

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO
MESTRADO EM CIÊNCIA DA INFORMAÇÃO
UNIVERSIDADE FEDERAL FLUMINENSE
INSTITUTO DE ARTE E COMUNICAÇÃO SOCIAL

LUCIANA LOPES MONTEIRO

**MENSAGENS TEXTUAIS NO CANAL DE ATENDIMENTO DO PORTAL
IBGEANDO: OBTENDO INSUMOS PARA A TOMADA DE DECISÃO
UTILIZANDO MINERAÇÃO DE TEXTOS**



Niterói
2017

LUCIANA LOPES MONTEIRO

**MENSAGENS TEXTUAIS NO CANAL DE ATENDIMENTO DO PORTAL
IBGEANDO: OBTENDO INSUMOS PARA A TOMADA DE DECISÃO
UTILIZANDO MINERAÇÃO DE TEXTOS**

Dissertação apresentada ao Programa de Pós-Graduação (stricto-sensu) em Ciência da Informação da Universidade Federal Fluminense como requisito parcial à obtenção do título de Mestre em Ciência da Informação.

Área de concentração: Dimensões Contemporâneas da Informação e do Conhecimento.

Linha de Pesquisa 2: Fluxos e Mediações Sócio-técnicas da Informação.

Orientador: Prof^o. Dr. Leonardo Cruz da Costa

**Niterói
2017**

M775 Monteiro, Luciana Lopes.
Mensagens textuais no Canal de Atendimento do Portal Ibgeando:
obtido insumos para a tomada de decisão utilizando mineração de
textos / Luciana Lopes Monteiro. – 2017.
163 f. ; il.
Orientador: Leonardo Cruz da Costa.
Dissertação (Mestrado em Ciência da Informação) – Universidade
Federal Fluminense, Departamento de Ciência da Informação, 2017.
Bibliografia: f. 148-155.

1. Comunidade virtual. 2. Análise do discurso. 3. Produção textual.
4. Mineração de texto. 5. Tomada de decisão. I. Costa, Leonardo Cruz
da. II. Universidade Federal Fluminense. Departamento de Ciência da
Informação. III. Título.

LUCIANA LOPES MONTEIRO

FOLHA DE APROVAÇÃO

Mensagens Textuais no Canal de Atendimento do Portal IBGEANDO: Obtendo insumos para a tomada de decisão utilizando Mineração de Textos

Dissertação apresentada ao Programa de Pós-Graduação (stricto-sensu) em Ciência da Informação da Universidade Federal Fluminense como requisito parcial à obtenção do título de Mestre em Ciência da Informação.

Área de concentração: Dimensões Contemporâneas da Informação e do Conhecimento.

Linha de Pesquisa 2: Fluxos e Mediações Sócio-técnicas da Informação.

Aprovado em: ___ / ___ / _____

BANCA EXAMINADORA:

Prof^o. Dr. Leonardo Cruz da Costa (Orientador)
Universidade Federal Fluminense (UFF)

Prof^o. Dr. Carlos Henrique Marcondes (Membro Titular Interno)
Universidade Federal Fluminense (UFF)

Prof^o. Dr. Claudio José Silva Ribeiro (Membro Titular Externo)
Universidade Federal do Estado do Rio de Janeiro (UNIRIO)

Prof^a. Dr^a. Ana Célia Rodrigues (Membro Suplente Interno)
Universidade Federal Fluminense (UFF)

Prof^a. Dr^a. Cicera Henrique da Silva (Membro Suplente Externo)
Instituto de Comunicação e Informação Científica e Tecnológica em Saúde (ICICT-FIOCRUZ)

**Niterói
2017**

AGRADECIMENTOS

A **Deus**, por estar sempre presente na minha vida.

Ao **IBGE**, por me conceder o afastamento do trabalho, investindo no meu desenvolvimento profissional e pessoal, além de autorizar o uso de informações primordiais para a minha pesquisa.

Ao Coordenador, **Bruno Malheiros**, e aos **servidores da CRH** que disponibilizaram o seu tempo para responder ao questionário desta pesquisa.

Às minhas amigas e companheiras da extinta DE/CRH/GEPRO, **Náthali Pereira, Paula Dias, Rachel Simões e Renata Baêta**, pelo carinho, auxílio e informações para o enriquecimento da pesquisa.

Ao meu orientador **Leonardo Cruz**, por sua paciência, confiança, disponibilidade e dedicação neste trabalho, deixando-me a vontade para trilhar os caminhos sugeridos por mim, mas sempre me conduzindo para a melhor direção para a pesquisa.

Aos **membros da Banca** por suas valiosas contribuições para o desenvolvimento teórico e metodológico deste trabalho.

Aos **professores do PPGCI-UFF** pelo constante apoio e aprendizado.

Ao **Vitor**, da Secretaria do PPGCI-UFF, que sempre me atendeu com cordialidade e presteza.

Aos meus **colegas da turma do mestrado e doutorado do PPGCI-UFF de 2015**, que me ajudaram, acolheram com carinho e tiveram paciência em ouvir as minhas diversas indagações nas aulas por não ser da área de Biblioteconomia e Arquivologia.

Ao Doutor, Pastor e amigo **Irenio Chaves**, por me incentivar a fazer a seleção do mestrado e acreditar na minha capacidade intelectual, além de me ajudar ao longo do percurso, emprestando livros e auxiliando nas revisões do trabalho.

Aos meus pais, **Lândi e Sandra**, que com seus princípios, ensinamentos e amor incondicional, sempre me incentivaram a buscar conhecimento e a estudar cada vez mais, investindo na minha educação desde pequena.

Ao meu querido irmão, **Leandro Monteiro**, por seus conhecimentos técnicos que me socorreram nos momentos críticos durante a dissertação.

Ao meu marido **Ricardo Leal**, por estar sempre ao meu lado, apoiando nos momentos necessários, e pela paciência e cuidados com a nossa filha enquanto estive ausente.

À minha amada filha **Luana Leal**, que foi concebida e nasceu durante o meu mestrado. Ela “roubou” um tempo dos meus estudos, mas eu também estive ausente em alguns preciosos momentos do seu desenvolvimento para finalizar esta dissertação. Ela deu um novo sentido para a minha vida e para ela, eu dedico este trabalho.

A todos, muito obrigada por tudo!

“nem tudo é verdadeiro, mas em todo lugar e a todo momento existe uma verdade a ser dita e a ser vista, uma verdade talvez adormecida, mas que, no entanto, está somente à espera de nosso olhar para aparecer, à espera de nossa mão para ser desvelada. A nós cabe achar as boas perspectivas, o ângulo correto, os instrumentos necessários, pois de qualquer maneira ela está presente aqui e em todo lugar”.

Foucault

RESUMO

Estudo que tem como objetivo geral analisar a viabilidade do uso de ferramenta de mineração de textos para processamento automático das mensagens textuais no Canal de Atendimento do Portal IBGEANDO de forma a obter insumos para auxiliar a tomada de decisão e melhorar a comunicação da área de Recursos Humanos. Apresenta uma análise comparativa entre ferramentas de mineração de textos prospectadas em estudos já realizados de mineração de textos em meios digitais, como canais de atendimento e *blogs*, com o intuito de selecionar uma ferramenta que se aproxima da ideal a ser implantada no IBGE para minerar as mensagens objeto deste estudo. Apresenta os resultados obtidos com a mineração das mensagens textuais por categoria, incluindo a sua visualização gráfica. Utiliza Prova de Conceito para comprovar o pressuposto com validação dos resultados pelos servidores da área de RH do IBGE.

PALAVRAS-CHAVE: Comunidade Discursiva; Mensagens Textuais; Mineração de Texto; Tomada de Decisão.

ABSTRACT

This study has as general objective to analyze the feasibility of using a text mining tool for automatic processing of textual messages in the service channel of the IBGEANDO Portal in order to obtain inputs to aid decision making and improve the communication of the Human Resources area. It presents a comparative analysis between text mining tools prospected in studies already done of text mining in digital media, such as service channels and blogs, with the purpose of selecting a tool that approaches the ideal to be implemented in IBGE to mine the textual messages object of this study. It presents the results obtained with the mining of textual messages by category, including their graphic visualization. It uses Proof of Concept to prove the assumption with validation of the results by the servants of the IBGE HR area.

KEY WORDS: Discourse Community; Textual messages; Text Mining; Decision Making.

LISTA DE FIGURAS

| | |
|--|-----|
| Figura 1: Visão holística do uso da informação – Choo (2003)..... | 51 |
| Figura 2: Etapas do processo de mineração de textos | 58 |
| Figura 3: Tela inicial da ferramenta <i>KH Coder</i> | 85 |
| Figura 4: Exemplo do arquivo de <i>stemming</i> do <i>KH Coder</i> | 86 |
| Figura 5: Exemplo da visualização da frequência de palavras do <i>KH Coder</i> | 86 |
| Figura 6: Exemplo de Grafo do <i>KH Coder</i> | 87 |
| Figura 7: Tela inicial da ferramenta <i>SOBEK Mining</i> | 90 |
| Figura 8: Grafo da ferramenta <i>SOBEK Mining</i> | 91 |
| Figura 9: Seleção de nodos da ferramenta <i>SOBEK Mining</i> | 91 |
| Figura 10: Tela da ferramenta <i>TagCrowd</i> com o resultado..... | 93 |
| Figura 11: Tela inicial da ferramenta <i>TextAlyser</i> | 94 |
| Figura 12: Exibição do Resultado Geral da ferramenta <i>TextAlyser</i> | 94 |
| Figura 13: Exibição das palavras com maior frequência da ferramenta <i>TextAlyser</i> | 95 |
| Figura 14: Exibição das sequências de palavras mais frequentes da ferramenta <i>TextAlyser</i> ... | 95 |
| Figura 15: Tela inicial da ferramenta <i>WordCounter</i> | 96 |
| Figura 16: Exibição do resultado da ferramenta <i>WordCounter</i> | 96 |
| Figura 17: Grafo - Categoria Promoção / Progressão Funcional – 15 termos..... | 111 |
| Figura 18: Grafo - Categoria Política e Diretrizes – 15 termos..... | 112 |
| Figura 19: Grafo - Categoria Avaliação de Desempenho – 15 termos..... | 113 |
| Figura 20: Grafo - Categoria Aposentadoria – 15 termos | 115 |
| Figura 21: Grafo - Categoria Sistemas de RH – 15 termos..... | 116 |
| Figura 22: Grafo - Categoria Sistemas de RH – 30 termos..... | 119 |
| Figura 23: Grafo - Categoria Outros – 15 termos..... | 122 |
| Figura 24: Grafo - Categoria Saúde – 15 termos..... | 124 |
| Figura 25: Grafo - Categoria Normas e Legislação – 15 termos..... | 126 |
| Figura 26: Grafo - Categoria Titulação/Qualificação – 15 termos..... | 127 |
| Figura 27: Grafo - Categoria Cadastro Pessoal/Funcional – 15 termos | 129 |
| Figura 28: Grafo - Categoria Ressarcimento de Saúde – 15 termos | 131 |
| Figura 29: Grafo - Categoria SECAF Greve – 15 termos | 132 |
| Figura 30: Grafo - Categoria Pagamento – 15 termos..... | 134 |
| Figura 31: Resumo do objetivo da pesquisa: Mensagens Textuais, Termos, Insumos e Decisões..... | 145 |

LISTA DE GRÁFICOS

| | |
|---|-----|
| Gráfico 1: Quantidade de pessoas por tempo de Serviço no IBGE e na CRH..... | 137 |
| Gráfico 2: Quantidade de pessoas respondendo à seção “Qual a sua Dúvida?” por tempo ... | 137 |
| Gráfico 3: Quantidade de pessoas que continuam respondendo à seção “Qual a sua Dúvida?” | 138 |
| Gráfico 4: Consolidação das respostas referentes às perguntas fechadas do 2ª parte do questionário | 138 |
| Gráfico 5: Análise das respostas da questão II-d do questionário | 139 |

LISTA DE QUADROS

| | |
|---|-----|
| Quadro 1: Mensagens textuais postadas por categoria..... | 22 |
| Quadro 2: Estrutura do relatório do “Qual a sua dúvida?” | 24 |
| Quadro 3: Exemplos de mensagens postadas na Categoria Aposentadoria | 25 |
| Quadro 4: Listagem das ferramentas utilizadas em trabalhos similares..... | 78 |
| Quadro 5: Método de Reeves e os critérios desejáveis para ferramenta de mineração de textos | 79 |
| Quadro 6: Critérios e parâmetros para uma ferramenta ideal para mineração de textos do IBGEANDO | 82 |
| Quadro 7: Comparativo entre as ferramentas selecionadas - Critério #1 | 98 |
| Quadro 8: Comparativo entre as ferramentas selecionadas – Critério #2 | 98 |
| Quadro 9: Comparativo entre as ferramentas selecionadas – Critério #3 | 99 |
| Quadro 10: Comparativo entre as ferramentas selecionadas – Critério #4 | 99 |
| Quadro 11: Comparativo entre as ferramentas selecionadas – Critério #5 | 100 |
| Quadro 12: Comparativo entre as ferramentas selecionadas – Critério #6 | 100 |
| Quadro 13: Comparativo entre as ferramentas selecionadas – Critério #7 | 101 |
| Quadro 14: Comparativo entre as ferramentas selecionadas – Critério #8 | 101 |
| Quadro 15: Comparativo entre as ferramentas selecionadas – Critério #9 | 102 |
| Quadro 16: Comparativo entre as ferramentas selecionadas – Critério #10 | 102 |
| Quadro 17: Comparativo entre as ferramentas selecionadas – Critério #11 | 103 |
| Quadro 18: Comparativo entre as ferramentas selecionadas – Critério #12 | 103 |
| Quadro 19: Comparativo entre as ferramentas selecionadas – Critério #13 | 104 |
| Quadro 20: Comparativo entre as ferramentas selecionadas – Critério #14 | 104 |
| Quadro 21: Comparativo entre as ferramentas selecionadas – Critério #15 | 105 |
| Quadro 22: Comparativo entre as ferramentas selecionadas – Critério #16 | 105 |
| Quadro 23: Comparativo entre as ferramentas selecionadas – Critério #17 | 106 |
| Quadro 24: Comparativo entre as ferramentas selecionadas – Critério #18 | 106 |
| Quadro 25: As categorias e suas respectivas quantidades de mensagens..... | 109 |

LISTA DE TABELAS

| | |
|--|-----|
| Tabela 1: Frequência de Palavras – Categoria Promoção / Progressão Funcional – 15 termos | 111 |
| Tabela 2: Frequência de Palavras – Categoria Política e Diretrizes – 15 termos..... | 112 |
| Tabela 3: Frequência de Palavras – Categoria Avaliação de Desempenho – 15 termos..... | 114 |
| Tabela 4: Frequência de Palavras – Categoria Aposentadoria – 15 termos | 115 |
| Tabela 5: Frequência de Palavras – Categoria Sistemas de RH – 15 termos | 117 |
| Tabela 6: Frequência de Palavras – Categoria Sistemas de RH – 30 termos | 120 |
| Tabela 7: Frequência de Palavras – Categoria Outros – 15 termos..... | 123 |
| Tabela 8: Frequência de Palavras – Categoria Saúde – 15 termos..... | 124 |
| Tabela 9: Frequência de Palavras – Categoria Normas e Legislação – 15 termos..... | 126 |
| Tabela 10: Frequência de Palavras – Categoria Titulação/Qualificação – 15 termos..... | 128 |
| Tabela 11: Frequência de Palavras – Categoria Cadastro Pessoal/Funcional – 15 termos | 129 |
| Tabela 12: Frequência de Palavras – Categoria Ressarcimento de Saúde – 15 termos..... | 131 |
| Tabela 13: Frequência de Palavras – Categoria SECAF Greve – 15 termos | 133 |
| Tabela 14: Frequência de Palavras – Categoria Pagamento – 15 termos..... | 134 |
| Tabela 15: Termos mais frequentes em mais de uma categoria analisada | 142 |

LISTA DE ABREVIATURAS E SIGLAS

| | |
|---------|---|
| ABRH-RJ | Associação Brasileira de Recursos Humanos – Rio de Janeiro |
| BRAPCI | Base de Dados Referencial de Artigos de Periódicos em Ciência da Informação |
| CAPES | Coordenação de Aperfeiçoamento de Pessoal de Nível Superior |
| CI | Ciência da Informação |
| CRH | Coordenação de Recursos Humanos |
| DCT | Descoberta de Conhecimento em Textos |
| DE | Diretoria-Executiva |
| ETC | Editor de Texto Coletivo |
| GQ | Gratificação de Qualificação |
| HTML | <i>HyperText Markup Language</i> |
| IBGE | Fundação Instituto Brasileiro de Geografia e Estatística |
| PDF | <i>Portable Document Format</i> |
| POC | <i>Proof of Concept</i> (Prova de Conceito) |
| RH | Recursos Humanos |
| SAC | Serviço de Atendimento ao Consumidor |
| SECAF | Sistema Eletrônico de Controle de Acesso e Frequência |
| SIAPE | Sistema Integrado de Administração de Recursos Humanos |
| UFF | Universidade Federal Fluminense |
| URL | <i>Uniform Resource Locator</i> |
| XML | <i>eXtensible Markup Language</i> |

SUMÁRIO

| | |
|--|------------|
| 1 INTRODUÇÃO | 15 |
| 1.1 Objeto de estudo: mensagens textuais da seção “Qual a sua dúvida?” do IBGEANDO | 21 |
| 1.2 Justificativa..... | 27 |
| 1.3 Objetivos..... | 31 |
| 2 METODOLOGIA..... | 32 |
| 3 REVISÃO DE LITERATURA..... | 36 |
| 3.1 Comunidade Discursiva..... | 37 |
| 3.2 Tipificação de Mensagem Textual e seu Uso Social..... | 41 |
| 3.3 Tomada de Decisão | 45 |
| 3.4 Mineração de Textos | 53 |
| 3.4.1 <i>Trabalhos relacionados sobre mineração de mensagens textuais</i> | 64 |
| 3.5 Conceituação de Insumos | 73 |
| 4 RESULTADOS ENCONTRADOS..... | 77 |
| 4.1 Análise e Seleção da Ferramenta de Mineração de Textos | 77 |
| 4.2 Processamento das Mensagens Textuais por Categoria | 107 |
| 4.2.1 <i>Categoria: Promoção / Progressão Funcional</i> | 110 |
| 4.2.2 <i>Categoria: Política e Diretrizes</i> | 112 |
| 4.2.3 <i>Categoria: Avaliação de Desempenho</i> | 113 |
| 4.2.4 <i>Categoria: Aposentadoria</i> | 114 |
| 4.2.5 <i>Categoria: Sistemas de RH</i> | 116 |
| 4.2.6 <i>Categoria: Outros</i> | 122 |
| 4.2.7 <i>Categoria: Saúde</i> | 124 |
| 4.2.8 <i>Categoria: Normas e Legislação</i> | 125 |
| 4.2.9 <i>Categoria: Titulação / Qualificação</i> | 127 |
| 4.2.10 <i>Categoria: Cadastro Pessoal / Funcional</i> | 128 |
| 4.2.11 <i>Categoria: Ressarcimento de Saúde</i> | 130 |
| 4.2.12 <i>Categoria: SECAF Greve</i> | 132 |
| 4.2.13 <i>Categoria: Pagamento</i> | 133 |
| 4.3 Validação dos Resultados..... | 135 |
| 4.4 Discussão sobre os Resultados | 141 |
| 5 CONSIDERAÇÕES FINAIS..... | 144 |
| REFERÊNCIAS | 148 |
| APÊNDICES | 156 |
| Apêndice A – Questionário de Análise do Grafo da Categoria “Sistemas de RH”..... | 157 |
| Apêndice B - Quadro único com o comparativo das ferramentas de mineração de textos | 161 |
| ANEXOS | 162 |
| Anexo A – Autorização de acesso aos dados do IBGEANDO | 163 |

1 INTRODUÇÃO

O IBGE – Fundação Instituto Brasileiro de Geografia e Estatística é um órgão da Administração Pública Indireta Federal, fundado no ano de 1936, com aproximadamente 11.760 profissionais (servidores do quadro permanente, estagiários e contratados temporários), estando presente em todos os estados e Distrito Federal e 579 agências.¹

No ano de 2008, o IBGE apresentou o seu Programa de Melhoria da Qualidade na Gestão Institucional para os anos de 2008 a 2011:²

O IBGE, no cumprimento da sua missão de “Retratar o Brasil com informações necessárias ao conhecimento de sua realidade e ao exercício da cidadania”, produz e dissemina informações estatísticas e geocientíficas que oferecem ao governo e à sociedade os elementos fundamentais para a compreensão da realidade nacional e elementos essenciais para o enfrentamento dos desafios do país, bem como a possibilidade de uma gestão pública mais eficiente para o planejamento nacional de longo prazo. Visando este objetivo, o IBGE está alinhado com o desafio do governo de implementar uma nova gestão pública: ética, transparente, participativa, descentralizada, com controle social e orientada para o cidadão. Nesse contexto, a Diretoria-Executiva - DE do IBGE vem implementando ações que permitam à instituição se manter na vanguarda, através da promoção de um canal de reflexão e de debates, bem como através da elaboração e execução de projetos que intentem imprimir melhoria significativa na gestão institucional, tendo como macro-objetivo estratégico “Assegurar a Qualidade da Gestão Institucional”. (IBGE, 2008, p. 7)

Este Programa se propôs a introduzir mudanças de valores e comportamentos individuais e organizacionais, e se constituiu como o principal instrumento para a transposição de uma cultura burocrática para uma cultura gerencial no IBGE, assumindo caráter estratégico no âmbito da instituição, na medida em que procurou implementar e institucionalizar boas práticas de gestão, sempre levando em consideração os fundamentos da Qualidade na Gestão Pública. Foi necessário rever todos os processos de trabalho com vistas à sua maior eficiência e eficácia, bem como assegurar a infraestrutura necessária a seu funcionamento e desenvolvimento.

¹ Números extraídos do sistema governamental SIAPE no mês de junho de 2015

² Todo o Programa de Melhoria da Qualidade na Gestão Institucional para o período de 2008 a 2011 está descrito num documento interno do IBGE, disponível na intranet da Diretoria-Executiva no link <http://portal.de.ibge.gov.br/web/cps1/programa-de-melhoria>, acesso em julho de 2014. As informações sobre este Programa descritas neste projeto de pesquisa foram extraídas deste documento.

Conforme descrito no Programa de Melhoria da Qualidade na Gestão Institucional, a sua implementação teve como base os seguintes princípios:

- Agilidade: Buscar a constante redução de etapas que não agregam valor. Reduzir ao mínimo o tempo de permanência dos documentos em trâmite nas unidades intermediárias envolvidas. Manter comunicação clara ponto a ponto. Manter-se apto a ter segurança na ação de forma a dar perfeita fluidez aos processos de trabalho.
- Qualidade da Informação: Garantir o máximo de precisão da informação. Fornecer segurança quanto a fidedignidade da fonte. Disponibilizar a informação no tempo esperado por seus usuários.
- Confiabilidade: Garantir qualidade de resultado, do produto ou serviço. Dar transparência e transmitir segurança quanto a legalidade das ações. Manter-se comprometido com todo o processo até a entrega do produto ou serviço no tempo esperado por seus clientes.
- Relacionamento: Formar e fortalecer continuamente a rede de relacionamento com grupos de interesse e aliados estratégicos. Manter interatividade com todas as fronteiras do processo de trabalho.
- Cordialidade: Dar tratamento justo e manter atitude permanente de civilidade, junto a pares, colaboradores, clientes, parceiros, fornecedores e demais elementos da ambiência interna e externa à Instituição. (IBGE, 2008, p. 24)

Tendo em vista estes princípios, e o papel da Diretoria-Executiva – DE, da qual a Coordenação de Recursos Humanos (CRH) faz parte, de elaboração e execução dos projetos deste Programa de Melhoria da Qualidade na Gestão Institucional, surgiu o projeto de reformulação da *intranet* da CRH com foco nas demandas do servidor “ibgeano”, público interno da instituição. A *intranet* fornecia informações com conteúdo estático feito em HTML³, sem uma ferramenta de busca de conteúdo, além de ser uma página meramente informativa e desatualizada por depender exclusivamente de técnicos de informática para fazer a atualização da página.

Assim, surgiu o projeto chamado Portal do Servidor com o intuito de ser mais do que uma simples *intranet*, mas um Portal com o conteúdo sendo atualizado de forma ágil, fácil e descentralizado, com ferramentas de busca de conteúdo, com espaços que permitissem a colaboração e integração entre os servidores, utilizando uma ferramenta que permitisse esta atualização fácil e descentralizada de conteúdo.

³ HTML (abreviação para a expressão inglesa *HyperText Markup Language*, que significa Linguagem de Marcação de Hipertexto) é uma linguagem de marcação utilizada para produzir páginas na *web*.

Desta forma, foi selecionada a ferramenta *Joomla!*⁴ que utiliza conceitos de *Web 2.0*⁵.

Então, em outubro de 2008 foi implantado o “IBGEANDO – Portal do Servidor”, nome este escolhido pelos próprios servidores por meio de um concurso interno. Este projeto permitiu aprimorar o canal de atendimento ao servidor, centralizar as informações pertinentes à área de Recursos Humanos (RH), padronizar os processos e procedimentos internos referentes a esta área e fomentar maior interação com o servidor com o objetivo de valorizar, desenvolver e reter os talentos do IBGE. Foi uma iniciativa da área de RH do IBGE que serviu também de modelo para outras instituições governamentais.

A expectativa da CRH era que a implantação deste Portal proporcionasse os seguintes benefícios:⁶

- a) através de uma linguagem fácil, torná-los usuários constantes de informações oferecidas pelo Portal;
- b) aprimorar o canal de atendimento da área de RH com o servidor;
- c) centralizar as informações pertinentes à área de RH;
- d) permitir o controle de acesso e dar transparência às informações;
- e) criar homogeneidade de atuação da área de RH em todo o Brasil, padronizando processos e procedimentos;
- f) criar canais de colaboração e integração entre os servidores, proporcionando a inclusão digital e fomentando a criatividade;
- g) facilitar o processo de digitalização das informações.

O IBGEANDO abrange uma grande variedade de conteúdos distribuídos em seções específicas de cada gerência de RH, utilizando uma ferramenta *Web 2.0*, como: Notícias e Quadro de Avisos, Guia do Servidor, Formulários e legislações vigentes referentes ao IBGE com foco em RH e Sistema de Busca de informações disponíveis no Portal.

⁴ *Joomla!* é um dos principais sistemas de gestão de conteúdo da atualidade (*Content Management System - CMS*).

⁵ *Web2.0* é um termo que designa uma segunda geração de comunidades e serviços na *web*, tendo a *web* como plataforma, envolvendo funcionalidades colaborativas, *blogs* e redes sociais.

⁶ A descrição do IBGEANDO com suas seções e conteúdos, as expectativas da implantação do Portal, os seus benefícios e desafios citados neste projeto, foram extraídos da documentação do Projeto do Portal do Servidor, um documento interno do IBGE, criado em maio de 2008 pela equipe de RH da Diretoria-Executiva.

