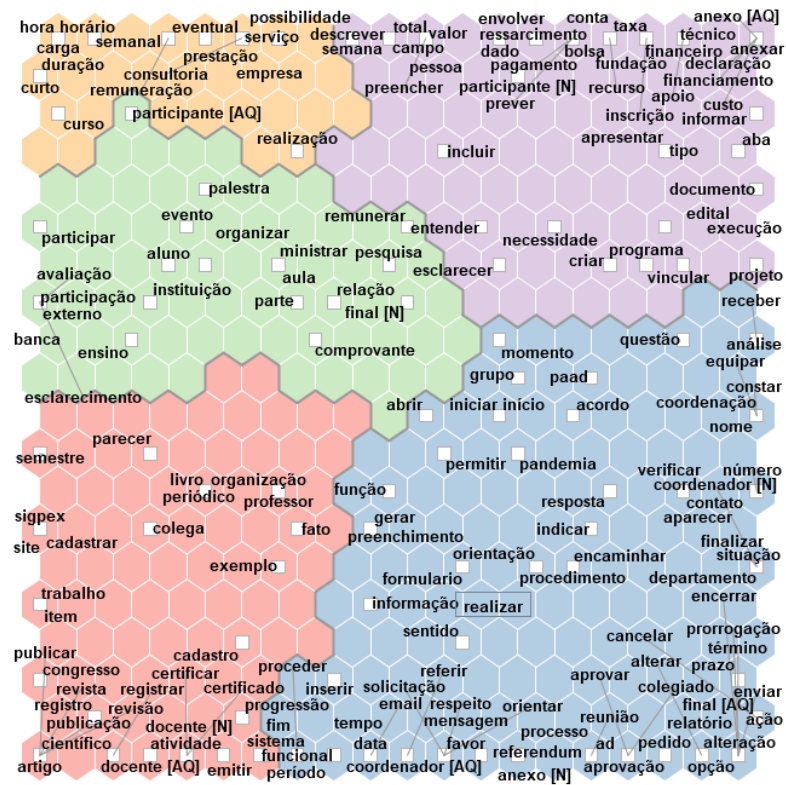


Fonte: Autora (2021) - KH Coder

5.3.1.6 Mapa auto-organizável

O mapa auto-organizável permitiu explorar associações entre palavras, organizadas em cluster através da matriz. Foram feitas duas consultas, uma com cinco clusters (Figura 17) e outra com 10 clusters (Figura 18).

Figura 20 - Mapa Auto-Organizável em 5 clusters - planilha PERGUNTAS



Fonte: Autora (2021) - KH Coder

Figura 21 - Mapa Auto-Organizável em 10 clusters - planilha PERGUNTAS



Fonte: Autora (2021) - KH Coder

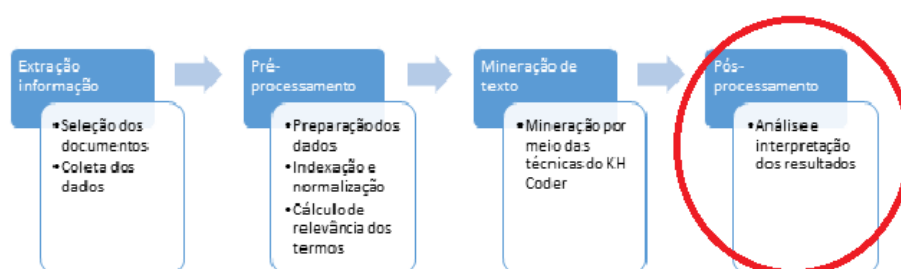
5.3.2 Mineração por meio das técnicas do KH Coder – planilha RESPOSTAS

As respostas aos chamados de atendimento são produzidas pelos atendentes da PROEX, sendo esperada uma conexão bem maior do que as perguntas. Assim, procurou-se verificar se ocorre uma padronização nas respostas.

As informações da planilha RESPOSTAS foram analisadas por meio das seguintes técnicas: rede de coocorrência de palavras, rede de coocorrência de palavras e variáveis, análise por cluster, análise hierárquica de cluster, mapa auto-organizável e classificador Naive Bayes. Os resultados encontram-se no Apêndice 7.

5.4 PÓS-PROCESSAMENTO

Figura 22 - Pós-processamento



Fonte: Autora (2021)

Esta seção apresenta a análise e interpretação dos resultados obtidos por meio da aplicação das técnicas de mineração de textos.

A partir da análise Rede de Coocorrência de palavras, Rede de Coocorrência de Palavras e Variáveis, Clusters e Análise Hierárquica de Cluster, apresenta-se a síntese dos **tópicos das perguntas mais frequentes identificadas** (quadro 38).

Quadro 38 - Síntese de tópicos das perguntas

	Tópicos de dúvidas identificados	Rede de Coocorrência de Palavras	Rede de Coocorrência de Palavras e Variáveis	Clusters	Análise Hierárquica de Cluster
01	Alteração de data início da ação				X
02	Analisar ou aplicar declaração de compatibilidade técnica e custos	X		X	X
03	Aporte financeiro	X			
04	Aprovação de programas e projetos	X	X	X	X

05	Aprovação por reunião de colegiado ou ad referendum	X			X
06	Banca externa	X	X		X
07	Bancas de mestrado e doutorado	X		X	
08	Criação de programa		X		
09	Cobrança de inscrição	X			X
11	Cursos de curta duração	X	X	X	X
12	Emissão de certificados	X		X	X
13	Envio e aprovação do relatório final			X	
14	Fundação de apoio	X		X	X
15	Grupo de estudos/pesquisa	X			X
16	Membro de comissão	X		X	
17	Organização periódico/livro				X
18	Participação em eventos/palestras		X	X	X
19	Prestação de serviço, serviço eventual e consultoria	X	X	X	X
20	Progressão funcional	X		X	X
21	Prorrogação	X			X
22	Publicações, artigos científicos, revisão de artigos, revistas	X	X	X	X
23	Recolhimento de ressarcimentos	X		X	X
24	Registro de atividades docentes	X			X
25	Registro de carga horária	X		X	X
26	Registro de projetos	X	X		
27	Relatório final, prazo				X
28	Remuneração			X	X
29	Vínculo de ações e projetos à programas		X		
30	Vínculo de bolsa, financiamento edital				X

Fonte: Autora (2021)

Foram identificados 30 tópicos de perguntas. Destaca-se que os seguintes tópicos, encontrados na análise de todas as técnicas:

- Aprovação de programas e projetos,
- Cursos de curta duração,
- Prestação de serviço, serviço eventual e consultoria,
- Publicações, artigos científicos, revisão de artigos, revistas.

A partir da análise Rede de Coocorrência de palavras, Rede de Coocorrência de Palavras e Variáveis, Clusters e Análise Hierárquica de Clusters, apresenta-se a síntese dos **tópicos das respostas mais frequentes identificadas** (Quadro 39).

Quadro 39 - Síntese de tópicos de respostas

	Tópicos de respostas	Rede de Coocorrência de palavras	Rede de Coocorrência de Palavras e Variáveis	Clusters	Análise Hierárquica de Cluster
01	Alteração da data final	X			
02	Alteração situação para revisão para alteração de data				X
03	Anexo de comprovantes	X		X	X
04	Aprovação ad referendum e colegiado			X	
05	Aprovação de projeto e programa	X	X	X	
06	Aprovação do relatório final	X		X	X
07	Bancas externas		X		
08	Cancelamento de ações	X			X
09	Carga horária semanal e remuneração de servidores	X		X	X
10	Cursos e eventos de curta duração		X		
11	Emissão de certificados	X			X
12	Fundação de apoio	X		X	X
13	Início e execução de ações	X			
14	Organização de evento e inscrição curso				X
15	Participação 2/3 membros da instituição	X		X	X
16	Portaria de substituição emitida pelo Centro de Ensino	X		X	X
17	Prazos de aprovação pelos órgãos responsáveis	X			
18	Prestação de serviço, serviço eventual e consultoria	X	X	X	X
19	Progressão funcional e CPPD	X		X	X
20	Realização de trabalho (acordo/convênio)	X			X
21	Recursos financeiros	X			
22	Registro de ações como projetos (tipificação)		X		
23	Registro de atividades docentes	X	X	X	X
24	Registro de ação como atividade docente, eventos e palestras (tipificação)		X		
25	Registro de publicações por semestres	X	X		X
26	Relação ensino e pesquisa	X			
27	Situação dos projetos (aprovada ou encerrada)		X		
28	Ressarcimento, valor e recolhimento	X		X	X
29	Vínculo de ações a um programa		X		

Fonte: Autora (2021)

Foram encontrados 29 tópicos de respostas. Destaca-se que os seguintes tópicos, encontrados na análise de todas as técnicas:

- Prestação de serviço, serviço eventual e consultoria,
- Registro de atividades docentes.

As legislações mais referenciadas nas respostas foram:

- Resolução Normativa nº 88/2016/CUn, que dispõe sobre as normas que regulamentam as ações de extensão na UFSC (117 vezes – 19,5% das respostas);
- Resolução Normativa nº 114/2017/CUn, que estabelece os critérios e os procedimentos para a concessão das progressões e promoções na carreira do magistério federal no âmbito da UFSC (10 vezes – 1,7% das respostas).

Os artigos mais referenciados foram o 8º, que trata do **registro e aprovação** das ações de extensão, e o 12, que trata dos **prazos para preenchimento do relatório final e aprovação** pelo coordenador de extensão do departamento.

Todos os Artigos da RN 88/2016/CUn referenciados nas respostas encontram-se no Quadro 40.

Quadro 40 - Artigos da RN 88/2016/CUn mais referenciados nas orientações

Resolução Normativa nº 88/2016/CUn – Dispõem sobre as normas que regulamentam as ações de extensão na Universidade Federal de Santa Catarina.		
Artigo	Nº de referências	Descrição
1º	2	A extensão universitária, sob o princípio constitucional da indissociabilidade entre ensino, pesquisa e extensão, é um processo interdisciplinar, educativo, cultural, científico e político que promove a interação transformadora entre a Universidade Federal de Santa Catarina (UFSC) e outros setores da sociedade.
2º	1	A extensão universitária visa: I – estimular e potencializar as relações de intercâmbio entre a Universidade e a sociedade em relação aos objetivos da instituição; II – propiciar mecanismos para que a sociedade utilize o conhecimento existente na realização de suas atividades; III – facilitar e melhorar a articulação e a operacionalização do conhecimento advindo do ensino e da pesquisa para a sociedade; IV – preservar o conhecimento produzido pela interação da Universidade com a sociedade; V – incentivar a participação tanto de alunos de graduação como de pós-graduação, além de professores e servidores técnico-administrativos em educação.
3º	4	A extensão universitária é realizada por meio de ações como: I – programa de extensão, que constitui um conjunto articulado de projetos e outras ações de extensão, tais como cursos, eventos, prestação de serviços e publicações, preferencialmente integrando as ações de extensão, pesquisa e ensino, tendo caráter orgânico-institucional, clareza de diretrizes e orientação para um objetivo comum, e sendo executado a médio e longo prazo; II – projeto de extensão, que constitui um conjunto de ações de caráter educativo, social, cultural, científico ou tecnológico, com objetivo específico e prazo determinado, podendo ser isolado ou vinculado a um programa; III – curso de extensão, que constitui uma ação pedagógica de caráter teórico e/ou prático, com participação de forma presencial, semipresencial ou a distância, com planejamento, organização e critérios de avaliação definidos; IV – evento de extensão, que consiste em ação que implica na apresentação, disseminação e/ou exibição pública, livre ou com público específico do conhecimento ou produto cultural, artístico, esportivo, científico ou tecnológico desenvolvido, conservado ou reconhecido pela Universidade;

		V – prestação de serviço, que consiste em realização de trabalho oferecido pela Universidade ou solicitado por terceiros, na forma de assessorias, consultorias e perícias.
4º	1	Os cursos de extensão serão executados em até cento e oitenta horas sob a forma de: I – iniciação, que consiste em curso com o objetivo de oferecer noções introdutórias em uma área específica do conhecimento; II – atualização, que consiste em curso com o objetivo de atualizar e ampliar conhecimentos, habilidades ou técnicas em uma área do conhecimento; III – treinamento, que consiste em curso com o objetivo de treinamento, qualificação e capacitação em atividades profissionais específicas. Parágrafo único. Excetua-se deste artigo o ensino de graduação e de pós-graduação (stricto e lato sensu), que, pelas suas próprias características, constituem modalidades específicas de formação.
6º	2	Cada ação de extensão terá um coordenador com comprovada qualificação na respectiva área, o qual será responsável por sua proposição e execução, observado o disposto nesta Resolução Normativa. § 1º Podem ser coordenadores de ações de extensão os servidores docentes ou técnico-administrativos em educação integrantes do quadro de pessoal efetivo da Universidade. § 2º A realização de ações de extensão por servidores da Universidade observará as limitações inerentes ao cargo e previstas nas legislações que o regulam. § 3º Cabe aos coordenadores das ações de extensão o acompanhamento e a verificação do aproveitamento dos bolsistas de extensão.
7º	3	Os servidores docentes poderão fazer constar no Planejamento e Acompanhamento de Atividades Docentes (PAAD) carga horária para realização de ações de extensão, observado o limite de até vinte horas semanais na média semestral e respeitados os limites impostos pela legislação pertinente em cada regime de trabalho. Parágrafo único. A alocação de carga horária regular no PAAD dos docentes deverá seguir critérios regulamentados no âmbito do departamento ou órgão equivalente no qual o docente se insere.
8º	14	Todas as ações de extensão deverão ser registradas pelo coordenador proponente no sistema de registro de ações de extensão e aprovadas pelo órgão responsável. § 1º O órgão responsável poderá ser qualquer órgão ou instância da Universidade, tais como departamentos, centros de ensino, órgãos administrativos ou órgãos suplementares. § 2º As ações de extensão deverão ser aprovadas antes do início de sua execução, podendo somente em casos excepcionais ser aprovadas durante o primeiro mês de sua execução. § 3º Para iniciar a tramitação da ação de extensão é necessária a aprovação da participação do coordenador, devendo a aprovação dos demais participantes seguir o disposto no § 2º deste artigo. § 4º Quando a ação de extensão envolver servidores de mais de um departamento, ou equivalente, deverá ser submetida à apreciação de cada órgão responsável envolvido.
10	4	A aprovação dos programas e projetos de extensão dar-se-á por prazo de até 5 (cinco) anos.
12	10	O coordenador terá prazo de até 30 (trinta) dias após o término da ação de extensão para preencher o relatório final no sistema de registro de ações de extensão, e o órgão responsável terá prazo de 45 (quarenta e cinco) dias para aprová-lo ou reprová-lo.
17	1	Compete ao coordenador-geral de extensão: I – aprovar a tramitação do registro das ações de extensão de sua unidade universitária, quando o proponente for o coordenador de extensão de departamento; II – aprovar a tramitação do registro das ações de extensão de sua unidade universitária, quando a ação envolver servidores de mais de um departamento, após a aprovação nos departamentos envolvidos; III – participar da câmara de extensão de sua unidade universitária, se houver; IV – representar sua unidade universitária na Câmara de Extensão da Universidade;

		V – outras atribuições, conforme regimento da unidade universitária.
20	3	<p>Compete ao coordenador de extensão do departamento:</p> <p>I – aprovar a tramitação do registro das ações de extensão, conforme deliberação do colegiado do departamento;</p> <p>II – representar seu departamento na câmara de extensão da unidade universitária, se houver;</p> <p>III – outras atribuições conforme regimento de seu departamento.</p> <p>Parágrafo único. No caso de unidades universitárias com departamento único, as atribuições mencionadas no caput poderão ser absorvidas pelo coordenador-geral de extensão, a critério da unidade universitária.</p>
21	1	<p>Cabe aos coordenadores proponentes de ações de extensão:</p> <p>I – elaborar propostas de ações de extensão, de acordo com o disposto nesta Resolução Normativa;</p> <p>II – efetuar o registro da proposta de ação de extensão no sistema de registro de ações de extensão e encaminhar ao setor encarregado da Universidade as ações de extensão que exigirem a celebração de convênios ou contratos para a sua execução;</p> <p>III – responsabilizar-se pela execução da ação de extensão;</p> <p>IV – supervisionar e avaliar o desempenho dos envolvidos na execução das atividades da ação de extensão;</p> <p>V – elaborar relatórios a respeito das ações de extensão realizadas, de acordo com as normas estabelecidas;</p> <p>VI – anexar aos relatórios os comprovantes da realização da ação de extensão, quando for o caso;</p> <p>VII – prestar contas dos recursos financeiros dentro dos prazos previstos e das normas vigentes;</p> <p>VIII – manter cadastro dos participantes para emissão de certificados, quando for o caso.</p>
22	5	<p>As ações de extensão da UFSC poderão ser desenvolvidas nas instalações da própria Universidade ou fora dela, com recursos humanos, materiais e financeiros próprios ou não.</p> <p>§ 1º Em qualquer ação de extensão desenvolvida pela UFSC, dois terços da equipe envolvida, preferencialmente, deverão ter ligação formal e em vigor com a instituição, respeitada a legislação vigente.</p> <p>§ 2º A captação de recursos financeiros para a viabilização das ações de extensão será de responsabilidade do coordenador proponente da ação de extensão.</p> <p>§ 3º Quando de interesse da Universidade, esta poderá buscar financiamento junto a organizações públicas e privadas.</p> <p>§ 4º Poderão ser fixadas taxas de inscrição nos cursos e eventos de extensão visando cobrir, parcial ou integralmente, os custos da respectiva ação de extensão.</p>
23	1	Quando a ação de extensão receber aporte financeiro, a fonte deste deverá estar explicitada.
24	4	<p>As ações de extensão poderão ser remuneradas.</p> <p>§ 1º A remuneração dos servidores envolvidos nas ações de extensão de que trata este artigo poderá ocorrer desde que sua participação:</p> <p>I – seja de caráter eventual, nos limites estabelecidos pela legislação vigente;</p> <p>II – ocorra em atividades ligadas a sua especialização ou atuação na Universidade, observando as limitações inerentes ao cargo e previstas nas legislações que o regulam.</p> <p>§ 2º Em ações de extensão com aporte financeiro, a carga horária remunerada dos servidores docentes em regime de dedicação exclusiva (DE) não poderá exceder 8 (oito) horas semanais ou 416 (quatrocentas e dezesseis) horas anuais, tal como estabelecido no § 4º do art. 21 da Lei nº 12.772/2012, com a modificação dada pela Lei nº 12.863/2013 e pela Lei nº 13.243/2016, ou conforme estabelecido na legislação vigente.</p>
25	1	<p>As ações de extensão, quando envolverem a captação de recursos financeiros, terão a sua gestão executada pela própria Universidade ou por uma das fundações de apoio devidamente credenciada.</p> <p>§ 1º Todo material permanente adquirido com recursos financeiros captados por meio de ações de extensão será incorporado ao patrimônio da Universidade.</p>