Dando continuidade aos princípios do Programa de Melhoria da Qualidade na Gestão Institucional, em maio de 2009 foi implantada uma das seções mais importantes do Portal, a seção “Qual é a sua Dúvida?”, onde os servidores postam dúvidas, reclamações e elogios a respeito da área de RH. Até o mês de abril de 2015, foram postadas mais de 40.000 dúvidas em todo o Brasil, sendo manualmente analisado somente o conteúdo de algumas respostas aleatórias para assegurar que as informações fornecidas sejam precisas e completas. Esta análise era feita por um Gestor que só tinha a preocupação de garantir que as dúvidas fossem respondidas corretamente. A média de perguntas respondidas na semana era em torno de 70%.⁷

Mas os desafios para esta implantação foram grandes, tendo em vista o comportamento, demanda e perfil do público-alvo, o servidor “ibgeano” do quadro permanente do IBGE, em torno de 6.100 pessoas:⁸

- a) grande parte dos servidores tinham nível intermediário: cerca de 4.500;
- b) somente 1.600 tinham nível superior;
- c) a maioria: 4.400 servidores possuíam mais de 20 anos de serviço, quase que 73% do quadro permanente do IBGE;
- d) grande parte sem familiaridade com novas tecnologias;
- e) público reticente a mudanças e acomodado;
- f) conheciam pouco das informações existentes.

No entanto, durante esses 8 (oito) anos de existência do IBGEANDO, a percepção é de que foi atingido o resultado esperado desta implantação, tendo em vista a maior participação dos servidores nas seções colaborativas, aos questionamentos mais fundamentados de acordo com as informações disponíveis no Portal e ao maior uso por servidores com mais tempo de serviço e de outras Unidades Estaduais do IBGE. Isto pode ser observado analisando a base de dados do Portal. Esta nova *intranet* permitiu à instituição reduzir a distância entre a sede (Rio de Janeiro) e as unidades estaduais do IBGE, acolher o servidor e dar transparência às informações de RH. Agilidade na resposta, facilitação do processo de digitalização, controle de acesso às informações de RH, inclusão digital, transparência são exemplos de resultados obtidos com o Portal do Servidor do IBGE, sendo considerado como uma boa prática de

⁷ Os números de dúvidas postadas e a média de perguntas respondidas são extraídos dos relatórios de atendimento disponíveis no módulo administrativo do Portal IBGEANDO referentes ao mês de abril de 2015.

⁸ Números extraídos do sistema governamental SIAPE no mês de dezembro de 2013

Gestão de Pessoas no Âmbito da Administração Pública. O IBGEANDO, inclusive, foi selecionado como um dos cinco *cases* finalistas do 32º Prêmio Ser Humano da ABRH-RJ (Associação Brasileira de Recursos Humanos – Rio de Janeiro) no ano de 2012 na categoria *Case Organizacional* no grupo Organização do Setor Público.⁹

No ano de 2011, o IBGE se deparou com o desafio de integrar a comunicação, em suas diferentes formas de processamento, aos objetivos e metas da instituição. Desta forma, foi criado um comitê para implementar a Política de Comunicação Integrada¹⁰ como um instrumento orientador e normativo, concebido para sistematizar as ações de comunicação do IBGE, maximizando seu desempenho de forma que a Comunicação Integrada fosse tratada como uma área estratégica, em que os comunicadores fossem ouvidos e participassem, quando pertinente, da tomada de decisões no IBGE, buscando soluções que visassem a dar transparência e agilidade ao fluxo de informações e que, conseqüentemente, fortalecessem a reputação e o reconhecimento institucional.

Nesta política existia um conjunto de ações a serem executadas que teve como subsídio os resultados da pesquisa interna sobre “Hábitos de Comunicação do Servidor” que foi respondida por 2.753 servidores do quadro permanente (equivalente a 45,1% do total de servidores), de 21 de maio a 15 de junho de 2012. Os resultados mostraram que 14,6% dos servidores nunca se sentiam informados sobre as questões organizacionais.

A análise dos resultados destacou que:

- a) o IBGEANDO era o canal mais utilizado, de uma forma geral, para os servidores obterem informações de pessoal, com 51% das indicações;
- b) quando desejavam esclarecer uma dúvida ou fazer uma consulta, os servidores utilizavam com mais frequência o IBGEANDO, com 37%;
- c) para fazer uma crítica ou sugestão, o IBGEANDO aparecia como a principal opção geral, com 28,7% das respostas. No entanto, a informação que surpreendeu foi que a resposta “NÃO SABE” aparecia em segundo lugar, com 28% dos servidores afirmando desconhecer o melhor canal para fazer uma crítica ou sugestão.

⁹ O Prêmio Ser Humano ABRH-RJ (Associação Brasileira de Recursos Humanos) é um reconhecimento aos profissionais que atuam na atividade de gestão com pessoas, às organizações do setor privado e do setor público cuja atuação diferenciada e práticas inovadoras nessa atividade tenham alcançado significativos resultados quantitativos e qualitativos, que possam ser consideradas uma referência no mercado. Mais informações disponíveis no site <http://www.abrhrj.org.br>.

¹⁰ Toda a Política de Comunicação Integrada está descrita num documento interno do IBGE, disponível na *intranet* da Presidência no link http://w3.presidencia.ibge.gov.br/comites_comissoes_grupos/interna/pdf/cpc/politica_comunicacao.pdf, acesso em maio de 2014. As informações sobre esta Política, os resultados da pesquisa interna sobre os hábitos de comunicação do servidor e a análise dos resultados foram extraídos deste documento.

- d) as três dificuldades de comunicação percebidas como principais pelos servidores foram a ausência de um canal eficiente de dúvidas (41,7%), a existência de muitos canais de informação (40%) e o atraso na informação “oficial”, praticamente empatadas. Merecendo destaque também o aspecto Falta de Informação, com 34,7% das respostas.

Esta pesquisa mostrou que o IBGEANDO atingiu o seu objetivo inicial, no entanto novas lacunas surgiram com a sua implantação, como a integração ao processo de tomada de decisões e à comunicação estratégica institucional.

Analisando os dados disponíveis na base de dados do IBGEANDO, a área de RH também se deparou com uma quantidade de informações diversas e primordiais que poderiam ser utilizadas para melhorar a gestão e a comunicação interna, mas que por desconhecimento teórico e técnico, não sabiam como obtê-los de forma eficiente.

Esta produção de quantidade de informações foi devido ao uso de ferramentas da *Web* 2.0 que propicia, de modo formal ou informal, que os servidores compartilhem ideias, dúvidas, complementando e disseminando informações entre os servidores, através dos canais de *blogs* e de atendimento. Isso acarreta também num aumento dos fluxos de informação dentro da organização e servem para melhorar a comunicação interpessoal, intrapessoal e organizacional.

No entanto, a falta de uma estratégia de comunicação, falta de ações focadas nos problemas demandados pelos servidores e muitas informações subutilizadas no Portal para subsidiar políticas de RH são alguns dos problemas que ainda precisam de solução na instituição e que um estudo mais aprofundado sobre estas informações disponíveis no Portal pode ser um caminho para encontrar soluções para estes problemas.

Uma questão identificada na instituição é que o IBGE instituiu o Programa de Melhoria da Qualidade da Gestão Institucional no ano de 2008, com o intuito de ter uma gestão transparente e participativa, e implementou uma Política de Comunicação Integrada no ano de 2012 de forma a ser uma área estratégica onde os servidores fossem ouvidos e participassem da tomada de decisões, e, no entanto, mesmo depois de oito anos de uso do Portal IBGEANDO (2008-2016), a área de RH não utiliza as mensagens postadas neste Portal para ouvir todos os servidores do Brasil, de forma sistemática e periódica, a fim de auxiliar a tomada de decisões e melhorar a comunicação institucional. Ao ler os relatórios emitidos pelo módulo de Administração do Portal IBGEANDO, foi possível observar que as mensagens postadas pelos servidores mostravam a avaliação dos próprios servidores em relação ao funcionamento dos processos de RH e seus sistemas, indicando os principais problemas

operacionais, assim como a percepção e visão dos servidores em relação à Gestão da Instituição num determinado período.

Desta forma, surge a proposta de identificar e avaliar, de uma forma automática, se as mensagens textuais postadas no IBGEANDO podem auxiliar à tomada de decisões e melhorar a comunicação institucional. Como o Portal possui seções bem abrangentes, o presente estudo se concentra somente nas mensagens textuais postadas na seção “Qual a sua Dúvida?”, onde os questionamentos individuais dos servidores do IBGE são respondidos pela equipe de RH.

Este campo de pesquisa foi escolhido levando em consideração que:

- a) existem várias seções estáticas no Portal, que são cadastradas pela equipe de RH, e outras seções que são colaborativas, com postagens dos servidores de todo o IBGE;
- b) este Portal é exclusivamente voltado para a área de RH;
- c) a pesquisa interna sobre “Hábitos de Comunicação do Servidor” aponta que os servidores percebem como principais dificuldades de comunicação, a ausência de um canal eficiente de dúvidas, o atraso na informação “oficial” e o aspecto de falta de informação.

Assim, a questão a ser investigada por esta pesquisa é a de identificar e avaliar se as mensagens textuais postadas pelos servidores no Canal de Atendimento, “Qual a sua Dúvida?”, do Portal IBGEANDO apresentam insumos para a tomada de decisão e melhoria da comunicação institucional na área de RH.

1.1 Objeto de estudo: mensagens textuais da seção “Qual a sua dúvida?” do IBGEANDO

A seção “Qual a sua Dúvida?” tem o objetivo de criar um canal único e centralizado de atendimento aos servidores de forma a manter o controle dos atendimentos e dúvidas solicitados via *intranet*. Os servidores podem postar dúvidas, reclamações e elogios a respeito da área de RH. Esta seção do IBGEANDO está dividida em 20 categorias onde para cada categoria existe uma equipe responsável, que possui o conhecimento sobre o assunto específico, de forma a responder corretamente aos questionamentos dos servidores. Estas categorias foram criadas para se ter um melhor direcionamento das solicitações dos servidores, mais agilidade nas respostas e posterior análise geral de atendimento por equipe.

Cabe ressaltar que a denominação das Categorias foi de acordo com os principais temas tratados na CRH, não houve nenhum estudo específico para esta categorização.

O servidor ao cadastrar a sua dúvida deve indicar em qual categoria a dúvida se encaixa e esta já estará vinculada à gerência e aos responsáveis pelo assunto. As dúvidas são encaminhadas através de formulários próprios e as respostas são enviadas por *e-mail* para os solicitantes. O responsável deverá responder à dúvida no próprio Portal num prazo de 7 (sete) dias úteis. Após esse tempo sem resposta, o próprio sistema dispara alertas para os responsáveis, através de *e-mail*. Um responsável pode redirecionar a dúvida para outro responsável de outra categoria, caso o usuário tenha indicado a categoria indevida para a dúvida, mas é mantido o mesmo prazo de resposta. Existe também um espaço para consulta das dúvidas mais frequentes por categoria. A publicação destas dúvidas frequentes é de responsabilidade da CRH, eleitas como de importância e alta relevância para todos os servidores. Os usuários poderão pesquisar esta seção antes de formularem suas dúvidas.

O Quadro 1 mostra o total de mensagens textuais postadas por categoria no período de Maio/2009 até Maio/2015, no total foram 40.165 mensagens postadas.

Quadro 1: Mensagens textuais postadas por categoria

(continua)

| Categorias | Quantidade de mensagens | Período | | Observação |
|------------------------------|-------------------------|----------------|------------|---|
| | | Data de início | Data final | |
| Aposentadoria | 530 | 19/05/2009 | 10/04/2015 | Servidores aposentados não possuem acesso ao IBGEANDO. As mensagens são de servidores ativos que tiram dúvidas sobre a aposentadoria. |
| Avaliação de Desempenho | 505 | 19/05/2009 | 08/04/2015 | |
| Cadastro Pessoal Funcional | 817 | 19/05/2009 | 07/04/2015 | |
| Cargos e Salários | 178 | 19/05/2009 | 08/04/2015 | |
| Concurso / Processo Seletivo | 173 | 19/05/2009 | 17/03/2015 | |

Quadro 1: Mensagens textuais postadas por categoria

(conclusão)

| Categorias | Quantidade de mensagens | Período | | Observação |
|---------------------------------|-------------------------|----------------|------------|--|
| | | Data de início | Data final | |
| Estágio de Estudante | 117 | 10/06/2009 | 31/03/2015 | |
| Normas e Legislação | 2.091 | 19/05/2009 | 09/04/2015 | |
| Outros | 817 | 19/05/2009 | 10/05/2015 | Qualquer assunto de RH que não se enquadre nas categorias específicas. |
| Pagamento | 1.703 | 19/05/2009 | 13/04/2015 | |
| Pensão | 52 | 27/05/2009 | 03/12/2014 | Pensionistas não possuem acesso ao IBGEANDO. |
| Políticas e Diretrizes de RH | 493 | 14/04/2009 | 02/04/2015 | |
| Programa Novo Tempo | 23 | 03/05/2011 | 06/04/2015 | Novo Tempo é o nome do Programa de preparação de Aposentadoria que começou no ano de 2011. |
| Promoção / Progressão Funcional | 432 | 19/05/2009 | 10/04/2015 | |
| Ressarcimento de Saúde | 1.342 | 11/01/2011 | 08/04/2015 | Categoria incluída em Janeiro de 2011 quando foi implantado o processo automatizado de Ressarcimento de Saúde. |
| Saúde | 1.244 | 29/04/2009 | 09/04/2015 | |
| SECAF | 24.480 | 19/05/2009 | 13/04/2015 | SECAF é o sistema Eletrônico de Controle de Acesso e Frequência. É o sistema mais utilizado na Instituição uma vez que controla as faltas e imp pontualidades dos servidores no trabalho e os respectivos cálculos dos descontos caso as imp pontualidades e faltas não sejam compensadas. |
| SECAF - Greve | 1.519 | 19/05/2009 | 13/04/2015 | Para melhor controle, resolveram criar uma categoria específica do SECAF para os períodos de greve. |
| Segurança do Trabalho | 23 | 12/11/2009 | 23/03/2015 | |
| Sistemas de RH | 650 | 22/05/2009 | 08/04/2015 | |
| Titulação / Qualificação | 2.976 | 19/05/2009 | 01/04/2015 | A partir de 21/01/2013 foram implantados os processos automatizados de Qualificação, até janeiro de 2013 só existiam 280 dúvidas e depois este número cresceu enormemente. |

Fonte: Elaborado pela autora

O módulo de Administração do IBGEANDO gera relatórios no formato “csv” com as mensagens postadas nestas categorias. O Quadro 2 mostra a estrutura destes relatórios com a descrição dos campos existentes.

Quadro 2: Estrutura do relatório do “Qual a sua dúvida?”

| CAMPO | DESCRIÇÃO |
|--------------------|---|
| UF | UF do Estado de lotação do servidor solicitante |
| Siape | Matrícula do servidor solicitante |
| Solicitante | Nome do servidor solicitante |
| Tipo | Tipo da Dúvida: Dúvida, Reclamação, Sugestão ou Elogio |
| Texto | Mensagem do servidor solicitante |
| Categoria | Categoria da Mensagem |
| Data - Pergunta | Data de postagem da mensagem |
| Hora - Pergunta | Hora de postagem da mensagem |
| Resposta | Mensagem de resposta do responsável da categoria |
| Data - Resposta | Data de resposta da mensagem |
| Hora - Resposta | Hora de resposta da mensagem |
| Tempo de Resposta | Tempo que a solicitação levou para ser respondida |
| Quem Respondeu | Nome do servidor responsável pela resposta |
| Data Encaminhada | Data que a mensagem foi encaminhada, se foi postada na categoria indevida |
| Hora Encaminhada | Hora que a mensagem foi encaminhada |
| Categoria Anterior | Nome da categoria que a mensagem foi postada inicialmente |
| Quem Encaminhou | Nome do servidor responsável que encaminhou a solicitação |

Fonte: Elaborado pela autora

Com estes campos do relatório, é possível fazer vários levantamentos, como:

- a) servidores que mais postam mensagens;
- b) responsáveis que mais demoram a responder, o que pode sinalizar algum problema na equipe;
- c) quais os Estados que mais possuem mensagens postadas e as que menos postam por categoria, o que pode indicar uma ação específica do RH num Estado;
- d) categorias que são mais encaminhadas, o que pode indicar um problema na denominação das categorias;
- e) perfil dos servidores que mais postam no canal, se servidores novos na Instituição ou os mais antigos e etc.

O Quadro 3 mostra algumas mensagens textuais postadas na Categoria Aposentadoria na estrutura do relatório com alguns campos ocultados.

Quadro 3: Exemplos de mensagens postadas na Categoria Aposentadoria

| UF | Siape | Solicitante | Tipo | Texto | Data - Pergunta |
|----|-------|-------------|------------|---|-----------------|
| PE | ** | ** | DÚVIDA | Tenho 54 anos de idade, 33 anos de contribuição, e o salário de nível médio, eu prestando o concurso de nível superior do IBGE, caso seja aprovado, gostaria de saber quanto tempo teria que trabalhar a mais nesta nova função de nível superior, diante da minha idade e contribuição acima mencionado? Grato. | 21/02/2011 |
| RJ | ** | ** | ELOGIO | Gostaria de elogiar o trabalho deste setor por ter resolvido com competência assunto tratado sobre contribuições simultâneas para o calculo da aposentadoria. | 26/09/2013 |
| GO | ** | ** | DÚVIDA | Pretendo me aposentar em 2014, mês de março. Tenho dúvida sobre o Art 12 da OS. CRH no. 05/13 (14/10/13), que trata sobre Indenização das Férias. Poderei ser indenizado se não gozar as férias, ou preciso tirá-las antes de me aposentar? | 22/10/2013 |
| MG | ** | ** | DÚVIDA | Quando o servidor já tem idade e tempo de contribuição para se aposentar pela PEC Paralela e falece em atividade, o pensionista tem direito ou não a paridade? | 30/10/2013 |
| RJ | ** | ** | RECLAMAÇÃO | Solicito saber o porque quando envio o pedido de formulário do NADA CONSTA, para fins de aposentadoria o mesmo não esta vindo em minha caixa do correio(email). Ja é o segundo pedido que faço a até o momento, nada. | 24/02/2014 |
| RJ | ** | ** | DÚVIDA | Prezados bom dia. Gostaria de saber quais são as regras de aposentadoria para mim que entrei no IBGE em 05/06/2010. Tenho 18,5 anos de contribuição no mercado privado. Pelo que entendi preciso de: 30 anos de contribuição (faltam 11,5 anos), 10 anos no serviço público e 55 anos para me aposentar com rendimentos iguais ao teto do INSS. Meu entendimento está correto? Eu poderia me aposentar assim que completasse os 30 anos de contribuição, ou seja, com 52 anos? Obrigada | 09/01/2015 |

Fonte: Elaborado pela autora

Este estudo está focado somente nas mensagens textuais postadas pelos servidores solicitantes com o intuito de identificar sobre “o quê?” os servidores estão falando. Analisando as mensagens também é possível fazer um recorte sobre o que se fala num determinado período, num determinado Estado, numa determinada Região. Pode-se fazer também uma análise das respostas dos responsáveis para saber se respondem ao solicitado, confrontando com o questionamento inicial. No entanto, inicialmente, este estudo se concentra apenas nas mensagens textuais dos servidores de forma a identificar o que eles estão dizendo numa determinada categoria num contexto geral, de forma a obter os termos destas mensagens para que sirvam de insumos para a tomada de decisão.

A idéia não é estabelecer um vocabulário controlado e, sim, identificar sobre o que os servidores estão discursando.

O objeto de estudo é complexo, uma vez que se trata de mensagens textuais escritas num formato livre, sem limite de caracteres para escrita. Não existe um vocabulário controlado e nem correções de ortografia, o que pode acarretar em eventuais falhas na identificação destes termos e sua posterior análise para obtenção dos insumos para a tomada de decisão. Além disso, o volume destas mensagens é grande, pois existem mais de 40.000 mensagens textuais, gerando arquivos, para cada categoria, de tamanho entre 150kb a 1Mb, com exceção da categoria SECAF, que só ela possui mais de 24.000 mensagens gerando um arquivo de 10Mb. Ou seja, tratar estas mensagens textuais manualmente é uma tarefa árdua e morosa, então se faz necessário adotar uma abordagem automática para obtenção destes insumos.

Cabe ressaltar que foi concedida autorização pelo Diretor-Executivo do IBGE para acesso a todas as mensagens textuais da seção “Qual a sua dúvida?” (vide ANEXO A), mantendo sigilo sobre as informações obtidas e analisadas, durante todo o estudo em questão, assim como a não identificação de qualquer informação sobre os servidores que postaram as mensagens, como nome, matrícula SIAPE ou número de processo administrativo. Todos os resultados obtidos com esta pesquisa serão disponibilizados ao IBGE para que possa utilizá-los visando à melhoria na tomada de decisões e comunicação interna.

1.2 Justificativa

Neste estudo, é discutido o conceito de comunidade discursiva, a tipificação de mensagens textuais, o uso social destas mensagens e o processo de mineração de textos de forma a contribuir para uma gestão da informação mais eficiente dentro do IBGE que acarrete em implantação de políticas, serviços ou sistemas que atendam aos resultados esperados dos servidores e da própria Instituição.

Como o IBGE possui o Programa de Melhoria da Qualidade da Gestão Institucional, com o intuito de ter uma gestão transparente e participativa, e a Política de Comunicação Integrada onde os servidores sejam ouvidos e participem da tomada de decisões, acredita-se que a instituição quer um envolvimento maior dos servidores, ou seja, estabelecer “um processo participativo que utiliza toda a capacidade dos funcionários e tem por objetivo estimular um comprometimento crescente com o sucesso da organização.” (ROBBINS, 2005, p. 164). Desta forma, os servidores se tornam mais motivados, mais comprometidos, mais produtivos e mais satisfeitos com a instituição.

A principal característica comum a todos os programas de gestão participativa é a utilização do processo decisório coletivo. Isto quer dizer que os subordinados realmente compartilham um grau relevante de poder decisório com seus chefes imediatos (...). A participação também aumenta o comprometimento com as decisões tomadas. Há menor probabilidade de que as pessoas sabotem uma decisão no momento de sua implementação se elas participarem de sua definição. Finalmente, a participação traz recompensas intrínsecas para os funcionários. (ROBBINS, 2005, p. 164).

Além disso, existe uma constante preocupação do IBGE em relação à melhoria da comunicação institucional pois:

[...] a função da comunicação se relaciona a seu papel como facilitadora de tomada de decisões. Ela proporciona as informações de que as pessoas e os grupos precisam para tomar decisões ao transmitir dados para que se identifiquem e avaliem alternativas (...) Os canais formais são estabelecidos pela organização e transmitem mensagens que se referem às atividades relacionadas com o trabalho de seus membros (...) Os canais informais são espontâneos e surgem como resposta às escolhas individuais. (...) O ruído é composto pelas barreiras à comunicação que distorcem a clareza da mensagem. Exemplos de possíveis fontes de ruído incluem problemas de percepção, excesso de informações, dificuldades semânticas ou diferenças culturais. (ROBBINS, 2005, p. 233).

Uma vez que a pesquisa interna sobre “Hábitos de Comunicação do Servidor” do IBGE teve, dentre os resultados sobre dificuldades de comunicação percebidas, a ausência de um canal eficiente de dúvidas (41,7%), o atraso na informação “oficial” (40%) e a falta de informação (34,7%), há que se investigar melhor este resultado para entender as suas causas, pois “a escolha do canal adequado, a escuta eficaz e a utilização do *feedback* podem ajudar muito a comunicação a se tornar mais eficaz”. (ROBBINS, 2005, p. 262)

Tendo em vista o exposto e, considerando que existe um canal de atendimento no IBGEANDO bastante utilizado pelos servidores, o “Qual a sua Dúvida?”, conforme resultado apontado na pesquisa interna mencionada anteriormente, onde os mesmos se expressam por mensagens de texto para esclarecer dúvidas, criticar ou elogiar os processos ou a própria área de RH, chega-se a conclusão da importância de se desenvolver estudos e técnicas que apresentem caminhos para analisar estas mensagens textuais, de forma periódica e sistemática, com o intuito de obter insumos para auxiliar a tomada de decisão e melhorar a comunicação. Estas mensagens textuais devem representar o discurso da comunidade e a fala institucional.

A comunicação escrita é tangível e verificável. Geralmente, tanto o emissor quanto o receptor mantêm registro das mensagens. Elas podem ficar armazenadas por muito tempo. Se houver dúvidas quanto a seu conteúdo, elas podem ser facilmente verificadas nos registros. Esse aspecto é particularmente importante quando se trata de mensagens complexas ou muito longas. (ROBBINS, 2005, p. 235)

Assim, parece claro perceber a importância da informação que pode ser obtida através da análise destas mensagens textuais como um elemento chave capaz de, quando identificado e interpretado, gerar conhecimento para os gestores da organização para a formulação de estratégias e tomada de decisões. “A informação é reconhecida como um recurso fundamental, tendo em vista que, se gerida eficientemente, promove desenvolvimento organizacional e valores estratégico e econômico. Nesse cenário, o diferencial competitivo é adquirido a partir da gestão eficiente dos insumos informacionais”. (CÂNDIDO, 2011, p. 193)

Outro ponto que reforça a importância do estudo destas mensagens textuais, é que:

As possibilidades de melhorar a comunicação interna e externa foram, assim, muitas vezes ampliadas, permitindo aos colaboradores, desde que possibilitadas as condições necessárias, participarem e colaborarem na vida organizacional, fazendo ouvir a sua voz e podendo introduzir ideias novas, que podem ir de encontro à resolução de problemas de uma forma mais

eficaz e eficiente, ao invés de recorrer a opiniões e pessoas externas à organização. Por vezes, o conhecimento existe dentro da própria organização, em um estado latente ou então de dormência. As ferramentas colaborativas permitem reavivar ou tornar útil esse conhecimento, partilhado e disponível, desta forma. (MARAVILHAS-LOPES, 2013, p. 132)

Assim, acredita-se que analisando as mensagens textuais postadas pelos servidores no canal de atendimento, é possível que a CRH possa obter insumos que auxiliem a tomada de decisões e na resolução de alguns problemas. E, caso ocorram ações efetivas na instituição a partir da análise destas mensagens, consequentemente, os servidores podem perceber que a seção “Qual a sua dúvida?” realmente é um canal de atendimento que funciona, gerando um sentimento de satisfação por saber que a instituição tem interesse em saber dos assuntos tratados individualmente, existindo a possibilidade de se tornarem ações coletivas, incentivando mais ainda o uso do canal de atendimento.

Muitos estudiosos abordaram o processo de comunicação com maior ou menor ênfase, o que, de certa forma, traduz a influência da teoria de informação na área (...) e que, no âmbito da ciência da informação, a comunicação pode ser entendida, mais apropriadamente, como transferência da informação. (PINHEIRO, 1995, p. 2)

No entanto, esta pesquisa pretende contribuir com a CI na medida em que avalia que a comunicação não se traduz somente em transferência da informação e que neste processo deve ser discutido o que é a informação e para quem, o contexto sócio-histórico cultural em que esta comunicação ocorre e o caráter seletivo da mensagem.