		<p>§ 2º Concluídas as ações de extensão, não havendo interesse da Universidade nos materiais permanentes adquiridos e havendo finalidade didática, pedagógica, cultural ou social, esses materiais poderão ser doados mediante solicitação do órgão interessado e submissão ao Conselho de Curadores.</p> <p>§ 3º Quando a ação de extensão for gerida por uma fundação de apoio:</p> <p>I – a gestão financeira das ações de extensão observará a legislação aplicável à espécie, obedecidos os termos de convênios ou contratos específicos celebrados com a Universidade;</p> <p>II – todo material permanente adquirido com recursos financeiros captados por meio de ações de extensão será incorporado ao patrimônio da Universidade, salvo o previsto no § 2º deste artigo;</p> <p>III – ao final da ação de extensão, a fundação deverá apresentar relatório financeiro ao setor competente da Universidade com a correspondente prestação de contas.</p>
26	4	<p>Nos convênios, contratos e instrumentos correlatos celebrados com entidades públicas ou privadas, assim como nos projetos financiados na forma de descentralização de recursos por entes governamentais para financiamento de ações de extensão, incidirão valores relativos ao ressarcimento institucional da Universidade pelo uso do capital intelectual, do nome e da imagem da instituição, bem como dos serviços e das instalações, conforme o ACÓRDÃO Nº 2731/2008 – TCU – Plenário, o art. 6º da Lei nº 8.958/1994, o inciso V do art. 1º - A da Portaria MEC/MCT 475/2008 e demais legislações pertinentes.</p> <p>§ 1º Como ressarcimento institucional especificado no caput, serão recolhidos os seguintes valores:</p> <p>I – 1% (um por cento) destinado à unidade universitária de origem do processo;</p> <p>II – 2% (dois por cento) destinados ao departamento de ensino ou a setores equivalentes (órgãos administrativos ou órgãos suplementares) de origem do projeto;</p> <p>III – 4% (quatro por cento) distribuídos da seguinte forma:</p> <p>a) 0,9% para incrementar os Programas de Bolsas de Extensão;</p> <p>b) 0,6% para incrementar os Programas de Bolsas de Monitoria e Estágio;</p> <p>c) 1% para a constituição do Fundo de Extensão (FUNEX), gerenciado pela PROEX para incrementar e viabilizar ações de extensão;</p> <p>d) 0,5% para incrementar ações de cultura gerenciadas pela Secretaria de Cultura e Arte;</p> <p>e) 0,5% para incrementar ações de inovação gerenciadas pela Secretaria de Inovação;</p> <p>f) 0,5% para incrementar Programas de Permanência gerenciados pela Pró-Reitoria de Assuntos Estudantis. § 2º Para as ações de extensão que envolverem mais de um departamento ou equivalente, o percentual de recolhimento previsto no inciso II deste artigo será dividido de forma proporcional ao envolvimento de cada participante.</p> <p>§ 3º Em caráter excepcional, o departamento de ensino e/ou a unidade universitária poderão, mediante justificativa circunstanciada e aprovada pelos seus órgãos colegiados, aumentar ou reduzir o percentual estabelecido nos incisos I e II do § 1º.</p> <p>§ 4º A Administração Central, representada pelo pró-reitor de extensão, poderá reduzir ou não cobrar o valor descrito no § 1º mediante justificativa circunstanciada nos seguintes casos:</p> <p>I – ações envolvendo recursos oriundos de fomento governamental, de aplicação compulsória por empresas, previstos em regulamentação específica, que não permitam descontos dessa natureza;</p> <p>II – ações envolvendo organizações sociais sem fins lucrativos de apoio à extensão e ao desenvolvimento tecnológico e social que, por restrições legais, normativas ou estatutárias, não permitam descontos dessa natureza;</p> <p>III – recursos oriundos de taxas de inscrição em congressos, seminários e cursos organizados pela UFSC, quando sem fins lucrativos.</p> <p>§ 5º Não estão previstas neste artigo eventuais taxas cobradas por fundação de apoio que venha a administrar os recursos captados pelas ações de extensão.</p>
27	1	<p>Durante o período de execução da ação de extensão, quando remunerada, as despesas de manutenção e utilização de equipamentos serão de responsabilidade do coordenador.</p>

28	2	Serão consideradas atividades de extensão, no sentido de pontuar para os critérios de progressão funcional do quadro docente, até sua incorporação em legislação específica, as seguintes atividades de curta duração sem caráter continuado, registradas no sistema de registro de ações de extensão: I – participação em bancas externas de concurso ou de formação acadêmica; II – participação em cursos de extensão de curta duração; III – participação em eventos e palestras; IV – prestação de serviços; V – produção de publicações e/ou produtos acadêmicos decorrentes das ações de extensão, para difusão e divulgação cultural, científica ou tecnológica; VI – revisão de artigos científicos e editoração externa de periódicos.
----	---	--

Fonte: Autora (2021) – RN 88/2016/CUn

Os links de páginas mais referenciados nas respostas foram:

- Orientações sobre movimentações financeiras (11 vezes – 1,8% das respostas)
- Orientações sobre atividade docentes (8 vezes – 1,3 % das respostas)
- Cartilha de perguntas e respostas da PROEX (6 vezes – 1% das respostas)
- Orientações sobre convênios e contratos (5 vezes – 0,8% das respostas)

Com relação às técnicas empregadas na mineração de textos, foram encontrados os seguintes resultados:

1) Classificador Naive Bayes:

- ✓ Permitiu classificar automaticamente as perguntas e respostas em assuntos, por meio da identificação dos termos dos documentos e de cálculo probabilístico de ocorrência baseado em um modelo pré-definido de aprendizagem.
- ✓ Possibilitou a conferência da precisão do modelo de aprendizagem por meio da matriz de confusão por validação cruzada.
- ✓ A partir do modelo de aprendizagem, os documentos foram classificadas automaticamente de acordo com um dos assuntos definidos no modelo de aprendizagem.
- ✓ Permitiu exportar uma matriz de tabulação, mostrando a frequência de palavras específicas em cada documento classificado por assunto. Esta matriz possibilita o uso de informações em diferentes softwares estatísticos, permitindo análises mais especializadas das informações.

2) Rede de Coocorrência de Palavras:

- ✓ Permitiu a visualização um diagrama de rede que mostrou as palavras com padrões de aparência semelhantes, ou seja, com alto grau de coocorrência.

- ✓ O tamanho dos círculos está ligado à frequência das palavras, facilitando a análise e interpretação dos dados.
- ✓ Linhas de conexão entre palavras identificam associação entre grupos de palavras, facilitando a análise e interpretação dos dados.

3) Rede de Coocorrência de Palavras e Variáveis:

- ✓ Permitiu a identificação detalhada das relações das palavras vinculadas a variáveis pré-determinadas.
- ✓ A utilização de estatística de localização permitiu a identificação das palavras que aparecem mais frequentemente antes e depois da palavra alvo, indicando uma forte relação.
- ✓ A concordância textual permitiu visualizar todos os documentos relacionados à variável selecionada.

4) Análise de Cluster:

- ✓ Permitiu a classificação das palavras, alocando-as em grupos internamente homogêneos, mas também em grupos heterogêneos (reúne o que é semelhante separando o que é diferente).
- ✓ Para definir a semelhança ou a diferença entre as palavras utiliza uma técnica estatística que determina a distância entre as palavras de um cluster.
- ✓ Permitiu identificar a probabilidade de relação entre as palavras de forma condicional e incondicional.

5) Análise hierárquica de cluster:

- ✓ Permitiu a visualização das combinações ou dos grupos de palavras que têm padrões de aparência semelhante usando análise de agrupamento hierárquico.
- ✓ As diferentes cores permitem distinguir os agrupamentos de palavras, os clusters, facilitando a análise e interpretação dos dados.
- ✓ As linhas intra e extra cluster permitiram a visualização da organização hierárquica das palavras, criando subgrupos de palavras dentro do próprio cluster, facilitando a análise e interpretação dos dados.
- ✓ As barras ligadas às palavras permitem a visualização da frequência das mesmas, quanto maior a barra, mais vezes a palavra foi encontrada.

6) Mapa auto-organizável:

- ✓ Permite o agrupamento de dados de forma que palavras semelhantes pertençam ao mesmo grupo e palavras pouco semelhantes pertençam a grupos diferentes.
- ✓ Seu gráfico bidimensional facilita a visualização e a interpretação de dados complexos.

Este capítulo apresentou os resultados obtidos de diferentes técnicas de mineração de textos, disponíveis por meio da ferramenta KH Coder, aplicadas no PAI/PROEX.

Com relação aos resultados obtidos, pode-se concluir que, o sistema PAI/PROEX tem informações relevantes para a Gestão Universitária e que foram bem exploradas pelas técnicas de mineração de textos.

6 Análise de Twitter e a Mineração de Textos aplicado a @UFSC

Este capítulo apresenta a aplicação da mineração de textos na rede social *Twitter*.

Para a análise de *Twitter* e mineração de textos aplicado a @UFSC, foi utilizado o tratamento proposto em #Twitter e Text Mining – Bitcoin, por Leandro Guerra (2021) (em [#https://www.outspokenmarket.com/blog](https://www.outspokenmarket.com/blog)).

O código R foi ajustado à mineração de *tweets* sobre a UFSC. O código trabalha utilizando os pacotes RTweet, Wordcloud e TM, respectivamente destinados a análise de *Twitter*, criação de nuvens de palavras e mineração de textos.

Para utilizar informações do *Twitter* é necessário possuir uma conta e obter autorização junto ao *Twitter*. Há uma limitação de busca de 18.000 *tweets* a cada quinze minutos.

O código de Guerra (2021) apresenta os seguintes procedimentos:

- Acessar ao *Twitter* definindo o critério de busca;
- Criação e limpeza de um Corpus;
- Uso de ferramentas de visualização;
- Limpeza e tratamento do texto com a Matiz de Termos;
- Aplicação de conceitos de Machine Learning – Clustering (Dendograma e K-Means).

A limpeza do Corpus utilizou-se de rotinas para padronizar o texto no formato UTF-8, transformar todo o texto em minúsculas, remoção de *stopwords* e de toda a pontuação.

Para fins de comparação, serão apresentados os dados obtidos com #UFSC e # Bitcoin (para o qual foi escrito o código).

6.1 RESULTADOS UFSC X BITCOIN

Serão apresentados os resultados para cada um dos procedimentos definidos por Guerra (2021), conforme Tabela 3.

Tabela 3 - Critérios de busca e ferramentas de visualização

Etapa	UFSC	Bitcoin
Critério de busca	Comando 1	Comando 2
Ferramentas de visualização		
Frequência de Tweets	Figura 1a	Figura 1b
Primeiro WordCloud	Figura 2a	Figura 2b
Gráfico de Termos Frequentes	Figura 3a	Figura 3b
WordCloud aprimorado (redução de termos esparsos)	Figura 4a	Figura 4b
Dendograma	Figura 5a	Figura 5b
Gráfico K-Means Tweet	Figura 6a	Figura 6b

Fonte - Autora (2021)

Comando 1:

```
bitcoin_tweets <- search_tweets("#UFSC", n = 15000, include_rts = FALSE, lang = "pt")
```

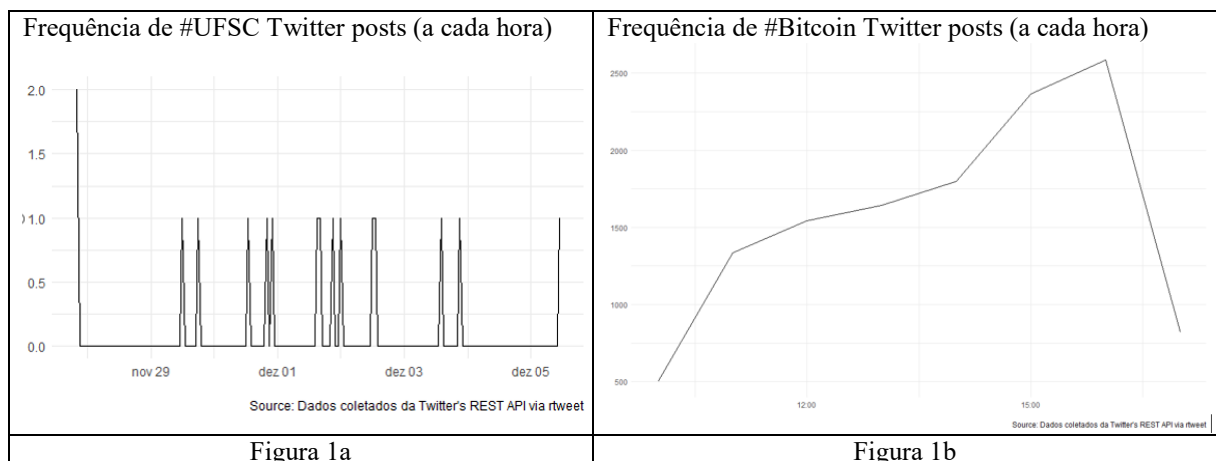
Comando 2:

```
bitcoin_tweets <- search_tweets("#bitcoin", n = 18000, include_rts = FALSE, lang = "en")
```

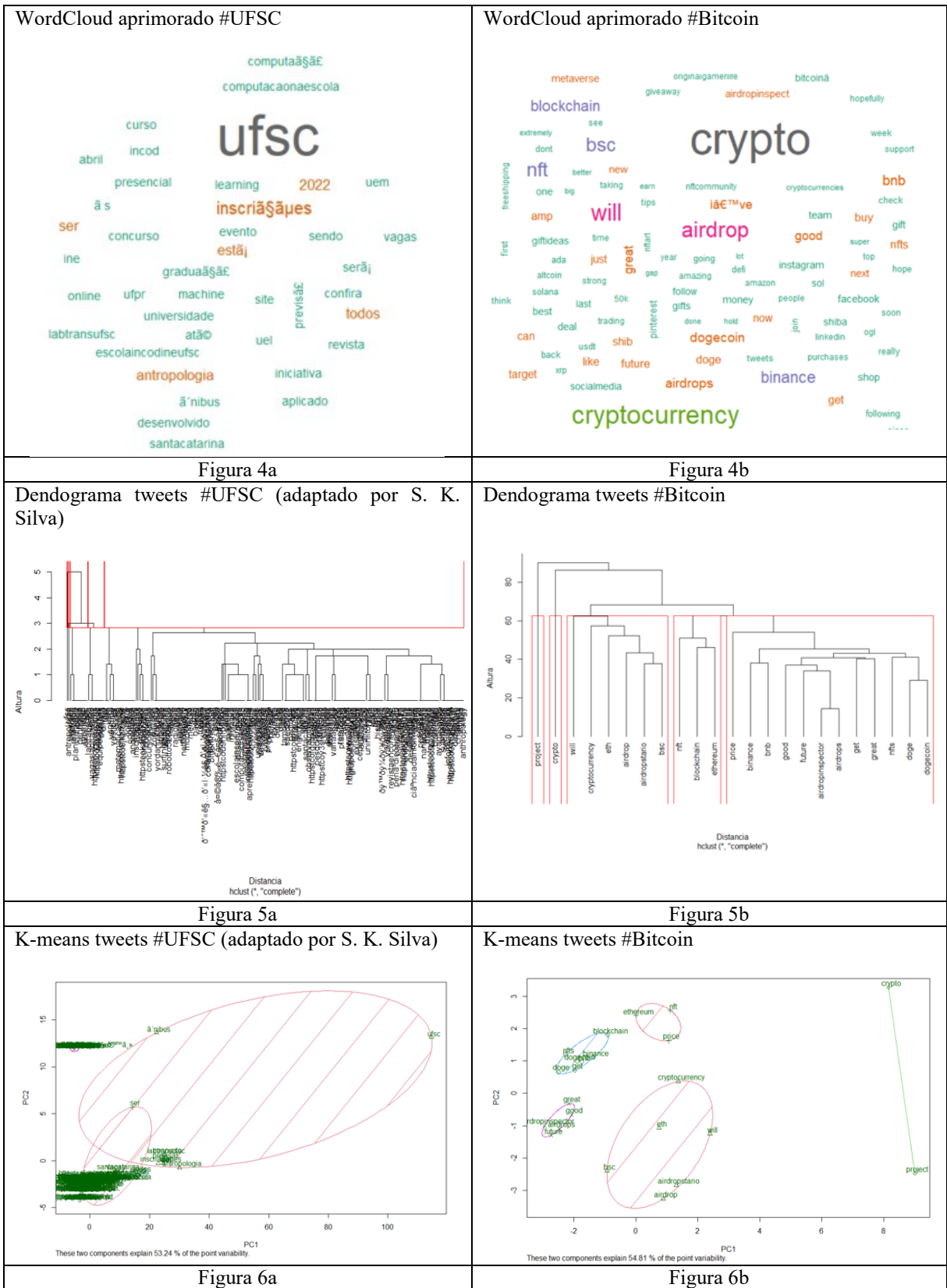
Portanto, em 05/12/21 foram buscados 15 mil *tweets* que apresentavam o texto “UFSC”, sem considerar *re-tweets* e sendo criados na língua portuguesa. Já para a aplicação *Bitcoin* foram buscados 18 mil *tweets* e foi considerada a língua inglesa.

Foram extraídas as partes de texto (uso de variável `_Tweet$text`) e iniciado o processo de análise.

Quadro 41 - Resultados obtidos



Quadro 42 - Resultados obtidos



A figura 6a (assim como a 5a) reafirmam a baixa utilização dessa ferramenta. Por outro lado, as figuras nomeadas com figura xb mostram a utilidade na obtenção de padrões e características a partir da análise de *tweets*.

Os resultados da aplicação de mineração de textos na rede social *Twitter*, levantaram a hipótese de a ferramenta ser utilizada mais para replicar mensagens do que para gerar textos. Assim, conclui-se que há poucos textos gerados úteis para o processo de tomada de decisão e Gestão Universitária.

7 CONCLUSÃO E TRABALHOS FUTUROS

Esta pesquisa teve por objetivo avaliar o uso de informações não estruturadas no processo de descoberta do conhecimento e de tomada de decisão. Para tanto, foi utilizada uma ferramenta de mineração de textos e realizado um estudo de caso no Portal de Atendimento Institucional da Pró-Reitoria de Extensão da Universidade Federal de Santa Catarina e uma busca de informações na rede social *Twitter*.

Os constantes avanços tecnológicos e o grande volume de dados armazenados em formato eletrônico, principalmente na forma de textos, linguagem natural, têm impulsionado diversos estudos voltados ao tratamento e transformação de informações em conhecimento.

Dentre as diversas ferramentas de Gestão do Conhecimento, a mineração de textos destaca-se quando o objeto de análise envolve bancos de informações textuais, em linguagem natural, pois é capaz de extrair conhecimento potencialmente relevante, muitas vezes desconhecido, de uma grande quantidade de dados não estruturados, como no caso do PAI desta pesquisa.

A partir da aplicação da ferramenta de mineração de textos no PAI/PROEX foi possível extrair informações potencialmente relevantes para a Gestão Universitária e que foram bem exploradas pelas técnicas de mineração de textos.

A identificação de informações relevantes armazenadas no PAI forneceu subsídios importantes aos gestores para a tomada de decisão e para a melhoria da qualidade da comunicação entre os usuários e a organização.

Para a seleção da ferramenta de mineração de textos utilizou-se os seguintes critérios: ferramenta gratuita, para mineração de textos em língua portuguesa, não oriunda de empresa comercial, com manual de aplicação, entrada de dados com importação de arquivos, saída de dados com exportação de arquivos, com visualização gráfica, sem exigência de conhecimento computacional e com o maior número de técnicas disponíveis. A ferramenta KH Coder, utilizada nesta pesquisa, permitiu a aplicação de diferentes técnicas de mineração de textos e a elaboração de mapas temáticos que se constituem na base da criação de ontologias.

Foram aplicadas as seguintes técnicas: (i) rede de coocorrência de palavras, (ii) rede de coocorrência de palavras e variáveis, (iii) análise por cluster, (iv) análise hierárquica de cluster, (v) mapa auto-organizável e (vi) classificador Naive Bayes. Com os resultados obtidos, pode-se concluir que o sistema PAI/PROEX tem informações relevantes para a Gestão Universitária. Importante ressaltar que cada técnica permitiu diferentes visualizações e aplicabilidades, permitindo uma ampla análise dos dados nesta pesquisa.

A aplicação de diferentes técnicas de mineração de textos possibilitou a descoberta de conhecimento importante como: a identificação de 30 tópicos perguntas e 29 tópicos de respostas; a identificação dos tópicos mais frequentes, sendo eles: aprovação de programas e projetos; cursos de curta duração; prestação de serviço, serviço eventual e consultoria; publicações, artigos científicos, revisão de artigos, revistas.

Outra importante descoberta foi a referência de artigos da Resolução Normativa nº 88/2016/CUn em cerca de 1/5 das respostas. Os artigos mais referenciados foram o 8º, que trata do registro e aprovação das ações de extensão, e o 12, que trata dos prazos para preenchimento do relatório final e aprovação pelo coordenador de extensão do departamento.

A análise de *tweets* sobre a UFSC na rede social *Twitter* apontou que a ferramenta é utilizada mais para replicar mensagens do que para gerar textos, assim, neste caso não foram identificadas informações relevantes que possam ser úteis na tomada de decisão. Contudo, as redes sociais têm recebido bastante atenção das organizações devido ao número de pessoas que tem acesso as mesmas e a quantidade de interesses e opiniões postados diariamente.

Considerando o fato de que a mineração de textos permitiu a descoberta de informações relevantes armazenadas no PAI, e que estas informações podem ser utilizadas pelos gestores na tomada de decisão para melhorar a qualidade da comunicação entre os usuários e a organização, conclui-se que o objetivo desta pesquisa foi alcançado.

Considerando ainda que, esta pesquisa coletou as informações de um determinado período, neste caso de 25/08/2019 a 31/07/2021, quando 2.417 chamados estavam registrados no PAI, porém abordou apenas 600 chamados; e considerando ainda que, na redação da conclusão desta pesquisa, em janeiro de 2022, encontravam-se registrados 3.290 chamados, um aumento de 36% em seis meses; verificou-se a importância de implementação de rotinas de monitoramento e análise para constante adequação das ações de tomadas de decisão.

Assim, para estudos futuros sugere-se a implementação de um sistema automatizado com análise estatística dos chamados. Sugere-se ainda a expansão da pesquisa para todos os chamados do PAI/PROEX e a aplicação em PAI de outros setores, projetando o estudo para toda a Universidade.

REFERÊNCIAS

- ABUBAKAR, Abubakar Mohammed *et al.* Knowledge management, decision-making style and organizational performance. **Journal of Innovation & Knowledge**, [s. l.], v. 4, n. 2, p. 104–114, 2019.
- ADWAN, Omar Y. *et al.* Twitter sentiment analysis approaches: A survey. **International Journal of Emerging Technologies in Learning**, [s. l.], v. 15, n. 15, p. 79–93, 2020. Disponível em: <https://doi.org/10.3991/ijet.v15i15.14467>
- ALAVI, M.; LEIDNER, D.E. Review: Knowledge Management and Knowledge Management Systems: Conceptual Foundations and Research Issues. **Management Information Systems Research Center**, [s. l.], v. 25, n. 1, p. 107–136, 2001.
- ALVES, Elton; LEAL, Adonis. Application of Kohonen self-organizing map for clustering negative cloud-to-ground lightning electric field waveforms. **IEEE Latin America Transactions**, [s. l.], v. 19, n. 11, p. 1959–1966, 2021. Disponível em: <https://doi.org/10.1109/TLA.2021.9475850>
- ANDRADE, Patrícia Helena Maia Alves. **Universidade de Brasília classificação de documentos : um estudo da automatização da triagem de denúncias na CGU**. 65 f. 2015. - Universidade de Brasília, [s. l.], 2015.
- BELLENZIER, Marina. **Mineração de dados no portal de atendimento e URA da FOCCO**. 93 f. 2013. - Universidade de Caxias do Sul, [s. l.], 2013.
- BRYLSBAERT, Marc; MANDERA, Paweł; KEULEERS, Emmanuel. The word frequency effect in word processing: an updated review. **Current Directions in Psychological Science**, [s. l.], v. 27, n. 1, p. 45–50, 2018. Disponível em: <https://doi.org/10.1177/0963721417727521>
- CALAZANS, Angélica Toffano Seidel. Qualidade da informação: conceitos e aplicações. **Transinformação**, [s. l.], v. 20, n. 1, p. 29–45, 2008. Disponível em: <https://doi.org/10.1590/s0103-37862008000100003>
- CAMILO, Cássio Oliveira. **Uma metodologia para mineração de regras de associação usando ontologias para integração de dados estruturados e não-estruturados**. 148 f. 2010. - Universidade Federal de Goiás, [s. l.], 2010.
- CAMPELO, Elaine Martins *et al.* Gestão do conhecimento e gestão de projetos como ferramentas complementares na aprendizagem organizacional. **Tópicos em Administração Volume 28**, [s. l.], p. 8, 2020.
- CARVALHO, Jonnathan; PLASTINO, Alexandre. **On the evaluation and combination of state-of-the-art features in Twitter sentiment analysis**. [S. l.]: Springer Netherlands, 2021. ISSN 15737462.v. 54 Disponível em: <https://doi.org/10.1007/s10462-020-09895-6>
- CASSARO, Antonio Carlos. **Sistemas de informações para tomada de decisões**. São Paulo: Pioneira, 1995.
- CAVALCANTI, Marcelo. A Internet como ferramenta na produção e gestão do conhecimento organizacional. **RECIMA21-Revista Científica Multidisciplinar-ISSN 2675-6218**, [s. l.], v. 1, n. 2, p. 8–20, 2020.
- CHAVES, Maurício Silveira. **Mapeamento e comparação de similaridade entre estruturas ontológicas**. 119 f. 2004. - Pontífica Universidade Católica do Rio Grande do Sul, [s. l.], 2004.
- CHIAVENATO, Idalberto. **Introdução à teoria geral da administração: uma visão**

abrangente da moderna administração das organizações. 7. ed. Rio de Janeiro: Elsevier, 2003.

CHISTOL, Mihaela. A Comparative study of parametric versus non-parametric text classification algorithms. *In:* , 2020, Suceava, Romania. **15th International Conference on Development and Application Systems.** Suceava, Romania: [s. n.], 2020. p. 208–213.

CHOO, Chun Wei. **A Organização do conhecimento: como as organizações usam a informação para criar significado, construir conhecimento e tomar decisões.** Oxford: Oxford University, 2003.

COSTA, Claudio Napolis *et al.* Descoberta de conhecimento em bases de dados. **Revista Eletrônica: Faculdade Santos Dumont**, [s. l.], v. 2, p. 20, 2019.

DANTAS, EB. **A importância da pesquisa para a tomada de decisões.** [S. l.: s. n.], 2013. Disponível em: [https://doi.org/ISSN: 1646-3137](https://doi.org/ISSN:1646-3137)

DIAS, Júlia de Figueiredo Pinheiro. **Big Data e direito da concorrência: a concorrência no mercado de dados pessoais à luz do RGPD.** 2021. - Universidade de Lisboa, [s. l.], 2021.

FAHEY, Liam; PRUSAK, Laurence. The eleven deadliest sins of knowledge management. **California Management Review**, [s. l.], n. 3, p. 265–276, 1998. Disponível em: <https://doi.org/10.2307/41165954>

FAYYAD, UM; PIATETSKY-SHAPIRO, G; SMYTH, P. Knowledge discovery and data mining: towards a unifying framework. **Int Conf on Knowledge Discovery and Data Mining**, [s. l.], p. 82–88, 1996. Disponível em: <http://www.aaai.org/Papers/KDD/1996/KDD96-014>

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From data mining to knowledge discovery in databases. **AI Magazine**, [s. l.], v. 17, n. 3, p. 37–53, 1996.

FERREIRA, Francieli Aparecida; MOURA, Fábio Longo de; BARROS, Victor Freitas de Azerêdo. Avaliação Da Qualidade Da Informação: Um Estudo De Caso. **Proceedings of International Conference on Engineering and Technology Education**, [s. l.], v. 13, n. 0, 2014. Disponível em: <https://doi.org/10.14684/intertech.13.2014.467-471>

FERREIRA, Márcio Henrique Wanderley; CORREA, Renato Fernandes. Mineração de textos científicos: análise de artigos de periódicos científicos brasileiros da área de Ciência da Informação. **Em Questão**, [s. l.], v. 27, n. 1, p. 237–262, 2021. Disponível em: <https://doi.org/10.19132/1808-5245271.237-262>

GUPTA, Kriti. **Fake News Analysis and Graph Classification on a COVID-19 - Twitter Dataset.** 51 f. 2021. - San José State University, [s. l.], 2021. Disponível em: <https://doi.org/10.1109/bigdataservice52369.2021.00013>

HASSANI, Hossein *et al.* Text mining in big data analytics. **Big Data and Cognitive Computing**, [s. l.], v. 4, n. 1, p. 1–34, 2020. Disponível em: <https://doi.org/10.3390/bdcc4010001>

HIGUCHI, Koichi. **KH Coder 3 Reference Manual.** [S. l.: s. n.], 2017.

LAROSE, D.T. **Discovering knowledge in data: an introduction to data mining.** [S. l.]: John Wiley and Sons, 2005.

LOH, Stanley; OLIVEIRA, José Palazzo M; GAMEIRO, Mauricio A. Knowledge discovery in texts for constructing decision support systems. **Applied Intelligence**, [s. l.], v. 18, p. 357–366, 2003. Disponível em: <https://doi.org/10.1023/A:1023258306854>

LOH, Stanley; WIVES, Leandro Krug; OLIVEIRA, José Palazzo M. Concept-based

knowledge discovery in texts extracted from the Web. **ACM SIGKDD Explorations Newsletter**, [s. l.], v. 2, n. 1, p. 29–39, 2000a. Disponível em: <https://doi.org/10.1145/360402.360414>

LOH, Stanley; WIVES, Leandro Krug; OLIVEIRA, José Palazzo M de. Descoberta proativa de conhecimento em coleções textuais: aplicações em inteligência competitiva. *In:* , 2000b. **IV Oficina de Inteligência Artificial**. [S. l.: s. n.], 2000. p. 1–16.

LOPES, Humberto Elias Garcia; GOSLING, Marlusa de Sevilha. Cluster analysis in practice : dealing with outliers in managerial research. **Revista de Administração Contemporânea**, [s. l.], v. 25, n. 1, p. 1–19, 2021. Disponível em: <https://doi.org/10.1590/1982-7849rac2021200081>

MACHADO, Aydano *et al.* Mineração de texto em redes sociais aplicada à educação a distância. **Colabor@ - A Revista Digital da CVA-RICESU**, [s. l.], v. 6, 2010.

MARTÍN, Angel; JULIÁN, Ana Belén Anquela; COS-GAYÓN, Fernando. Analysis of Twitter messages using big data tools to evaluate and locate the activity in the city of Valencia (Spain). **Cities**, [s. l.], v. 86, p. 37–50, 2019.

MONTEIRO, Luciana Lopes. **Mensagens textuais no canal de atendimento do portal IBGEANDO: obtendo insumos para a tomada de decisão utilizando mineração de textos**. 164 f. 2017. - Universidade Federal Fluminense, [s. l.], 2017.

MORAIS, Edison Andrade Martins; AMBRÓSIO, Ana Paula L. **Mineração de textos: relatório técnico**. Goiania: [s. n.], 2007a.

MORAIS, Edison Andrade Martins; AMBRÓSIO, Ana Paula L. **Mineração de Textos**. [S. l.: s. n.], 2007b.

NASEEM, Usman *et al.* **Transformer based deep intelligent contextual embedding for Twitter sentiment analysis**. [S. l.: s. n.], 2020. Disponível em: <https://doi.org/10.1016/j.future.2020.06.050>

NONAKA, Ikujiro; KONNO, Noboru. The concept of “Ba”: building a foundation for knowledge creation. **California Management Review**, [s. l.], v. 40, n. 3, p. 40–54, 1998.

OJO, O. O.; AKINNULI, B. O.; FARAYIBI, P. K. Data mining and statistical analysis for available budget allocation pre-procurement of manufacturing equipment. **Journal of Engineering Research and Reports**, [s. l.], v. 5, n. 3, p. 1–13, 2019. Disponível em: <https://doi.org/10.9734/jerr/2019/v5i316926>

OLETO, Ronaldo Ronan. Percepção da qualidade da informação. **Ciência da Informação**, [s. l.], v. 35, n. 1, p. 57–62, 2006. Disponível em: <https://doi.org/10.1590/s0100-19652006000100007>

OLIVEIRA, Denis Renato *et al.* Gestão Do Conhecimento, Cultura Organizacional E Gestão De Pessoas Com a Gestão De Processos E Questões Organizacionais Emergentes: Uma Análise Crítica Da Dinâmica Subjetiva Em Gestão Por Processos (Bp). **Revista Gestão em Análise**, [s. l.], v. 9, n. 1, p. 154, 2020. Disponível em: <https://doi.org/10.12662/2359-618xregea.v9i1.p154-167.2020>

PERROCA, Márcia Galan; GAIDZINSKI, Raquel Rapone. Avaliando a confiabilidade interavaliadores de um instrumento para classificação de pacientes--coeficiente kappa. **Revista da Escola de Enfermagem da USP**, [s. l.], v. 37, n. 1, p. 72–80, 2003. Disponível em: <https://doi.org/10.1590/s0080-62342003000100009>

PEZZINI, Anderson. Mineração de textos: conceito, processo e aplicações. **Revista Eletrônica**

do Alto Vale do itajaí, [s. l.], v. 5, n. 8, p. 58–61, 2016. Disponível em: <https://doi.org/10.5965/2316419005082016058>

ROBBINS, Stephen P. **Comportamento Organizacional**. 11. ed. São Paulo: Pearson Prentice Hall, 2005.

ROBREDO, Jaime; CUNHA, Murilo Bastos da. Aplicação de técnicas infométricas para identificar a abrangência do léxico básico que caracteriza os processos de indexação e recuperação da informação. **Ciência da Informação**, [s. l.], v. 27, n. 1, p. 11–27, 1998. Disponível em: <https://doi.org/10.1590/s0100-19651998000100003>

SABINO, M. M. F. L. **Diretrizes estratégicas para o compartilhamento do Conhecimento Tradicional visando a sustentabilidade cultural: um estudo de caso do projeto Ilha Rendada**. 2019. - Universidade Federal de Santa Catarina, [s. l.], 2019.

SANTOS, Neri; RADOS, Gregório Jean Varvakis. **Fundamentos Teóricos da Gestão do Conhecimento**. [S. l.: s. n.], 2020.

SANTOS, Ronnie E S *et al.* Técnicas de processamento de linguagem natural aplicadas ao processo de mineração de textos: resultados preliminares de um mapeamento sistemático. **Revista de Sistemas e Computação**, [s. l.], v. 4, p. 116–125, 2014.

SERAPIÃO, Paulo Roberto Barbosa; SUZUKI, Kátia Mitiko Firmino; MARQUES, Paulo Mazzoncini de Azevedo. Uso de mineração de texto como ferramenta de avaliação da qualidade informacional em laudos eletrônicos de mamografia. **Radiologia Brasileira**, [s. l.], v. 43, n. 2, p. 103–107, 2010. Disponível em: <https://doi.org/10.1590/s0100-39842010000200010>

SERRAT, Olivier. Notions of knowledge management. **Knowledge solutions**, [s. l.], v. 18, n. November, p. 1–11, 2008. Disponível em: <https://doi.org/10.1080/13600860410001674733>

SHRIHARI R, Chauhan; DESAI, Amish. A review on knowledge discovery using text classification techniques in text mining. **International Journal of Computer Applications**, [s. l.], v. 111, n. 6, p. 12–15, 2015. Disponível em: <https://doi.org/10.5120/19542-0784>

SHUHAI, Fan *et al.* Overview of comprehensive information quality management under mass customization. *In:* , 2019. **International Computer Science and Applications Conference**. [S. l.: s. n.], 2019. p. 42–45. Disponível em: <https://doi.org/10.33969/eecs.v2.010>

SILVA, Marcio Ponciano da; VIERA, Angel Freddy Godoy. Descoberta de conhecimento com uso de técnicas de mineração de textos aplicadas em documentos textuais da investigação policial brasileira. **Investigación bibliotecológica**, [s. l.], v. 35, n. 88, p. 161–183, 2021.

SILVA, Edilberto Magalhães. **Descoberta de conhecimento com o uso de text mining: cruzando o Abismo de Moore**. 174 f. 2002. - Universidade Católica de Brasília, [s. l.], 2002. Disponível em: http://textmining.xpg.uol.com.br/Dissertacao_Edilberto.pdf

SILVA, Thales N. **Uma arquitetura para descoberta de conhecimento a partir de bases textuais**. 78 f. 2012. - Universidade Federal de Santa Catarina, [s. l.], 2012.

TAKEUCHI, Hirotaka; NONAKA, Ikujiro. **Gestão do Conhecimento [recurso eletrônico]**. Porto Alegre: Bookman, 2008.

TUOMI, Ilkka. Data Is More than Knowledge: Implications of the Reversed Knowledge Hierarchy for Knowledge Management and Organizational Memory. **Journal of Management Information Systems**, [s. l.], v. 16, n. 3, p. 103–117, 1999. Disponível em: <http://www.jstor.org/stable/40398446>

URIARTE, Filemon A. **Introduction to knowledge management**. Jakarta: Asean Foundation, 2008. *E-book*.

VALE, Marcos Neve. **Agrupamentos de dados: avaliação de métodos e desenvolvimento de aplicativo para análise de grupos**. 2005. - Pontifícia Universidade Católica Do Rio De Janeiro, [s. l.], 2005.

VICENTE, Ana Carolina Galvão; DA CUNHA, Pedro Henrique Braz. Gestão do Conhecimento: Como Facilitar o Compartilhamento de Conhecimento em Equipes Remotas. **Boletim do Gerenciamento**, [s. l.], v. 24, n. 24, p. 50–60, 2021.

WANG, Richard Y; STRONG, Diane M. Beyond accuracy: what data quality means to data consumer. **Journal of Management Information Systems**, [s. l.], v. 12, n. 4, p. 5–33, 1996.

WIVES, Leandro Krug. **Um estudo sobre agrupamento de documentos textuais em processamento de informações não estruturadas usando técnicas de Clustering**. 84 f. 1999. - Universidade Federal do Rio Grande do Su, [s. l.], 1999.

WOLSKI, Marcin; GOMOLIŃSKA, Anna. Data meaning and knowledge discovery: Semantical aspects of information systems. **International Journal of Approximate Reasoning**, [s. l.], v. 119, p. 40–57, 2020.

ZHU, Hongwei *et al.* Data and information quality research: Its evolution and future. **Computing Handbook, Third Edition: Information Systems and Information Technology**, [s. l.], p. 16-1-16–20, 2014. Disponível em: <https://doi.org/10.1201/b16768>

Apêndice 1

Lista de nomes de referência à legislação padronizadas na preparação dos dados:

- CF1988 = Constituição Federal de 1988
- D31971999 = Decreto Federal nº 3.197 de 1999
- D74162010 = Decreto Federal nº 7.416 de 2010
- D74232010 = Decreto Federal nº 7.423 de 2010
- D103152020 = Decreto Federal nº 10.315 de 2020
- D104262020 = Decreto Federal nº 10.426 de 2020
- L81921990 = Lei nº 8.192 de 1990
- L127722012 = Lei nº 12.772 de 2012
- L128632013 = Lei nº 12.863 de 2013
- L130052014 = Lei nº 13.005 de 2014
- L132432016 = Lei nº 13.243 de 2016
- MC102017PROEX = Memorando Circular nº 10/2017/PROEX
- OC422020PROEX = Ofício Circular nº 42/2020/PROEX
- RN92010 = Resolução Normativa nº 9/2010/CUn
- RN122011 = Resolução Normativa nº 12/CUn/2011
- RN132011 = Resolução Normativa nº 13/CUn/2011
- RN672015 = Resolução Normativa nº 67/2015/CUn
- RN1142017 = Resolução Normativa nº 114/CUn/2017
- RN72018 = Resolução Normativa nº 7/2018/CUN
- RN631019 = Resolução Normativa nº 63/2019/CPG

Apêndice 2

Lista de nomes de referência para os endereços de páginas da WEB padronizados na preparação dos dados:

- atividadesdocentes = <https://proex.ufsc.br/atividades-docentes-no-sigpex/>
- cartilha = https://proex.paginas.ufsc.br/files/2019/04/CARTILHA-PERGUNTAS-PROEX_web_i.pdf
- conveniosproad = <http://dpc.proad.ufsc.br/convenios/>
- cursocoordenadores = <https://proex.ufsc.br/curso-de-formacao-de-coordenadores-de-extensao/>
- documentorenex = <https://www.ufmg.br/proex/renex/images/documentos/Organizacao-e-Sistematizacao.pdf>
- movimentacoesfinanceiras = <https://proex.ufsc.br/movimentacoes-financeiras/>
- orientaçõesatividadesdocentes = <https://proex.ufsc.br/outras-orientacoes-sobre-atividades-docentes/>
- orientaçõesgerais = <https://proex.ufsc.br/orientacoes-gerais/>
- resoluçãoextensão = https://proex.ufsc.br/files/2016/11/Resolu%C3%A7%C3%A3oNormativa_88_Extens%C3%A3o.pdf
- sistemacertificados = <https://certificados.ufsc.br/>
- sistemasigpex = <https://sigpex.sistemas.ufsc.br/>
- sucupiracapes = <https://sucupira.capes.gov.br/sucupira/public/consultas/coleta/veiculoPublicacaoQualis/listaConsultaGeralPeriodicos.jsf>
- tramitafacil = <https://tramitafacil.ufsc.br/>

Apêndice 3

Lista de palavras de parada (*stopwords*):

a, à, agora, ainda, alguém, algum, alguma, algumas, alguns, ampla, amplas, amplo, amplos, ano, ante, antes, ao, aos, apenas, após, aquela, aquelas, aquele, aqueles, aquilo, as, até, através, cada, coisa, coisas, com, como, contra, contudo, da, daquele, daqueles, das, de, dela, delas, dele, deles, depois, dessa, dessas, desse, desses, desta, destas, deste, deste, destes, deve, devem, devendo, dever, deverá, deverão, deveria, deveriam, devia, deviam, dia, disse, disso, disto, dito, diz, dizem, do, dos, duvida, dúvida, e, é, e', ela, elas, ele, eles, em, enquanto, entre, era, essa, essas, esse, esses, esta, está, estamos, estão, estas, estava, estavam, estávamos, este, estes, estou, extensão, eu, fazendo, fazer, feita, feitas, feito, feitos, foi, for, foram, fosse, fossem, grande, grandes, há, http, isso, isto, já, la, lá, lhe, lhes, lo, mas, me, mês, mesma, mesmas, mesmo, mesmos, meu, meus, minha, minhas, muita, muitas, muito, muitos, na, não, nas, nem, nenhum, nessa, nessas, nesta, nestas, ninguém, no, nos, nós, nossa, nossas, , osso, nossos, num, numa, nunca, o, os, ou, outra, outras, outro, outros, para, pela, pelas, pelo, pelos, pequena, pequenas, pequeno, pequenos, per, perante, pode, pôde, podendo, poder, poderia, poderiam, podia, podiam, pois, por, porém, porque, posso, pouca, poucas, pouco, poucos, primeiro, primeiros, problema, própria, próprias, próprio, próprios, quais, qual, quando, quanto, quantos, que, quem, são, se, seja, sejam, sem, sempre, sendo, será, serão, seu, seus, si, sido, sim, só, sob, sobre, sua, suas, talvez, também, tampouco, te, tem, tendo, tenha, ter, teu, teus, ti, tido, tinha, tinham, toda, todas, todavia, todo, todos, tu, tua, tuas, tudo, última, últimas, último, últimos, um, uma, umas, uns, vendo, ver, vez, vindo, vir, vos, vós.