[...] a partir do questionamento dos aspectos decisoriais e seletivos das práticas e ações de informação, consideraremos, de um modo mais direto, que alguma forma de seleção, individual e social, de caráter emocional, cultural, prático e gnosiológico, participa da emergência de um valor de informação.” (GONZÁLEZ DE GOMEZ, 1999, p. 3)

Outra justificativa para a relevância teórica e social do problema é que:

Não basta que algo esteja gravado ou registrado em qualquer suporte material, faz-se necessário avaliar a qualidade, a utilidade, a relevância, o uso, a pertinência, a potencialidade de gerar conhecimento que essa informação carrega. Entretanto, uma informação particular ou singular pode ser a mais bem equipada em todos esses quesitos, mas caso não seja registrada, armazenada, organizada, selecionada e recuperada por suportes materiais e tecnológicos, não cumprirá no ser humano a realização da sua ideia. Porém, qualquer informação particular depende do entendimento de

qualquer humano, ou seja, sem entendimento não se pode denominar qualquer informação particular de informação. (XAVIER, 2010, p. 82)

Esta pesquisa tem o objetivo de apresentar um caminho para viabilizar a construção de uma prática sistematizada para que os gestores tenham acesso a informações primordiais para a tomada de decisões e para o desenvolvimento de ações voltadas para o público interno do IBGE, ressaltando que “toda investigação em qualquer área do conhecimento necessita ter relevância científica e relevância social para justificar seu desenvolvimento”. (FUJITA, 2013, p. 147)

1.3 Objetivos

O objetivo geral desta pesquisa é analisar a viabilidade do uso de ferramenta de mineração de textos para processamento automático das mensagens textuais no Canal de Atendimento do Portal IBGEANDO de forma a obter insumos para auxiliar a tomada de decisão e melhorar a comunicação da área de RH.

Com base nas mensagens textuais postadas pelos servidores no Portal, é possível obter informações importantes que, se recuperadas, mostram indicadores e subsídios para direcionar e atender de forma mais eficaz aos anseios dos servidores, melhorando a satisfação dos servidores no trabalho e mostrando questões primordiais a serem focadas pela área de RH ou da própria instituição. Além dos gestores passarem, também, a conhecer melhor a realidade institucional de forma a formularem e implementarem estratégias de comunicação interna mais direcionados.

Os objetivos específicos desta pesquisa são:

- a) Revisar a literatura correspondente à Ciência da Informação (CI) que trata sobre comunidade discursiva, mensagens textuais e mineração de textos;
- b) Pesquisar trabalhos realizados ou em andamento sobre mineração de textos em meios digitais, em ambiente *web*, como *blogs* e canais de atendimento;
- c) Prospeccionar ferramentas de mineração de textos que possam processar automaticamente as mensagens textuais, identificando os termos e suas relações contidos nas mensagens;
- d) Definir critérios para a escolha de uma ferramenta de mineração de texto para analisar e processar as mensagens textuais;
- e) Apresentar os termos e suas relações de forma sistematizada de maneira auxiliar na obtenção de insumos para a tomada de decisão;
- f) Validar os resultados com servidores da CRH responsáveis pela seção “Qual a sua dúvida?”.

2 METODOLOGIA

Parte-se do pressuposto que as mensagens textuais postadas no Portal IBGEANDO apresentam insumos para a tomada de decisão, tendo em vista que os servidores do IBGE funcionam como uma comunidade discursiva, que possuem objetivos públicos em comum apesar das aspirações individuais que podem compreender o coletivo e acarretar algumas vezes em conflitos. Estas mensagens devem estar carregadas do discurso institucional e possuem uma mesma linguagem, e apesar de ser um canal de atendimento da área de RH, muitas vezes a pretensão da mensagem é dirigida para os Diretores ou Presidente do IBGE. Ao usar e analisar estas mensagens textuais de uma forma sistemática, automática e criteriosa, pode-se obter insumos para auxiliar o processo de tomada de decisões, assim como melhorar a comunicação interna e ouvir o que realmente os servidores da instituição estão falando, possibilitando que a Fundação desenvolva ações que atendam ou minimizem os problemas já identificados pela pesquisa interna sobre “Hábitos de Comunicação do Servidor”.

Para isso, foi necessário entender o conceito de Comunidade Discursiva, caracterizar estas mensagens textuais que são o objeto do estudo e identificar uma forma de analisar estas mensagens automaticamente com o intuito de obter insumos para a tomada de decisão.

O estudo foi realizado no âmbito da Fundação IBGE, analisando as mensagens textuais postadas por cada servidor no canal de atendimento “Qual a sua dúvida?” do IBGEANDO, contendo o seu questionamento individual dentro de um contexto coletivo.

A forma de abordagem para esta pesquisa foi qualitativa, uma vez que foi proposta uma análise automática das mensagens textuais, utilizando uma ferramenta de Mineração de Textos, de forma a apoiar a tomada de decisão e a melhoria da comunicação interna da CRH. O que se desejou foi identificar na fala dos servidores insumos para a tomada de decisão, pois “[...] nada pode ser intelectualmente um problema se não tiver sido, em primeiro lugar, um problema da vida prática. As questões da investigação estão, portanto, relacionadas a interesses e circunstâncias socialmente condicionadas”. (MINAYO, 2011, p.16). Utilizando a pesquisa qualitativa foi trabalhado “o universo dos significados, dos motivos, das aspirações, das crenças, dos valores e das atitudes”. (MINAYO, 2011, p. 21).

Tendo em vista a necessidade de estudos para maior compreensão do uso da informação em prol de uma tomada de decisão mais eficaz e um alinhamento da comunicação

interna da CRH de forma a atender também aos anseios dos servidores públicos da Fundação, se fez necessário uma pesquisa bibliográfica aprofundada, como livros, teses e artigos de origem nacional ou internacional, sobre Tomada de Decisão, Comunidade Discursiva, definição e uso de Mensagens Textuais e Mineração de Textos. Como Marconi e Lakatos (1996, p. 23-24) enfatiza, “a pesquisa bibliográfica é um apanhado geral sobre os principais trabalhos já realizados, revestidos de importância, por serem capazes de fornecer dados atuais e relevantes relacionados com o tema”. Além disso, também houve a necessidade de uma busca sobre trabalhos já realizados sobre Mineração de Textos em meios digitais ou em ambiente *web*, de forma a complementar e recuperar o conhecimento científico sobre o assunto.

A proposta foi realizar a modalidade de Pesquisa Exploratória, pois “tem o objetivo de reunir dados, informações, padrões, ideias ou hipóteses sobre um problema ou questão de pesquisa com pouco ou nenhum estudo anterior”. (BRAGA, 2007, p. 25), em conjunto com a Descritiva, que “tem o objetivo de identificar as características de um determinado problema ou questão e descrever o comportamento dos fatos e fenômenos”. (BRAGA, 2007, p. 25).

Com esta metodologia, pretendeu-se cobrir todos os aspectos necessários para a análise das mensagens textuais postadas que estão armazenadas na base de dados do Portal IBGEANDO, e de viabilizar que esta análise apresente ou não insumos para auxiliar o processo de tomada de decisão e melhorar a comunicação da área da CRH. Desta forma, o campo empírico desta pesquisa foram as mensagens textuais postadas no canal de atendimento chamado “Qual a sua dúvida?”, no período de Maio/2009 a Maio/2015 pelos servidores ativos de todas as Unidades Federativas do Brasil que estão disponíveis na base de dados do Portal IBGEANDO.

Este estudo tentou identificar o que está sendo falado nas mensagens textuais e quais os assuntos principais destas mensagens de acordo com as categorias. Para isso, foi utilizada uma ferramenta de mineração de textos para identificar os termos mais relevantes contidos nestas mensagens.

De uma forma geral, este trabalho fez uso da Prova de Conceito (POC), que tem como propósito verificar se um conceito ou teoria tem potencial aplicação num mundo real.¹¹ Neste trabalho, o conceito a ser provado foi que a utilização de uma ferramenta de mineração de textos para processar as mensagens textuais do canal de atendimento serve para obter insumos

¹¹ Definição de POC disponível no site <https://www.techopedia.com/definition/4066/proof-of-concept-poc>

para a tomada de decisão. A prova de conceito trouxe alguns benefícios para esta proposta como: evitar o desperdício de investimento financeiro e humano em algo que não é tecnicamente viável, obter provas ou evidências tangíveis de que a ideia é de valor e que não poderá facilmente ser descartada com argumentos contrários e traz confiança para a equipe que irá trabalhar com a ferramenta, uma vez que passa a visualizar que a ferramenta pode trazer resultados satisfatórios.

A prova de conceito é muito utilizada em desenvolvimento de *softwares*. Como Shoaib Ahmed, formado em computação e gerente de serviços, explica, a prova de conceito é um pequeno exercício para testar um pressuposto. Para ser deliberado, uma POC deve claramente indicar o que é para ser provado e em que grau.¹²

A POC tem como característica se limitar a uma pequena parte do trabalho a fim de minimizar tempo e custo. Considerando esta característica, neste trabalho, a ideia não foi estabelecer requisitos funcionais para o desenvolvimento de uma ferramenta de Mineração de Textos idealizada para o processamento de mensagens textuais de um canal de atendimento para o IBGE, e sim, escolher uma ferramenta já pronta e disponível para uso que seja a mais próxima do que seria a ferramenta ideal.

Assim, esta pesquisa abrangeu as seguintes fases:

1. Pesquisa bibliográfica que aborde os conceitos de Comunidade Discursiva, Gêneros Textuais, Mineração de Textos e Tomada de Decisão;
2. Pesquisa sobre trabalhos que aborde mensagens textuais, sejam em canais de atendimento, *e-mails*, *blogs*, redes sociais, fóruns de discussão, entre outros, e a utilização de ferramentas de mineração de textos para identificar os termos das mensagens textuais. Esta pesquisa foi feita nas bases BRAPCI¹³, de teses do Programa de Pós-Graduação em Ciência da Informação da UFF¹⁴, no Portal de Periódicos da CAPES¹⁵ e, de forma adicional, através da ferramenta de busca *Google Acadêmico*¹⁶, utilizando as palavras-chave: mineração de textos, *text mining*, mensagens textuais;

¹² Explicação de Shoaib Ahmed sobre POC disponível no site <https://www.projectmanagement.com/blog-post/6136/Proof-of-Concept--Prototype--Pilot--Agile---confused->

¹³ <http://www.brapci.inf.br>

¹⁴ <http://www.ci.uff.br/ppgci/index.php/dissertacoes/>

¹⁵ <http://www.periodicos.capes.gov.br/>

¹⁶ <http://scholar.google.com.br>

3. Realização da Prova de Conceito que consistiu nas seguintes etapas:
- a. Prospectar ferramentas de mineração de texto para analisar e processar as mensagens textuais. As ferramentas prospectadas foram as obtidas a partir dos trabalhos relacionados na fase 2;
 - b. Estabelecer critérios para a escolha de uma ferramenta entre aquelas identificadas na etapa anterior;
 - c. Escolher uma ferramenta aplicando os critérios definidos na etapa anterior;
 - d. Processar as mensagens textuais por categoria utilizando a ferramenta escolhida;
 - e. Validar junto aos servidores da CRH, responsáveis pela seção “Qual a sua dúvida?”, os resultados obtidos pelo processamento na etapa anterior.

Apesar do processo de tomada de decisão não ser fácil e dada à complexidade, da quantidade e qualidade da informação, esperou-se com este projeto de pesquisa viabilizar o uso de uma ferramenta de mineração de textos para processamento automático das mensagens textuais postadas disponíveis no canal de atendimento do Portal IBGEANDO para que no IBGE este processo não seja tão penoso, danoso e burocrático e que a comunicação utilizada seja mais transparente e eficaz.

3 REVISÃO DE LITERATURA

Para melhor delimitar e embasar teoricamente o presente projeto de pesquisa, se faz necessário abordar os seguintes conceitos:

- a) definição de comunidade discursiva, para entender o funcionamento e papel dos servidores dentro da instituição IBGE, da comunidade “ibgeana” como se autodenominam internamente o grupo de servidores que trabalham no IBGE;
- b) tipificação das mensagens textuais e seu uso social, para definir e caracterizar as mensagens textuais postadas pelos servidores no canal de atendimento “Qual a sua dúvida?”;
- c) mineração de textos, como um processo de identificação dos termos e as suas relações contidos nas mensagens textuais para obtenção dos insumos para a tomada de decisão;
- d) entendimento do processo de tomada de decisão;
- e) conceituação do termo insumos sendo equivalente ao termo Informação na CI.

Todo este estudo visa entender o funcionamento e importância das comunidades discursivas, representados aqui pelos servidores do IBGE, que através das mensagens escritas no Portal IBGEANDO podem revelar insumos para tomada de decisão e melhorar a comunicação da CRH. Assim, pode-se afirmar que:

[...] a informação responde a condições daquilo acerca do que informa, estabelecendo relações com uma ordem cultural, cognitiva, ética e estética, na qual estão enraizadas suas referências semânticas e de conteúdo; remete, neste sentido, a uma formação discursiva e a seus universos de referência. Pode-se denominar a esta linha de articulação simplesmente “informação”. Finalmente, toda ação de informação que constrói um novo valor de informação age a partir de algo que a precede e que reúne uma memória de ações de informação — ora intencionais, ora anônimas; ora institucionalizadas, ora não institucionalizadas —, junto com todos os instrumentos e meios disponibilizados pelo ambiente cultural. (GONZÁLEZ DE GOMEZ, 1999, p. 5).

3.1 Comunidade Discursiva

O conceito de comunidade discursiva vem sendo utilizado por linguistas e pesquisadores que adotam uma visão de discurso como prática social, e de produção textual como atividade socialmente situada, realizada dentro de comunidades que possuem convenções específicas sobre a forma e o conteúdo dos textos.

Hjørland (2002a, 2002b) diz que o problema essencial na Ciência da Informação é a forma como as pessoas interpretam os textos a serem organizados e procurados, de maneira que satisfaça as necessidades de informações dos usuários. Para ele, as necessidades de informação dependem dos problemas a serem resolvidos, da natureza do conhecimento disponível e das qualificações dos usuários, pois a maioria dos problemas de informação são altamente complexas e as necessidades de informação tendem a ser socializadas por várias influências teóricas e paradigmas.

Para Hjørland e Albrechtsen (1995), a análise do domínio, que é um paradigma social da Ciência da Informação, é a melhor maneira de entender a informação, estudando os domínios do conhecimento como pensamento e as comunidades discursivas como parte da divisão social do trabalho na sociedade. Organização do conhecimento, estrutura, padrões, linguagem e formas de comunicação, sistemas de informação são exemplos dos objetos de trabalho dessas comunidades e seu papel na sociedade.

Hjørland (1997) define a análise de domínio como uma abordagem sociocognitiva focando a interação do usuário individual e do ambiente social/organizacional, onde a informação seria melhor compreendida sendo estudada a partir dos domínios de conhecimento (*domain analysis*) relacionados a suas comunidades discursivas (*discourse communities*). As comunidades discursivas são aquelas formadas pelo pensamento, linguagem e conhecimento sincronizados de grupos sociais distintos que fazem parte da sociedade moderna; são construções sociais constituídas por indivíduos e suas dimensões culturais, sociais e históricas.

As características principais da análise do domínio são:

- a) prioridade em entender as necessidades dos usuários pela perspectiva social, as necessidades de informação são consideradas como sendo causadas por fatores sociais e culturais;
- b) abordagem coletiva;

- c) preocupada com a natureza do conhecimento;
- d) autonomia e clareza dos textos no discurso.

Para Capurro (1985), a abordagem da análise de domínio entende o conhecimento como resultado da interação do sujeito com o meio, o que apresenta como desafio metodológico para a Ciência da Informação o deslocamento do individualismo metodológico para o seu oposto, o coletivismo metodológico.

Um ponto central na abordagem de Hjørland (2002b) é a afirmação de que as ferramentas, conceitos, significados, estruturas de informação, as necessidades de informação e critérios de relevância são moldados em comunidades discursivas. A comunidade discursiva é uma comunidade em que ocorre um processo de comunicação de forma ordenada e limitada, estruturada por um contexto institucional. Este ponto de vista muda o foco dos indivíduos (ou computadores) para o mundo social, cultural e científico. Hjørland (2002a) diz que tais comunidades discursivas podem ser muito diferentes em tamanho e estrutura e que os símbolos e seus significados são formados por grupos sociais como parte da divisão social do trabalho na sociedade. Um grande número de grupos pode desenvolver sistemas de símbolos e compartilhar conhecimento que eles não compartilham com o resto da sociedade.

A análise de domínio reconhece que as comunidades discursivas compõem-se de atores com pontos de vista distintos, estruturas de conhecimento individuais, predisposições, critérios de relevância subjetivos, estilos cognitivos particulares. Mas se fazem presentes no jogo entre as estruturas de domínio e o conhecimento individual e na interação entre o nível individual e social. A história do indivíduo, inserida dentro de uma história coletiva, apresenta suas variáveis e diferenças, e são estas que caracterizam as possibilidades de diferentes percepções, trajetórias, propósitos e apreciações em cada domínio de conhecimento. (NASCIMENTO; MARTELETO, 2004, p.9)

Swales (1990) é outro linguista que estuda as comunidades discursivas. Para ele a comunidade discursiva e seus membros reconhecem as estruturas informacionais produzidas já que revelam, num conjunto de eventos ou propósitos comunicativos, a forma dos moldes de discursos, a influência e restrição de conteúdo, a escolha de estilo e a pretensão da audiência. Ele aponta seis características que podem definir uma comunidade discursiva:

- a) o conjunto de objetivos públicos comuns;
- b) a existência de mecanismos para comunicação entre os participantes;

- c) a participação da comunidade nos mecanismos de comunicação principalmente para fornecer informação e *feedback*;
- d) a capacidade que a comunidade tem para desenvolver seus próprios gêneros de comunicação;
- e) o uso de um léxico específico;
- f) a existência de membros que possuem um conhecimento profundo do discurso e dos conteúdos que circulam na comunidade.

Analisando estas características apontadas por Swales, no contexto do IBGE, com o intuito de identificar se o grupo de servidores do IBGE pode ser denominado de comunidade discursiva, verifica-se que:

- a) os servidores do IBGE são regidos pelo estatuto dos servidores públicos federais, que é a Lei 8.112/91, que define os direitos e deveres dos servidores. Além disso, o IBGE possui um estatuto que define a sua natureza, finalidade, forma de funcionamento, atribuições dos dirigentes e competências das áreas da instituição, além de várias outras regulamentações. Desta forma, os servidores que ingressam no IBGE possuem um conjunto de objetivos públicos comuns a todos;
- b) o IBGE investe em sistemas de informação e de comunicação entre os servidores, seja estimulando os mesmos a participarem em reuniões, comitês e eventos, como em usar ferramentas de *e-mails*, videoconferência, canais de atendimento, e etc.
- c) para vários assuntos importantes da instituição que envolve mais de uma Diretoria ou Unidades Estaduais, por exemplo, são criados comitês ou grupos de trabalhos para coletar informações ou *feedback*, como na elaboração de Edital de Concurso Público, no Planejamento Estratégico da instituição ou no mapeamento de processos. Além disso, também existe a seção do “Qual a sua Dúvida?” do IBGEANDO, onde não se registram apenas dúvidas, mas reclamações e elogios também de todos os servidores do IBGE do Brasil.
- d) é fácil identificar que o IBGE desenvolve o seu próprio gênero de comunicação, isto se observa claramente quando há o ingresso de novas pessoas que passam no Concurso Público vindas principalmente das empresas privadas, ou numa reunião entre os servidores do IBGE e os funcionários de outras empresas, pois as pessoas não entendem o que os servidores do IBGE falam por utilizarem termos específicos de uma fundação pública ou abreviações de sistemas ou palavras

utilizadas na instituição ou até mesmo por adotarem comportamentos de acordo com a sua função adquiridos naquele contexto institucional;

- e) léxico específico também é comum encontrar, pois existem muitas siglas e abreviaturas de leis, sistemas e processos, além de usar termos, como por exemplo, “ibgeano” para se referir ao servidor que trabalha no IBGE ou “casa” para se referir ao próprio IBGE;
- f) no IBGE, 73% dos servidores do quadro permanente possuem mais de 20 anos de serviço, de acordo com o levantamento feito em abril de 2015. Estes são os membros que têm o conhecimento técnico e funcional da instituição, são os que conhecem profundamente o discurso interno e as suas particularidades, além de toda a parte técnica e metodológica das pesquisas do IBGE. Existe, inclusive, nos discursos, uma separação interna explícita entre os novos e os antigos servidores. Os antigos consideram o IBGE como sua própria família, e a nova geração de servidores considera o IBGE como uma instituição de trabalho, criando, como consequência, toda uma resistência na transferência do conhecimento interno e na absorção de novos conhecimentos.

Assim, parece que o conjunto de servidores do IBGE se enquadra nas características definidas por Swales (1990) para uma comunidade discursiva. Além disso, há que se considerar que os servidores do IBGE trabalham num mesmo ambiente institucional, cercados por regras e uma linguagem própria da Fundação, que definem o funcionamento do coletivo, mas que também se inserem os conhecimentos individuais e as suas diferentes percepções que compõem todo um discurso. Neste sentido, parece apropriado adotar as definições e características de uma comunidade discursiva para o conjunto de servidores que trabalham no IBGE, com o intuito de identificar os termos contidos nas mensagens textuais postadas pelos servidores para obtenção de insumos para a tomada de decisão.

Com esta abordagem, pretende-se entender que as mensagens textuais postadas no IBGEANDO devem conter o discurso da comunidade “ibgeana” em relação à área de RH e da própria instituição, alinhados com as perspectivas e necessidades individuais de cada servidor que compõem esta comunidade. O intuito de analisar estas mensagens com uma abordagem sociocognitiva é contribuir para melhorar a gestão organizacional do IBGE.

3.2 Tipificação de Mensagem Textual e seu Uso Social

Como o objeto desta pesquisa é a análise de mensagens escritas, é necessário entender o que significam estes textos que compõem estas mensagens, o que elas podem representar e o que podem contribuir para a Fundação IBGE.

Foi realizado um levantamento sobre autores que estudam a linguagem escrita e seu uso social e a concepção do autor Marcuschi (2002a) se enquadra bem no contexto deste trabalho, uma vez que ele define a noção de gênero textual, se concentrando nos últimos anos nos estudos sobre gêneros emergentes do ambiente digital. De uma forma geral:

[...] para a noção de gênero textual, predominam os critérios de ação prática, circulação sócio-histórica, funcionalidade, conteúdo temático, estilo e composicionalidade, sendo que os domínios discursivos são as grandes esferas da atividade humana em que os textos circulam. Importante perceber que os gêneros não são entidades formais, mas sim entidades comunicativas. Gêneros são formas verbais de ação social relativamente estáveis realizadas em textos situados em comunidades de práticas sociais e em domínios discursivos específicos. (MARCUSCHI, 2002a, p. 24 e 25)

Assim, as mensagens escritas podem ser consideradas como gênero textual uma vez que são textos escritos por servidores dentro de um contexto institucional e num canal de comunicação, ou seja, é uma prática institucional e social de comunicação. Marcuschi (2002a, p.35) também afirma que os “gêneros textuais não são frutos de invenções individuais, mas formas socialmente maturadas em práticas comunicativas”. Essa definição de gênero textual de Marcuschi se alinha ao conceito de comunidade discursiva.

Marcuschi (2002a) define o termo Gêneros como entidades sociodiscursivas e formas de ação social incontornáveis em qualquer situação comunicativa que contribuem para ordenar e estabilizar as atividades comunicativas do dia-a-dia, mas que isso não significa que os gêneros sejam instrumentos estanques ou enrijecedores da ação criativa. Esta afirmação vem do fato de que os gêneros textuais surgem, situam-se e integram-se funcionalmente nas culturas em que se desenvolvem, caracterizando-se muito mais por suas funções comunicativas, cognitivas e institucionais do que por suas peculiaridades linguísticas e estruturais. Assim, os gêneros textuais são de difícil definição formal, devendo ser contemplados em seus usos e condicionamentos sociopragmáticos caracterizados como práticas sociodiscursivas. Os gêneros textuais possuem uma diversidade de formas e de denominações e, assim como surgem, podem desaparecer.

Embora Marcuschi (2002a) afirme que os gêneros textuais não se caracterizam e nem se definam por aspectos formais, sejam eles estruturais ou linguísticos, e sim por aspectos sociocomunicativos e funcionais, ele não despreza a forma, pois em muitos casos são as formas que determinam o gênero, e em outros tantos serão as funções ou o próprio suporte ou o ambiente em que os textos aparecem que determinam o gênero presente.

Assim, Marcuschi (2002a) usa a expressão “gênero textual como uma noção propositalmente vaga para referir os textos materializados que encontramos em nossa vida diária e que apresentam características sociocomunicativas definidas por conteúdos, propriedades funcionais, estilo e composição característica”, e discurso como “aquilo que um texto produz ao se manifestar em alguma instância discursiva. Assim, o discurso se realiza nos textos. Em outros termos, os textos realizam discursos em situações institucionais, históricas, sociais e ideológicas”.

Marcuschi (2002a) diz que já se tornou trivial “a ideia de que os gêneros textuais são fenômenos históricos, profundamente vinculados à vida cultural e social”, surgindo novos gêneros textuais de acordo com as necessidades e atividades socioculturais, bem como na relação com inovações tecnológicas, o que é facilmente perceptível ao se considerar a quantidade de gêneros textuais hoje existentes em relação a sociedades anteriores à comunicação escrita.

Resumindo, Marcuschi (2002b, p.6) diz “que os gêneros textuais são frutos de complexas relações entre um meio, um uso e a linguagem”. Desta forma, o surgimento do ambiente digital oferece peculiaridades específicas para usos sociais, culturais e comunicativos que não se oferecem nas relações interpessoais face a face. E aqui o Portal IBGEANDO se insere por ser uma ferramenta de *intranet* que inseriu uma nova forma de comunicação com seus servidores através da seção “Qual a sua Dúvida?”, um canal de atendimento entre os servidores e a CRH.

Barbosa, Severo e Reategui (2009, p.1) afirmam que “a internet tem demonstrado um grande potencial para diversificadas manifestações linguísticas, onde surgem novas linguagens com base em tecnologias que formam gêneros textuais emergentes, tais como: *sites*, *blogs*, *e-mail*, mensagens instantâneas, etc.” Noblia (1998, p.2) também diz que a “comunicação mediada por computador, através da linguagem escrita, transformou-se em uma ferramenta que torna possível a construção de um novo tipo da interação social para além das barreiras espaciais”.

Marcuschi (2002b) diz que esses gêneros textuais emergentes têm características próprias e são mediados pela tecnologia computacional que oferece um programa de base, sendo diversificados em seus formatos e funções, mas que todos dizem respeito a interações entre indivíduos reais, embora suas relações sejam realizadas num ambiente virtual.