Apêndice 4

Cálculo de relevância dos termos para a planilha RESPOSTAS

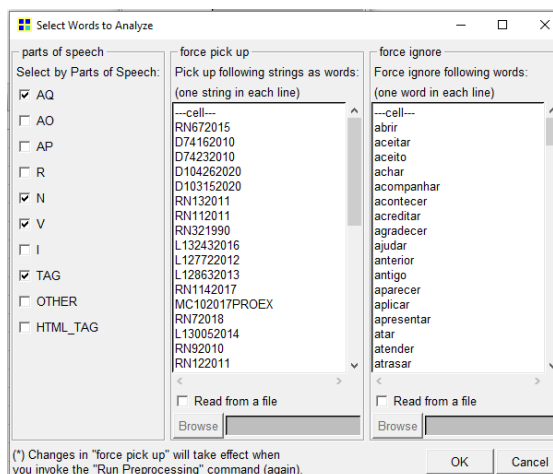
Utilizando a técnica WORD FREQUENCY LIST (WFL) do KH Coder, realizou-se a extração das palavras da planilha RESPOSTAS. A técnica gerou uma lista em ordem decrescente de frequência de palavras contendo os seguintes dados: palavra, classificação gramatical e número de vezes que a palavra aparece nos textos (frequência).

O resultado da WFL foi extraído para análise em planilha Excel e por meio de uma leitura dos resultados identificou-se palavras de menor importância para o objeto deste estudo.

As palavras identificadas foram relacionadas no software na técnica FORCE IGNORE (Quadro 43). Conforme a preparação dos dados, a lista de nomes de referência à legislação e páginas da WEB foram relacionados no software na técnica FORCE PICK UP (Quadro 43).

O software utiliza a seguinte classificação de palavras: adjetivos (AQ), numerais ordinais (AO) advérbios, preposições e conjunções (R), substantivos (N), verbos (V) e interjeição (I). Foram subtraídas da análise palavras das classes gramaticais advérbios, preposições e conjunções (R) e interjeição (I).

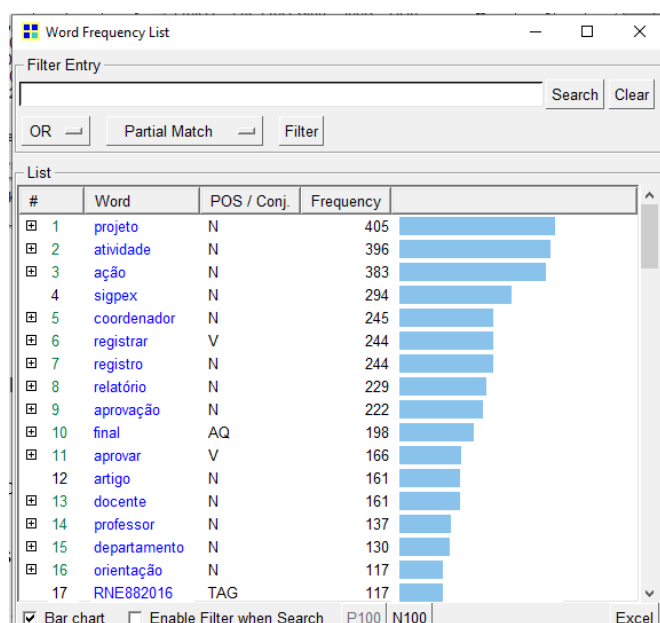
Quadro 43 - Seleção das palavras para análise - planilha RESPOSTAS



Fonte: Autora (2021) - KH Coder

Definida a seleção de palavras de análise, foi gerada uma nova lista de frequência de palavras (Quadro 44).

Quadro 44 - Lista de frequência de palavras - planilha RESPOSTAS



Fonte: Autora (2021) - KH Coder

O resultado da lista de frequência de palavras foi extraído para análise em planilha Excel (Quadro 45) para nova conferência.

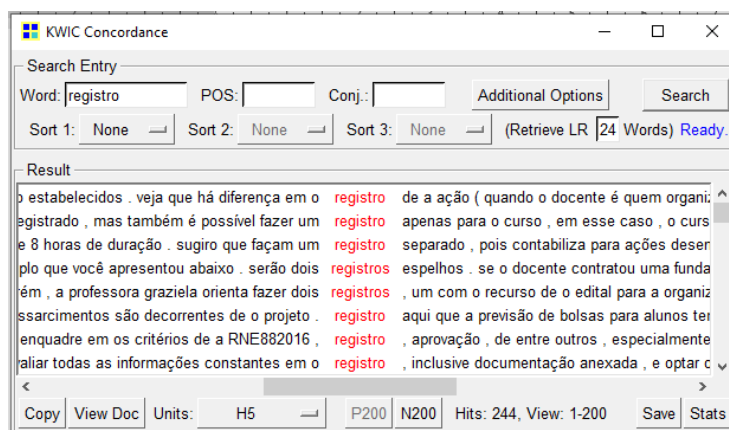
Quadro 45 - Dados WFL da planilha RESPOSTAS em Excel

	A	B	C	D	E	F	G	H
1	AQ		N		V		TAG	
2	final	198	projeto	405	registrar	244	RNE882016	117
3	docente	110	atividade	396	aprovar	166	movimentacaoe	11
4	horário	74	ação	383	anexar	83	RN1142017	10
5	financeiro	57	sigpex	294	receber	82	cartilha	6
6	externo	42	coordenador	245	preencher	77	L132432016	6
7	semanal	33	registro	244	encerrar	73	conveniosproac	5
8	eventual	31	relatório	229	alterar	62	L127722012	5
9	participante	26	aprovação	222	emitir	59	L128632013	5
10	institucional	25	artigo	161	enviar	44	atividadesdoce	4
11	efetivo	22	docente	161	incluir	33	D104262020	4
12	anexo	21	professor	137	cancelar	32	RN132011	3
13	científico	21	departamento	130	certificar	30	resoluçãooexter	2
14	responsável	20	orientação	117	vincular	28	sistemasigpex	2
15	servidor	20	hora	113	cadastrar	26	tramitafacil	2

Fonte: Autora (2021)

Através da técnica KWIC - Concordance, as palavras extraídas na WFL puderam ser analisadas dentro do seu contexto e sua concordância textual, esse recurso possibilitou uma conferência da classificação das palavras (Quadro 46).

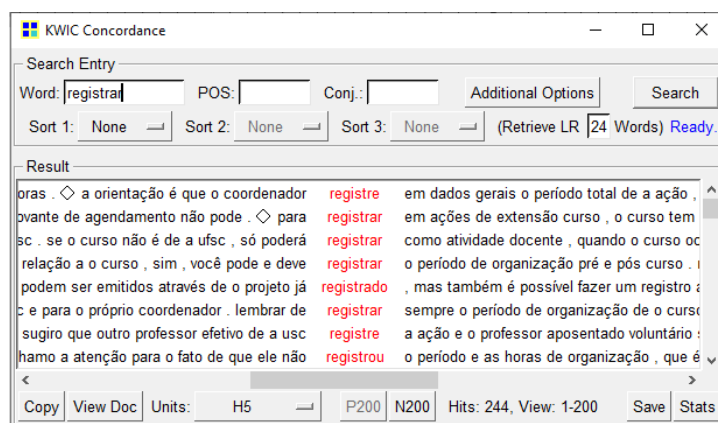
Quadro 46 - Concordância textual da palavra “registro” - planilha RESPOSTAS



Fonte: Autora (2021) - KH Coder

A mesma técnica permitiu checar o processo automático de *Stemming* do software (Quadro 47).

Quadro 47 - Stemming do verbo “registrar” - planilha RESPOSTAS

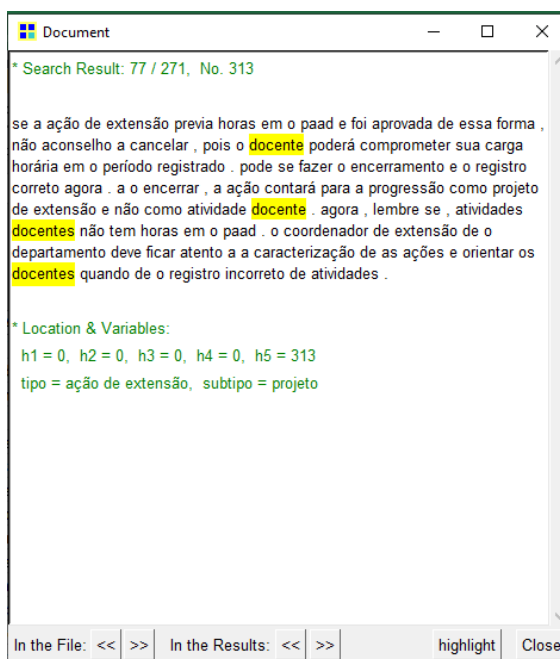


Fonte: Autora (2021) - KH Coder

Na conferência da extração e classificação das palavras, foram analisadas 30 palavras de forma aleatória utilizando a técnica KWIC - Concordance. Verificou-se que a extração e classificação das palavras foi satisfatória na amostra utilizada, bem como o processo de *stemming* automático.

Confirmou-se a identificação de palavras classificadas em duas classes gramaticais na WFL. Como no caso da palavra DOCENTE, que aparece como substantivo e adjetivo (Quadro 48).

Quadro 48 - Checagem da classificação da palavra “docente” - planilha RESPOSTAS



Fonte: Autora (2021) - KH Coder

Apêndice 5

Mineração por meio das técnicas do KH Coder – planilha PERGUNTAS.**Análise de rede de coocorrência de palavras e variáveis.**

a) Variável PROGRAMA

Por meio da concordância textual da variável PROGRAMA (Quadro 49) e a estatística de localização (Quadro 50). Apresenta-se a seguinte análise de resultados:

Quadro 49 - Concordância textual palavra “programa”-
planilha PERGUNTAS

The screenshot shows the KWIC Concordance window with the search entry 'programa'. The results display several lines of text with the word 'programa' highlighted in red. The text includes phrases like 'de extensão, mas sim para os projetos indi', 'com várias ações vinculadas, a ação de e', 'de extensão deve possuir sua conta individu', 'que os recursos de o programa foram utiliza', 'foram utilizados para a ação específica ? < >', 'de extensão que tinha como prazo final 31 1', 'de extensão entre o design e o hu, ricardo t', 'de extensão (hu+d), para que os projetos', 'de extensão, a coordenação necessariame', 'de extensão em 27 07 2020, contudo, atrin', 'de extensão com financiamento em 10 06 2', 'serviço social e organização popular, direit', and 'de extensão e que geram certificado não der'.

Fonte: Autora (2021) - KH Coder

Quadro 50 - Estatística de localização palavra “programa”
- planilha PERGUNTAS

The screenshot shows the Collocation Stats window for the word 'programa'. It displays a table with the following data:

N	Word	POS	Total	LT	RT	L5	L4	L3	L2	L1	R1	R2	R3	R4	R5	The Score
1	projeto	N	10	6	4	0	2	2	2	0	0	3	0	1	0	3.917
2	criar	V	6	5	1	0	0	1	4	0	1	0	0	0	0	3.333
3	vincular	V	10	6	4	0	1	5	0	0	0	0	0	1	3	2.767
4	pós-graduação	N	5	0	5	0	0	0	0	0	0	5	0	0	0	2.500
5	ação	N	10	8	2	4	3	1	0	0	0	0	0	1	1	2.333
6	referir	V	2	2	0	0	0	0	0	2	0	0	0	0	0	2.000
7	virar	V	2	2	0	0	0	0	0	2	0	0	0	0	0	2.000
8	serviço	N	2	0	2	0	0	0	0	0	1	1	0	0	0	1.500
9	cadastrar	V	4	2	2	0	0	0	2	0	0	0	0	0	2	1.400
10	registrar	V	4	3	1	1	0	1	1	0	0	0	1	0	0	1.367
11	coordenador	N	4	2	2	0	0	2	0	0	0	0	2	0	0	1.333
12	institucional	AQ	2	0	2	0	0	0	0	0	1	0	1	0	0	1.333

Fonte: Autora (2021) - KH Coder

O quadro de estatística (Quadro 50) mostra que a palavra “projeto” aparece 10 vezes, sendo 6 vezes em posição à esquerda (LT) da palavra PROGRAMA e 4 vezes em posição à direita (RT). O verbo “criar” aparece 6 vezes, sendo 5 vezes em posição LT e 1 vez em posição RT. O verbo “vincular” aparece 10 vezes, sendo 6 vezes em posição LT e 4 vezes em posição RT. A palavra “ação” aparece 10 vezes, sendo 8 vezes em posição LT e 2 vezes em posição RT. Esta análise revela que as dúvidas mais frequentes são relacionadas ao vínculo de ações ou projetos em programas e a criação de programas.

b) Variável PUBLICAÇÃO

Por meio da concordância textual da variável PUBLICAÇÃO (Quadro 51) e a estatística de localização (Quadro 52). Apresenta-se a seguinte análise de resultados:

Quadro 51 - Concordância textual palavra “publicação” - planilha PERGUNTAS

Search Entry
Word: publicação POS: Conj.: Additional Options Search
Sort 1: None Sort 2: None Sort 3: None (Retrieve LR 24 Words) Ready.

Result

for um ou outro . ◊ a o tentar incluir duas **publicações** que foram publicadas em o semestre pass
de tirar uma dúvida quanto a o registro de **publicações** (revisão de artigo científico para eventos e
de saber como proceder para cadastro de **publicações** com retratatividade , para fins de progressã
fins de progressão . gostaria de cadastrar **publicações** para que contem para minha progressão at
e para o registro de as atividades de o tipo **publicações** , o sigpex exhibe a seguinte mensagem , pri
tovo fluxo em registro de atividade docente **publicações** , a partir de agora após o registro de a ativ
relatório final . assim , o comprovante de a **publicação** deverá ser anexado em o momento de o re
envio para aprovação . entretanto , para as **publicações** registradas anteriormente (em duas etapas
es , como por exemplo , para o registro de **publicações** conforme o mês em que o item efetivament
lo , o registro de uma atividade docente de **publicação** poderia ter um prazo diferenciado para as d
a em o sigpex . ◊ a professora registrou a **publicação** de o artigo publicado indicando o qualis cor
curta duração . ◊ o docente registrou uma **publicação** em 23 05 2020 cujo artigo foi publicado em
007331 pode ser registrada em a categoria **publicações** e subcategoria outros . trata se de a publici

Copy View Doc Units: H5 P200 N200 Hits: 111, View: 1-111 Save Stats

Fonte: Autora (2021) - KH Coder

Quadro 52 - Estatística de localização palavra “publicação” - planilha PERGUNTAS

Node Word
Word: publicação POS: Conj.: Hits: 111

N	Word	POS	Total	LT	RT	L5	L4	L3	L2	L1	R1	R2	R3	R4	R5	The Score
1	artigo	N	22	2	20	1	0	0	1	0	1	5	6	6	2	8.100
2	registrar	V	14	8	6	0	1	0	7	0	2	0	1	1	2	6.733
3	data	N	13	8	5	0	0	3	5	0	1	1	2	0	1	5.867
4	registro	N	13	11	2	0	1	1	9	0	0	0	0	1	1	5.533
5	cadastrar	V	8	5	3	0	0	0	4	1	0	0	2	0	1	3.867
6	atividade	N	12	12	0	3	3	3	3	0	0	0	0	0	0	3.850
7	docente	AQ	7	7	0	1	2	0	2	2	0	0	0	0	0	3.700
8	site	N	7	1	6	1	0	0	0	2	1	1	0	2	0	3.433
9	revisão	N	6	2	4	0	2	0	0	0	1	3	0	0	0	3.000
10	científico	AQ	7	0	7	0	0	0	0	0	1	0	3	0	3	2.600
11	opção	N	3	2	1	0	0	0	0	2	0	0	0	1	0	2.250
12	semestre	N	7	5	2	1	1	3	0	0	0	1	0	0	1	2.150
13	categoria	N	2	2	0	0	0	0	0	2	0	0	0	0	0	2.000

Copy Filter Sort: The Score Window span: L5 R5

Fonte: Autora (2021) - KH Coder

O quadro de estatística (Quadro 52) mostra que a palavra “artigo” aparece 22 vezes, sendo 2 vezes em posição à esquerda (LT) da palavra PUBLICAÇÃO e 20 vezes em posição à direita (RT). O verbo “registrar” aparece 14 vezes, sendo 8 vezes em posição LT e 6 vezes em posição RT. A palavra “registro” aparece 13 vezes, sendo 11 vezes em posição LT e 2 vezes em posição RT. A palavra “data” aparece 13 vezes, sendo 8 vezes em posição LT e 5 vezes em posição RT. Esta análise revela que as dúvidas mais frequentes são relacionadas ao registro de publicação de artigo e data de publicação.

c) Variável BANCA

Por meio da concordância textual da variável BANCA (Quadro 53) e a estatística de localização (Quadro 54). Apresenta-se a seguinte análise de resultados:

Quadro 53 - Concordância textual palavra “banca” - planilha PERGUNTAS

Search Entry
Word: banca POS: Conj.: Additional Options Search
Sort 1: None Sort 2: None Sort 3: None (Retrieve LR 24 Words) Ready.