Marcuschi (2002b) define características dos gêneros textuais emergentes no contexto da tecnologia digital:

[...] três aspectos tornam a análise desses gêneros relevante: (1) seu franco desenvolvimento e um uso cada vez mais generalizado; (2) suas peculiaridades formais e funcionais, não obstante terem eles contrapartes em gêneros prévios; (3) a possibilidade que oferecem de se rever conceitos tradicionais, permitindo repensar nossa relação com a oralidade e a escrita. (MARCUSCHI, 2002b, p. 01)

Assim, como estas mensagens postadas estão armazenadas num ambiente digital do Portal IBGEANDO, e, com o aumento cada vez maior do número de postagens, se faz necessário entender o funcionamento social deste gênero textual. A princípio, as mesmas se caracterizam como um *e-mail* (MARCUSCHI, 2002b, p.22) que é uma forma de comunicação escrita normalmente assíncrona de remessa de mensagens entre usuários do computador, que podem interagir de um emissor para um receptor ou a vários receptores simultaneamente. No IBGEANDO as mensagens são postadas por categoria onde existe um grupo de pessoas responsáveis para responder ao questionamento. Quanto ao formato textual, é normal compará-lo com uma carta que podem ter respostas ou não. No geral, todas as mensagens textuais postadas no IBGEANDO são respondidas. Sua linguagem é no geral não-monitorada, podendo ser, porém, muito bem elaborada e escrita em separado. Uma das vantagens dos *e-mails* é sua transmissão instantânea encurtando o tempo de recebimento. E isso faz com que o uso deste canal de atendimento seja cada vez mais utilizado. Os *e-mails* efetivamente se constituíram num novo gênero tendo-se em vista suas peculiaridades formais e discursivas. Esta análise da característica do gênero textual também é muito importante para a investigação se as mensagens textuais postadas podem indicar insumos para a tomada de decisões.

O estudo de Noblia (1998) também corrobora com as características apresentadas por Marcuschi (2002b) sobre o *e-mail*. Para Noblia (1998) “o *e-mail* está intimamente ligada a estilos epistolares, sendo a sua estrutura semelhante ao de uma carta”, desta forma é utilizada a linguagem escrita que é assíncrona, existindo uma diferença temporal entre sua produção e consumo, tendo como vantagem o controle sobre o que está sendo dito, mas não na forma

como está sendo interpretado, ou seja, impossível ter controle sobre o que foi enviado e nem interagir com quem leu para ter a possibilidade de corrigir ou modificar uma interpretação errada.

Fairclough (2001) é outro estudioso que corrobora com a concepção de Marcuschi, que diz que a linguagem é vista como uma forma de prática social. Para ele, isso significa que: primeiro a linguagem é uma parte da sociedade, e não uma forma externa a ela; segundo, que a linguagem é um processo social; e em terceiro lugar, que a linguagem é um processo socialmente condicionado, condicionado pelas outras partes não linguísticas da sociedade. Ele argumenta que:

[...] não há uma relação externa ‘entre’ linguagem e sociedade, mas uma relação dialética interna. Os fenômenos linguísticos são sociais na medida em que, sempre que alguém fala ou ouve ou escreve ou lê, essas ações são feitas de formas socialmente condicionadas, e provocam efeitos sociais. Por outro lado, os fenômenos sociais são linguísticos na medida em que as atividades linguísticas que ocorrem em contextos sociais não são um mero reflexo ou expressão de processos e práticas sociais, na verdade elas são partes desses processos e práticas. (FAIRCLOUGH, 2001, p. 23)

Com isso, aplicando esses conceitos na presente questão da pesquisa, pretende-se entender que as mensagens escritas do IBGEANDO fazem parte de um processo social e que o que está escrito deve refletir os processos e práticas de RH dentro do IBGE.

Quando pensamos sobre o impacto que a Internet produziu e ainda produz em nossas vidas, encontramos uma ampla gama de possibilidades que representa esse impacto. Uma dessas possibilidades é o fato de considerarmos que as mudanças sociais ocorrem devido à nossa tentativa de experimentar novas formas de interação social e cultural, que, conseqüentemente, resultam na necessidade de reestabelecer novas questões políticas, sociológicas ou culturais. (NOBLIA, 1998, p. 1)

O Portal IBGEANDO mudou a forma de comunicação entre os servidores e a área de RH do IBGE, e como toda mudança, houve resistência e reclamações. O IBGEANDO foi uma forma de centralizar as informações de RH que possuem constantes atualizações. Além disso, com a criação do canal de atendimento “Qual a sua Dúvida?”, houve uma centralização no atendimento dos servidores que, tendo em vista tudo o que o foi exposto neste capítulo, registram mensagens escritas como uma prática sociocomunicativa no contexto de uma comunidade discursiva específica, no IBGE ele se denomina como comunidade “ibgeana”, gerando efeitos sociais. Estes efeitos sociais podem ser realizados quando há uma resposta aos

questionamentos ou uma ação por parte da instituição. Em relação aos efeitos sociais, como próprio Noblia (1998) diz, acaba surgindo a necessidade de se reestabelecer novas questões políticas e ações efetivas por parte da instituição. O que se observa no IBGE é que esse reestabelecimento de novas políticas e ações é que não é realizado levando em consideração o que esta comunidade ibgeana registra nas suas mensagens textuais postadas no canal. Assim, percebe-se a necessidade de experimentar novas formas de obtenção de informações para a tomada de decisões no IBGE, revendo as formas tradicionais, e analisando estas mensagens postadas pode ser um dos caminhos para “ouvir” o que os servidores estão falando.

3.3 Tomada de Decisão

Na área de Administração, a tomada de decisão é um processo que envolve aspectos cognitivos, pois depende da leitura que o tomador de decisão faz da situação, baseado nas suas emoções e experiências vividas, pelo qual se escolhe uma solução ou um plano de ação, a partir de variados cenários, ambientes, análises e fatores para resolver uma situação ou problema.

“Um problema fundamental das organizações é, portanto, definir as premissas que orientam a tomada de decisões e constituem o ambiente organizacional”. (CHOO, 2003, p. 42) Para Choo, além disso, deveriam ser aplicadas rotinas, regras e princípios para simplificar o processo decisório e reduzir a incerteza e a complexidade deste processo.

Para Cassarro (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”.

Na área administrativa, pode-se perceber que para uma empresa funcionar, decisões sempre devem ser tomadas, como estabelecer critérios para recrutamento e seleção dos funcionários, definir metas para os setores de trabalho da empresa, abrir um processo disciplinar, entre outros. Os processos administrativos, na verdade, acabam sendo processos de decisões, onde, de acordo com a situação ou problema, são reunidos elementos para que as autoridades da empresa tomem as suas decisões.

As decisões são tomadas por pessoas capacitadas ou autorizadas pela empresa, normalmente, de acordo com a sua alocação na hierarquia na estrutura organizacional ou a sua

função ou por seus conhecimentos especializados na questão a ser decidida. Assim, a decisão cabe a uma pessoa ou a um grupo de pessoas que fazem as suas escolhas a partir de suas percepções.

A forma como as pessoas tomam as decisões e a qualidade de suas escolhas finais dependem muito de suas percepções. [...] Todas as decisões requerem interpretação e avaliação de informações. Os dados costumam vir de diversas fontes e precisam ser selecionados, processados e interpretados. Quais dados, por exemplo, são relevantes para uma determinada decisão? A resposta fica por conta da percepção de quem toma a decisão. (ROBBINS, 2005, p. 111)

Para que estas decisões não sejam tão subjetivas e limitadas apenas a percepções individuais, ao longo do tempo surgiram vários estudos sobre a tomada de decisão para torná-lo mais racional e objetivo e, principalmente, reforçando a importância na busca e coleta de informações.

[...] 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)

Buchanan e O'Connell (2006) fizeram um levantamento sobre a história do estudo da tomada de decisão e chegaram a conclusão que a tomada de decisão é uma mescla de várias disciplinas do saber, como matemática, sociologia, psicologia, economia e ciência políticas. A filosofia, para refletir sobre o que uma decisão revela sobre o nosso eu e nossos valores. A história, para dizer a decisão tomada por líderes em momentos críticos. O estudo do risco e do comportamento organizacional nascendo de um desejo mais prático: ajudar o administrador a obter melhores resultados. E, para Buchanan e O'Connell, embora uma boa decisão não garanta um bom resultado, tal pragmatismo em geral compensa. A crescente sofisticação da gestão de risco, a compreensão das variações do comportamento humano e o avanço tecnológico, que respalda e simula processos cognitivos, melhoraram, em muitas situações, a

tomada de decisão. Mas, apesar disso, a história da tomada de decisão não é a de puro progresso rumo a um perfeito racionalismo.

Assim, vários modelos foram propostos para a tomada de decisões para tornar este processo mais racional e embasado em informações relevantes.

O uso de modelos de tomada de decisão permite aos gestores compreender a estrutura organizacional e as relações complexas inerentes aos processos desenvolvidos nesse âmbito. Portanto, há a crescente relevância no que tange a investigar e construir modelos, que proporcionem uma melhor aplicabilidade de métodos e técnicas no processo de tomada de decisão organizacional, cuja base é a informação, visto que se constitui em recurso fundamental para o referido processo. (LOUSADA; VALENTIM, 2008, p. 148)

O modelo de tomada de decisões vê a organização como um sistema decisório racional. O comportamento decisório é provocado pelo reconhecimento de um problema. Os que decidem buscam alternativas, avaliam as consequências e escolhem resultados aceitáveis, de acordo com seus objetivos e preferências. Como os indivíduos são limitados por sua capacidade de processar informações, rotinas que orientam a busca de alternativas e a tomada de decisões simplificam o processo decisório. O resultado do processo é a seleção de cursos de ação capazes de levar a um comportamento racional e orientado para os objetivos. (CHOO, 2003, p. 46)

Existem muitos modelos propostos para a tomada de decisões, neste estudo são apresentados brevemente apenas dois modelos para entendimento de como funciona o processo de tomada de decisão.

Segundo Robbins (2005, p. 111), existem seis passos do modelo de tomada de decisões racionais:

1. Definir o problema, quando existe uma discrepância entre o estado existente e um estado desejável;
2. Identificar os critérios para a decisão, determinando o que é relevante para decidir;
3. Atribuir pesos específicos a cada um desses critérios para identificar a prioridade correta nas decisões;
4. Desenvolver alternativas, apenas listando sem qualquer tentativa de avaliá-las;
5. Avaliar as alternativas, classificando-as de acordo com os critérios estabelecidos anteriormente e seus respectivos pesos;
6. Escolher a melhor alternativa pelo cálculo da decisão ótima, ou seja, selecionar aquela que tiver pontuação maior.

Já o economista e pesquisador Herbert Simon (1997), propôs um modelo para o processo de tomada de decisão que se inicia com a busca da informação, o que exige o uso da inteligência, já que é preciso analisar que informação é relevante. O segundo passo é o *design* da informação, ou seja, o seu tratamento, de modo a facilitar a elaboração de alternativas. Para Choo (2003, p. 304), este passo “envolve buscar informações com o objetivo de inventar, criar ou desenvolver cursos de ação que possam resolver uma situação. Nesse caso, as necessidades de informação são de localizar, elaborar e analisar alternativas em termos de seus resultados e sua contribuição para os objetivos da organização”. O terceiro e último passo é a escolha da alternativa que se afigura como a mais interessante para a decisão a ser tomada que para Choo “começa quando um determinado curso de ação é escolhido entre os vários que foram planejados. A escolha pode ser influenciada pelas informações sobre o contexto no qual a decisão deve ser tomada, tais como o conjunto de outras decisões e problemas que estão sendo considerados ao mesmo tempo, e os fatos e prazos futuros que poderão afetar o sucesso ou a percepção da decisão” (CHOO, 2003, p. 304). Esse processo é cíclico, o que significa que após a escolha da alternativa inicia-se um novo ciclo de busca da informação, que, por sua vez, dá início a um novo processo que para Choo implica "ter acesso aos resultados de ações passadas, como parte de um ciclo repetitivo que conduz a novas decisões. Aprender com ações passadas requer ter informações para inferir relações causais entre decisões e resultados que possam ser separados no tempo e no espaço”. (CHOO, 2003, p. 304)

Por isso, para Choo é muito importante que as empresas criem rotinas, pois estas refletem o que a organização aprendeu com a experiência diante de situações recorrentes, gerando uma memória procedimental da organização. Por exemplo, procedimentos orçamentários, de planejamento e de avaliação de projetos permitem que grupos internos disputem os recursos com base em critérios e procedimentos claros e justos. Além disso, Choo (2003) lista várias razões que ele considera importantes para que uma empresa crie regras e rotinas:

- a) permitem que a organização projete a legitimidade externamente, para a sua comunidade, já que uma organização que obedece a rotinas para a tomada de decisões revela um comportamento responsável;
- b) refletem práticas sensatas ou aceitáveis de escolha que a organização aprendeu com o tempo;

- c) oferecem uma racionalidade procedimental interna, no sentido de que deixam claros os passos e critérios para se chegar a uma decisão, permitindo que os grupos dentro da organização disputem os recursos de uma maneira justa;
- d) comprometem-se com cursos de ação determinados;
- e) esclarecem o necessário processamento de informação diante de problemas complexos;
- f) incorporem técnicas eficientes e confiáveis aprendidas com a experiência e;
- g) coordenem ações e resultados dos diferentes grupos organizacionais.

Seguir rotinas e procedimentos pode institucionalizar certas visões de mundo, formar hábitos de aquisição e transmissão de informações, e estabelecer valores e normas capazes de influenciar a maneira como a organização lida com a escolha e a incerteza. O resultado que se espera dessa combinação de cultura, comunicação e consenso é uma maior eficiência das decisões e um comportamento decisório mais racional. O resultado não pretendido é a rigidez das rotinas decisórias e dos valores que orientam a decisão, assim como o desejo coletivo de manter o sistema interligado de cultura e comunicação construído ao longo do tempo. (CHOO, 2003, p. 254)

Estes modelos apresentados, e todo o discurso de Choo (2003) sobre regras e rotinas, deixam claro a tentativa de tornar o processo de tomada de decisões o mais racional possível.

No entanto, Lousada e Valentim observam:

[...] que há consenso na literatura analisada quanto ao entendimento de que tomar uma decisão totalmente racional é uma tarefa praticamente impossível, pois o tomador de decisão não tem condições de possuir conhecimento sobre todas as variáveis influenciadoras do processo; isso porque, no momento da coleta de informações, já se pressupõe a análise inicial das alternativas e, também, das prováveis consequências que cada uma pode causar. (LOUSADA; VALENTIM, 2008, p. 150)

De qualquer forma, independente de quão racional ou não é o processo de tomada de decisão, pode-se notar que a informação é um elemento importante durante todo o processo, seja para definir especificamente o problema/situação ou para gerar alternativas para uma decisão.

Durante o processo de tomada de decisões, a busca de informação é guiada pelos hábitos e princípios que o indivíduo adquiriu em decorrência de treinamento, educação ou experiência. Ao mesmo tempo, as organizações criam e institucionalizam regras e rotinas para estruturar os comportamentos

de busca e de escolha com base nos objetivos organizacionais. Portanto, a busca de informação é fruto das preferências individuais, dos valores institucionais e dos atributos da situação de escolha. A busca de informação é uma atividade motivada por problemas: começa quando se percebe um problema (inclusive o problema de como aproveitar uma oportunidade) e se reconhece que ele exige decisão e ação. A busca parece respeitar uma hierarquia de fontes de informação, que é ordenada pela proximidade em relação a um problema ou a seus sintomas, e pelas características das fontes, como sua acessibilidade ou credibilidade. (CHOO, 2003, p. 310)

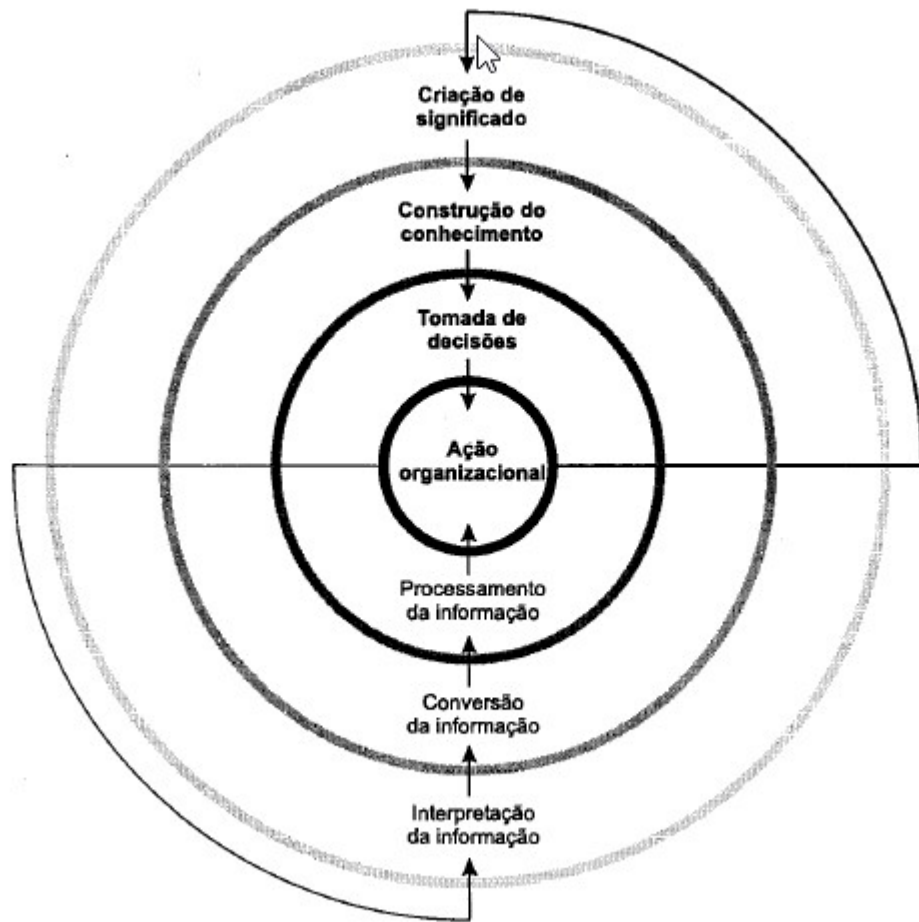
Para Cassarro (1995), a informação é o insumo básico do processo decisório e que permeia todas as fases do ciclo de tomada de decisões. O autor defende que é indispensável aos decisores dispor de informação confiável, adequada, em tempo certo, para que possam tomar decisões eficazes e eficientes.

[...] informação é insumo do processo decisório empresarial, por isso tem papel fundamental para qualquer modelo de gestão empresarial. Para tomar a decisão certa é preciso que o gestor possua informações internas e externas selecionadas, tratadas, organizadas e acessíveis, de forma que propicie a redução das incertezas. Portanto, é de suma importância que o acesso à informação seja no tempo certo, que a informação seja confiável, bem como seja consistente. (LOUSADA; VALENTIM, 2008, p. 156)

Choo (2003) criou uma visão holística do uso da informação, conforme a Figura 1, onde, numa organização de conhecimento,

[...] a criação de significado, a construção do conhecimento e a tomada de decisões são como três camadas concêntricas, em que cada camada interna produz os fluxos de informação para a camada externa adjacente. A informação flui do ambiente exterior (fora dos círculos) e é progressivamente assimilada para permitir a ação da empresa. Primeiro, é percebida a informação sobre o ambiente da organização; então, seu significado é construído socialmente. Isso fornece o contexto para toda a atividade da empresa e, em particular, orienta os processos de construção do conhecimento. O conhecimento reside na mente dos indivíduos, e esse conhecimento pessoal precisa ser convertido em conhecimento que possa ser compartilhado e transformado em inovação. Quando existe conhecimento suficiente, a organização está preparada para a ação e escolhe seu curso racionalmente, de acordo com seus objetivos. A ação organizacional muda o ambiente e produz novas correntes de experiência, às quais a organização terá de se adaptar, gerando assim um novo ciclo. (CHOO, 2003, p. 30)

Figura 1: Visão holística do uso da informação – Choo (2003)



Fonte: (CHOO, 2003, p. 31)

Na criação de significado, a principal atividade envolvida é a interpretação de notícias e mensagens sobre o ambiente para que os membros da organização possam decidir qual informação é relevante. Na construção do conhecimento, a principal atividade é a conversão do conhecimento, onde por meio do diálogo e do discurso e uso de canais mais formais de comunicação, os membros partilham seus conhecimentos. Na tomada de decisões, a principal atividade é o processamento e a análise da informação a partir das alternativas disponíveis, cujas vantagens e desvantagens são avaliadas. “Os três modos de uso da informação - interpretação, conversão e processamento - são processos sociais dinâmicos, que continuamente constituem e reconstituem significados, conhecimentos e ações”. (CHOO, 2003, p. 30)

A organização do conhecimento possui informações e conhecimentos que a tornam bem-informada e capaz de percepção e discernimento. Suas ações baseiam-se numa compreensão correta de seu ambiente e de suas necessidades, e são alavancadas pelas fontes de informação disponíveis e pela competência de seus membros. A organização do conhecimento possui informações e conhecimentos que lhe conferem uma especial vantagem, permitindo-lhe agir com inteligência, criatividade e, ocasionalmente, esperteza. (CHOO, 2003, p. 31)

Assim, a tomada de decisões na organização requer informações capazes de reduzir a incerteza para que seja possível:

- a) estruturar uma situação de escolha, determinando os tipos e o conteúdo das informações que serão necessárias para a tomada de decisão;
- b) definir preferências e selecionar regras, comparando informações que descrevem situações conhecidas e reações aprendidas;
- c) visualizar alternativas viáveis e suas possíveis consequências, identificando, desenvolvendo e avaliando diversos cursos de ação.

“As necessidades de informação, então, dependem de vários fatores: se as alternativas já existem, se as soluções podem ser customizadas ou modificadas, ou se novas soluções precisam ser encontradas”. (CHOO, 2003, p. 303)

Na teoria, toda decisão deve ser tomada racionalmente, com base em informações completas sobre os objetivos da empresa, alternativas plausíveis, prováveis resultados dessas alternativas e importância desses resultados para a organização. Na prática, a racionalidade da decisão é atrapalhada pelo choque de interesses entre sócios da empresa, pelas barganhas e negociações entre grupos e indivíduos, pelas limitações e idiosincrasias que envolvem as decisões, pela falta de informações e assim por diante. Apesar dessas complicações, uma organização deve manter ao menos a aparência de racionalidade, para manter a confiança interna e, ao mesmo tempo, preservar a legitimidade externa. Embora a tomada de decisões seja um processo complexo, não há dúvida de que ela é uma parte essencial da vida da organização: toda ação da empresa é provocada por uma decisão, e toda decisão é um compromisso para uma ação. (CHOO, 2003, p. 29)

Apesar de existirem vários estudos e modelos sobre a tomada de decisões, “na prática, a tomada de decisões caracteriza-se pela limitação da racionalidade, pelos erros e vieses humanamente comuns e pelo uso da intuição” (ROBBINS, 2005, p. 118). Os modelos costumam não ser utilizados, ou a empresa sequer tem o conhecimento deles. “As escolhas costumam ficar na área dos sintomas do problema e na proximidade da alternativa mais óbvia”. (ROBBINS, 2005, p. 114)

Por isso, a importância de unir os estudos científicos com as necessidades mercadológicas e empresariais, para que os estudos auxiliem e ajudem as organizações, neste caso, a tomarem melhores decisões baseadas em informações relevantes e que atinjam aos objetivos esperados, minimizando o peso subjetivo existente nas decisões.

Qualquer que seja o modo de decisão, o ambiente organizacional no qual ela ocorre é definido no mínimo por duas propriedades: a estrutura e a clareza dos objetivos organizacionais, que têm um impacto sobre as preferências e escolhas, e a incerteza ou quantidade da informação sobre os métodos e processos pelos quais as tarefas devem ser cumpridas e os objetivos devem ser atingidos. (CHOO, 2003, p. 275)

No caso do IBGE, o que se espera com este estudo, é que, se as mensagens textuais postadas pelos servidores apresentarem um resultado, após o processo de mineração de textos, que possam se tornar insumos para a tomada de decisão no âmbito da área de RH, que estes insumos sejam efetivamente utilizados, gerando ações por parte da CRH e dando clareza ao processo decisório. Conseqüentemente é possível que seja criado um ambiente organizacional mais participativo e que os servidores se sintam realmente ouvidos.

3.4 Mineração de Textos

A área de Mineração de Textos ou *Text Mining* auxilia neste projeto de pesquisa, sendo importante entender como esta área surgiu e para que serve.

[...] o meio mais simples de externalização é registrar, em textos livres, pensamentos, ideias, sentimentos e opiniões de pessoas. Nas organizações, há muito conhecimento deste tipo disponível na forma de: - sugestões e reclamações de clientes em pesquisas, e-mails e serviços de atendimento; - descrições de defeitos, causas e soluções aplicadas por funcionários; - manuais, normas e procedimentos definidos como padrão; - e-mails oriundos de listas de discussão; - memorandos e comunicações formais, distribuídos através de meios eletrônicos; etc. Entretanto, as organizações e as pessoas têm dificuldade para tratar adequadamente este tipo de informação por não estar estruturada. A área de *Text Mining* surgiu para minimizar este problema, ajudando a explorar conhecimento armazenado em meios textuais. (LOH ET AL, 2001)

O texto acima aborda exatamente o problema da pesquisa em questão de como obter os insumos para a tomada de decisão com base nas mensagens textuais que os servidores

registram no canal de atendimento do Portal IBGEANDO, mensagens em textos livres sem existir nenhuma padronização ou vocabulário controlado.

Brito (2016, p.19) afirma que “de fato, estamos vivenciando o crescimento acelerado de informações não estruturadas (textos), com isso, a Mineração de Textos ganha espaço não somente no meio acadêmico, mas também no mundo dos negócios”.

Torres menciona a questão de como localizar de forma rápida e precisa uma informação específica quando existe uma grande quantidade de documentos, e ele afirma que:

[...] sem o auxílio das atuais ferramentas de busca, a tarefa de encontrar a informação de forma eficiente e eficaz não seria possível. Se levarmos em conta que muitas destas páginas possuem documentos de texto anexados, como arquivos em PDF, RTF, DOC, ODS, dentre uma infinidade de outros formatos, o problema se tornaria ainda maior. É neste contexto que popularizou a mineração de texto, que pode ser genericamente definida como um processo que busca extrair informação útil de coleções de documentos de texto através da identificação e exploração de padrões. (TORRES, 2012, p. 350)

A disciplina de Mineração costuma ser dividida em duas vertentes: Mineração de Dados (*Data Mining*) e Mineração de Textos (*Text Mining*). A mineração de dados surgiu devido ao

[...] explosivo crescimento do volume de dados que gerou uma urgente necessidade de novas técnicas e ferramentas capazes de transformar, de forma inteligente e automática, *terabytes* de dados em informações significativas e em conhecimento. Essas informações, de grande valia para o planejamento, gestão e tomadas de decisão, estavam, na verdade, implícitas e/ou escondidas sob uma montanha de dados, e não podiam ser descobertas ou, no mínimo, facilmente identificadas utilizando-se sistemas convencionais de gerenciamento de banco de dados. (CORRÊA; SFERRA, 2003, p. 20)

Então a Mineração de Dados, entre as mais diversas definições, “pode ser entendido como o processo de extração de informações, sem conhecimento prévio, de um grande banco de dados e seu uso para tomada de decisões, definindo um processo automatizado de captura e análise de grandes conjuntos de dados para extrair um significado, sendo usado tanto para descrever características do passado como para prever tendências para o futuro”. (CORRÊA; SFERRA, 2003, p. 22)

Já a Mineração de Textos, segundo Hearst (2003), tem como objetivo descobrir informações desconhecidas presentes num texto elaborado para pessoas lerem, sendo uma variação da mineração de dados, que tenta encontrar padrões interessantes de grandes bancos

de dados. Para ele, a diferença entre mineração de dados e mineração de textos é que na mineração de texto os padrões são extraídos do texto em linguagem natural e não de bancos de dados estruturados.

Torres (2012, p. 350) afirma que a mineração de texto sofreu forte influência da mineração de dados, e que existem várias similaridades nos processos de ambas, sendo a principal diferença, o fato que a mineração de dados utiliza informações que se encontram em formato estruturado, enquanto boa parte do esforço despendido na mineração de textos tem como objetivo a construção de um modelo de dados estruturado, a partir de documentos não estruturados, escritos em linguagem natural.

Ao pesquisar sobre a definição de Mineração de Textos, foram encontradas as mais diversas definições, no entanto a exposta pelos autores Feldman e Sanger (2006) parece ser a mais completa para este estudo:

Mineração de texto pode ser amplamente definido como um processo de conhecimento intensivo em que um usuário interage com uma coleção de documentos ao longo do tempo, utilizando um conjunto de ferramentas de análise. [...] Mineração de texto visa extrair informações úteis a partir de fontes de dados através da identificação e exploração de padrões. [...] Na sua forma mais simples, uma coleção de documentos pode ser qualquer agrupamento de documentos baseados em texto. (FELDMAN; SANGER, 2006, p. 1 e 2)

Em todo o material pesquisado, percebe-se que a mineração de texto é um processo que utiliza métodos e técnicas de vários campos de estudo como afirma o autor Tan (1999, p. 71) que diz que a “Mineração de texto é um campo multidisciplinar que envolve recuperação da informação, análise de textos, extração da informação, agrupamento, categorização, visualização, base de dados, aprendizado de máquina e mineração de dados”. Gupta (2009, p. 60) complementa ainda que a mineração de texto também se baseia na estatística e linguística computacional.

[...] considerada uma evolução da área de Recuperação de Informações, a Mineração de Textos 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. (MORAIS; AMBROSIO, 2007, p.2)

Em vários estudos sobre a mineração de textos foi observado que este processo pode ter várias denominações ou sinônimos. Em Morais e Ambrósio (2007, p. 6) afirma-se que “atualmente, a mineração de textos pode ser considerada sinônimo de descoberta de conhecimento em textos ou Mineração de Dados em Textos (*Text Data Mining*) ou Descoberta de Conhecimento a partir de Bancos de Dados Textuais (*Knowledge Discovery from Textual Databases*)”. Gupta (2009) também afirma que pode ser chamado de Análise de Texto Inteligente, Mineração de Dados Textuais ou Descoberta de Conhecimento em Textos (DCT), referindo-se genericamente ao processo de extração de informação útil e não-trivial de texto não estruturado.

Outro ponto observado nas pesquisas sobre a mineração de textos é que os autores e estudiosos na área definem o texto como dados estruturados, semi-estruturados ou não estruturados. Para Brito (2016, p. 18), “o principal objetivo da Mineração de Textos (MT) consiste na extração de características em uma grande quantidade de dados não estruturados”. Gupta (2009, p. 60) afirma que a “mineração de texto pode trabalhar com dados estruturados ou semi-estruturados como *e-mails*, documentos de textos e arquivos HTML etc.” Já Feldman e Sanger (2006) afirmam que apesar de rotularem o texto como dados não estruturados, um texto pode ser visto, a partir de muitas perspectivas, como um objeto estruturado, e apresentam a seguinte explicação:

De uma perspectiva linguística, um documento apresenta uma rica quantidade de estrutura semântica e sintática, embora esta estrutura é implícita e, até certo ponto está escondida em seu conteúdo textual. Além disso, os elementos tipográficos como sinais de pontuação, numéricos e caracteres especiais – particularmente quando combinados com elementos de layout, como espaçamento em branco, sublinhamento, asteriscos, tabelas, colunas, e assim por diante - muitas vezes podem servir como uma espécie de "marcação" da linguagem, fornecendo pistas para ajudar a identificar subcomponentes de documentos importantes tais como parágrafos, títulos, datas de publicação, os nomes dos autores, registros da tabela, cabeçalhos e notas de rodapé. Documentos que têm relativamente poucos elementos tipográficos ou indicadores de marcação para demarcar a estrutura - como a maioria dos trabalhos de pesquisa científica, relatórios empresariais, memorandos jurídicos - são muitas vezes referidos como formato livre ou documentos fracamente estruturados. Por outro lado, documentos com extensos e consistentes elementos de formatação que podem ser facilmente inferidos campos para metadados como alguns *e-mails*, páginas da *Web* em HTML, arquivos PDF e arquivos de processamento de texto com *templates* de documentos ou restrições de folha de estilo - são ocasionalmente descritos como documentos semi-estruturados. (FELDMAN; SANGER, 2006, p. 3 e 4)

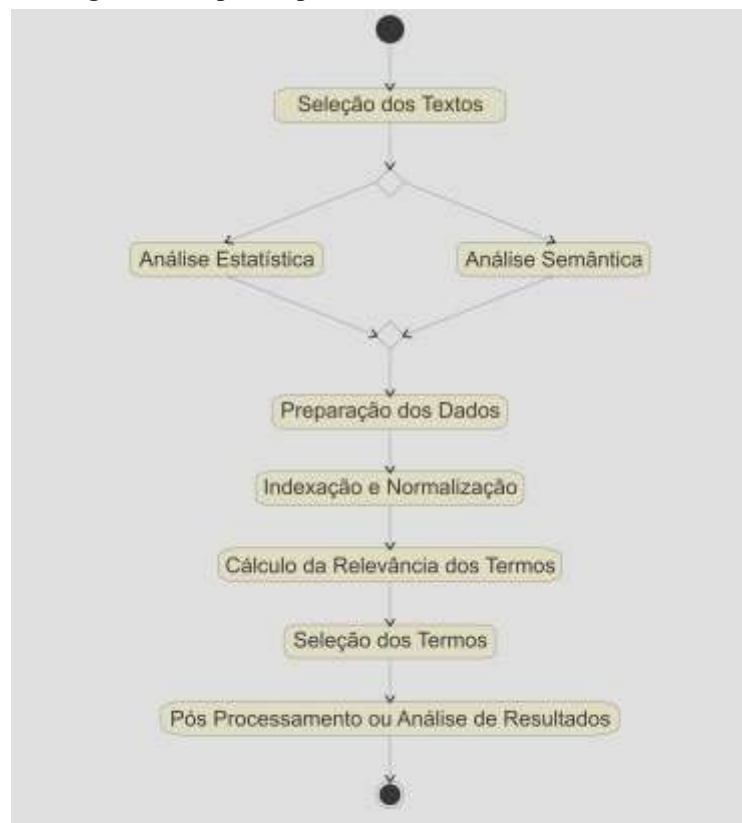
Esta discussão sobre a estruturação ou não dos dados que se apresentam nos textos é feita em vários estudos, no entanto, uma boa observação foi feita no estudo de Torres (2012, p. 351) que diz “que os conceitos de documento estruturado ou não estruturado apresentados dizem respeito unicamente a análise automática, a uma operação pautada no processamento do texto por uma máquina. Visualmente, uma página HTML ou um documento PDF possuem uma estruturação de tópicos, seções e uma série de elementos que facilitam o entendimento humano do documento, de forma que, para a nossa leitura, estes documentos são claramente estruturados”.

Este projeto de pesquisa não está tratando de dados, e sim de textos, utilizando o processo de mineração de textos para identificar os termos e suas relações nas mensagens textuais de forma a obter insumos, ou informações, que auxiliem a tomada de decisão e melhorem a comunicação entre os servidores e a área de RH. Assim, neste estudo, é adotada a abordagem de Morais e Ambrósio (2007) que afirma que:

[...] as principais contribuições da Mineração de Textos estão relacionadas à busca de informações específicas em documentos, à análise qualitativa e quantitativa de grandes volumes de textos, e à melhor compreensão de textos disponíveis em documentos. Textos estes que podem estar representados das mais diversas formas, dentre elas: *e-mails*; arquivos em diferentes formatos (pdf, doc, txt, por exemplo); páginas *Web*; campos textuais em bancos de dados; textos eletrônicos digitalizados a partir de papéis. (...) 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)

Assim, para explicar melhor o processo de mineração de textos, será utilizada a abordagem adotada em Morais e Ambrósio (2007) conforme as etapas da Figura 2:

Figura 2: Etapas do processo de mineração de textos



Fonte: (MORAIS; AMBROSIO, 2007, p.7)

De uma forma geral, o processo de mineração de textos se divide nas seguintes etapas:

1. Seleção dos Textos ou de Documentos: etapa em que o usuário define qual o conjunto de textos ou documentos textuais deseja ser analisado;
2. Definição do tipo de abordagem do texto, que podem ser utilizados separadamente ou em conjunto:
 - a. Análise Estatística: baseada na frequência dos termos encontrados nos textos;
 - b. Análise Semântica: baseada na funcionalidade dos termos encontrados nos textos, empregando técnicas que avaliam a sequência dos termos no contexto dos textos. Sua base está na utilização de técnicas em Processamento de Linguagem Natural (PLN).
3. Preparação dos Dados: Envolve a seleção do núcleo que melhor expressa o conteúdo dos textos selecionados, tentando identificar similaridades em função da morfologia ou do significado dos termos nos textos.

4. Indexação e Normalização (Pré-processamento): etapa que facilita a identificação de similaridade de significado entre suas palavras, considerando as variações morfológicas e problemas de sinonímia, tendo como resultado a geração de um índice. Esta é uma etapa automática que envolve as principais fases:
 - a. Identificação de Termos: objetivo principal é a identificação dos termos contidos no texto, sejam eles simples (aplicando um analisador léxico ou *parsing* ou *tokenization*) ou compostos (utilizando um dicionário de expressões ou identificando os termos que co-ocorrem com muita frequência em uma coleção de documentos).
 - b. Remoção de *Stopwords*: fase que envolve a eliminação de algumas palavras que não devem ser consideradas no texto, por não serem relevantes na análise dos textos ou por não traduzirem a essência do texto. Normalmente fazem parte desta lista os advérbios, preposições, artigos, pronomes e outras classes de palavras auxiliares.
 - c. Normalização morfológica (*stemming*): eliminação das variações morfológicas de uma palavra, através da identificação do radical de uma palavra, retirando prefixos e sufixos. Essa técnica de identificação de radicais é denominada lematização ou *stemming*, que em inglês significa reduzir uma palavra ao seu radical (ou raiz).
5. Cálculo da Relevância dos Termos: como nem todas as palavras contidas num texto possuem a mesma relevância, deve-se fazer o cálculo de relevância dos termos que se baseiam na frequência dos termos (a mais utilizada por ser mais simples), na análise estrutural do documento ou na posição sintática de uma palavra.
6. Seleção dos Termos: esta etapa de seleção dos termos ocorre após os processos de pré-processamento e cálculo da relevância. Esta técnica pode ser baseada no peso dos termos ou na sua posição sintática em relação ao texto. As principais técnicas de seleção de termos são: filtragem baseada no peso do termo, seleção baseada no peso do termo, seleção por análise de co-ocorrência, seleção por *Latent Semantic Indexing* e seleção por análise de linguagem natural.
7. Pós-processamento ou análise de resultados: aplicação de técnicas de análise de resultados com os resultados obtidos com o processo de mineração.

A Ciência da Informação possui vários estudos métricos da informação com análise de frequência de palavras em determinado texto, como a bibliometria. Cassettari et al (2015, p. 158) afirma que a bibliometria “tem uma função relevante para a CI, visando analisar a qualidade e quantidade de itens utilizados, consultados, de visibilidade e de produtividade”. A Bibliometria indica o tratamento quantitativo da comunicação escrita. A Lei de Zipf é uma teoria bastante utilizada para este fim, uma base matemática-linguística que analisa a frequência e distribuição das palavras contidas em um texto, seja científico ou não, possibilitando a representação por meio de gráficos e análises estatísticas de quais são os termos que mais se repetem, de modo que seja possível criar um ranking de palavras-chave.

Em seu trabalho, Cassettari et al (2015, p. 159) afirma que “a Lei de Zipf pode ser aplicada de diversas maneiras em análise linguística. A sua aplicação está voltada a textos escritos, onde é realizada uma análise sobre as ocorrências das palavras”. Eles também informam que na área da ciência da informação, iniciativas com a Lei de Zipf serviram para simular representações temáticas mais adequadas, inclusive para sustentar aspectos de classificações em unidades de informação e para a probabilidade da recuperação da informação, com base também na infometria. Mais recentemente, esta lei foi aplicada para as questões semânticas, utilizando a distribuição de frequência, o comportamento de escala temporal e o fator de decaimento.

As Leis de Zipf vêm sendo utilizadas com sucesso como ferramenta estatística, em diferentes áreas do conhecimento, tais como: lingüística, urbanismo, física, medicina, economia, engenharia, química, entre outras. Os estudos que utilizam a frequência de ocorrência de palavras, como ferramenta de representação temática da informação, têm evoluído para tentativas de desenvolvimento de algoritmos, visando contribuir para automatização, em parte ou no todo, da indexação temática da informação. (GUEDES; BORSCHIVER; 2005)

Guedes (1994) diz que “os primeiros estudos bibliométricos fundamentados em frequência de ocorrência de palavras, como medida de conteúdo temático, foram os de Luhn (1957)”. Luhn (1958) desenvolve uma abordagem estatística, com vistas à classificação e busca automáticas de documentos, e um método automático probabilístico, com vistas à criação de *abstracts*. Ele propôs uma técnica para extrair as palavras mais significativas do texto, através da eliminação das mais comuns e frequentes e, também, das menos comuns. Para ele, a frequência de um termo em um documento revela a sua significância no documento como um todo, sendo que as palavras com baixíssimas e altíssimas frequências não são representativas do texto. Alvares e De Araújo Júnior (2012) afirmam que “no que se

refere à recuperação de informação, Luhn foi o responsável por muitas técnicas, tais como o processamento de textos completos (usando concordância como índices), a indexação através de palavra-chave do contexto (KWIC) e a indexação automatizada”.

O processo descrito na Figura 2 é a base de qualquer processo de mineração de textos, sendo que as operações de pré-processamento desempenham um papel fundamental de transformar o conteúdo não estruturado de um texto num nível de representação mais tratável por um sistema. Os sistemas de mineração de texto dependem de abordagens algorítmicas e heurísticas para considerar distribuições, conjuntos de frequência e vários tipos de processamentos.

Uma etapa que não é mencionada no trabalho de Moraes e Ambrósio (2007) é a etapa de Visualização de Resultados, esta etapa consiste na representação gráfica dos resultados obtidos com o processo de mineração de textos e considerada fundamental para esta pesquisa, uma vez que serão usuários leigos ao assunto que analisarão o resultado gerado pelo processo de mineração.

Ferramentas de visualização são frequentemente utilizadas por sistemas de mineração de texto para facilitar a navegação e exploração de padrões de conceito. Estas usam várias abordagens gráficas para expressar relações de dados complexos. No passado, ferramentas de visualização para mineração de texto geravam apenas mapas estáticos ou gráficos que eram essencialmente rígidos na exibição dos resultados. O estado da arte dos sistemas de mineração de texto, no entanto, melhorou as representações gráficas interativas que permitem que um usuário possa manipular os resultados de pesquisa como arrastar, puxar, clicar, ou qualquer outra forma de interação direta com a representação gráfica dos padrões de conceito. (FELDMAN; SANGER, 2006, p. 10)

Torres (2012, p. 373) também diz que, “a área de mineração visual de textos provê ferramentas gráficas que obtém vantagem das habilidades visuais dos usuários para dar suporte ao processo de aquisição do conhecimento”. E Tan (1999, p. 73) complementa que “há um bom número de produtos de mineração de texto que se preocupam com a visualização do resultado tendo como abordagem geral a organização dos documentos com base em suas semelhanças, apresentando os grupos ou aglomerados de documentos numa representação gráfica”.

Cabe ressaltar que a mineração de textos também engloba um conjunto de estratégias e métodos que são utilizadas de acordo com a finalidade do estudo, a mineração pode ser aplicada para diversas finalidades sendo que as três principais aplicações são:

a) Sumarização

É uma técnica na qual o computador simula a atividade humana na elaboração de sumários para documentos de texto. O desafio nesta aplicação é ter programas que analisem e interpretem semanticamente o texto.

A grande vantagem desta aplicação é o resumo de grandes textos, mantendo os principais pontos e significado do texto inteiro. Esta técnica vem sendo estudada com bastante ênfase nos últimos anos com o objetivo de desenvolver aplicativos que sumarizem a diversidade de informações contidas no texto retornando informações mais precisas já sumarizadas.

Com a sumarização de textos, o usuário otimiza o tempo de busca e leitura de um texto longo, avaliando, por exemplo, se vale a pena ou não ler o documento inteiro.

b) Classificação

O processo de classificação de textos é a identificação dos principais assuntos dos documentos textuais com o objetivo de classificá-los em categorias já previamente definidas, utilizando normalmente Tesouros. Na Classificação, o que se faz é a contagem das palavras, procurando por termos gerais, termos mais próximos, sinônimos e termos relacionados, através de métodos e técnicas de ranqueamento de documentos.

Para fazer esta classificação, é necessário utilizar algoritmos, baseados em aprendizado de máquina ou em regras estatísticas, que definem automaticamente a categoria do documento a ser classificado. Exemplos destas técnicas são a técnica do Vizinho mais próximo (*k-nearest neighbor*), Máquinas de Vetores de Suporte (*Support Vector Machines*) e o algoritmo *Naive Bayes*.

Esta é a aplicação de mineração de texto mais conhecida e utilizada.

c) Agrupamento (*Clustering*)

O agrupamento, ou *clustering*, consiste em dividir uma coleção de documentos texto em grupos (*clusters*) de acordo com algum relacionamento de similaridade entre os mesmos. A organização dos documentos em cada *cluster* deve ser feita de forma que haja:

- Alta similaridade entre os documentos pertencentes a um mesmo *cluster*;
- Baixa similaridade entre elementos que pertencem a *clusters* diferentes.

No agrupamento, os documentos podem aparecer em vários subtópicos, garantindo que num processo de busca, por exemplo, um documento útil não seja omitido na exibição dos resultados. A tecnologia utilizada nos agrupamentos pode ser útil na organização de sistemas de informação de gestão, que normalmente contém milhares de documentos.

Sendo assim, a mineração de textos é um processo indicado para ser utilizado na análise das mensagens textuais postadas pelos servidores no canal de atendimento do IBGEANDO e a utilização de uma ferramenta de mineração de textos serve para descobrir elementos importantes nestas mensagens. Gupta (2009, p. 60) afirma que “a mineração de textos é a melhor solução para empresas”.

Surge então a necessidade de se escolher uma ferramenta de mineração de textos no idioma português para aplicar a esse estudo. Tan (1999, p. 75) em seu artigo alerta que “a maioria das ferramentas de mineração de texto possui o foco em processamento de documentos em inglês, e que a mineração a partir de documentos em outros idiomas permite o acesso a informações anteriormente inexplorado, oferecendo uma nova série de oportunidades”. Lima (2014) em seu artigo também defende que:

[...] o grande volume de informação registrada em meio digital e o desenvolvimento acelerado de novos estoques de informação cobram o desenvolvimento de ferramentas específicas, e é neste ambiente renovado pelas tecnologias, com o avanço da internet e uma disponibilidade cada vez maior de informação, que a extração automática de informações surge como ferramenta capaz de assegurar a identificação das informações potencialmente relevantes de um texto, através de um conjunto de processos e técnicas que garante acesso e interpretação destas informações, visto que, na sua grande maioria, estas informações não apresentam regularidade quanto à formatação, estrutura e conteúdo. (LIMA, 2014, p. 956)

Assim, utilizando a mineração de textos, pode-se ter como resultado a identificação dos termos e as suas relações nas mensagens textuais que após analisadas podem ser utilizadas como insumos para a tomada de decisão.

No subcapítulo 3.4.1 são elencadas ferramentas de mineração de textos analisadas para este trabalho.

O conteúdo apresentado neste capítulo não tem a pretensão de esgotar o tema, mineração de textos, que é vasto, mas serve para contextualizar de uma forma geral quanto à sua definição, processo e aplicação.

3.4.1 Trabalhos relacionados sobre mineração de mensagens textuais

Com o intuito de verificar os trabalhos já realizados sobre mineração de textos em meios digitais, ambientes *web* e canais de atendimento, foram realizados levantamentos nas bases BRAPCI, de teses no Programa de Pós-Graduação em Ciência da Informação da UFF, no Portal de Periódicos da CAPES e, de forma adicional, através da ferramenta de busca *Google Acadêmico*, utilizando as palavras-chave: mineração de textos, *text mining*, mensagens textuais. A maioria dos trabalhos encontrados na pesquisa está vinculada à área de Computação ou de Sistemas de Informação ou de Tecnologia, uma vez que o tema Mineração de Textos é muito abordado nestas áreas. Dos trabalhos pesquisados, doze devem ser destacados por existir alguma similaridade ou por fornecer subsídios ao presente projeto, sendo que três são da área de Tecnologia da Educação, dois da Ciência da Computação, dois de Sistemas de Informação, dois de Ciências Sociais e Engenharia, um de Matemática Aplicada, um de Engenharia Civil e, apenas um da Ciência da Informação.

Bastos (2006), da área de Engenharia Civil, descreve o desenvolvimento e a implementação de uma metodologia eficiente e prática de construção de um ambiente de descoberta de conhecimento que contempla a busca pela informação existente em sites da *web* escritos em português. A aplicação desempenha as seguintes tarefas de mineração aplicadas sobre documentos: identificação de similaridades e diferenças entre conteúdos de páginas e *sites*, classificação e agrupamentos de documentos. Segundo Bastos (2006, p.1) o processo manual de avaliação de conteúdo textual e de classificação de documentos, como aquele efetuado em grandes ambientes de informação, torna-se humanamente difícil de ser realizado, resultando em uma visão limitada dos conteúdos dos documentos. Assim, se faz necessário uma aplicação que contemple um conjunto de ferramentas automáticas que avaliem o conteúdo de documentos, páginas *web* ou *sites*, através da utilização de ferramentas de mineração de textos, que são importantes na extração de informações e grandes massas de dados, podendo retornar resultados com alto nível de qualidade na informação obtida, aumentando as chances de descoberta de informações relevantes ou inesperadas. A motivação de Bastos, a época dos estudos, era que existia um grande número de ferramentas de mineração de textos, proprietárias ou disponíveis na *web* (de forma gratuita), capazes de manipular grandes conjuntos de dados textuais em inglês, e quase não existia no idioma português. Dez anos após os estudos de Bastos, ainda existem poucas ferramentas eficientes de mineração de textos em língua portuguesa. Outro ponto já destacado nos seus estudos é a

necessidade de ferramentas automáticas. No trabalho de Bastos (2006, p. 52 a 65), são elencadas cinco ferramentas de mineração de textos, a saber:

1. TMSK – *Text-Miner Software Kit*: é um pacote de software para mineração preditiva de textos. Possui funcionalidades para pré-processar documentos de textos formato XML e provê implementações para as tarefas de Pré-processamento, Predição, Recuperação de informação, *Clustering* e Extração de informações;
2. RIKTEXT – *Rule Induction Kit for Text*: é um pacote de software completo para categorização de documentos baseado em regras de decisão. O objetivo é determinar o melhor conjunto de regras para a predição e a classificação, onde o melhor é o menor número de regras com o erro mínimo. O RIKTEXT complementa o TMSK, disponibilizando métodos para construção e uso de regras para classificação de documentos;
3. *Text Mine*: é um conjunto de ferramentas de mineração de textos escritas em Perl, linguagem apropriada para o desenvolvimento de *scripts* para servidores *web*;
4. *Intext Software*: é uma ferramenta de domínio público para análise de textos. Possui várias funcionalidades que permitem analisar conteúdo de documentos em inglês e alemão. *Intext* é considerado uma ferramenta, com várias funcionalidades de análise do texto;
5. Vivísimo/*Clusty*: é uma ferramenta de busca e de agrupamento de documentos da *web*. *Clusty* consulta vários motores de busca, combina os resultados, e gera uma lista ordenada baseada no *ranking* comparativo extraído desses motores de busca. Essa abordagem de meta-busca ajuda na melhoria dos resultados, fazendo com que as melhores URL's (endereço virtual de uma rede) apareçam primeiro, ficando os piores resultados para o fim da lista.

Morais (2007), da área da Ciência da Computação, propõe implementar um sistema que utiliza técnicas de mineração de textos para associar semanticamente documentos a domínios representados por ontologias. A metodologia proposta usa técnicas estatísticas de mineração de textos, a partir da análise de documentos, atribuindo um grau de similaridade (ou relevância) desse documento ao domínio representado pela ontologia. No estudo, foi realizada a mineração de textos com os documentos contendo jurisprudências do Tribunal de Justiça de Goiás.

[...] já existem várias ferramentas que auxiliam na busca de informações na internet. No entanto estas ferramentas ainda apresentam problemas principalmente pela dificuldade que os sistemas têm de entender a semântica contida nas páginas. Uma proposta para amenizar tais problemas, é realizar uma análise automática do contexto em que os termos das buscas são usados, permitindo uma "compreensão" do conteúdo, reduzindo ambigüidades e aumentando a relevância dos documentos recuperados". (MORAIS, 2007, p. 95)

Assim, este estudo corrobora com a necessidade de se ter uma análise automática para aumentar a relevância dos documentos recuperados. O presente estudo pretende desde o início utilizar uma ferramenta que faça uma análise automática das mensagens textuais para recuperar insumos para auxiliar a tomada de decisão e melhorar a comunicação.