Result

. por exemplo , se seleciono o relatório para **banca** externas , o relatório gerado não contém a inf
gerado não contém a informação de o tipo de **banca** (mestrado , doutorado , etc) . ◊ é possível e
catarina para fazer parte de uma comissão (**banca**) de avaliação de docente cuja tarefa é creden
o referido programa . acontece que em a aba **banca** externas não existe essa possibilidade , apen
mas não existe essa possibilidade , apenas **banca** de concurso público , processo seletivo , mes
estrado , doutorado e tcc . faltam , portanto , **banca** de credenciamento e reconheciment
a programas de pós-graduação , bem como **banca** de seleção para o curso de mestrado e douto
o curso de mestrado e doutorado e também **banca** de avaliação externa de curso de graduação (
terna de curso de graduação (inep mec) ou **banca** externa de (re) credenciamento a a pós-grad
pergunta , onde cadastro essa atividade de **banca** externa após a sua realização . o registro con
o professor registrou a participação em **banca** externa após a sua realização . o registro con
consta com data de início em o dia em que a **banca** ocorreu e término em o dia seguinte (em cor
laração de participação) . a participação em **banca** deve ser registrada antes de o evento ? se sin

Copy View Doc Units: H5 P200 N200 Hits: 44, View: 1-44 Save Stats

Fonte: Autora (2021) - KH Coder

Quadro 54 - Estatística de localização palavra “banca” - planilha PERGUNTAS

Node Word
Word: banca POS: Conj.: Hits: 44

N	Word	POS	Total	LT	RT	L5	L4	L3	L2	L1	R1	R2	R3	R4	R5	The Score
1	externo	AQ	18	0	18	0	0	0	0	0	17	0	1	0	0	17.333
2	participação	N	13	12	1	2	1	1	8	0	0	0	0	1	0	5.233
3	avaliação	N	5	1	4	1	0	0	0	0	3	1	0	0	0	2.033
4	atividade	N	6	5	1	3	0	1	1	0	0	0	0	0	1	1.633
5	exemplo	N	4	3	1	1	0	0	2	0	0	0	1	0	0	1.533
6	tcc	N	3	0	3	0	0	0	0	0	3	0	0	0	0	1.500
7	mestrado	N	4	1	3	1	0	0	0	0	1	0	2	0	0	1.200
8	tese	N	3	0	3	0	0	0	0	0	2	0	0	1	0	1.200
9	palestra	N	3	1	2	0	0	0	1	0	0	2	0	0	0	1.167
10	concurso	N	3	0	3	0	0	0	0	0	1	1	0	1	0	1.033
11	professor	N	4	3	1	1	1	1	0	0	0	0	0	1	0	1.033
12	aba	N	1	1	0	0	0	0	0	1	0	0	0	0	0	1.000

Copy Filter Sort: The Score Window span: L5 R5

Fonte: Autora (2021) - KH Coder

O quadro de estatística (Quadro 54) mostra que a palavra “externo” aparece 18 vezes, sendo 0 vez em posição à esquerda (LT) da palavra BANCA e 18 vezes em posição à direita (RT). Para a palavra “externo”, a posição R1 permite concluir que em 17 vezes forma-se a expressão “banca externa”. A palavra “participação” aparece 13 vezes, sendo 12 vezes em posição LT e 1 vez em posição RT. Esta análise revela que as dúvidas mais frequentes são relacionadas à participação em bancas externas.

d) Variável CURSO

Por meio da concordância textual da variável CURSO (Quadro 55) e a estatística de localização (Quadro 56). Apresenta-se a seguinte análise de resultados:

Quadro 55 - Concordância textual palavra “curso” - planilha PERGUNTAS

The screenshot shows the KWIC Concordance interface. The search entry is 'curso'. The results display several lines of text with the word 'curso' highlighted in red. The text includes phrases like '... (história) e ele me informou que não aparece...', '... de extensão de curta duração (participante)', '... de extensão, o docente terá registro de carga...', '... de curta duração, mas todos são referentes a...', '... de extensão, sim, ele recebe carga horária...', '... de curta duração acima de 30 horas e o siste...', '... com maior carga horária não são contabilizadi...', '... gostaria de saber como cadastrar um cui...', '... de 40 horas em a sigpex. os cursos de curta...', '... de curta duração são até 30h. em nova ati...', '... de curta duração em que participei, nem sem...', '... e por isso não consigo cadastrar a data certa...', '... (limite de um mês anterior). tentei também t...

Fonte: Autora (2021) - KH Coder

Quadro 56 - Estatística de localização palavra “curso” - planilha PERGUNTAS

The screenshot shows the Collocation Stats interface. The word is 'curso' and there are 223 hits. The table below shows the results:

N	Word	POS	Total	LT	RT	L5	L4	L3	L2	L1	R1	R2	R3	R4	R5	The Score
1	curto	AQ	16	0	16	0	0	0	0	0	10	1	5	0	0	6.583
2	ministrar	V	14	9	5	1	0	0	7	1	0	1	0	2	2	6.100
3	realizar	V	9	3	6	0	0	0	1	2	2	0	3	1	0	5.750
4	duração	N	18	0	18	0	0	0	0	0	0	11	2	5	0	5.167
5	registrar	V	14	9	5	1	1	3	4	0	0	2	0	0	3	5.050
6	cadastrar	V	10	9	1	1	1	0	6	1	0	1	0	0	0	4.950
7	atividade	N	12	8	4	1	1	5	1	0	0	2	1	1	0	4.200
8	registro	N	13	10	3	3	2	3	2	0	0	1	0	0	2	4.000
9	oferecer	V	8	3	5	0	0	0	3	0	1	1	1	0	2	3.733
10	evento	N	8	1	7	0	0	0	1	0	0	5	1	1	0	3.583
11	projeto	N	8	4	4	0	1	0	2	1	0	1	1	1	1	3.533
12	docente	AQ	7	5	2	1	0	1	2	1	0	0	0	1	1	2.983
13	docente	N	8	7	1	0	2	2	3	0	0	0	0	0	1	2.867

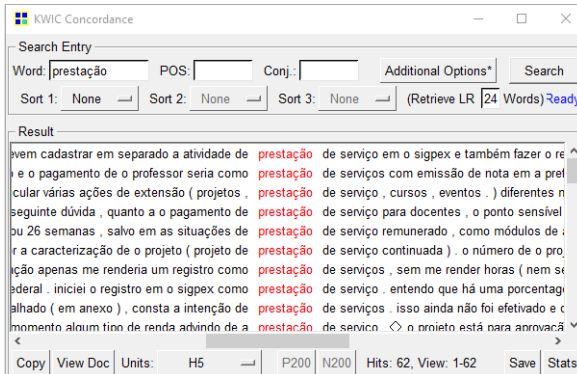
Fonte: Autora (2021) - KH Coder

O quadro de estatística (Quadro 56) mostra que a palavra “curto” aparece 16 vezes, sendo 0 vez em posição à esquerda (LT) da palavra CURSO e 16 vezes em posição à direita (RT). A palavra “duração” aparece 18 vezes, sendo 0 vezes em posição LT e 18 vezes em posição RT. O verbo “ministrar” aparece 14 vezes, sendo 9 vezes em posição LT e 5 vezes em posição RT. O verbo “realizar” aparece 9 vezes, sendo 3 vezes em posição LT e 6 vezes em posição RT. O verbo “registrar” aparece 14 vezes, sendo 9 vezes em posição LT e 5 vezes em posição RT. O verbo “cadastrar” aparece 10 vezes, sendo 9 vezes em posição LT e 1 vezes em posição RT. Esta análise revela que as dúvidas mais frequentes são relacionadas ao registro/cadastro de cursos de curta duração.

e) Variável PRESTAÇÃO DE SERVIÇO

Por meio da concordância textual da variável PRESTAÇÃO DE SERVIÇO (Quadro 57) e a estatística de localização (Quadro 58). Apresenta-se a seguinte análise de resultados:

Quadro 57 - Concordância textual palavra “prestação”, condicional serviço em R2 - planilha PERGUNTAS



Fonte: Autora (2021) - KH Coder

Quadro 58 - Estatística de localização palavra “prestação” - planilha PERGUNTAS

N	Word	POS	Total	LT	RT	L5	L4	L3	L2	L1	R1	R2	R3	R4	R5	The Score
1	serviço	N	73	0	73	0	0	0	0	0	62	3	7	1		33.950
2	docente	AQ	6	6	0	1	0	0	1	4	0	0	0	0	0	4.700
3	atividade	N	9	8	1	1	0	2	5	0	0	0	0	1	0	3.617
4	eventual	AQ	13	0	13	0	0	0	0	0	0	0	3	3	7	3.150
5	cadastrar	V	5	5	0	0	0	0	5	0	0	0	0	0	0	2.500
6	registrar	V	8	7	1	3	2	0	2	0	0	0	0	0	1	2.300
7	projeto	N	5	5	0	1	1	0	3	0	0	0	0	0	0	1.950
8	registro	N	4	4	0	2	0	0	2	0	0	0	0	0	0	1.400
9	sigpex	N	4	3	1	2	0	0	1	0	0	0	0	0	1	1.100
10	categoria	N	1	1	0	0	0	0	1	0	0	0	0	0	0	1.000

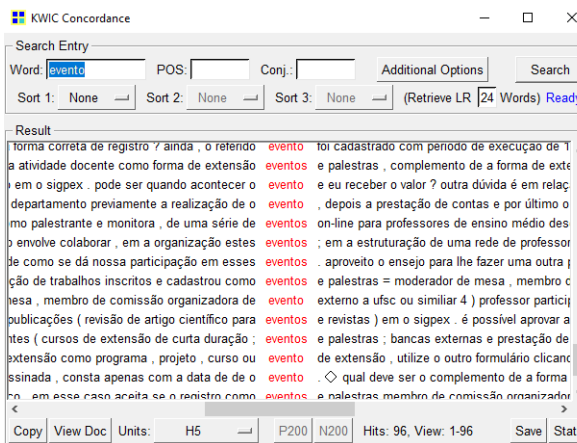
Fonte: Autora (2021) - KH Coder

O quadro de estatística (Quadro 58) mostra que a palavra “serviço” aparece 73 vezes, sendo 73 vezes em posição à esquerda (LT) da palavra PRESTAÇÃO e 0 vez em posição à direita (RT). Para a palavra “serviço”, a posição L2 permite concluir que em 62 vezes formase a expressão “prestação de serviço”. A palavra “eventual” aparece 13 vezes, sendo 0 vez em posição LT e 13 vezes em posição RT. Os verbos “cadastrar” e “registrar” aparecem 5 e 7 vezes em posição LT respectivamente. Esta análise revela que as dúvidas mais frequentes são relacionadas ao registro da atividade docente prestação de serviço, serviço eventual.

f) Variável EVENTO

Por meio da concordância textual da variável EVENTO (Quadro 59) e a estatística de localização (Quadro 60). Apresenta-se a seguinte análise de resultados:

Quadro 59 - Concordância textual palavra “evento” - planilha PERGUNTAS



Fonte: Autora (2021) - KH Coder

Quadro 60 - Estatística de localização palavra “evento” - planilha PERGUNTAS

N	Word	POS	Total	LT	RT	L5	L4	L3	L2	L1	R1	R2	R3	R4	R5	The Score
1	palestra	N	8	1	7	0	1	0	0	0	1	5	0	1	0	4.000
2	curso	N	8	7	1	0	1	1	5	0	1	0	0	0	0	3.583
3	participação	N	8	8	0	0	1	2	5	0	0	0	0	0	0	3.417
4	registrar	V	5	4	1	0	1	1	2	0	0	1	0	0	0	2.083
5	ação	N	7	6	1	2	2	1	1	0	0	0	1	0	0	2.067
6	realizar	V	4	3	1	0	0	0	3	0	0	1	0	0	0	2.000
7	referir	V	2	2	0	0	0	0	0	2	0	0	0	0	0	2.000
8	projeto	N	5	3	2	1	1	0	1	0	0	1	0	1	0	1.700
9	registro	N	4	4	0	0	0	2	2	0	0	0	0	0	0	1.667
10	sigpex	N	5	2	3	0	0	2	0	0	0	0	3	0	0	1.667
11	docente	N	6	4	2	1	2	1	0	0	0	0	1	1	0	1.617
12	cadastrar	V	3	1	2	0	0	0	1	0	0	2	0	0	0	1.500

Fonte: Autora (2021) - KH Coder

O quadro de estatística (Quadro 60) mostra que a palavra “palestra” aparece 8 vezes, sendo 1 vez em posição à esquerda (**LT**) da palavra EVENTO e 7 vezes em posição à direita (**RT**). A palavra “curso” aparece 8 vezes, sendo 7 vezes em posição LT e 1 vez em posição RT. A palavra “participação” aparece 8 vezes, sendo 8 vezes em posição LT e 0 vez em posição RT. O verbo “registrar” aparece 5 vezes, sendo 4 vezes em posição LT e 1 vez em posição RT. Esta análise revela que as dúvidas mais frequentes são relacionadas à participação em eventos e palestras.

Apêndice 6

Mineração por meio das técnicas do KH Coder – planilha PERGUNTAS.**Análise de cluster.**

a) Análise Cluster 2

Verificando o conteúdo do cluster 2, pode-se listar o número de documentos classificados (Quadro 61) e a lista de associação de palavras (Quadro 62).

Quadro 61 - Documentos classificados no cluster 2 - planilha

PERGUNTAS

Search Documents

Search Entry: #direct

Coding Rule File: Browse No File Selected

#direct: and <->_cluster_tmp-->2

AND no sort Unit: H5 Run

Result:

a docente registrou um curso de extensão que vai ter aulas teóricas de forma remota sou professora de o colégio de aplicação e através de uma ação de extensão vou real gostaria de saber se o coordenador de extensão pode aprovar legalmente alguma açã gostaria de saber como os chefes de departamento podem acompanhar o recebiment qual o procedimento para a exclusão de uma ação de extensão registrada incorretam sobre aprovação ad-referendum , em o meu centro temos o sistema de câmara de ext o docente registrou como evento de extensão , entretanto , não será organizado por ... gostaria apenas de confirmar uma informação , por gentileza , atividades que estão ... tenho um projeto de extensão cadastrado em o sigpex e que está ativo .recentement peço auxilio com instruções de como proceder com alteração de o período de a ativid. desejo pedir prorrogação de a atividade de extensão em virtude de a pandemia . a or...

Copy View Doc P200 N200 Hits: 121, View: 1-121 Ready.

Fonte: Autora (2021) - KH Coder

Quadro 62 - Lista de associação de palavras cluster 2 -

planilha PERGUNTAS

Word Association

Search Entry: #direct

Coding Rule File: Browse No File Selected

#direct: and <->_cluster_tmp-->2

AND Unit: H5 Run

Result:

N	word	POS	unconditional	conditional	Jaccard
1	atividade	N	261 (0.435)	90 (0.744)	0.3082
2	artigo	N	64 (0.107)	39 (0.322)	0.2671
3	publicação	N	55 (0.092)	31 (0.256)	0.2138
4	docente	AQ	105 (0.175)	39 (0.322)	0.2086
5	registrar	V	193 (0.322)	50 (0.413)	0.1894
6	docente	N	137 (0.228)	41 (0.339)	0.1889
7	publicar	V	25 (0.042)	23 (0.190)	0.1870
8	sigpex	N	252 (0.420)	52 (0.430)	0.1620
9	registro	N	161 (0.268)	37 (0.306)	0.1510
10	progressão	N	55 (0.092)	23 (0.190)	0.1503

Copy KWIC Sort: Jaccard Filter Network Hits: 12 Ready.

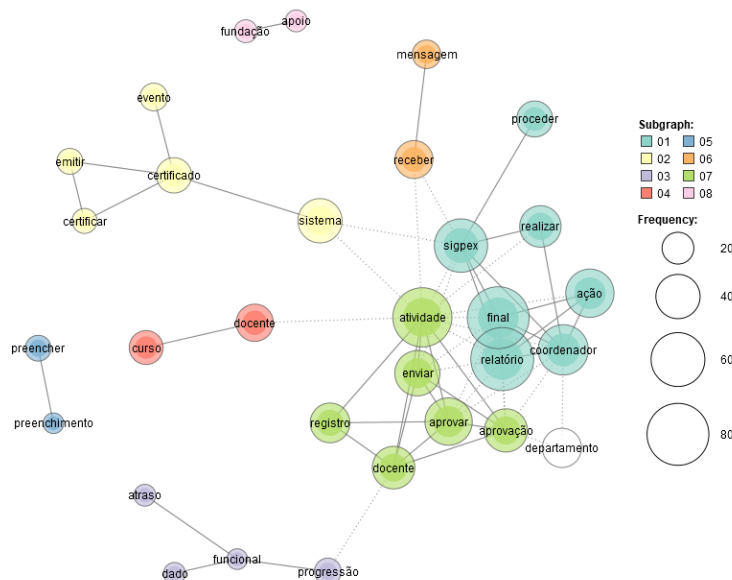
Fonte: Autora (2021) - KH Coder

Foram classificados 121 documentos no cluster 2. A palavra “atividade” aparece em 74,4% dos documentos. O verbo “registrar” e a palavra “sigpex” aparecem em 41,3% e 43% dos documentos respectivamente. As palavras “docente” (adjetivo) e “docente” (substantivo) aparecem em 32,2% e 33,9% dos documentos respectivamente. A palavra “publicação” aparece em 25,6% dos documentos. A rede de coocorrência do Cluster 2 pode ser observada na Figura 19, com destaque para a rede na cor verde.

Para o cluster 2, pode-se identificar que as principais dúvidas estão relacionadas ao registro de atividades docentes no SIGPEX, publicação em revista de artigo científico, revisão de artigo científico, trabalho em congresso e anais.

documentos. A rede de coocorrência do cluster 3 pode ser observada na Figura 20, com destaque para as redes nas cores azul e verde.

Figura 24 - Rede de coocorrência do cluster 3 - planilha PERGUNTAS



Fonte: Autora (2021) - KH Coder

Para o cluster 3, pode-se identificar que as principais dúvidas estão relacionadas com ao envio e aprovação do relatório final no SIGPEX.

c) Análise Cluster 4

Verificando o conteúdo do cluster 4, pode-se listar o número de documentos classificados (Quadro 65) e a lista de associação de palavras (Quadro 66).

Quadro 65 - Documentos classificados no cluster 4 - planilha PERGUNTAS

Search Documents

Search Entry: #direct

Coding Rule File: Browse [No File Selected]

#direct: and <-_cluster_tmp->4

AND no sort Unit: H5 Run

Result:

gostaria de saber porque não é possível cadastrar os cursos de curta duração com ca
um docente a o cadastrar uma atividade em o sigpex errou em o cadastro de a carga
o coordenador registrou um evento de extensão com 2 horas totais (2 palestras de 1.
estou avaliando um pedido de evento de extensão , o período de realização é de apen
tenho uma demanda de um professor que gostaria de saber se seria possível atender
em relação a a carga horária , a divisão de a carga horária total de o projeto por ...
a RNE882016 diz que ações de extensão observarão as limitações inerentes a o carg
dúvida , a professora kalina fez registro de o programa de extensão em 27 07 2020 , c
estou com dúvida sobre o limite de carga horária de 8 horas semanais para docentes
os projetos em que os coordenadores e ou integrantes receberem remuneração obrig
estou criando um novo projeto de extensão , como semestre passado (um grupo de é

Copy View Doc P200 N200 Hits: 38, View: 1-38 Ready.

Fonte: Autora (2021) - KH Coder

Quadro 66 - Lista de associação de palavras cluster 4 - planilha PERGUNTAS

Word Association

Search Entry: #direct

Coding Rule File: Browse [No File Selected]

#direct: and <-_cluster_tmp->4

AND Unit: H5 Run

Result:

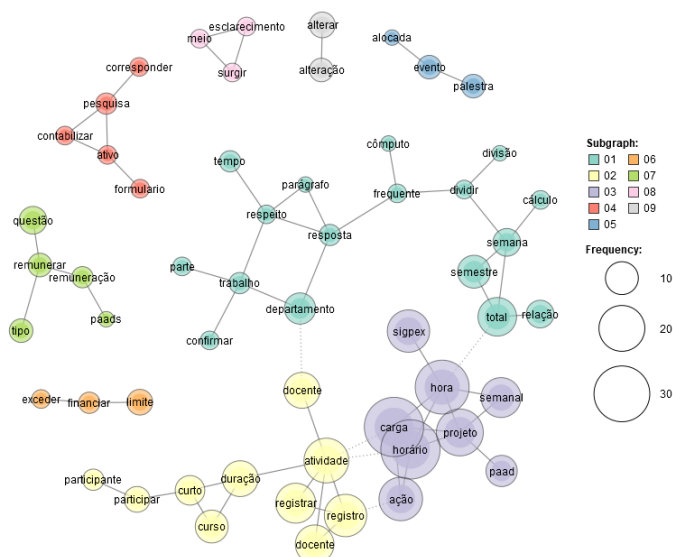
N	word	POS	unconditional	conditional	Jaccard
1	horário	AQ	53 (0.088)	34 (0.895)	0.5965
2	carga	N	54 (0.090)	34 (0.895)	0.5862
3	hora	N	63 (0.105)	28 (0.737)	0.3836
4	semanal	AQ	23 (0.038)	15 (0.395)	0.3261
5	total	AQ	22 (0.037)	14 (0.368)	0.3043
6	duração	N	27 (0.045)	12 (0.316)	0.2264
7	participante	N	50 (0.083)	13 (0.342)	0.1733
8	curto	AQ	17 (0.028)	8 (0.211)	0.1702
9	semestre	N	38 (0.063)	10 (0.263)	0.1515
10	limite	N	12 (0.020)	6 (0.158)	0.1364

Copy KWIC Sort: Jaccard Filter Network Hits: 38 Ready.