Schiessl (2007), da área da Ciência da Informação, em seu trabalho de dissertação de Mestrado, analisou um Serviço de Atendimento ao Consumidor (SAC) de uma instituição financeira que centraliza, em forma textual, os questionamentos, as reclamações, os elogios e as sugestões, verbais ou escritas, de clientes. Ele aplicou técnicas de mineração de texto numa base do SAC, a fim de demonstrar a utilidade da Descoberta de Conhecimento em Textos (DCT) para a criação de agrupamentos e modelo de classificação automática de textos a partir de coleção de documentos existentes. Ele demonstrou que, com a utilização de ferramentas e metodologias adequadas, é possível maximizar a descoberta e a utilização de informações úteis ainda não identificadas. Como resultado, também foram gerados indicadores de desempenho do grau de satisfação do cliente referente a produtos e serviços oferecidos que subsidiam a gestão na política de atendimento. A ferramenta utilizada para todo processo foi o SAS *Enterprise Miner* "que é capaz de fornecer uma *interface* amigável para a construção de modelos e, além disso, uma eficiente linguagem de programação para resolução de problemas específicos". (SCHISSL, 2007, p. 101) A base explorada foi extraída de um sistema de ouvidoria que armazena todas as mensagens enviadas pelos clientes desde o ano 2000, mas ele selecionou apenas as mensagens do ano 2006 que foram 52.646 mensagens textuais. No entanto, dada a complexidade de processamento textual versus recursos tecnológicos, ele decidiu fazer uma amostragem com apenas 7.895 registros. Na sua conclusão, ele afirma que a metodologia proposta foi útil e aplicável na transformação de textos em informações organizadas, na extração de conhecimento e na automatização de processos que dependem de leitura de pessoas dedicadas a essa tarefa. No âmbito acadêmico, ele considerou que a pesquisa obteve êxito, porém sua aplicação no âmbito profissional precisa de equipe dedicada, requerendo integração de profissionais, do atendente, dos especialistas de domínio, da equipe de tecnologia e dos pesquisadores dedicados à elaboração de modelos e de indicadores.

A ciência é grande parceira da atividade empreendedora. No mundo empresarial, pesquisas têm sentido se promovem negócios de acordo com a missão de cada organização. Portanto, é fundamental promover a aproximação contínua entre equipe técnica e gestores para transformar ações de gestão em produtos e serviços. Melhor ainda se essas atividades são beneficiadas com técnica e ciência, daí o papel essencial da academia nas empresas. (SCHIESSL, 2007, p. 110)

O objeto de pesquisa do Schiessl é bem similar ao objeto de estudo do presente projeto, os dois trabalhos são bem semelhantes. Uma das motivações deste estudo é justamente encontrar uma forma de aliar a ciência às necessidades de gestão de uma instituição.

Barbosa, Severo e Reategui (2009), da área de Tecnologia na Educação, apresentam um estudo sobre aplicação de ferramentas para mineração de textos em *blogs*. O estudo fez um comparativo entre duas ferramentas, SOBEK e *TagCrowd*. O objetivo do trabalho foi descobrir padrões nos termos resultantes da mineração, bem como, verificar se os resultados obtidos podem auxiliar no entendimento do conteúdo dos textos com foco educacional. A maior contribuição do artigo, baseando-se em um contexto educacional, foi que o SOBEK possui uma capacidade maior para extrair informações relevantes, uma vez que relaciona termos e exibe resultados mais completos, além de ser mais fácil a sua utilização. O presente estudo também analisa as ferramentas SOBEK e *TagCrowd*, mas aplicando a uma seção de um portal que se assemelha mais ao gênero textual digital *e-mail* e com fins gerenciais.

A tese de Doutorado de Macedo (2010), da área de Tecnologia da Educação, teve como objetivo principal identificar possibilidades de práticas pedagógicas a partir de informações geradas por uma ferramenta denominada Rede de Conceitos. Tais informações são extraídas de textos produzidos pelos participantes/alunos no Editor de Texto Coletivo (ETC) que foi desenvolvido pelo NUTED (Núcleo de Tecnologia Digital Aplicada à Educação) da Universidade Federal do Rio Grande do Sul (UFRGS). Este editor é um espaço para a escrita colaborativa à distância. A Rede de Conceitos surgiu como uma necessidade de apoiar o gerenciamento e acompanhamento do professor sobre as produções dos alunos no ETC. Como Macedo tem formação educacional, ela buscou auxílio na área da Ciência da Computação para criar uma ferramenta com base na técnica de Mineração de Textos, visto não ter encontrado nenhuma ferramenta disponível que favorecesse o acompanhamento da produção aliada à otimização do tempo de análise do professor. A primeira versão desta ferramenta surgiu no ano de 2007 e foi denominada SOBEK. A ferramenta Rede de Conceitos é a integração entre o SOBEK e o ETC e indica tanto o tema abordado quanto oferece

indicadores sobre a produção textual para que o professor possa agir com foco nas potencialidades e necessidades dos alunos, sem necessariamente ter que ler o texto produzido. As principais contribuições deste trabalho foram: a significativa diminuição de tempo de leitura exigido do professor, a ampliação do tempo de interação direta entre alunos e professor, proporcionando qualificação no processo ensino-aprendizagem, e obtenção de indicadores que auxiliam os professores a terem foco nas suas ações.

Molina e Steinberger-Elias (2011) realizaram uma ampla pesquisa que explora informações de um corpus de notícias sobre desastres climáticos e reconhece automaticamente, com apoio de uma ferramenta de Processamento de Linguagem Natural (PLN), palavras que denotem os atores envolvidos e suas principais ações na prestação de socorro às vítimas. Parte-se da hipótese de que relatos noticiosos de desastres não só permitem identificar as entidades que participam dos socorros, como também oferecem condições para caracterizar redes discursivas associadas a cada tipo de evento. Molina apresenta as etapas de composição e descrição de um corpus sobre o terremoto do Haiti em 2010 com apoio do pacote *Natural Language Toolkit* (NLTK). O uso do pacote foi bem-sucedido, com retornos em frequência de itens lexicais, *collocations*, aplicação de filtros para gerar listas com propriedades pré-determinadas léxico-gramaticalmente, manipulação do corpus para obter dados que permitissem validar estatisticamente retornos trazidos pelo pacote e análise de itens e combinações de itens em categorias. A descrição do corpus permitiu ver sua distribuição, o que deu margem para entender que a hipótese de partida poderia ser útil para obter uma rede discursiva sobre o evento estudado.

Souza et al (2015) descreve o Projeto *Media Cloud*, que surgiu frente à necessidade do desenvolvimento de metodologias – e tecnologias associadas – para se enfrentarem os muitos desafios que surgem quando lidamos com massivas quantidades de dados textuais, como nas bibliotecas e arquivos digitais, ou na *World Wide Web*, notadamente quando estes precisam ser regularmente organizados e pesquisados, visando à recuperação em tempo hábil de informações relevantes para algum objetivo específico. O projeto *Media Cloud* foi lançado em março de 2009 e constitui-se de uma plataforma para estudar ecossistemas de mídia, ou seja, as relações entre as instituições e os profissionais criadores de mídia - impressa e digital - e os cidadãos. Através do monitoramento programático de milhões de notícias, publicadas *online* ou transmitidas em canais de televisão, o sistema permite aos pesquisadores monitorar a disseminação de notícias, conceitos e memes, além de permitir a descoberta das redes de atores que pautam a mídia, através da genealogia das notícias. Também permite que se façam

análises geográficas - através da identificação da cobertura midiática nas diversas regiões, e políticas, monitorando a abordagem específica dos diversos atores, politicamente identificados com as principais correntes partidárias, segundo os diversos temas de interesse, identificando-se vieses. A plataforma de software é livre, e foi projetada como um substrato para diversos projetos correlatos, centrados em processos de comunicação. No tocante à tecnologia, um dos resultados é a disponibilização de análises e incorporação das funcionalidades da ferramenta PyPLN de engenharia textual, oferecendo capacidades analíticas avançadas baseadas em processamento de linguagem natural. O projeto *Media Cloud*, e sua plataforma analítica PyPLN, prometem contribuir sobremaneira para o estado da arte das técnicas de análise de assunto e de conteúdo, tanto pelas tecnologias inovadoras para tratamento de grandes massas de dados textuais, quanto pelas confluências interdisciplinares, que agregam, em termos de teoria e metodologia, os campos da Ciência da Informação, Ciência da Computação e Linguística; e em termos de contexto e escopo, o arcabouço das ciências sociais, como Direito, Comunicação, Política e Sociologia.

Na área de Tecnologia da Educação, também foi encontrada a tese de Doutorado do pesquisador Damasceno (2015) que investigou como a Mineração Textual pode contribuir no processo de resposta de teleconsultorias a profissionais solicitantes, aprimorando a recuperação de informação dentro da Plataforma de Telessaúde do Ministério da Saúde. O minerador SOBEK foi utilizado para extrair conceitos de um conjunto de solicitações de profissionais presentes na plataforma. Os resultados permitiram concluir que a lista de tópicos extraída pelo minerador SOBEK pôde auxiliar na localização de informações pertinentes para a construção de respostas, na aceleração do processo de resposta e na construção de respostas mais informativas e precisas. Os teleconsultores que participaram da pesquisa também consideraram que a consulta de solicitações e respostas prévias pode contribuir na educação permanente de profissionais de saúde. Cabe ressaltar aqui que “Telessaúde é o suporte à distância através de tecnologias de comunicação e informação para responder às necessidades dos profissionais de saúde, incluindo médicos, em sua prática assistencial. Além de suporte assistencial, pode ser um meio para disseminar atividades de desenvolvimento profissional contínuo ou, como chamada no Brasil, educação permanente em saúde”. (DAMASCENO, 2015, p.31)

A pesquisa acima é bem similar ao estudo deste projeto de mestrado em questão, uma vez que existe basicamente um ambiente onde o solicitante registra o seu questionamento e um responsável deve respondê-lo, só que o objetivo principal da tese do pesquisador

Damasceno é melhorar a resposta do atendimento das teleconsultorias, tanto em seu conteúdo quanto no tempo de resposta, vislumbrando também a contribuição na educação permanente dos profissionais de saúde. No presente estudo, o objetivo é identificar e analisar sistematicamente as mensagens textuais de forma a obter insumos para auxiliar a tomada de decisão gerencial e melhorar a comunicação interna na instituição.

Jain e Kumar (2015), da área da Ciência da Computação, desenvolveram um método para detectar surtos de Influenza-A (H1N1) por meio da análise de dados gerados por usuários do *Twitter* e a consciência geral da população em relação à doença. O objetivo do trabalho é mostrar que é possível detectar surtos de doenças através da análise de dados gerados pelos meios de comunicação social. Eles utilizaram uma técnica baseada na dinâmica das palavras-chave de RSS¹⁷ *feeds* para recuperar os *tweets*. Para obter as palavras-chave a partir das notícias, eles analisaram os principais RSS dos jornais utilizando técnicas baseadas em *N-gram*. Também foram aplicadas técnicas de classificação como SVM (*Support Vector Machine*), *Naive Bayes*, *Random Forest* e *Árvore de Decisão*.

Os três estudos interessantes mais recentes encontrados são da área de Sistemas de Informação e de Ciências Sociais e Engenharia, são estudos que ainda não foram totalmente finalizados, mas que vale ressaltar aqui tendo em vista o objeto de estudo e as ferramentas utilizadas.

Hemsley e Palmer (2016) escreveram um artigo sobre o uso racional do *Twitter* como uma ferramenta de comunicação e troca de informações entre adultos que possuem Esclerose Lateral Amiotrófica (ELA) ou Doenças do Neurônio Motor (DNM), detalhando o uso de métodos qualitativos e quantitativos para analisar o conteúdo das postagens no *Twitter*. Para eles, o *Twitter* é uma boa ferramenta para troca de informações, sendo um método muito importante de comunicação baseado em textos escritos para pessoas com problemas severos de comunicação oral. Eles fizeram dois estudos: em um, avaliaram a viabilidade e utilidade de reunir dados do *Twitter* do perfil de pessoas que tenham ELA ou DNM e, no outro, buscaram no *Twitter* as postagens que utilizaram as *hashtags* das doenças, que em inglês são #ALS e #MND.

¹⁷ A sigla RSS tem mais de uma denominação: *RDF Site Summary*, *Really Simple Syndication*, *Rich Site Summary*. RSS é um padrão desenvolvido em linguagem XML que permite divulgar notícias ou novidades de *sites* e *blogs*. Para isso, o link e a notícia são armazenados em um arquivo, conhecido como *feed* ou *feed RSS*.

Em ambos estudos, eles utilizaram a ferramenta *KH Coder* para analisar os textos das postagens coletadas pois esta ferramenta permite variados métodos para analisar os textos e permitir a visualização dos resultados por meio de gráficos. Os resultados com o *KH Coder* mostraram, por exemplo, que 40% das atividades nos perfis destas pessoas são de conversas com outras pessoas, seguidos de visualização de conteúdos na Internet; e que a utilização do *hashtag* engloba três grandes grupos: arrecadação para ajudar os familiares, questões de tratamento e cura da doença e relatos e notícias de pessoas com a doença. Como conclusão, eles sugerem que o *Twitter* é uma importante plataforma de comunicação para estas pessoas, mas que é subutilizado como um instrumento para facilitar as discussões com este grupo.

Brito (2016) apresentou como requisito parcial para a obtenção do título de Mestre em Sistemas de Informação e Gestão do Conhecimento um trabalho que tem como objetivo o estudo e a implementação de uma ferramenta que conterà um algoritmo de classificação de sentimentos, sendo ele capaz de avaliar a polaridade de textos extraídos de mídias sociais, baseando-se em técnicas da mineração de textos. Ele propõe montar um arcabouço conceitual de um sistema para detecção automática de sentimentos em bases textuais, que utilize os conceitos da aprendizagem de máquina e que seja fácil sua adaptação a vários domínios de negócio. A ferramenta utilizada no trabalho de Brito foi o NLTK (*Natural Language Toolkit*) “pela sua curva de aprendizagem, sua sintaxe transparente e na facilidade de manipular funções, através da linguagem de programação *Python*”. (BRITO, 2016, p. 29) Cabe ressaltar que o trabalho do Brito é da área de Sistemas de Informação então, acredita-se que o mesmo deva ter conhecimentos de programação.

Neves e Zaccaro (2016) no Projeto de Pesquisa apresentado ao Instituto Federal de Educação, Ciência e Tecnologia Fluminense, como requisito parcial para conclusão do Curso de Pós-graduação *Lato Sensu* em Análise e Gestão de Sistemas de Informação, apresentam como objetivo principal analisar as interações entre professores e alunos dentro da rede social acadêmica da UCAM (Universidade Cândido Mendes), utilizando técnicas de mineração de textos para extrair padrões importantes nos textos redigidos nas interações entre docentes e discentes. Eles acreditam que obtendo conhecimentos e informações úteis através da análise de fóruns de discussão da rede social acadêmica da UCAM, utilizando a tecnologia de mineração de textos, estes elementos podem ser aproveitados pelos professores para melhorar o processo de ensino-aprendizagem. Eles utilizaram a ferramenta *MineraFórum*, “que realiza a análise qualitativa das mensagens postadas pelos alunos em um fórum de discussão. Esta ferramenta é capaz de apresentar ao docente uma visão sobre as contribuições escritas pelos

discentes, organizando e agrupando as mensagens de cada aluno. Entende-se que, a partir dos resultados apresentados pelo MineraFórum, o professor pode direcionar seu apoio aos alunos que colocaram poucas contribuições relevantes ao tema do fórum”. (NEVES, ZACCARO, 2016, p. 43)

A ferramenta MineraFórum é um minerador de textos para fóruns de discussão. A ferramenta extrai os principais conceitos abordados no debate e oferece a opção de calcular a relevância do conteúdo postado em cada mensagem. Esta ferramenta é uma produção técnica do grupo de Pesquisa do NUTED (Núcleo de Tecnologia Digital Aplicada à Educação) da Universidade Federal do Rio Grande do Sul (UFRGS). Ela é bem similar à ferramenta SOBEK que também foi desenvolvida por um grupo de pesquisa da UFRGS, a diferença é que a MineraFórum é específica para mineração de textos em fóruns e não está disponível para ser baixada gratuitamente em seu sítio.

Estes estudos mostram a necessidade de se ter uma ferramenta automática que faça a mineração de textos e sua análise nos diversos canais de comunicação social, sejam na *web*, em *blogs*, *e-mails*, redes sociais, fóruns de discussão e bases textuais, como mensagens postadas em um ambiente de atendimento a clientes, profissionais ou usuários. Esta necessidade surge devido ao grande volume de informação para ser analisada e que, se feita manualmente, demanda muito tempo que poderia ser utilizado em outras atividades, podendo comprometer até a qualidade da análise. O objetivo desses estudos sempre é de identificar informações úteis, verificar se os resultados obtidos auxiliam no entendimento dos textos, gerar indicadores, melhorar resposta de atendimento, tanto no seu conteúdo quanto no tempo de resposta, visando ações de gestão, criações de políticas de atendimento e melhorias dos processos, tanto no contexto educacional, empresarial, social ou da saúde.

Na pesquisa realizada, também foi encontrado o trabalho de Klemann, Reategui e Rapkiewicz (2011) que comparava quatro diferentes ferramentas de análise e mineração de textos, tendo-se como princípio a utilização destas ferramentas como recurso educacional, com foco no apoio a produção textual. As ferramentas analisadas foram *TextAlyser*, *WordCounter*, *TagCrowd* e SOBEK. Elas foram analisadas a partir de diferentes critérios, tais como: facilidade de operação, visualização de termos relevantes e disponibilidade na *web*. Nas considerações, eles dizem que a atual tendência é de se utilizar computação em nuvem, pois as ferramentas *desktop* são menos flexíveis. No entanto, para eles, esta flexibilidade pouco acrescenta ao usuário se outras características importantes estiverem ausentes, particularmente a possibilidade de visualização de relacionamento entre os termos e

respectiva visualização gráfica, tendo ainda a possibilidade de manipular o grafo apresentado diretamente na forma gráfica. “As estatísticas apresentadas pelas ferramentas *TextAlyser*, *Wordcounter* e *TagCrowd* podem contribuir na produção, revisão e avaliação de textos. A ferramenta SOBEK apresenta resultados mais completos, além de ser de fácil manejo”. (KLEMMANN et al, 2011, p. 1103)

Num período de 10 anos, de acordo com os estudos relatados acima, o que se identifica é que ainda se busca uma ferramenta que forneça resultados cada vez mais próximos de uma determinada necessidade e com o intuito de obter otimização do tempo, principalmente no idioma Português.

3.5 Conceituação de Insumos

Um elemento importante e necessário no processo decisório para diminuir a incerteza da escolha da decisão é a informação.

A informação tornou-se o diferencial não apenas para manter as organizações, mas também para auxiliá-las na organização das tarefas do dia-a-dia. Vale lembrar que o sucesso da organização não depende somente das informações disponíveis, mas sim de saber coletar, organizar, analisar e implementar as mudanças com base nas informações que serão utilizadas para a melhoria contínua de suas atividades. O crescimento das empresas e o necessário uso das informações, entretanto, geraram uma nova dificuldade: controlar os estoques informacionais, para que as informações sejam recuperadas e possam contribuir para a tomada de decisões. (DANTAS, 2013, p. 4)

Para Dantas (2013, p.6), “um dos principais motivos para se obter o máximo de informações dentro de um projeto em que se está envolvido é o alcance dos objetivos, o que faz da informação um insumo de especial importância”.

No âmbito desta pesquisa, os insumos são provenientes da análise dos resultados obtidos com a mineração dos textos das mensagens textuais postadas pelos servidores, ou seja, provenientes dos termos e as suas relações. Aqui neste estudo, pode-se conceituar insumos, como sendo as informações resultantes da análise dos resultados do processo de mineração de textos.

O termo informação na CI possui diversas conceituações dentro de visões diferenciadas para a área que na maioria das vezes parecem rupturas entre elas, no entanto, esta pesquisa mostra também que existem coesões entre elas.

A literatura da CI que trata das origens, fundamentos e concepções na CI apresentam-nos numa perspectiva de relações no tempo, onde novas visões são em geral lidas como críticas ou questionamentos das existentes, supondo-se, portanto, um desenrolar histórico da CI. Nem incomensurabilidade, nem perfeita tradução, as visões da CI podem dialogar, com maior ou menor conflito, a partir dos mesmos interesses expressos por suas questões coesoras. (FERNANDES, 2006, p. 3)

Estes insumos também podem indicar que o IBGE funciona como uma Comunidade Discursiva, produzindo sentido somente para quem integra esta comunidade.

Como o objetivo desta pesquisa é analisar a viabilidade do uso de ferramenta de mineração de textos para processamento automático das mensagens textuais no Canal de Atendimento do Portal IBGEANDO de forma a obter insumos para auxiliar a tomada de decisão e melhorar a comunicação da área de RH, é necessário abranger concepções objetivas para recuperar, identificar e analisar a informação em si, uma vez que serão processados textos armazenados numa base de dados, e, posteriormente utilizar concepções subjetivistas de forma a obter insumos para auxiliar a tomada de decisão e melhorar a comunicação da área de RH, ou seja, a informação para alguém ou para um público específico.

Neste sentido, é abordada uma visão documentalista que distingue conhecimento e informação e sublinha, de um lado, seu campo de trabalho como sendo este conhecimento objetivado, donde se devem extrair as informações, organizá-las e disponibilizá-las e, de outro, o efeito dinamizador destas informações sobre aqueles que produzem conhecimento teórico ou para tomar decisões. Esta visão considera que seus conhecimentos devem gerar produtos e serviços que atendam às necessidades informacionais do homem ou organizações.

Informação não é conhecimento. Ela depende do conhecimento, de onde é extraída e, enquanto unidade, indica, dá um roteiro para acesso aos documentos ou parte de documentos que conteriam a informação desejada. Mas é esta disponibilização ou a disseminação (após sua seleção e reorganização) que proporcionaria a dinamização do conhecimento. Sem os mapas dos SRIs as organizações da sociedade estariam cegas, sem meios para encontrar o que precisam para produzir e tomar decisões. (FERNANDES, 2006, p. 6)

Analisando as mensagens textuais postadas pelos servidores no Portal, a CRH e, até mesmo a Alta Direção da organização, tem como avaliar as demandas e necessidades dos

servidores podendo gerar esclarecimentos, políticas de RH ou qualquer outro produto ou serviço que atendam estas necessidades.

Por outro lado, esta pesquisa também utiliza concepções da visão matemática que trata a informação como um conhecimento potencial e é definida como um redutor de incertezas. Utiliza a criação de ferramentas estatísticas e analíticas para administrar melhor a ciência. Considera que estudar o comportamento da informação é uma maneira de estudar o comportamento do Homem. “A atual visão matemática fará laços com a “inteligência competitiva” e utilizará as técnicas do *data mining* (garimpagem em dados em bases); das redes neurais (descobrir padrões em dados aleatórios) e o *data warehousing*.” (FERNANDES, 2006, p. 12) Além disso, o foco da pesquisa está no agente que é o tomador de decisões, na CRH, enfim, qualquer que seja sua função, naquele que busca informação, movido por suas incertezas ou pelo desejo de criar conhecimento.

Para analisar as mensagens do Portal, como já mencionado, é necessário utilizar uma ferramenta de mineração de textos, tratar os textos de forma a obter informações que possam produzir um sentido ou conhecimento para uso da CRH para tomada de decisões e melhoria da comunicação.

Assim, a pesquisa também está fortemente embasada no paradigma social, no construtivismo social, onde há a necessidade de produção de sentido por alguém, que é criado e recriado na interação social. A comunicação consistirá em interação de indivíduos.

É produção de sentido, por alguém, mas de um sentido possível, existente nos discursos circulantes, inteligível aos demais e, portanto, já existente. Uma vez que a linguagem não é entendida fora de seu uso cotidiano (a língua), então, para a visão ela não é apenas um conjunto de regras (sintaxe) e significados (semântica), mas carrega sentidos (ordenações de mundo, valores, interdições etc) que antecedem aqueles que a utilizam e que são criados e recriados na interação social. Dito de outro modo, a informação é uma produção de sentido a partir de uma estrutura cognitiva construída por elementos históricos, sociais, econômicos, culturais. (FERNANDES, 2006, p. 21)

Ou seja, a tentativa de criar um sentido ou conhecimento dos termos contidos nas mensagens postadas no Portal, para uso da CRH, de interação entre todos os servidores e a equipe de RH, onde há uma comunicação constante para tirar dúvidas, reclamar ou elogiar, é estudada também por uma visão Construtivista Social.

[...] o que estamos vendo é uma mudança de paradigmas que está ocorrendo não apenas no âmbito da ciência, mas também na arena social, em proporções ainda mais amplas. Para analisar essa transformação cultural, generalizei a definição de Kuhn de um paradigma científico até obter um

paradigma social, que defino como “uma constelação de concepções, de valores, de percepções e de práticas compartilhados por uma comunidade, que dá forma a uma visão particular da realidade, a qual constitui a base de maneira como a comunidade se organiza. (CAPRA, 2001, p. 24 e 25)

Para Nascimento e Marteleto (2004), a informação é entendida como fenômeno social coletivo, estruturas de conhecimento e instituições de memória das comunidades. O objeto de trabalho das comunidades encontra-se refletido nos padrões de cooperação, nas formas de linguagem e comunicação, nas estruturas e organizações do conhecimento, nos sistemas de informação, na literatura (e suas formas de distribuição) e nos critérios de relevância.

Para este trabalho, pode-se “considerar como “social” qualquer processo de produção/organização/consumo de informação, uma vez que ele acontece entre grupos [...] – ou seja, a geração e apropriação de informações só ocorrem no âmbito da sociedade, das relações sociais”. (CARDOSO, 1994, p. 107-108) Assim, os insumos obtidos a partir destas mensagens textuais podem ser considerados como um processo de consumo de informação que foi produzido num âmbito de relações sociais dentro de uma instituição.

Surge então o questionamento de como obter estes insumos nas mais de 40 mil mensagens postadas em oito anos, considerando que cresce a cada ano o número destas mensagens registradas no Portal. Desta forma, se faz necessário pesquisar como analisar estas mensagens textuais, uma vez que o volume destas mensagens é grande para ser analisado manualmente, exigindo uma solução de processamento automático.

Assim, as mensagens textuais são processadas numa ferramenta de mineração de textos, identificando os termos e suas relações contidos nas mensagens textuais, que após serem analisadas pelos gestores de RH do IBGE, são obtidos os insumos para auxiliar na tomada de decisões e definir as ações a serem executadas. A obtenção dos insumos indica que o IBGE funciona como uma comunidade discursiva, pois somente servidores da Fundação conseguem de maneira efetiva dar um sentido para os termos das mensagens textuais e suas relações após o processamento da ferramenta de mineração de textos.

4 RESULTADOS ENCONTRADOS

Este capítulo engloba todas as etapas da realização da Prova de Conceito: a prospecção de ferramentas de mineração de texto para análise e seleção da ferramenta mais adequada para processar as mensagens textuais, de acordo com os critérios estabelecidos; o processamento das mensagens textuais por categoria, utilizando a ferramenta escolhida; e a validação dos resultados deste processamento com os servidores da CRH com o intuito de verificar a possibilidade de obtenção de insumos para a tomada de decisão.

Cabe ressaltar, novamente, que a proposta deste trabalho não é estabelecer requisitos funcionais para o desenvolvimento de uma ferramenta de mineração de textos idealizada para o processamento de mensagens textuais, e sim, escolher uma ferramenta já pronta e disponível para uso que seja a mais próxima do que seria a ferramenta ideal para utilização no IBGE.