Fonte: Autora (2021) - KH Coder

Foram classificados 38 documentos no Cluster 4. As palavras “carga” e “horário” e aparece em 89,5% dos documentos. A palavra “hora” aparece em 73,7% dos documentos. As palavras “semanal” e “total” aparecem em 39,5% e 36,8% dos documentos respectivamente. A palavra “participante” aparece em 34,2% dos documentos. A rede de coocorrência do cluster 4 pode ser observada na Figura 21, com destaque para as redes na cor roxa.

Figura 25 - Rede de coocorrência do cluster 4 - planilha PERGUNTAS



Fonte: Autora (2021) - KH Coder

Para o cluster 4, pode-se identificar que as principais dúvidas estão relacionadas ao registro da carga horária semanal e total dos projetos.

d) Análise Cluster 5

Verificando o conteúdo do cluster 5, pode-se listar o número de documentos classificados (Quadro 67) e a lista de associação de palavras (Quadro 68).

Quadro 67 - Documentos classificados no cluster 5 - planilha PERGUNTAS

Search Documents

Search Entry: #direct

Coding Rule File: Browse [No File Selected]

#direct: and <->_cluster_tmp->5

AND no sort Unit: H5 Run

Result:

o projeto de extensão cadastrado em a plataforma sigpex sob o número 201702850 e...
estou em vias de sair para um pós-doutorado dia 11 03 e a secretaria municipal de e...
ano passado fui coordenadora de uma ação de extensão e ministrei um curso de a ár...
ações de extensão relacionadas a cursos de capacitação, em a modalidade ead, qu...
em o cadastramento de cursos de extensão, em a aba de dados gerais, em período...
estou escrevendo uma proposta de curso de curta duração para o edital rota 2030 de...
pretendemos oferecer um curso com direito a certificado, e minha dúvida é a seguinte...
estou analisando um curso de extensão que contará com dois tipos de recursos finan...
gostaria de saber se é possível fazer um curso de extensão via proex, cobrar a ins...
sou docente de a pós em linguística e, por intermédio de um colega de a ufscar, f...
estou avaliando o relatório de prestação de contas de o curso de extensão 202101788

Copy View Doc P200 N200 Hits: 229, View: 1-200 Ready

Fonte: Autora (2021) - KH Coder

Quadro 68 - Lista de associação de palavras cluster 5 - planilha PERGUNTAS

Word Association

Search Entry: #direct

Coding Rule File: Browse [No File Selected]

#direct: and <->_cluster_tmp->5

AND Unit: H5 Run

Result:

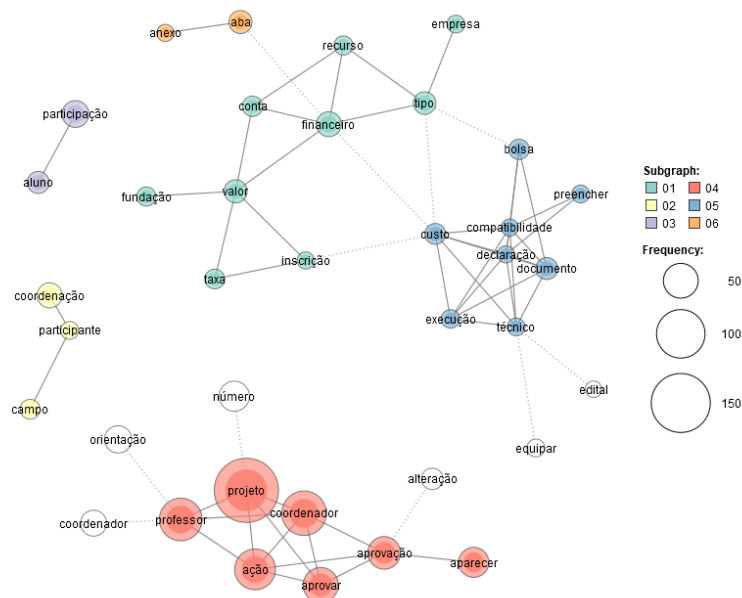
N	word	POS	unconditional	conditional	Jaccard
1	projeto	N	278 (0.463)	179 (0.782)	0.5457
2	coordenador	N	169 (0.282)	84 (0.367)	0.2675
3	professor	N	193 (0.322)	77 (0.336)	0.2232
4	ação	N	164 (0.273)	70 (0.306)	0.2167
5	aprovar	V	132 (0.220)	51 (0.223)	0.1645
6	aprovação	N	106 (0.177)	44 (0.192)	0.1512
7	aparecer	V	67 (0.112)	37 (0.162)	0.1429
8	número	N	72 (0.120)	37 (0.162)	0.1402
9	orientação	N	67 (0.112)	30 (0.131)	0.1128
10	participação	N	74 (0.123)	29 (0.127)	0.1058

Copy KWIC Sort: Jaccard Filter Network Hits: 22! Ready

Fonte: Autora (2021) - KH Coder

Foram classificados 229 documentos no cluster 5. A palavra “projeto” aparece em 78,2% dos documentos. As palavras “coordenador” e “professor” aparecem em 36,7% e 33,6% dos documentos respectivamente. As palavras “aprovar” e “aprovação” aparecem em 22,3% e 19,2% dos documentos respectivamente. A rede de coocorrência do cluster 5 pode ser observada na Figura 22, com destaque para as redes na cor vermelha.

Figura 26 - Rede de coocorrência do cluster 5 - planilha PERGUNTAS



Fonte: Autora (2021) - KH Coder

Para o cluster 5, pode-se identificar que as principais dúvidas estão relacionadas à aprovação de projetos pelo coordenador.

a) Variável PROJETO

Por meio da concordância textual da variável PROJETO (Quadro 69) e a estatística de localização (Quadro 70). Apresenta-se a seguinte análise de resultados:

Quadro 69 - Concordância textual palavra “projeto” - planilha RESPOSTAS

The screenshot shows the KWIC Concordance interface. The search entry is 'projeto'. The results display several lines of text with the word 'projeto' highlighted in red. The text includes phrases like 'apenas clique em a aba relatório final', 'ainda encontra se aprovado', 'deve ser sempre comprovante de pagam', 'já registrado, mas também é possível fazer u', 'mas o professor em a realidade contratou a', 'e o recurso de o edital será gerenciado por i', 'registro aqui que a previsão de bolsas para;', 'de ensino; pesquisa; extensão e de desenvc', 'é de a fundação de apoio, em este momento', 'o projeto deve ser aprovado em o (s) depar', 'deve ser aprovado em o (s) departamento (s', 'os campos de valores e nomenclatura de rui', and 'é encerrado em o sigpex não é possível efetu'.

Fonte: Autora (2021) - KH Coder

Quadro 70 - Estatística de localização palavra “projeto” - planilha RESPOSTAS

The screenshot shows the Collocation Stats interface. The word 'projeto' is entered, and the results are displayed in a table. The table has columns for N, Word, POS, Total, LT, RT, L5, L4, L3, L2, L1, R1, R2, R3, R4, R5, and The Score. The data is as follows:

N	Word	POS	Total	LT	RT	L5	L4	L3	L2	L1	R1	R2	R3	R4	R5	The Score
1	encerrar	V	22	3	19	0	0	3	0	2	7	5	5	0	0	9.917
2	sigpex	N	26	12	14	0	1	7	4	0	0	10	2	2	0	8.817
3	coordenador	N	27	20	7	0	2	17	1	0	0	2	2	3	0	8.433
4	aprovar	V	20	2	18	0	0	1	1	0	0	5	8	3	2	7.150
5	registrar	V	17	13	4	4	0	1	8	0	1	1	0	1	1	7.083
6	tipo	N	11	1	10	0	0	0	1	0	5	0	3	2	0	7.000
7	revisão	N	19	3	16	0	0	3	0	0	0	7	4	2	3	6.933
8	programa	N	17	16	1	2	5	1	7	1	0	0	0	0	1	6.683
9	prorrogar	V	8	8	0	0	0	3	5	0	0	0	0	0	0	6.500
10	final	AQ	21	13	8	3	0	10	0	0	0	1	1	6	0	5.717
11	pesquisa	N	12	1	11	0	1	0	0	0	10	0	1	0	0	5.500
12	projeto	N	14	7	7	1	2	3	0	1	1	0	3	2	1	5.400
13	situação	N	17	10	7	1	0	9	0	0	1	2	2	2	0	5.267

Fonte: Autora (2021) - KH Coder

O quadro de estatística (Quadro 70) mostra que a palavra “SIGPEX” é relacionada 26 vezes, 12 vezes em posição à esquerda (LT) da palavra PROJETO e 14 vezes em posição à direita (RT), trata-se do Sistema Integrado de Gerenciamento de Projetos de Pesquisa e de Extensão, onde todas as ações de extensão e atividades docentes são gerenciadas.

A palavra “coordenador” aparece 27 vezes, sendo 20 vezes em posição LT e 7 vezes em posição RT. O verbo “encerrar” aparece 22 vezes, sendo 3 vezes em posição LT e 19 vezes em posição RT, com destaque para a posição R2 (Quadro 71).

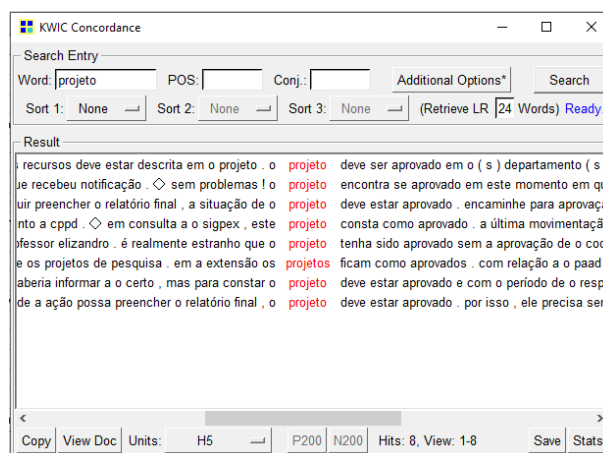
Quadro 71 - Concordância textual das palavras “projeto” e “encerrar” em posição R2 - planilha RESPOSTAS

The screenshot shows the KWIC Concordance interface. The search entry is 'projeto'. The results display several lines of text with the words 'projeto' and 'encerrar' highlighted in red. The text includes phrases like 'é encerrado em o sigpex não é possível efetuar', 'foi encerrado adequadamente, com inclusão de', 'é encerrado. o sigpex não aceita carga horá', 'seja encerrado. documentos trocados devem s', 'está encerrado. agora, para dar continuidade.', 'foi encerrado e de esta forma, não há necessid', and 'foi encerrado. mas o sistema é ligado a a prog'.

Fonte: Autora (2021) - KH Coder

O verbo “aprovar” aparece 20 vezes, sendo 2 vezes em posição LT e 18 vezes em posição RT. No Quadro 72, observamos a concordância textual para a posição de destaque R3.

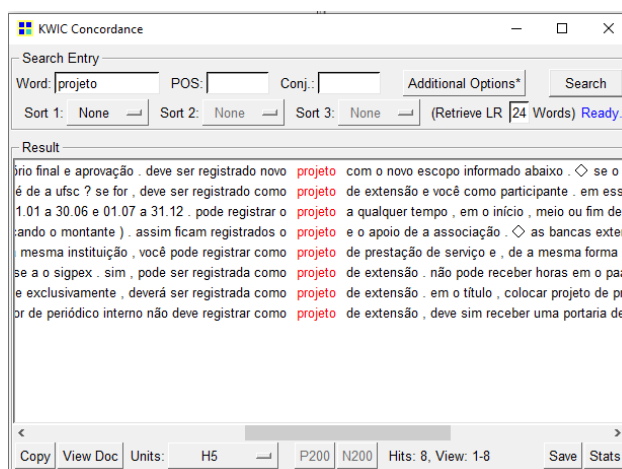
Quadro 72 - Concordância textual das palavras “projeto” e “aprovar” em posição R3 - planilha RESPOSTAS



Fonte: Autora (2021) - KH Coder

O verbo “registrar” aparece 17 vezes, sendo 13 vezes em posição LT e 4 vezes em posição RT. No Quadro 73, observamos a concordância textual para a posição de destaque L2.

Quadro 73 - Concordância Textual das palavras “projeto” e “registrar” em posição L2 - planilha RESPOSTAS



Fonte: Autora (2021) - KH Coder

Esta análise revela que as respostas mais frequentes estão relacionadas à orientação de registro de ações como projetos e à orientação da situação dos projetos (aprovada ou encerrada).

b) Variável PROGRAMA

Por meio da concordância textual da variável PROGRAMA (Quadro 74) e a estatística de localização (Quadro 75). Apresenta-se a seguinte análise de resultados:

Quadro 74 - Concordância textual palavra “programa” - planilha

RESPOSTAS

Search Entry
Word: programa POS: Conj.: Additional Options Search
Sort 1: None Sort 2: None Sort 3: None (Retrieve LR 24 Words) Ready.

Result

e que seria ministrado por os docentes de o programa , com o lucro , produziríamos bolsas - já que c
ária é realizada por meio de ações como , i - programa de extensão , que constitui um conjunto articu
ado , podendo ser isolado ou vinculado a um programa ; iii - curso de extensão , que constitui uma a
regues até o dia 30 de junho para a proex e o programa deverá estar aprovado em o sigpex em a data
e constam em relatórios institucionais . o programa é um conjunto de projetos e outras ações de e
ais , por o que foi registrado , não parece ser programa , o coordenador não coloca objetos , metas e
mestrado não existe pois não faz parte de o programa de pós-graduação de a ufsc . a RN631019 disj
as atividades de pesquisa realizadas junto a programa de pós-graduação de a ufsc por portador de o
mento está correto . não cadastre valor em o programa , apenas em os projetos , conforme fontes de
nificação . mesmo assim cadastrando um programa

Copy View Doc Units: H5 P200 N200 Hits: 45, View: 1-45 Save Stats

Fonte: Autora (2021) - KH Coder

Quadro 75 - Estatística de localização palavra “programa”

- planilha RESPOSTAS

Node Word
Word: programa POS: Conj.: Hits: 45

Result

N	Word	POS	Total	LT	RT	L5	L4	L3	L2	L1	R1	R2	R3	R4	R5	The Score
1	projeto	N	17	1	16	1	0	0	0	0	1	7	1	5	2	6.683
2	execução	N	4	4	0	0	0	0	4	0	0	0	0	0	0	2.000
3	trabalho	N	4	0	4	0	0	0	0	0	0	4	0	0	0	2.000
4	vincular	V	6	6	0	0	6	0	0	0	0	0	0	0	0	2.000
5	cadastrar	V	4	4	0	1	1	0	2	0	0	0	0	0	0	1.450
6	registrar	V	3	1	2	1	0	0	0	0	0	1	1	0	1.033	
7	aberto	AQ	1	0	1	0	0	0	0	0	1	0	0	0	0	1.000
8	aprovação	N	3	3	0	0	3	0	0	0	0	0	0	0	0	1.000
9	diferir	V	1	0	1	0	0	0	0	0	1	0	0	0	0	1.000
10	objetivo	N	4	4	0	4	0	0	0	0	0	0	0	0	0	1.000

Copy Filter Sort: The Score Window span: L5 R5

Fonte: Autora (2021) - KH Coder

O quadro de estatística (Quadro 75 mostra que a palavra “projeto” aparece 17 vezes, sendo 1 vez em posição à esquerda (LT) da palavra PROGRAMA e 16 vezes em posição à direita (RT), sendo 7 vezes em R2 (Quadro 76). O verbo “vincular” aparece 6 vezes, todas em posição L3 (Quadro 77). A palavra “execução” aparece 4 vezes, todas em posição R2. A palavra “trabalho” aparece 4 vezes, todas em posição R2.

Quadro 76 - Concordância textual das palavras “programa” e “projeto” em posição R2 - planilha RESPOSTAS

Search Entry
Word: programa POS: Conj.: Additional Options* Search
Sort 1: None Sort 2: None Sort 3: None (Retrieve LR 24 Words) Ready.

Result

ta catarina , as definições e características de programas e projetos estão em o artigo 3º , i - programa de
le bolsas de extensão deverá estar prevista em programa ou projeto que preencha os seguintes requisitos
a de o público . parágrafo único . em o caso de programas e projetos realizados em conjunto por mais de u
vai variar conforme o tipo (atividade docente , programas , projetos e convênios) . não temos um procedi
o artigo 10 de a RNE882016 a aprovação de os programas e projetos de extensão dará se por prazo de até
e a RNE882016 , artigo 10 . a aprovação de os programas e projetos de extensão dará se por prazo de até
n a RNE882016 , artigo 10 , a aprovação de os programas e projetos de extensão dará se por prazo de até

Copy View Doc Units: H5 P200 N200 Hits: 7, View: 1-7 Save Stats

Fonte: Autora (2021) - KH Coder

Quadro 77 - Concordância textual das palavras “programa” e “vincular” em posição L3 - planilha RESPOSTAS

Search Entry
Word: programa POS: Conj.: Additional Options* Search
Sort 1: None Sort 2: None Sort 3: None (Retrieve LR 24 Words) Ready.

Result

ninado , podendo ser isolado ou vinculado a um programa ; iii - curso de extensão , que constitui uma açã
jeto individual cadastra um projeto e vincula a o programa . o projeto que é vinculado a o programa em o momento de o registro , e não o contrário .
ágina de a proex . o projeto que é vinculado a o programa em o momento de o registro , e não o contrário .
ninado , podendo ser isolado ou vinculado a um programa . o fato de o coordenador colocar como projeto p
ninado , podendo ser isolado ou vinculado a um programa . envie seu questionamento para a prof graziela
ninado , podendo ser isolado ou vinculado a um programa ; iii - curso de extensão , que constitui uma açã

Copy View Doc Units: H5 P200 N200 Hits: 6, View: 1-6 Save Stats

Fonte: Autora (2021) - KH Coder

Esta análise revela que as respostas mais frequentes estão relacionadas à aprovação de programas e projetos e ao vínculo de ações a um programa.

c) Variável PUBLICAÇÃO

Por meio da concordância textual da variável PUBLICAÇÃO (Quadro 78) e a estatística de localização (Quadro 79). Apresenta-se a seguinte análise de resultados:

Quadro 78 - Concordância textual palavra “publicação” - planilha RESPOSTAS

The screenshot shows the KWIC Concordance interface. The search entry is 'publicação'. The results list various occurrences of the word in different contexts, such as 'publicações preferencialmente integrando as ações' and 'publicação em programas de tv e rádio'.