4.1 Análise e Seleção da Ferramenta de Mineração de Textos

Considerando os objetivos específicos da pesquisa de prospectar ferramentas de mineração de textos que possam processar automaticamente as mensagens textuais, identificando os termos e suas relações contidos nas mensagens e; apresentar os termos e suas relações de forma sistematizada de maneira auxiliar na obtenção de insumos para a tomada de decisão, se faz necessária a utilização de uma ferramenta de mineração de textos para verificar a possibilidade de identificação dos termos das mensagens textuais e, caso seja possível esta identificação, analisar de que forma esta ferramenta pode contribuir para a obtenção dos insumos para a tomada de decisão.

Francis e Flynn (2010) alertam que uma das dificuldades em trabalhar com mineração de textos é a aquisição de ferramentas para este fim, pois muitos dos *softwares* disponíveis são caros ou de difícil utilização. Neste cenário, as soluções *open source* ganham forte incentivo de comunidades de pesquisa e comerciais.

Desta forma, considerando os trabalhos expostos no Capítulo 3.4.1, foi elaborado o Quadro 4 com o objetivo de listar as ferramentas utilizadas nestes trabalhos, uma vez que as mesmas já foram usadas em projetos similares de mineração de textos.

Quadro 4: Listagem das ferramentas utilizadas em trabalhos similares

| Ferramenta | Sítio | Grupo de Pesquisa / Responsável pelo Desenvolvimento |
|---------------------------------------|---|---|
| Intext Software | http://www.intext.com.br/ | Empresa InText |
| KH Coder | http://sourceforge.net/projects/khc/ | Koichi Higuchi Ritsumeikan University, Kyoto, Japão |
| MineraFórum | http://www.nuted.ufrgs.br/?page_id=1386 | Grupo de Pesquisa do NUTED (Núcleo de Tecnologia Digital Aplicada à Educação) UFRGS - Universidade Federal do Rio Grande do Sul |
| NLTK – Natural Language Toolkit | http://www.nltk.org/ | Steven Bird e Edward Loper University of Pennsylvania |
| PyPLN | http://pypln.org/ | Grupo de Pesquisa do NAMD (Núcleo de Análise e Modelagem de Dados) Escola de Matemática Aplicada da Fundação Getúlio Vargas |
| RIKTEXT – Rule Induction Kit for Text | http://www.data-miner.com/ | Empresa AI-Data-Miner LLC |
| SAS Enterprise Miner | http://www.sas.com/en_us/software/analytics/enterprise-miner.html | SAS Institute Inc. |
| SOBEK Mining | http://sobek.ufrgs.br/ | Grupo de Pesquisa GTech.Edu UFRGS - Universidade Federal do Rio Grande do Sul |
| TagCrowd | http://tagcrowd.com/ | Daniel Steinbock University of Stanford |
| Text Mine | http://textmine.sourceforge.net/ | Manu Konchady George Mason University (Em outubro de 2015 a ferramenta parou de ter manutenção, conforme mensagem no sítio) |
| TextAlyser | http://textalyser.net/ | sem fonte |
| TMSK – Text-Miner Software Kit | http://www.data-miner.com/ | Empresa AI-Data-Miner LLC |
| Vivísimo/Clusty | http://www.clusty.com | Empresa Vivísimo (Ao entrar no site, é redirecionado para outro site http://yippy.com/) |
| Wordcounter | http://www.wordcounter.com/ | Steven Morgan Friedman University of Pennsylvania |

Fonte: Elaborado pela autora

Para o estudo do projeto de pesquisa de mestrado em questão, foram selecionadas para análise, dentre as ferramentas prospectadas no quadro 4, somente sete ferramentas, de acordo com as seguintes características:

- a) gratuitas;
- b) para uso da mineração de textos no idioma português;
- c) disponíveis para *download* da ferramenta no *site* ou execução *online*;
- d) não sejam oriundas de empresas comerciais.

O intuito é escolher apenas uma ferramenta para usar neste projeto de pesquisa, aquela que seja a mais viável e se aproxima com os objetivos de minerar os textos das mensagens do portal IBGEANDO. Assim, as ferramentas selecionadas para serem analisadas foram: KH Coder, NLTK, PyPLN, SOBEK Mining, TagCrowd, TextAlyser e Wordcounter.

Assim, para este estudo foi adotada a abordagem de avaliação da qualidade de *software* referente aos aspectos da interação humano-computador do professor de Aprendizado, Design e Tecnologia, Thomas Reeves (1998), por ter sido esse dentre os métodos pesquisados o que mais se adequou ao escopo do projeto. Com base nos fundamentos dos 10 aspectos elencados por Reeves, esta pesquisadora definiu critérios desejáveis com nomenclatura mais específica para este trabalho tendo em vista a particularidade das ferramentas de mineração de textos e para que não se gastasse muito tempo avaliando ferramentas que não se adequariam às necessidades do IBGE. Desta forma, o Quadro 5 apresenta os aspectos elencados por Reeves e os critérios desejáveis definidos por esta pesquisadora.

Quadro 5: Método de Reeves e os critérios desejáveis para ferramenta de mineração de textos

(continua)

| MÉTODO DE REEVES Aspectos da Interface com o Usuário (Interação Humano-Computador) | CRITÉRIOS | DESCRIÇÃO DOS CRITÉRIOS |
|---|------------------------|--------------------------------|
| 1. Facilidade de utilização Refere-se ao grau de facilidade percebido pelo usuário, quando ele interage com a aplicação. | Facilidade de Operação | se é fácil usar a ferramenta. |
| 2. Navegação Refere-se à habilidade de acessar os conteúdos do software, de um tópico a outro. | | |

Quadro 5: Método de Reeves e os critérios desejáveis para ferramenta de mineração de textos
(continua)

| MÉTODO DE REEVES Aspectos da Interface com o Usuário (Interação Humano-Computador) | CRITÉRIOS | DESCRIÇÃO DOS CRITÉRIOS |
|---|---|--|
| 3. Carga Cognitiva Refere-se ao esforço mental requerido durante a execução das tarefas no software, como exploração dos conteúdos, uso da estrutura, respostas demandadas, etc. | Conhecimento prévio de programação computacional | se para utilizar a ferramenta para mineração de texto é necessário conhecer programação para desenvolver algoritmos ou programas para a mineração. |
| 4. Mapeamento Refere-se à habilidade do software em rastrear e representar, de forma clara para o usuário, os caminhos por ele percorridos ao usar o software. | Não se aplica | |
| 5. Design de tela Compreende aspectos como aparência e disposição dos elementos nas telas do software incluindo texto, ícones, gráficos, cores, etc. | Linguagem da <i>Interface</i> | idioma da tela da ferramenta. |
| 6. Compatibilidade Espacial do conhecimento Refere-se à rede de conceitos e relacionamentos que compõem o esquema mental que um usuário possui sobre determinado tema ou fenômeno, que deve ser considerada pelo software. | Análise de texto no Idioma Português | se analisa texto no idioma português. |
| | Considera palavras acentuadas | se considera e mantém a acentuação das palavras em português. |
| 7. Apresentação da Informação Está relacionada a se a informação contida no espaço de conhecimento incorporado no software é apresentado de maneira entendível. | Entrada de Dados (Manual / Importação de Arquivos) | se o texto a ser analisado é inserido manualmente ou se aceita arquivo com o texto para a ferramenta analisar. |
| | Saída de Dados (Tela / Exportação de Arquivos) | se o resultado da análise é exibido em tela e/ou é exportado em arquivo. |
| 8. Integração de Mídias Refere-se à quão bem o software combina diferentes mídias (áudio, texto, imagem, vídeo), para produzir um todo que atenda aos objetivos do software. | Existe Integração com outras ferramentas | se existe a possibilidade tecnológica de compartilhar informações entre a ferramenta e outros sistemas. |
| 9. Estética Refere-se a aspectos artísticos do software, no sentido destes expressarem beleza, estilo, elegância, etc. | Visualização Gráfica do Resultado | se o resultado final da análise é exibido numa forma gráfica. |
| | Visualização Gráfica dos relacionamentos entre palavras | se gera uma rede de relacionamentos das palavras mais frequentes. |

Quadro 5: Método de Reeves e os critérios desejáveis para ferramenta de mineração de textos (conclusão)

| MÉTODO DE REEVES Aspectos da Interface com o Usuário (Interação Humano-Computador) | CRITÉRIOS | DESCRIÇÃO DOS CRITÉRIOS |
|---|--|---|
| <p>10. Funcionalidade Geral Representa uma dimensão mais abrangente, relacionada à utilidade do software e atendimento dos objetivos pretendidos.</p> | Instalação (Local / Online) | se a ferramenta tem que ser instalada numa máquina local ou pode ser executada na Internet. |
| | Possui documentação sobre a ferramenta | se possui documentação sobre a instalação e uso da ferramenta. |
| | Processamento do teste executou satisfatoriamente | se o teste com as mensagens textuais da categoria Aposentadoria do IBGEANDO foi executado de forma a obter um resultado satisfatório. |
| | Permite lista de <i>stopwords</i> | se permite inserir uma lista de palavras a serem removidas na análise do texto. |
| | Permite lista de <i>stopwords</i> em Português | se possui lista de <i>stopwords</i> no idioma português. |
| | Entrada da lista de <i>stopwords</i> (Manual / Importação de Arquivos) | se é possível inserir a lista manualmente ou importar um arquivo com a lista das palavras a serem removidas. |
| | Usa <i>stemming</i> ou lematização | se usa a técnica de lematização (identificação do radical de uma palavra) para calcular a frequência das palavras. |
| | Exibe o número da Frequência das Palavras | se mostra a frequência das palavras analisadas. |

Fonte: Elaborado pela autora

Considerando os critérios elaborados no Quadro 5 acima, foram definidos os parâmetros de uma ferramenta ideal de mineração de textos para as mensagens textuais do IBGEANDO, tendo em vista a pretensão futura de utilização da mesma no IBGE. Assim, surgiram os parâmetros de comparação para cada critério entre as ferramentas selecionadas e sua devida justificativa, conforme o Quadro 6.

Quadro 6: Critérios e parâmetros para uma ferramenta ideal para mineração de textos do IBGEANDO

(continua)

| | | FERRAMENTA IDEAL PARA O IBGE | |
|---|--|-------------------------------|--|
| # | CRITÉRIOS | PARÂMETRO | MOTIVO |
| 1 | Instalação (Local / <i>Online</i>) | Local | Como a pretensão futura é a utilização da ferramenta no IBGE e que seja possível a integração com o Portal IBGEANDO e outros eventuais sistemas, o ideal é que a ferramenta seja instalada localmente numa máquina servidora do IBGE. |
| 2 | Possui documentação sobre a ferramenta | Sim | Facilidade de aprendizado e otimização no tempo de instalação da ferramenta pelos técnicos e usuários do IBGE, minimizando a resistência para a implantação da mesma, uma vez que exista uma documentação sobre a instalação e uso da ferramenta.. |
| 3 | Facilidade de Operação | Fácil | Facilidade de utilização pelos usuários do IBGE, minimizando a resistência para o seu uso. |
| 4 | Processamento do teste executou satisfatoriamente | Sim | A ferramenta deve gerar um resultado que seja compreensível para os usuários do IBGE, assim a ferramenta analisada deve apresentar um resultado satisfatório nos testes com as mensagens textuais da categoria Aposentadoria do IBGEANDO. |
| 5 | Linguagem da <i>Interface</i> | Português | Qualquer usuário do IBGE pode utilizar a ferramenta, mesmo que não tenha domínio em outro idioma que não seja o português. |
| 6 | Entrada de Dados (Manual / Importação de Arquivos) | Importação de Arquivo | Os relatórios com todas as mensagens textuais do IBGEANDO serão geradas em arquivos o que facilita a entrada de dados por importação de arquivos e ainda viabiliza a sua sistematização. |
| 7 | Saída de Dados (Tela / Exportação de Arquivos) | Tela e Exportação de Arquivos | Os usuários do IBGE podem querer visualizar os resultados da mineração dos textos somente na tela ou gerar arquivos para integração entre ferramentas ou geração de relatórios formais. |

Quadro 6: Critérios e parâmetros para uma ferramenta ideal para mineração de textos do IBGEANDO
(continua)

| FERRAMENTA IDEAL PARA O IBGE | | | |
|------------------------------|---|-----------------------|--|
| # | CRITÉRIOS | PARÂMETRO | MOTIVO |
| 8 | Análise de texto no Idioma Português | Sim | Todas as mensagens postadas no canal de atendimento são escritas no idioma português. |
| 9 | Permite lista de <i>StopWords</i> | Sim | Primordial que sejam retiradas as palavras não relevantes na análise das mensagens textuais ou por não traduzirem a essência das mesmas, de forma a gerar um resultado mais objetivo para obter insumos para a tomada de decisão. |
| 10 | <i>Stopwords</i> em Português | Sim | Ideal que a ferramenta já possua internamente uma lista de <i>stopwords</i> com os artigos, pronomes, advérbios, entre outros, no idioma português. |
| 11 | Entrada de <i>stopwords</i> (Manual / Importação de Arquivos) | Importação de Arquivo | Muito importante que a ferramenta permita a entrada de arquivos com as palavras a serem removidas em sua configuração. Além disso, em alguns casos pode ser necessário a introdução de novas palavras para remoção para gerar um melhor resultado. Viabiliza a integração entre ferramentas. |
| 12 | Usa <i>stemming</i> ou lematização | Não | No caso das mensagens textuais do IBGEANDO, inicialmente não seria interessante utilizar a técnica de lematização, pois pode ocultar ou gerar um resultado indevido. Por exemplo, no caso das mensagens da categoria Aposentadoria, pode conter as palavras "Aposentadoria", "Aposentou", "Aposentei", "Aposentar", e, caso estas palavras sejam consideradas similares, no resultado final não será possível identificar que estão falando sobre servidores que já se aposentaram ou que ainda vão se aposentar, e isso, é vital no momento de tomar uma decisão, pois a ação é direcionada para públicos diferentes. |

Quadro 6: Critérios e parâmetros para uma ferramenta ideal para mineração de textos do IBGEANDO
(conclusão)

| FERRAMENTA IDEAL PARA O IBGE | | | |
|------------------------------|---|-----------|---|
| # | CRITÉRIOS | PARÂMETRO | MOTIVO |
| 13 | Considera palavras acentuadas | Sim | Primordial que sejam consideradas as palavras acentuadas corretamente para gerar um resultado compreensível e apresentável. |
| 14 | Exibe o número da Frequencia das Palavras | Sim | Para ter noção do percentual de ocorrência da palavra num contexto geral das categorias. |
| 15 | Visualização Gráfica do Resultado | Sim | A visualização gráfica do resultado permite fazer a leitura de um resultado num contexto geral mais rápido. |
| 16 | Visualização Gráfica dos relacionamentos entre palavras | Sim | A visualização gráfica facilita o entendimento das relações entre as palavras. |
| 17 | Existe Integração com outras ferramentas | Sim | O ideal é ter uma proposta tecnológica que permita a obtenção automática dos insumos a partir das mensagens textuais, para isso a ferramenta de mineração de textos deve permitir integração com outras ferramentas de modo a facilitar a sistematização da solução pelos técnicos do IBGE. |
| 18 | Conhecimento prévio de programação computacional | Não | O ideal é que a ferramenta seja utilizada por qualquer servidor do IBGE, sem demandar em conhecimentos técnicos ou treinamentos adicionais. |

Fonte: Elaborado pela autora

Desta forma, a análise das ferramentas selecionadas consistiu na instalação e tentativa de uso das ferramentas com algumas mensagens textuais da seção “Qual a sua Dúvida?” do IBGEANDO. Tentativa, pois, em algumas situações, não foi possível instalar a ferramenta por necessitar de conhecimentos de computação, ou executá-la devido a erros que ocorreram na ferramenta. Foram escolhidas as mensagens da Categoria Aposentadoria da seção “Qual a sua Dúvida?” por ter apenas 530 mensagens, considerada uma categoria com menos postagens e de grande importância para os servidores do IBGE. Cabe ressaltar que este trabalho não foi exaustivo na prospecção de todas as ferramentas gratuitas e em português disponíveis, se

limitando apenas aquelas que foram mencionadas nos trabalhos e artigos descritos no Capítulo 3.4.1.

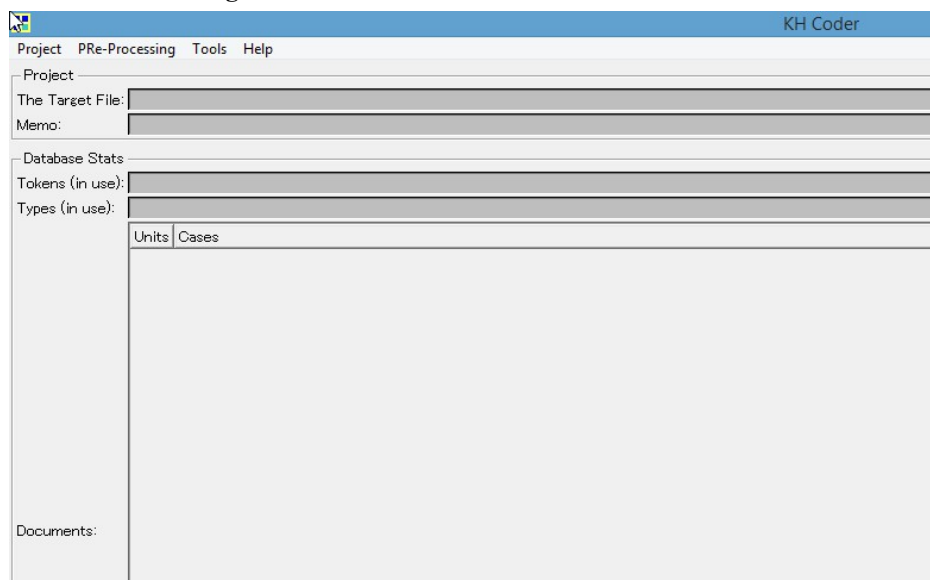
- a) *KH Coder*: ferramenta livre para análise quantitativa de conteúdo ou mineração de textos. Ela também é utilizada para a linguística computacional. É possível analisar os idiomas Japonês, Inglês, Francês, Alemão, Italiano, Português e Espanhol. Oferece vários tipos de funções de pesquisa e análise estatística, utilizando ferramentas de *back-end*, como *Stanford POS Tagger*, *stemmer Snowball*, MySQL e R.

Características utilizadas para Palavras: Lista de Frequência, *Searching*, KWIC Concordância, *Collocation Stats*, Análise de Correspondência, Escalonamento Multi-Dimensional, Rede de Co-Ocorrência e Análise de Agrupamento Hierárquico.

Características utilizadas para Documentos: *Searching*, *Clustering* e Classificador *Naive Bayes*.

A Figura 3 mostra a tela inicial da ferramenta.

Figura 3: Tela inicial da ferramenta *KH Coder*



Fonte: Elaborado pela autora

No teste com as mensagens textuais da categoria Aposentadoria do IBGEANDO, a ferramenta fez primeiro a lematização das palavras de forma a ser pouco compreensível para análise, conforme Figura 4. Um dos problemas identificados é que a ferramenta não considera a acentuação das palavras no idioma português.

Figura 4: Exemplo do arquivo de stemming do KH Coder

| | | | | | |
|---------------|---------------|-------------|---------|-----|---|
| EOS | | | | | |
| Prezados | Prezados | prez | ALL | . | . |
| srs. | srs. | srs | ALL | . | . |
| Estou | Estou | estou | ALL | . | . |
| em | em | em | ALL | . | . |
| processo | processo | processo | process | ALL | . |
| de | de | de | ALL | . | . |
| aposentadoria | aposentadoria | aposentador | | ALL | . |
| e | e | e | ALL | . | . |
| tenho | tenho | tenh | ALL | . | . |
| dvidas | dvidas | dvid | ALL | . | . |
| sobre | sobre | sobr | ALL | . | . |
| o | o | o | ALL | . | . |
| procedimento | procedimento | proced | ALL | . | . |
| de | de | de | ALL | . | . |
| como | como | com | ALL | . | . |
| como | como | com | ALL | . | . |
| fazer | fazer | faz | ALL | . | . |
| a | a | a | ALL | . | . |
| transferncia | transferncia | transfernc | | ALL | . |
| da | da | da | ALL | . | . |
| linha | linha | linh | ALL | . | . |
| telefonica | telefonica | telefn | ALL | . | . |
| . | . | . | ALL | . | . |
| jf | jf | jf | ALL | SP | . |
| Se | Se | se | ALL | . | . |
| tenho | tenho | tenh | ALL | . | . |
| que | que | que | ALL | . | . |
| transferir | transferir | transfer | | ALL | . |
| para | para | par | ALL | . | . |
| o | o | o | ALL | . | . |
| meu | meu | meu | ALL | . | . |
| superior | superior | superior | | ALL | . |
| imediat | imediat | imediat | ALL | . | . |
| (| (| (| ALL | . | . |
| gerente | gerente | gerent | ALL | . | . |

Fonte: Elaborado pela autora

Após fazer a lematização, a ferramenta fez o cálculo da frequência de palavras, gerando como resultado o exibido na Figura 5.

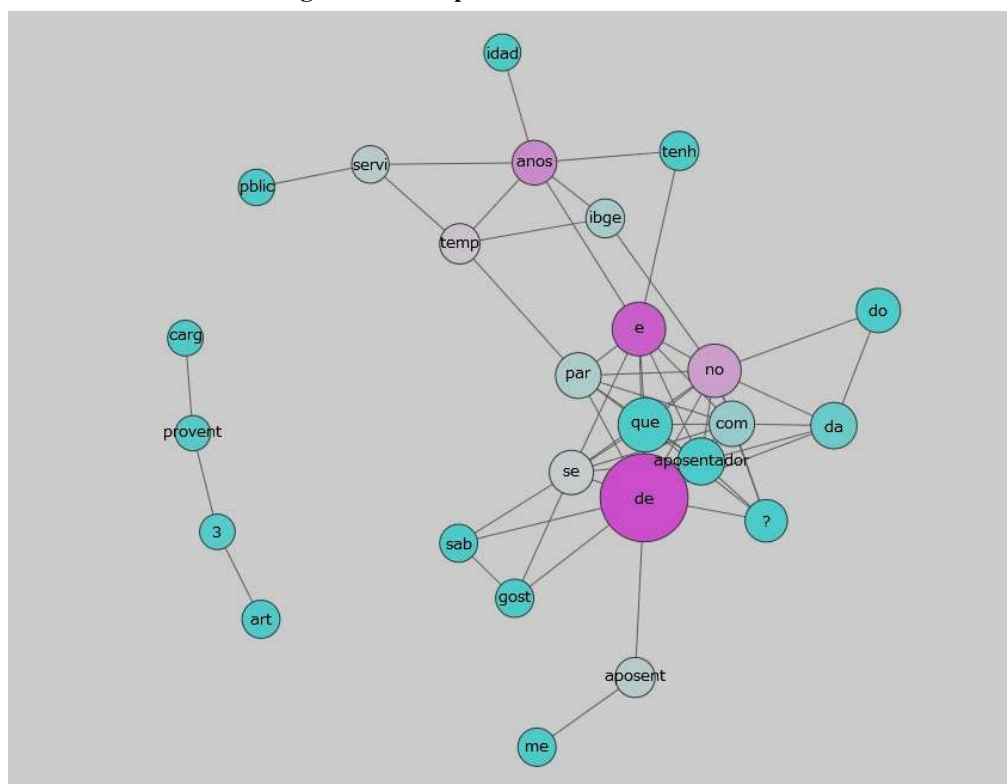
Figura 5: Exemplo da visualização da frequência de palavras do KH Coder

| | A | B | C |
|---|-------------|-----|---|
| 1 | ALL | | |
| 8 | e | 838 | |
| 9 | no | 826 | |
| 0 | aposentador | 573 | |
| 1 | em | 566 | |
| 2 | da | 541 | |
| 3 | par | 515 | |
| 4 | com | 507 | |
| 5 | anos | 473 | |
| 6 | do | 458 | |
| 7 | se | 458 | |
| 8 | ??? | 419 | |
| 9 | ? | 386 | |
| 0 |) | 313 | |
| 1 | temp | 299 | |
| 2 | aposent | 287 | |
| 3 | na | 258 | |
| 4 | os | 249 | |
| 5 | tenh | 248 | |
| 6 | ibge | 236 | |
| 7 | pel | 232 | |
| 8 | (| 222 | |
| 9 | servidor | 219 | |
| 0 | me | 215 | |
| 1 | gost | 210 | |
| 2 | art | 205 | |
| 3 | sab | 201 | |
| 4 | - | 200 | |
| 5 | por | 197 | |
| 6 | j | 195 | |
| 7 | ser | 190 | |
| 8 | minh | 188 | |

Fonte: Elaborado pela autora

Depois gerou um grafo com base na lematização que não fica compreensível, conforme pode-se observar na figura 6.

Figura 6: Exemplo de Grafo do *KH Coder*



Fonte: Elaborado pela autora

Apesar de dispor de diversas funcionalidades úteis, algumas não foram executadas corretamente nesta ferramenta e nem geraram um resultado final.

A avaliação geral é que demanda um tempo maior para entendimento da utilização da ferramenta. Acredita-se que o problema principal de uso desta ferramenta foi não trabalhar adequadamente com o idioma português, fazendo a lematização de uma forma pouco compreensível para o português, e não permitir a importação de arquivo com a lista de *stopwords*.

- b) NLTK – *Natural Language Toolkit*: é um pacote de código aberto para a construção de programas em Python que traz uma gama de ferramentas para diversos tipos de análise de texto, incluindo cálculo de métricas comuns, extração de informações e processamento de linguagem natural. O pacote foi projetado

[...] para disponibilizar coleções de módulo independentes, em que cada um define uma estrutura de dados específica, sendo os principais módulos:

parser, que especifica uma interface representando os textos através de árvores; o rotulador, que é responsável por expandir as características e informações de um determinado *token* com dados adicionais; o classificador, que se utiliza de uma interface para a classificação de textos em categorias, sendo realizado através do algoritmo *Naive Bayes*. (BRITO, 2016, p. 29)

De acordo com Bird et. al. (2009), o NLTK foi concebido em 2001 como parte de um curso de Linguística Computacional no Departamento de Ciência da Computação e Informação da Universidade da Pensilvânia. Devido a seu caráter de *software* aberto e gratuito, tem sido desenvolvido e ampliado com a ajuda de dezenas de colaboradores, pelo seu uso e concepção de módulos de análise linguística.

Para instalar o NLTK, você precisa primeiro ter instalado o Python. No site está disponível toda a documentação necessária para o seu uso.