Fonte: Autora (2021) - KH Coder

Quadro 79 - Estatística de localização palavra “publicação” - planilha RESPOSTAS

The screenshot shows the Collocation Stats interface. It displays a table with the following data:

N	Word	POS	Total	LT	RT	L5	L4	L3	L2	L1	R1	R2	R3	R4	R5	The Score
1	atividade	N	30	27	3	3	3	11	9	1	0	0	2	1	0	11.433
2	docente	AQ	16	15	1	0	0	2	5	8	0	0	0	1	0	11.417
3	semestre	N	32	25	7	1	5	9	10	0	0	0	2	4	1	11.317
4	registrar	V	34	15	19	12	2	0	1	0	1	0	11	6	1	9.767
5	efetivo	AQ	8	8	0	0	1	2	0	5	0	0	0	0	0	5.917
6	revisão	N	6	0	6	0	0	0	0	0	4	1	1	0	0	4.833
7	registro	N	12	8	4	1	1	2	4	0	0	1	1	2	0	4.450
8	docente	N	9	9	0	0	3	0	6	0	0	0	0	0	0	3.750
9	artigo	N	11	3	8	1	0	0	2	0	0	0	4	3	1	3.483
10	data	N	8	6	2	0	0	1	5	0	0	0	0	1	1	3.283

Fonte: Autora (2021) - KH Coder

O quadro de estatística (Quadro 79) mostra que a palavra “atividade” aparece 30 vezes, sendo 27 vezes em posição à esquerda (LT) da palavra PUBLICAÇÃO e 3 vezes em posição à direita (RT), com destaque para a posição L3. A palavra “docente” aparece 16 vezes, sendo 15 vezes em posição LT e 1 vez em posição RT, com destaque para a posição L1. O verbo “registrar” aparece 34 vezes, sendo 15 vezes em posição LT e 19 vezes em posição RT, com destaque para as posições L5 e R3. A palavra “semestre” aparece 32 vezes, sendo 25 vezes em posição LT e 7 vezes em posição RT, com destaque para a posição L2 (Quadro 80).

Quadro 80 - Concordância textual das palavras “publicação” e “semestre” em posição L2 - planilha RESPOSTAS

The screenshot shows the KWIC Concordance interface with search results for 'publicação' and 'semestre' in position L2. The results list various occurrences of the words in different contexts, such as 'publicação são computadas por semestre e uma vez' and 'publicação já com os comprovantes anexados a o si'.

Fonte: Autora (2021) - KH Coder

Esta análise revela que respostas mais frequentes são relacionadas ao registro da atividade docente publicações dentro do semestre de publicação.

d) Variável BANCA

Por meio da concordância textual da variável BANCA (Quadro 81) e a estatística de localização (Quadro 82). Apresenta-se a seguinte análise de resultados:

Quadro 81 - Concordância textual palavra “banca” -
planilha RESPOSTAS

The screenshot shows the KWIC Concordance interface with the search entry 'banca'. The results list various occurrences of the word in different contexts, such as 'banca externa', 'banca interna', and 'banca de extensão'.

Fonte: Autora (2021) - KH Coder

Quadro 82 - Estatística de localização palavra “banca” -
planilha RESPOSTAS

The screenshot shows the Collocation Stats interface for the word 'banca'. It displays a table with columns for N, Word, POS, Total, LT, RT, L5, L4, L3, L2, L1, R1, R2, R3, R4, R5, and The Score.

N	Word	POS	Total	LT	RT	L5	L4	L3	L2	L1	R1	R2	R3	R4	R5	The Score
1	externo	AQ	6	0	6	0	0	0	0	0	6	0	0	0	0	6.000
2	desconhecer	V	1	1	0	0	0	0	0	1	0	0	0	0	0	1.000
3	interno	AQ	1	0	1	0	0	0	0	0	1	0	0	0	0	1.000
4	participação	N	2	2	0	0	0	0	2	0	0	0	0	0	0	1.000
5	atividade	N	4	3	1	2	0	1	0	0	0	0	0	0	1	0.933
6	docente	AQ	2	2	0	0	1	0	1	0	0	0	0	0	0	0.750
7	término	N	2	2	0	0	0	2	0	0	0	0	0	0	0	0.667
8	evento	N	1	0	1	0	0	0	0	0	1	0	0	0	0	0.500
9	publicação	N	1	1	0	0	0	0	1	0	0	0	0	0	0	0.500
10	tipo	N	1	1	0	0	0	0	1	0	0	0	0	0	0	0.500

Fonte: Autora (2021) - KH Coder

O quadro de estatística (Quadro 82) mostra que a palavra “externo” aparece 6 vezes, sendo 0 vez em posição à esquerda (LT) da palavra BANCA e 6 vezes em posição à direita (RT). Esta análise revela que as respostas mais frequentes são relacionadas a orientações sobre bancas externas.

e) Variável CURSO

Por meio da concordância textual da variável CURSO (Quadro 83) e a estatística de localização (Quadro 84). Apresenta-se a seguinte análise de resultados:

Quadro 83 - Concordância textual palavra “curso” -
planilha RESPOSTAS

The screenshot shows the KWIC Concordance interface with the search entry 'curso'. The results list various occurrences of the word in different contexts, such as 'curso de extensão', 'curso de graduação', and 'curso de pós-graduação'.

Fonte: Autora (2021) - KH Coder

Quadro 84 - Estatística de localização palavra “curso” -
planilha RESPOSTAS

The screenshot shows the Collocation Stats interface for the word 'curso'. It displays a table with columns for N, Word, POS, Total, LT, RT, L5, L4, L3, L2, L1, R1, R2, R3, R4, R5, and The Score.

N	Word	POS	Total	LT	RT	L5	L4	L3	L2	L1	R1	R2	R3	R4	R5	The Score
1	evento	N	9	1	8	0	0	0	1	0	1	7	0	0	0	5.000
2	curto	AQ	7	0	7	0	0	0	0	0	0	4	0	3	0	2.750
3	docente	AQ	5	5	0	0	2	1	1	1	0	0	0	0	0	2.333
4	duração	N	8	0	8	0	0	0	0	0	0	0	4	0	4	2.133
5	ação	N	5	5	0	1	1	2	0	1	0	0	0	0	0	2.117
6	externo	AQ	4	0	4	0	0	0	0	0	1	0	2	1	0	1.917
7	registrar	V	6	4	2	1	1	0	2	0	0	0	0	1	1	1.900
8	participação	N	5	4	1	1	0	1	2	0	0	0	1	0	0	1.867
9	certificado	N	6	1	5	1	0	0	0	0	0	2	0	1	2	1.850
10	capacitação	N	4	0	4	0	0	0	0	0	0	3	0	0	1	1.700

Fonte: Autora (2021) - KH Coder

O quadro de estatística (Quadro 84) mostra que a palavra “evento” aparece 9 vezes, sendo 1 vez em posição à esquerda (LT) da palavra CURSO e 8 vezes em posição à direita (RT). A palavra “curto” aparece 7 vezes, todas em posição RT. A palavra “duração” aparece 8 vezes, todas em posição RT. Esta análise revela que as respostas mais frequentes são relacionadas a cursos e eventos de curta duração.

f) Variável PRESTAÇÃO DE SERVIÇO

Por meio da concordância textual da variável PRESTAÇÃO DE SERVIÇO (Quadro 85) e a estatística de localização (Quadro 86). Apresenta-se a seguinte análise de resultados:

Quadro 85 - Concordância textual palavra “prestação de serviço” - planilha RESPOSTAS

The screenshot shows the KWIC Concordance interface. The search entry is 'prestação'. The results list various occurrences of the word in different contexts, such as 'prestação de serviço', 'prestação de serviço eventual', and 'prestação de serviço - serviço eventual - assessoria'.

Fonte: Autora (2021) - KH Coder

Quadro 86 - Estatística de localização palavra “prestação de serviço” - planilha RESPOSTAS

The screenshot shows the Collocation Stats interface. The word 'prestação' is selected. The table below shows the results of the collocation analysis.

N	Word	POS	Total	LT	RT	L5	L4	L3	L2	L1	R1	R2	R3	R4	R5	The Score
1	serviço	N	98	0	98	0	0	0	0	0	0	76	11	11	0	44.417
2	docente	AQ	21	21	0	1	0	1	10	9	0	0	0	0	0	14.533
3	atividade	N	25	22	3	1	2	10	9	0	0	0	2	1	0	9.450
4	eventual	AQ	24	0	24	0	0	0	0	0	0	0	3	10	11	5.700
5	registrar	V	14	13	1	5	3	1	4	0	0	0	0	0	1	4.283
6	pagamento	N	5	5	0	1	0	1	3	0	0	0	0	0	0	2.033
7	evento	N	3	3	0	0	0	0	3	0	0	0	0	0	0	1.500
8	incluir	V	2	2	0	0	0	0	1	1	0	0	0	0	0	1.500
9	v	N	3	3	0	0	0	0	3	0	0	0	0	0	0	1.500
10	docente	N	4	3	1	1	0	1	1	0	0	0	0	1	0	1.283

Fonte: Autora (2021) - KH Coder

O quadro de estatística (Quadro 86) mostra que a palavra “serviço” aparece 76 vezes em posição R2 da palavra PRESTAÇÃO DE SERVIÇO, levando a concluir a formação da expressão “prestação de serviço” em 76 sentenças. A palavra “atividade” aparece 25 vezes, sendo 22 vezes em posição à esquerda (LT) da palavra PRESTAÇÃO e 3 vezes em posição à direita (RT). A palavra “docente” aparece 21 vezes, todas em posição LT. A palavra “eventual” aparece 24 vezes, todas em posição RT. O verbo “registrar” aparece 14 vezes, sendo 13 vezes em posição LT e 1 vez em posição RT. Esta análise revela que as respostas mais frequentes são relacionadas ao registro da atividade docente prestação de serviço, serviço eventual.

g) Variável EVENTO

Por meio da concordância textual da variável EVENTO (Quadro 87) e a estatística de localização (Quadro 88). Apresenta-se a seguinte análise de resultados:

Quadro 87 - Concordância textual palavra “evento” -
planilha RESPOSTAS

Search Entry
Word: evento POS: Conj.: Additional Options* Search
Sort 1: None Sort 2: None Sort 3: None (Retrieve LR 24 Words) Ready

Result

recurso de o edital para a organização de o **evento** , e outro com as inscrições , sugiro que siga ^
fixadas taxas de inscrição em os cursos e **eventos** de extensão visando cobrir , parcial ou integr
poderá atuar em atividades de extensão e **eventos** de capacitação . de esta forma , o voluntário
de gru . as aquisições para realização de o **evento** devem se submeter a o processo de licitaçã
registre o período e carga horária total de o **evento** incluindo a organização . mesmo que o even
nto incluindo a organização . mesmo que o **evento** seja de um dia , seu registro deve contempla
tador de a ação registre o período total de o **evento** , incluindo o período de planejamento , organ
e sob a perspectiva de que ele registrou um **evento** de um dia , em fevereiro de 2019 , agora que
tais de um ano . será que está correto ? ? o **evento** já não aconteceu ? qual a justificativa para a
tras ações de extensão . tais como cursos **eventos** prestação de serviços e publicações . refere

Copy View Doc Units: H5 P200 N200 Hits: 58, View: 1-58 Save Stats

Fonte: Autora (2021) - KH Coder

Quadro 88 - Estatística de localização palavra “evento” -
planilha RESPOSTAS

Node Word
Word: evento POS: Conj.: Hits: 58

N	Word	POS	Total	LT	RT	L5	L4	L3	L2	L1	R1	R2	R3	R4	R5	The Score
1	docente	AQ	11	11	0	1	0	0	8	2	0	0	0	0	0	6.200
2	curso	N	9	8	1	0	0	0	7	1	0	1	0	0	0	5.000
3	palestra	N	10	0	10	0	0	0	0	0	10	0	0	0	0	5.000
4	atividade	N	14	11	3	0	1	8	2	0	0	0	0	0	3	4.517
5	registrar	V	8	5	3	3	1	0	1	0	0	2	1	0	0	2.683
6	organização	N	7	4	3	1	0	3	0	0	0	1	1	0	1	2.233
7	externo	AQ	2	0	2	0	0	0	0	0	2	0	0	0	0	2.000
8	interesse	N	4	0	4	0	0	0	0	0	0	4	0	0	0	2.000
9	projeto	N	4	4	0	0	0	0	4	0	0	0	0	0	0	2.000
10	incluir	V	2	0	2	0	0	0	0	0	1	1	0	0	0	1.500

Copy Filter Sort: The Score Window span: L5 R5

Fonte: Autora (2021) - KH Coder

O quadro de estatística (Quadro 88) mostra que a palavra “docente” aparece 11 vezes, todas em posição à esquerda (LT) da palavra EVENTO. A palavra “atividade” aparece 14 vezes, 11 em posição LT e 3 em posição RT. A palavra “palestra” aparece 10 vezes, todas em posição RT (Quadro 89). A palavra “curso” aparece 9 vezes, sendo 8 vezes em posição LT e 1 vez em posição RT (Quadro 90). O verbo “registrar” aparece 8 vezes, sendo 5 vezes em posição LT e 3 vezes em posição RT.

Quadro 89 - Concordância textual das palavras “evento” e “palestra” em posição R2 - planilha RESPOSTAS

Search Entry
Word: evento POS: Conj.: Additional Options* Search
Sort 1: None Sort 2: None Sort 3: None (Retrieve LR 24 Words) Ready

Result

a atividade não se encaixa como ouvinte de **evento** ou palestra . essa atividade não deve ser reg ^
de ser cadastrado como atividade docente - **eventos** e palestras - palestrantes . ◇ cursos com m
ção , podes registrar em atividade docente , **eventos** e palestras , conferencista palestrante . ◇ e
isto deve ser feito como atividade docente - **eventos** e palestras . não deve ter receber carga em (
de a data de início como atividade docente - **eventos** e palestras . anexar carta convite , cronogar
como atividade docente . atividade docente - **eventos** e palestras - membro organização . não deve
de a data de início como atividade docente - **eventos** e palestras . anexar carta convite , cronogar
◇ o registro correto é atividade docente - **eventos** e palestras . observei que ele colocou local fl
rad@contato.ufsc.br ◇ atividades docentes **eventos** e palestras membro de a comissão organiza
s atividades docentes (publicação banca **evento** palestra) podem ser registradas em o sinq

Copy View Doc Units: H5 P200 N200 Hits: 10, View: 1-10 Save Stats

Fonte: Autora (2021) - KH Coder

Quadro 90 - Concordância textual das palavras “evento” e “curso” em posição L2 - planilha RESPOSTAS

Search Entry
Word: evento POS: Conj.: Additional Options* Search
Sort 1: None Sort 2: None Sort 3: None (Retrieve LR 24 Words) Ready

Result

er fixadas taxas de inscrição em os cursos e **eventos** de extensão visando cobrir , parcial ou integral
utras ações de extensão , tais como cursos , **eventos** , prestação de serviços e publicações , prefer
utras ações de extensão . tais como cursos , **eventos** , prestação de serviços e publicações , prefer
em certificados emitidos por a ufsc (cursos , **eventos** , etc) não devem ser registradas em o sigpex
centes referente a participação em cursos ou **eventos** em que foram emitidos certificados por o siste
utras ações de extensão , tais como cursos , **eventos** , prestação de serviços e publicações , prefer
em certificados emitidos por a ufsc (cursos , **eventos** , etc) não devem ser registradas em o sigpex

Copy View Doc Units: H5 P200 N200 Hits: 7, View: 1-7 Save Stats

Fonte: Autora (2021) - KH Coder

Esta análise revela que as respostas mais frequentes são orientações para o registro de ação como atividade docente, eventos e palestras.

A análise por meio da rede de coocorrência e da concordância textual das palavras e variáveis permitiu a identificação dos seguintes **tópicos de respostas**:

- Aprovação de programas e projetos
- Bancas externas.
- Cursos e eventos de curta duração
- Registro da atividade docente prestação de serviço, serviço eventual
- Registro da atividade docente publicações dentro do semestre de publicação
- Registro de ação como atividade docente, eventos e palestras.
- Registro de ações como projetos
- Situação dos projetos (aprovada ou encerrada)
- Vínculo de ações a um programa

3) Análise por Cluster

Para a análise, utilizou-se o agrupamento por variáveis em cinco clusters e adotou-se o coeficiente de similaridade Jaccard, para adicionar peso às palavras mais relevantes que aparecem em apenas alguns documentos (Quadro 91).

Quadro 91 - Análise de Cluster do Documento - planilha RESPOSTAS

cluster	documents
n/a	26
Cluster1	129
Cluster2	41
Cluster3	359
Cluster4	33
Cluster5	123

stage	cluster 1	cluster 2	coefficients
1	-163	-925	0.000
2	-216	-502	0.000
3	-526	2	0.000
4	-238	-523	0.000
5	-254	-256	0.000
6	-340	-996	0.000
7	-405	-409	0.000
8	-611	-632	0.000

a) Análise Cluster 1

Verificando o conteúdo do cluster 1, pode-se listar o número de documentos classificados (Quadro 92) e a lista de associação de palavras (Quadro 93), que mostra as palavras que aparecem com maiores probabilidades nos documentos contidos no cluster ou nas palavras que representam as características do cluster.

Quadro 92 - Documentos classificados no cluster 1 - planilha RESPOSTAS

Search Documents

Search Entry: #direct

Coding Rule File: Browse No File Selected

#direct: and <->_cluster_tmp->1

AND no sort Unit: Sentences Run

Result:

1 não é necessário solicitar a reabertura de o projeto apenas clique em a aba relatô...
 depois que o relatório final é aprovado a situação de a ação de extensão passa a se...
 foi informado um valor de financiamento em a ordem de 19.420,00 , em a aba financei...
 a princípio é possível alterar os dados de o projeto enquanto não estiver encerrado
 todos os membros de a equipe de o projeto , bem como os departamentos envolvidos ...
 em o caso de o professor humberto , você deve avaliar o pedido de ele sob a perspec...
 ele deve preencher o relatório final e , se for o caso , anexar documentos , que ju...
 então , analisando a movimentação de o projeto 201707087 e pelo que pude perceber

Copy View Doc P200 N200 Hits: 129, View: 1-129 Ready.

Fonte: Autora (2021) - KH Coder

Quadro 93 - Lista de associação de palavras cluster 1 - planilha RESPOSTAS

Word Association

Search Entry: #direct

Coding Rule File: Browse No File Selected

#direct: and <->_cluster_tmp->1

AND Unit: Sentences Run

Result:

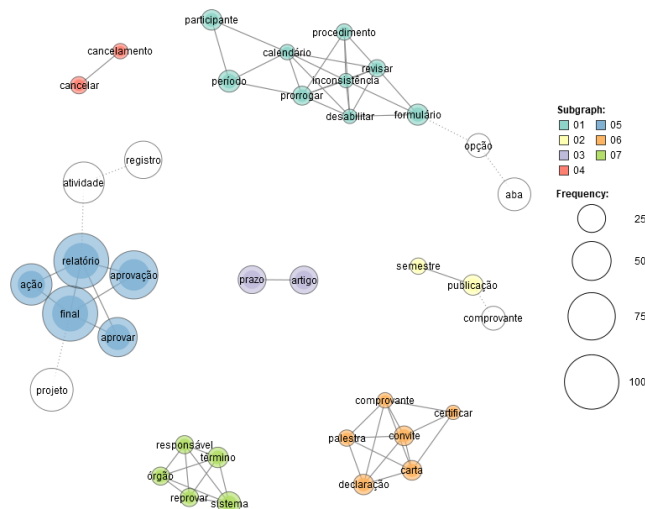
N	word	POS	unconditional	conditional	Jaccard
1	final	AQ	132 (0.186)	103 (0.798)	0.6519
2	relatório	N	140 (0.197)	102 (0.791)	0.6108
3	aprovação	N	137 (0.193)	78 (0.605)	0.4149
4	preencher	V	67 (0.094)	43 (0.333)	0.2810
5	aprovar	V	116 (0.163)	51 (0.395)	0.2629
6	coordenador	N	170 (0.239)	57 (0.442)	0.2355
7	projeto	N	218 (0.307)	61 (0.473)	0.2133
8	data	N	66 (0.093)	34 (0.264)	0.2112
9	situação	N	44 (0.062)	30 (0.233)	0.2098
10	ação	N	205 (0.288)	57 (0.442)	0.2058

Copy KWIC Sort: Jaccard Filter Network Hits: 129 Ready.

Fonte: Autora (2021) - KH Coder

Foram classificados 129 documentos no cluster 1. As palavras “relatório” e “final” aparecem em 79,1% e 79,8%, respectivamente, dos documentos deste cluster. O verbo “aprovar” e a palavra “aprovação” aparecem em 39,5% e 60,5%, respectivamente, dos documentos. A rede de coocorrência do Cluster 1 pode ser observada na Figura 25, com destaque para a rede na cor azul.

Figura 29 - Rede de coocorrência do cluster 1 - planilha RESPOSTAS



Fonte: Autora (2021) - KH Coder

Para o cluster 1, pode-se identificar que as principais respostas estão relacionadas à aprovação do relatório final.

b) Análise Cluster 2

Verificando o conteúdo do cluster 2, pode-se listar o número de documentos classificados (Quadro 94) e a lista de associação de palavras (Quadro 95).

Quadro 94 - Documentos classificados no cluster 2 - planilha

RESPOSTAS

Fonte: Autora (2021) - KH Coder

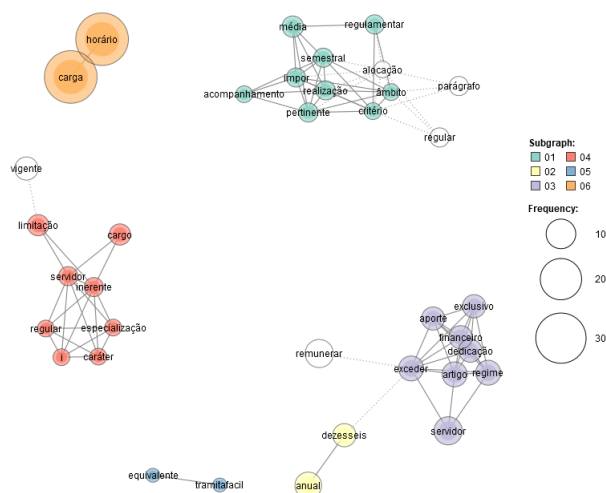
Quadro 95 - Lista de associação de palavras cluster 2 - planilha RESPOSTAS

N	word	POS	unconditional	conditional	Jaccard
1	horário	AQ	45 (0.063)	33 (0.805)	0.6226
2	carga	N	49 (0.069)	33 (0.805)	0.5789
3	semanal	AQ	25 (0.035)	21 (0.512)	0.4667
4	hora	N	64 (0.090)	29 (0.707)	0.3816
5	semana	N	12 (0.017)	9 (0.220)	0.2045
6	servidor	AQ	16 (0.023)	9 (0.220)	0.1875
7	anual	AQ	12 (0.017)	8 (0.195)	0.1778
8	limite	N	12 (0.017)	8 (0.195)	0.1778
9	remunerar	V	20 (0.028)	9 (0.220)	0.1731
10	exceder	V	8 (0.011)	7 (0.171)	0.1667

Fonte: Autora (2021) - KH Coder

Foram classificados 41 documentos no cluster 2. As palavras “carga” e “horário” aparecem em 80,5% dos documentos. A palavra “hora” aparece em 70,7% dos documentos. A palavra “semanal” aparece em 51,2% dos documentos. A rede de coocorrência do Cluster 2 pode ser observada na Figura 26, com destaque para a rede na cor laranja.

Figura 30 - Rede de coocorrência do Cluster 2 - planilha RESPOSTAS



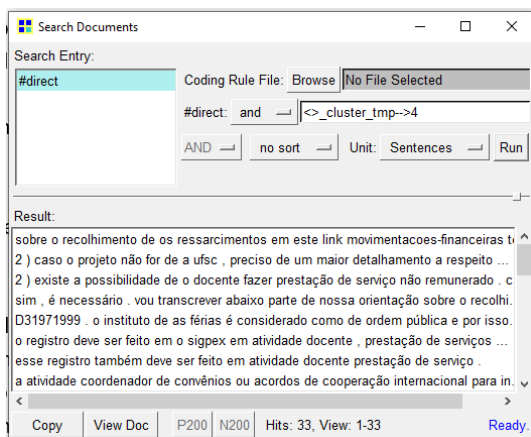
Para o cluster 3, pode-se identificar que as respostas estão relacionadas com a aprovação do projeto pelo coordenador do departamento.

d) Análise Cluster 4

Verificando o conteúdo do cluster 4, pode-se listar o número de documentos classificados (Quadro 98) e a lista de associação de palavras (Quadro 99).

Quadro 98 - Documentos classificados no cluster 4 - planilha

RESPOSTAS



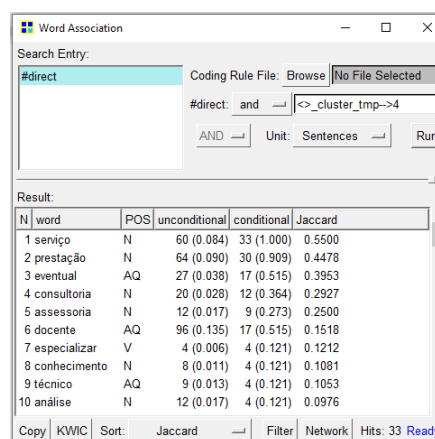
Result:

sobre o recolhimento de os ressarcimentos em este link movimentacoes-financeiras t
 2) caso o projeto não for de a ufsc , preciso de um maior detalhamento a respeito ...
 2) existe a possibilidade de o docente fazer prestação de serviço não remunerado . c
 sim , é necessário . vou transcrever abaixo parte de nossa orientação sobre o recolhi
 D31971999 o instituto de as férias é considerado como de ordem pública e por isso
 o registro deve ser feito em o sigpex em atividade docente , prestação de serviços ...
 esse registro também deve ser feito em atividade docente prestação de serviço .
 a atividade coordenador de convênios ou acordos de cooperação internacional para in

Copy View Doc P200 N200 Hits: 33, View: 1-33 Ready.

Fonte: Autora (2021) - KH Coder

Quadro 99 - Lista de associação de palavras cluster 4 - planilha RESPOSTAS



Result:

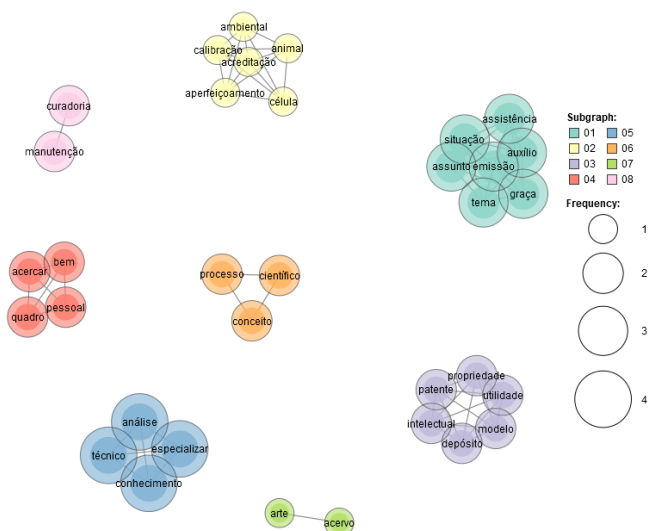
N	word	POS	unconditional	conditional	Jaccard
1	serviço	N	60 (0.084)	33 (1.000)	0.5500
2	prestação	N	64 (0.090)	30 (0.909)	0.4478
3	eventual	AQ	27 (0.038)	17 (0.515)	0.3953
4	consultoria	N	20 (0.028)	12 (0.364)	0.2927
5	assessoria	N	12 (0.017)	9 (0.273)	0.2500
6	docente	AQ	96 (0.135)	17 (0.515)	0.1518
7	especializar	V	4 (0.006)	4 (0.121)	0.1212
8	conhecimento	N	8 (0.011)	4 (0.121)	0.1081
9	técnico	AQ	9 (0.013)	4 (0.121)	0.1053
10	análise	N	12 (0.017)	4 (0.121)	0.0976

Copy KWIC Sort: Jaccard Filter Network Hits: 33 Ready.

Fonte: Autora (2021) - KH Coder

Foram classificados 33 documentos no cluster 4. As palavras “prestação” e “serviço” aparecem em 100% e 90,9%, respectivamente, dos documentos. A palavra “eventual” aparece em 51,5% dos documentos. A palavra “consultoria” aparece em 36,4% dos documentos. A palavra “assessoria” aparece em 27,3% dos documentos. A rede de coocorrência do Cluster 4 pode ser observada na Figura 28.

Figura 32 - Rede de coocorrência do Cluster 4 - planilha RESPOSTAS



Fonte: Autora (2021) - KH Coder

“docente” aparece em 29,3% dos documentos. As palavras “progressão” e “cppd” aparecem em 27,6% e 22,8% dos documentos respectivamente. A rede de cocorrência do Cluster 5 pode ser observada na Figura 29, com destaque para as redes na cor vermelha.

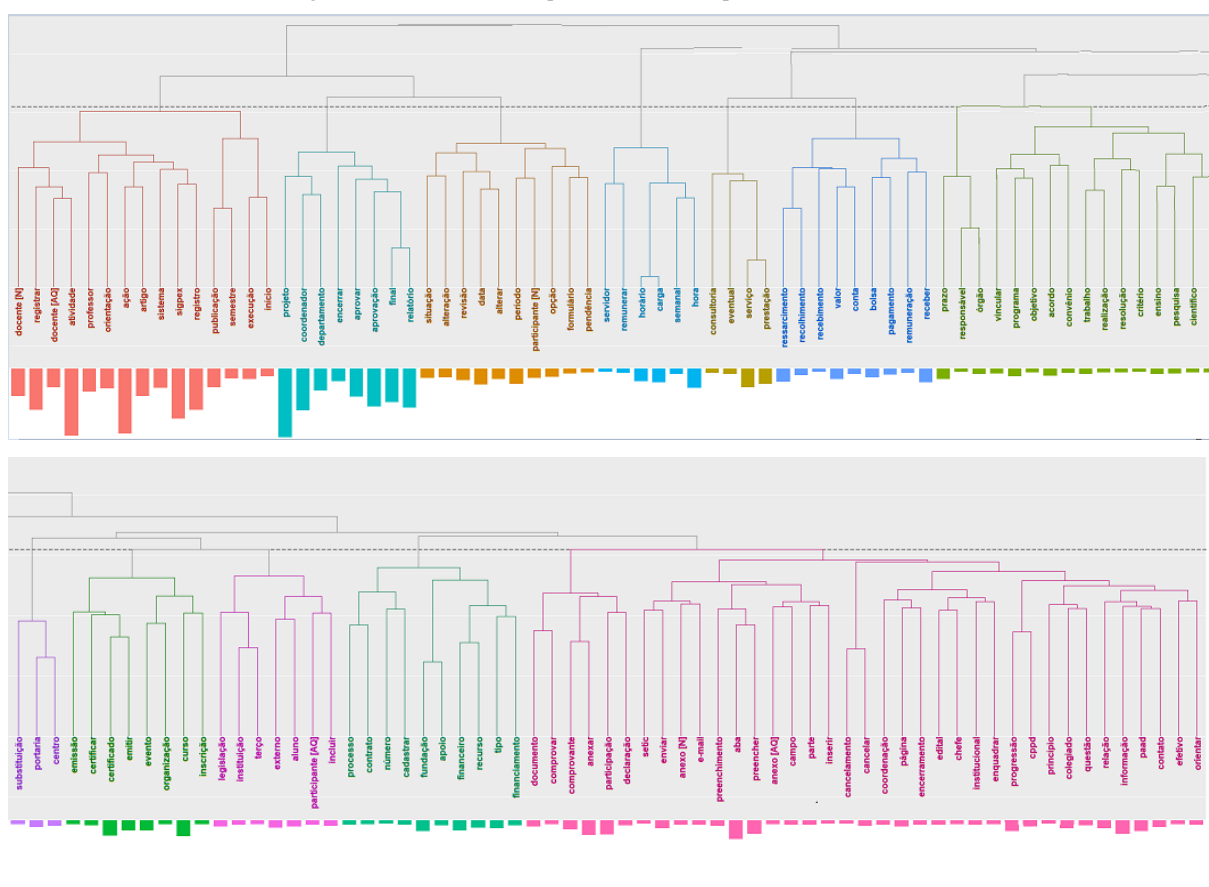
Para o cluster 5, pode-se identificar que as principais respostas estão relacionadas ao registro de atividades docentes para progressão junto à CPPD.

4) Análise hierárquica de cluster

A análise hierárquica de cluster (Figura 30) permitiu visualizar as combinações ou grupos de palavras que têm padrões de aparência semelhante usando análise de agrupamento hierárquico. O dendograma gerado mostra as palavras que aparecem no documento e suas combinações com as variáveis indicam as posições e comprimentos de documentos.

As diferentes cores do dendograma permitem distinguir os clusters. As linhas intra e extra cluster permitem a visualização da organização hierárquica das palavras, criando subgrupos de palavras dentro do próprio cluster, o que facilita a análise do mesmo.

Figura 34 - Análise Hierárquica de Cluster - planilha RESPOSTAS

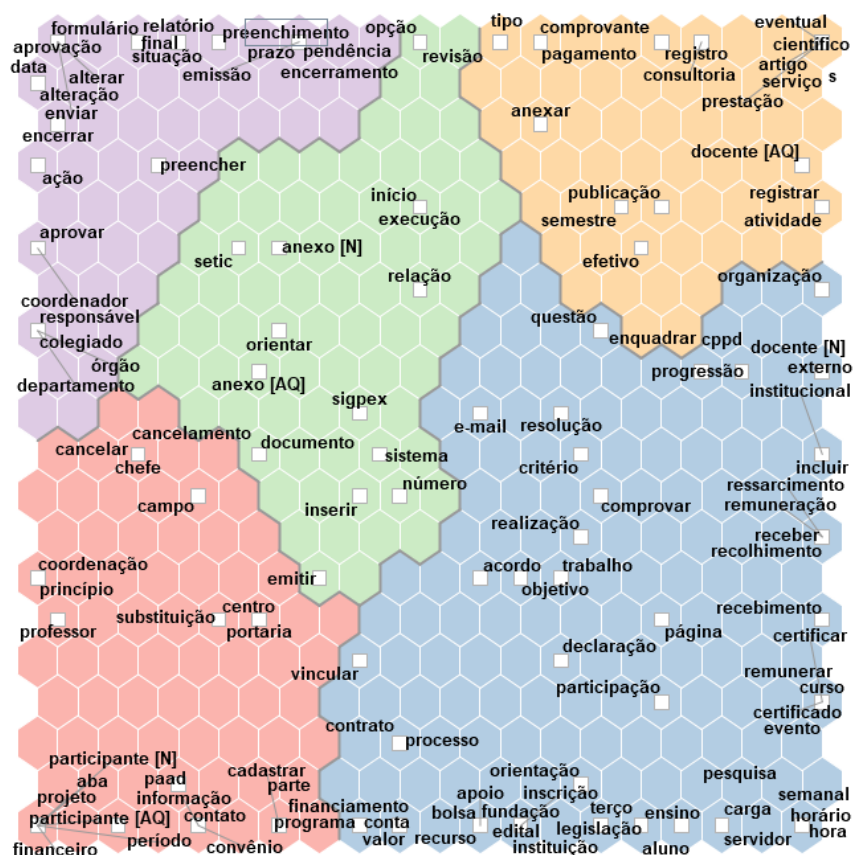


Fonte: Autora (2021) - KH Coder

5) Mapa auto-organizável

O mapa auto-organizável (Figura 31) permitiu explorar associações entre palavras, organizadas em cluster através da matriz com as variáveis para posições e comprimentos de documento removidos e criando um mapa auto-organizável usando distâncias euclidianas.

Figura 35 - Mapa auto-organizável - planilha RESPOSTAS



Fonte: Autora (2021) - KH Coder

6) Classificador Naive Bayes

A classificação automática das respostas em assuntos, seguiu o modelo de aprendizagem aplicado na planilha PERGUNTAS.

Assim, as 600 respostas foram classificadas automaticamente de acordo com um dos seguintes assuntos: **programa, projeto, curso, evento, prestação de serviço, curso de curta duração, publicação, banca externa, evento e palestra**, conforme o Quadro 102.

Quadro 102 - Classificação dos documentos - planilha RESPOSTAS

unit	variable	value	label	frequency
h5	Heading5	banca externa		46
h5	modelo_naive	curso		56
h5	tipo	curso curta duração		87
h5	subtipo	evento		40
h5	respostas	evento e palestra		18
		prestação de serviço		56
		programa		27
		projeto		235
		publicação		35

Fonte: Autora (2021) – KH Coder

Também foi possível exportar uma matriz de tabulação (Quadro 103), mostrando quantas vezes palavras específicas aparecem em cada documento, possibilitando o uso de informações em diferentes softwares estatísticos, permitindo análises mais especializadas dos dados

Quadro 103 - Matriz de tabulação - planilha RESPOSTAS

banca externa		curso		curso curta duração	
banca	0.1702	artigo	0.2520	registrar	0.2759
federal	0.1373	universidade	0.2083	atividade	0.2313
opção	0.1316	curso	0.1957	certificado	0.2252
listado	0.1224	ação	0.1814	docente	0.2099
progressão	0.1205	envolver	0.1690	progressão	0.2072
cppd	0.1081	equipar	0.1587	curso	0.1849
externo	0.1081	preferencial	0.1525	cppd	0.1604
consultar	0.1071	desenvolver	0.1493	docente	0.1592
funcional	0.1053	instituição	0.1406	sistema	0.1544
atividade	0.1031	ser	0.1364	ser	0.1510

evento e palestra		prestação de serviço		programa	
carta	0.5000	serviço	0.6667	programa	0.1667
palestra	0.4783	prestação	0.6056	encerrar	0.1410
convite	0.4583	eventual	0.3443	graziela	0.1190
declaração	0.2941	consultoria	0.2712	executar	0.1053
comprovante	0.2400	docente	0.2479	prof	0.0976
lives	0.2222	ressarcimento	0.2184	enviar	0.0909
palestrante	0.2222	recolhimento	0.2029	permitir	0.0909

guia	0.2000	pagamento	0.1912	lembrar	0.0893
evento	0.1837	remuneração	0.1719	haver	0.0891
certificar	0.1667	valor	0.1711	vincular	0.0889

publicação		evento		projeto	
semestre	0.4630	horário	0.4068	projeto	0.4764
publicação	0.4615	carga	0.3810	coordenador	0.3627
retroativo	0.1667	semanal	0.3333	aprovação	0.3626
momento	0.1304	hora	0.2658	final	0.3394
efetivo	0.1277	semana	0.2195	relatório	0.3274
comprovante	0.1176	seguir	0.1538	estar	0.3224
questão	0.1154	período	0.1512	sigpex	0.2634
registrar	0.1129	servidor	0.1429	ação	0.2407
cppd	0.1094	limite	0.1333	ir	0.2232
registro	0.1089	planejamento	0.1277	departamento	0.2138

Fonte: Autora (2021)

7) Concordância textual das palavras

Por meio da concordância textual das palavras foi realizada uma busca para descobrir quais legislações e endereços de páginas *Web* são mais referenciados nas respostas aos chamados. Para tanto, utilizou-se a lista de nomes de referência utilizados nos passos 4 (legislação – Apêndice 1) e 5 (endereços páginas *Web* – Apêndice 2) da preparação de dados.

Foram identificados 117 documentos que mencionam artigos da RN 88/2016/CUn como orientação (Quadro 104).

Quadro 104 - Concordância textual da palavra “RNE882016” - planilha RESPOSTAS

The screenshot shows the KWIC Concordance window. The search entry is 'RNE882016'. The results are displayed in a list view, showing text excerpts with the search term highlighted in red. The interface includes search options, sorting, and a status bar at the bottom showing 'Hits: 117, View: 1-117'.

Fonte: Autora (2021) - KH Coder

Abaixo, o número de vezes que cada legislação foi referenciada nas respostas dos atendimentos:

- 117 vezes - RN 88/2016/CUn = Resolução Normativa nº 88/2016/CUn
- 10 vezes - RN1142017 = Resolução Normativa nº 114/CUn/2017
- 6 vezes - L132432016 = Lei nº 13.243 de 2016
- 5 vezes - L128632013 = Lei nº 12.863 de 2013
- 5 vezes - L127722012 = Lei nº 12.772 de 2012
- 4 vezes - D104262020 = Decreto Federal nº 10.426 de 2020
- 3 vezes - RN132011 = Resolução Normativa nº 13/CUn/2011

As demais legislações foram referenciadas uma vez cada.

Abaixo, o número de vezes que cada página da *Web* foi referenciada nas respostas dos atendimentos:

- 11 vezes – Movimentações financeiras
- 8 vezes – Atividades docentes
- 6 vezes – Cartilha
- 5 vezes – Convênios PROAD
- 2 vezes – Tramita fácil
- 2 vezes – Sistema SIGPEX
- 2 vezes – Resolução extensão

As demais páginas de *Web* foram referenciadas uma vez cada.