[...] ainda que o NLTK não seja necessariamente o que há de mais avançado quando comparado às opções comerciais e aos recursos do meio acadêmico, ele oferece uma estrutura sólida e ampla. (...) Caso seu projeto seja tão sofisticado que a qualidade e a eficiência do NLTK não satisfaçam suas necessidades, você terá aproximadamente, três opções, dependendo da quantidade de tempo e dinheiro que estiver disposto a investir: explorar as opções do código aberto em busca de uma alternativa mais adequada realizando experimentos e testes de desempenhos comparativos, pesquisar textos técnicos e produzir sua própria biblioteca, ou adquirir um produto comercial. Nenhuma dessas opções é barata (presumindo que você acredite que tempo é dinheiro) ou fácil. (RUSSEL, 2011, p. 29)

Para este estudo, não houve tempo disponível para exploração da programação e, por isso, esta ferramenta foi considerada difícil por necessitar de conhecimentos técnicos de programação, não sendo de fácil uso para usuários leigos. Desta forma, ela não foi instalada e nem testada.

- c) PyPLN: é uma plataforma, em código aberto, e disponível para a comunidade, para processamento e extração de informações úteis a partir do texto. Ele integra muitas ferramentas de mineração de texto e de processamento de linguagem natural, que pode ser acessado através de uma interface *web* fácil de usar, onde você pode gerenciar documentos e interagir com suas análises/visualizações. Foi desenvolvido no ano de 2012, por um grupo de pesquisa chamado Núcleo de Análise e Modelagem de Dados (NAMD), localizado na Escola de Matemática Aplicada da Fundação Getúlio Vargas no Rio de Janeiro.

“Esta plataforma, 100% aberta e gratuita, foi desenvolvida em uma linguagem de programação não proprietária - Python - orientada a objeto e adotada em cursos de computação das principais universidades do mundo, o que garante sua acessibilidade e benefício para a comunidade”. (SOUZA et al, 2015, p. 6)

Assim como a ferramenta NLTK, existe uma necessidade de um conhecimento técnico de programação, não sendo de fácil uso por usuários leigos. Desta forma, esta ferramenta foi considerada difícil e não foi instalada e nem testada.

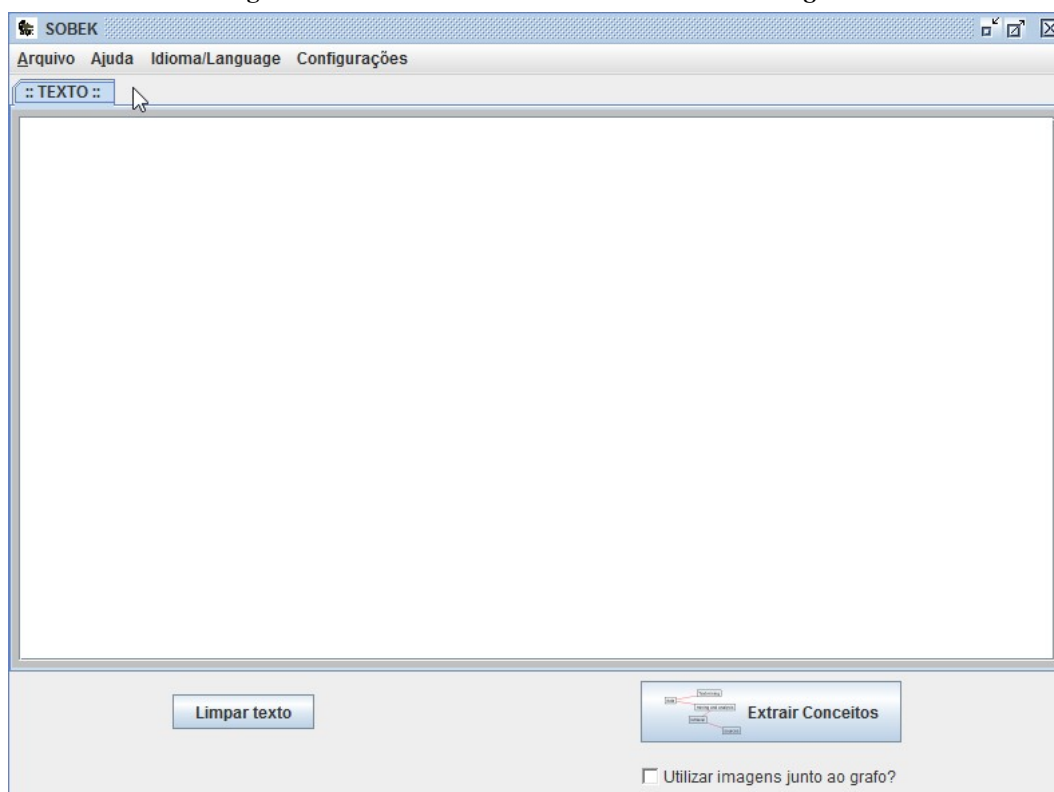
- d) *SOBEK Mining*: foi desenvolvida no Programa de Pós-Graduação em Informática na Educação, na Universidade Federal do Rio Grande do Sul (UFRGS) pelo Grupo de Pesquisa GTech.Edu, liderado pelo Professor Eliseo Berni Reategui. Ela é capaz de identificar os termos relevantes em um texto a partir da análise de frequência destes termos no material textual.

SOBEK foi criada no ano de 2007 como uma ferramenta de mineração de texto para auxiliar os professores do ensino a distância a avaliarem o trabalho dos alunos feito colaborativamente. No ano de 2009, SOBEK começou a ser utilizada para auxiliar os alunos na compreensão da leitura e tarefas de resumo de texto e no ano de 2010, foi incorporada em outros sistemas, tais como: avaliação de postagens dos alunos em fóruns de discussão, jogos digitais para promover a narrativa escrita, ferramenta de aprendizagem baseada em projetos com recomendação de conteúdo. A ferramenta SOBEK foi construída a partir de um algoritmo inicialmente definido por Schenker (2003) e modificado depois pelo Grupo de Pesquisa GTech.Edu a fim de torná-lo mais adequado às aplicações educacionais.

O SOBEK é um minerador de textos desenvolvido para fins pedagógicos, cuja uma de suas principais características é a facilidade de uso, não sendo necessário nenhum treinamento prévio e sendo bastante flexível para diferentes aplicações. Apesar de simples de utilizar, possui uma ampla gama de configurações, possibilitando a personalização da ferramenta para diferentes fins e ambientes. (EPSTEIN E REATEGUI, 2015, p. 4)

O minerador SOBEK, quando executado, apresenta uma interface para receber o texto a ser minerado e ajustar as suas configurações, como mostra a Figura 7.

Figura 7: Tela inicial da ferramenta SOBEK Mining

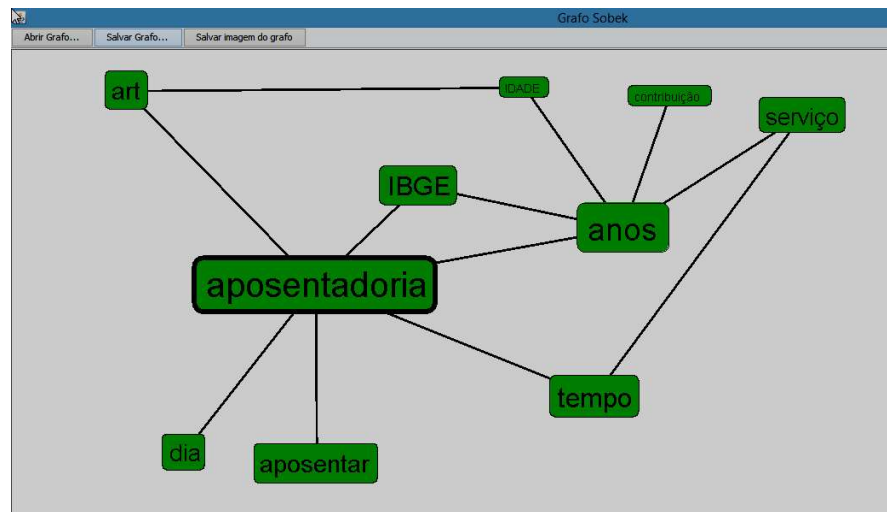


Fonte: Elaborado pela autora

Com o texto presente no minerador, basta clicar no botão “Extrair Conceitos”¹⁸ para que o minerador SOBEK comece o seu processo de análise para a construção do grafo dos termos frequentes presentes no texto. Após o tempo de análise (que varia conforme o tamanho do texto sendo minerado) o grafo é apresentado ao usuário, como mostra a Figura 8, com o resultado do teste com as mensagens textuais da categoria Aposentadoria do IBGEANDO.

¹⁸ Cabe ressaltar que o termo Conceito é utilizado pela ferramenta SOBEK como um *label* para denominar as palavras que são extraídas, sem ter nenhuma referência à Teoria do Conceito da CI ou outras teorias. Ao questionar aos desenvolvedores do SOBEK a referência de uso desta nomenclatura, os mesmos informaram que não existia e que eles atualmente chamam de “termos” ao invés de “conceitos” mas que ainda não fizeram uma nova versão do sistema para alteração desse *label* na interface.

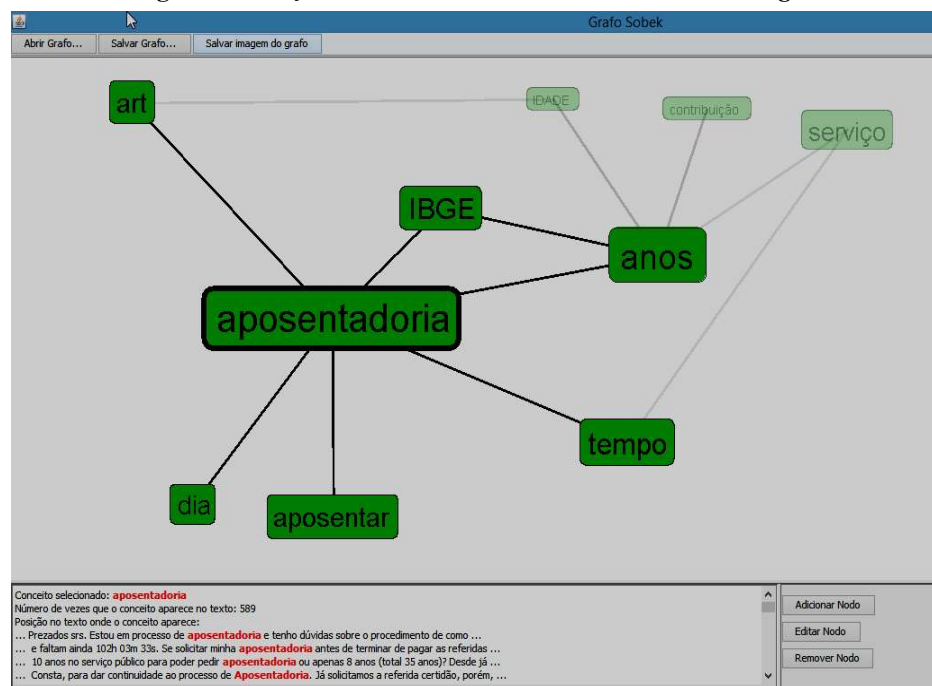
Figura 8: Grafo da ferramenta SOBEK Mining



Fonte: Elaborado pela autora

Nos grafos apresentados pelo SOBEK, os nodos maiores representam termos mais frequentes no texto analisado. Quando um destes nodos é selecionado, os nodos adjacentes a ele são destacados, além de algumas informações a respeito dele serem apresentadas na caixa de texto inferior da interface, mostrando o número de vezes que ele apareceu no texto minerado, por exemplo, conforme exibido na Figura 9.

Figura 9: Seleção de nodos da ferramenta SOBEK Mining



Fonte: Elaborado pela autora

A avaliação geral é que esta é uma ferramenta fácil de ser utilizada e bem documentada, mostrando um grafo que exibe os termos e suas relações de forma clara e objetiva que pode ser usada para o estudo aprofundado da presente pesquisa sem dispêndio de tempo de instalação e entendimento da ferramenta.

- e) *TagCrowd*: é uma aplicação *web* para a visualização de frequência de palavras em qualquer texto, criando o que é conhecido popularmente como uma nuvem de palavra, nuvem texto ou nuvem de *tags*. Foi criado em julho de 2006 por Daniel Steinbock, um estudante de doutorado em *Design* e Educação na Universidade de Stanford.

Esta nuvem de *tags*, geralmente reúne 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.

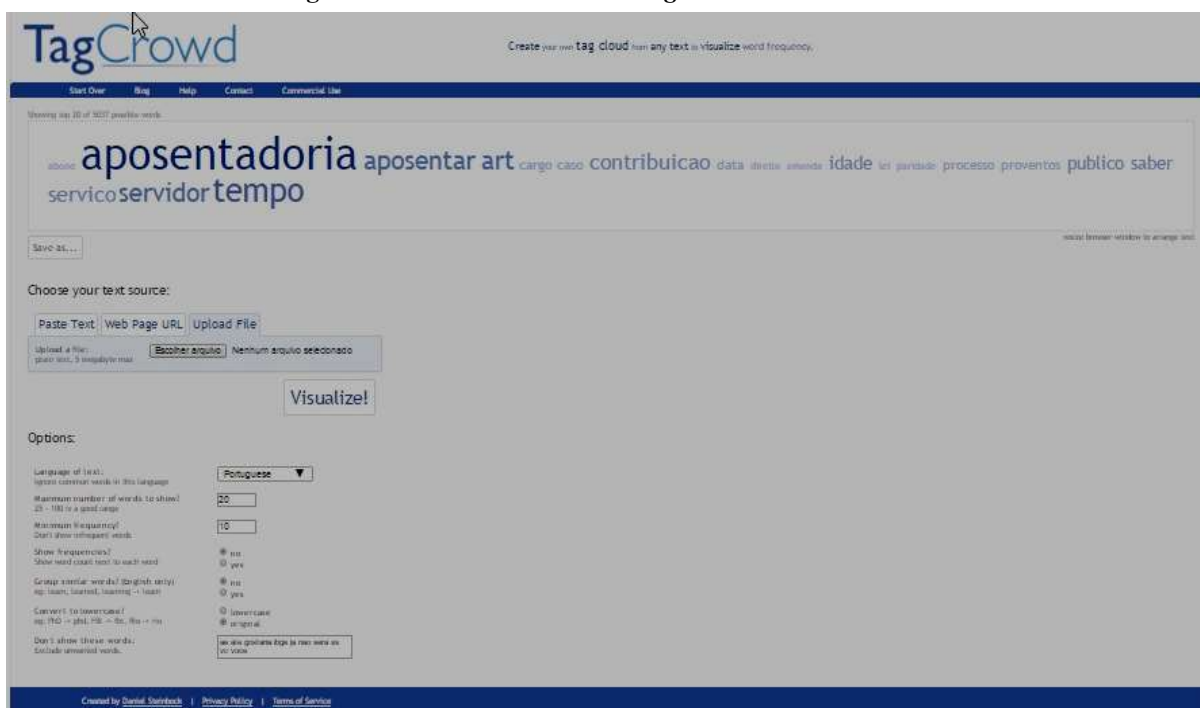
Uma nuvem de palavra é uma imagem exibida de forma agradável e informativa ao usuário. *TagCrowd* é especializada na criação de nuvens de palavras que sejam fáceis de ler, analisar e comparar, tendo diversos propósitos:

- como resumos para discursos e trabalhos escritos;
- como ferramenta de *blog* ou de análise de *site*;
- para a análise visual de dados qualitativos;
- como nuvens de marcas que permitem que empresas vejam como elas são percebidas pelo mundo;
- para ajudar escritores e estudantes refletirem sobre seu trabalho;
- como etiquetas para conferências, festas ou onde haja novas colaborações.

Na Figura 10, é possível visualizar a tela do *TagCrowd* e a exibição do resultado do teste com as mensagens textuais da categoria Aposentadoria do IBGEANDO.

A ferramenta é muito fácil de ser utilizada, mas existem algumas limitações como não permitir inserir uma lista de *stopwords* e gerar resultado graficamente.

Figura 10: Tela da ferramenta *TagCrowd* com o resultado



Fonte: Elaborado pela autora

- f) *TextAlyser*: é uma ferramenta de análise de texto *online*, exibindo estatísticas detalhadas e servindo para descobrir o assunto de um texto.

Inicialmente, 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. (KLEMMANN et al, 2011, p. 1101)

A tela inicial da ferramenta é exibida na Figura 11.

Figura 11: Tela inicial da ferramenta *TextAlyser*

Enter your text to analyze here :

or analyze a website :

or select file from local hdd : Nenhum arquivo selecionado

Analysis options :

Minimum characters per word :

Special word or expression to analyze :

Number of words to be analyzed :

Ignore numbers :

Log the query (only for websites) :

Apply stoplist :

Apply own stoplist (separe with blanks) :

Make a link analysis :

Exhaustive polyword phrases :

Fonte: Elaborado pela autora

Ao executar o teste com as mensagens textuais da categoria Aposentadoria do IBGEANDO, foi apresentado o seguinte resultado geral, conforme Figura 12.

Figura 12: Exibição do Resultado Geral da ferramenta *TextAlyser*

Textalyser The complete results, including compexity factor, and
Results other features

| | |
|--|-------|
| Total word count : | 875 |
| Number of different words : | 545 |
| Complexity factor (Lexical Density) : | 62.3% |
| Readability (Gunning-Fog Index) : (6-easy 20-hard) | 10.2 |
| Total number of characters : | 16628 |
| Number of characters without spaces : | 12998 |
| Average Syllables per Word : | 2.07 |
| Sentence count : | 223 |
| Average sentence length (words) : | 11.39 |
| Max sentence length (words) : | 161 |

Fonte: Elaborado pela autora

A exibição das palavras mais frequentes é demonstrada na Figura 13.

Figura 13: Exibição das palavras com maior frequência da ferramenta *TextAlyser*

Frequency and top words :

| Word | Occurrences | Frequency | Rank |
|---------------|-------------|-----------|------|
| aposentadoria | 20 | 2.3% | 1 |
| para | 13 | 1.5% | 2 |
| anos | 11 | 1.3% | 3 |
| abono | 10 | 1.1% | 4 |
| paridade | 10 | 1.1% | 4 |
| aposentar | 10 | 1.1% | 4 |
| permanência | 9 | 1% | 5 |
| ký | 9 | 1% | 5 |
| servidor | 9 | 1% | 5 |
| idade | 8 | 0.9% | 6 |

Fonte: Elaborado pela autora

A ferramenta também trabalha com sequência de palavras, a Figura 14 exibe o resultado de frases com 5 (cinco) palavras.

Figura 14: Exibição das sequências de palavras mais frequentes da ferramenta *TextAlyser*

5 word phrases frequency :

| Expression | Expression count | Frequency | Prominence |
|--------------------------------------|------------------|-----------|------------|
| total de acordo com o | 2 | 0.1% | 13.9 |
| paridade total de acordo com | 2 | 0.1% | 14 |
| e paridade total de acordo | 2 | 0.1% | 14 |
| integrais e paridade total de | 2 | 0.1% | 14.1 |
| proventos integrais e paridade total | 2 | 0.1% | 14.2 |
| com proventos integrais e paridade | 2 | 0.1% | 14.3 |
| com proventos integrais e | 2 | 0.1% | 14.3 |
| com proventos integrais | 2 | 0.1% | 14.4 |
| de com | 2 | 0.1% | 14.6 |
| aposentar a partir de | 2 | 0.1% | 14.8 |
| com o artigo da | 2 | 0.1% | 18.3 |
| acordo com o artigo | 2 | 0.1% | 18.4 |
| de acordo com o artigo | 2 | 0.1% | 18.5 |

Fonte: Elaborado pela autora

É uma ferramenta de fácil uso, mas com uma visualização pouco agradável e com algumas limitações, como não aceitar lista de *stopword* e não exportar o resultado.

- g) *WordCounter*: é uma ferramenta *online* gratuita que apresenta a relação das palavras mais utilizadas em um texto. O resultado é uma lista das palavras que mais se repetem no texto e sua respectiva frequência. Ela não disponibiliza ferramentas gráficas mais complexas para visualização dos resultados.

A tela inicial da ferramenta é apresentada na Figura 15.

Figura 15: Tela inicial da ferramenta *WordCounter*

WORDCOUNTER

Wordcounter ranks the most frequently used words in any given body of text. Use this to see what words you overuse (is everything a "solution" for you?) or maybe just to find some keywords from a document.

Wordcounter is useful for writers, editors, students, and anyone who thinks that they might be speaking redundantly or repetitively -- and it's free! Eventually, I'm going to expand it so that you can upload documents, but not yet.

Enter the body of text here (to count & rank the word frequency):

Include Small Words ("the", "it", etc)? No -- exclude them

Use Only Roots (group variations together)? Yes (beta)

How Many Words should I list?

Go >>

Fonte: Elaborado pela autora

No teste com as mensagens textuais da categoria Aposentadoria do IBGEANDO, a ferramenta apresentou o resultado conforme exibido na Figura 16.

Figura 16: Exibição do resultado da ferramenta *WordCounter*

WORDCOUNTER

Wordcounter ranks the most frequently used words in any given body of text. Use this to see what words you overuse (is everything a "solution" for you?) or maybe just to find some keywords from a document.

Wordcounter is useful for writers, editors, students, and anyone who thinks that they might be speaking redundantly or repetitively -- and it's free! Eventually, I'm going to expand it so that you can upload documents, but not yet.

Here are your results...

| Word | Frequency |
|---------------|-----------|
| de | 2239 |
| o | 876 |
| que | 861 |
| e | 838 |
| aposentadoria | 580 |
| em | 566 |
| da | 541 |
| para | 514 |
| anos | 473 |
| se | 450 |
| com | 328 |
| tempo | 298 |
| na | 258 |
| os | 249 |
| ibge | 219 |
| tenho | 214 |
| servidor | 211 |
| não | 206 |
| gostaria | 205 |
| art | 205 |
| saber | 203 |
| por | 197 |
| ao | 187 |
| já | 183 |
| ou | 180 |

Fonte: Elaborado pela autora

A ferramenta é muito simples e fácil de utilizar, no entanto, tem algumas desvantagens como não permitir importar o arquivo com o texto a ser analisado assim como, exportar o resultado, e o mais importante, não permitir inserir uma lista de *stopword* que acaba comprometendo o resultado final.

A escolha da ferramenta a ser utilizada no projeto de pesquisa em questão, foi aquela que teve os parâmetros mais próximos ao da ferramenta ideal, de acordo com o Quadro 6, sobre os critérios desejáveis considerados importantes numa ferramenta ideal de mineração de textos para analisar as mensagens textuais da seção “Qual a sua Dúvida?”.

Cabe ressaltar que estes foram critérios desejáveis definidos para o estudo do projeto em questão, mas que isso não significa que a ferramenta escolhida seja a ideal ou perfeita. No estudo realizado por Barbosa et al (2009, p. 9), utilizando as ferramentas SOBEK e *TagCrowd*, eles já concluíam “que ainda não há uma ferramenta completa, e o ideal é que o usuário leia sobre as características de cada uma e opte por escolher aquela que está mais de acordo com os seus objetivos ou necessidades”. E foi exatamente isso que foi feita nessa análise das ferramentas, identificaram-se as características básicas necessárias para uma ferramenta de mineração de textos para ser utilizado nas mensagens textuais do IBGEANDO e foi feito um comparativo entre elas.

De acordo com o comparativo dos quadros 7 a 24, a ferramenta SOBEK *Mining* foi a que mais se aproxima dos parâmetros de uma ferramenta ideal de mineração de textos para o IBGEANDO. Sendo assim, a ferramenta SOBEK *Mining* é a ferramenta utilizada para analisar as mensagens da seção “Qual a sua Dúvida?”. Cabe ressaltar que no Apêndice B deste trabalho pode ser visualizado um quadro único com o comparativo das ferramentas.

Quadro 7: Comparativo entre as ferramentas selecionadas - Critério #1

| FERRAMENTAS | Critério #1 Instalação (Local / Online) |
|---------------------|--|
| KH CODER | Local |
| NLTK | Local |
| PyPLN | Local |
| SOBEK <i>MINING</i> | Local |
| TAGCROWD | Online |
| TEXTALYZER | Online |
| WORDCOUNTER | Online |
| FERRAMENTA IDEAL | Local |

Fonte: Elaborado pela autora

Quadro 8: Comparativo entre as ferramentas selecionadas – Critério #2

| FERRAMENTAS | Critério #2 Possui documentação sobre a ferramenta |
|---------------------|---|
| KH CODER | Sim |
| NLTK | Sim |
| PyPLN | Sim |
| SOBEK <i>MINING</i> | Sim |
| TAGCROWD | Não |
| TEXTALYZER | Não |
| WORDCOUNTER | Não |
| FERRAMENTA IDEAL | Sim |

Fonte: Elaborado pela autora

Quadro 9: Comparativo entre as ferramentas selecionadas – Critério #3

| FERRAMENTAS | Critério #3 Facilidade de Operação |
|---------------------|---|
| KH CODER | Médio |
| NLTK | Difícil |
| PyPLN | Difícil |
| SOBEK <i>MINING</i> | Fácil |
| TAGCROWD | Fácil |
| TEXTALYZER | Fácil |
| WORDCOUNTER | Fácil |
| FERRAMENTA IDEAL | Fácil |

Fonte: Elaborado pela autora

Quadro 10: Comparativo entre as ferramentas selecionadas – Critério #4

| FERRAMENTAS | Critério #4 Processamento do teste executou satisfatoriamente |
|---------------------|--|
| KH CODER | Médio |
| NLTK | Não instalado e nem testado |
| PyPLN | Não instalado e nem testado |
| SOBEK <i>MINING</i> | Sim |
| TAGCROWD | Sim |
| TEXTALYZER | Sim |
| WORDCOUNTER | Sim |
| FERRAMENTA IDEAL | Sim |

Fonte: Elaborado pela autora

Quadro 11: Comparativo entre as ferramentas selecionadas – Critério #5

| FERRAMENTAS | Critério #5 Linguagem da Interface |
|---------------------|---|
| KH CODER | Inglês/Espanhol |
| NLTK | Não instalado e nem testado |
| PyPLN | Não instalado e nem testado |
| SOBEK <i>MINING</i> | Português |
| TAGCROWD | Inglês |
| TEXTALYZER | Inglês |
| WORDCOUNTER | Inglês |
| FERRAMENTA IDEAL | Português |

Fonte: Elaborado pela autora

Quadro 12: Comparativo entre as ferramentas selecionadas – Critério #6

| FERRAMENTAS | Critério #6 Entrada de Dados (Manual / Importação de Arquivos) |
|---------------------|---|
| KH CODER | Importação de Arquivo (XLS, CSV, TXT) |
| NLTK | Não instalado e nem testado |
| PyPLN | Não instalado e nem testado |
| SOBEK <i>MINING</i> | Importação de arquivos (DOC, PDF, TXT) |
| TAGCROWD | Manual / Importação de Arquivo |
| TEXTALYZER | Manual / Importação de Arquivo |
| WORDCOUNTER | Manual |
| FERRAMENTA IDEAL | Importação de Arquivo |

Fonte: Elaborado pela autora

Quadro 13: Comparativo entre as ferramentas selecionadas – Critério #7

| FERRAMENTAS | Critério #7 Saída de Dados (Tela / Exportação de Arquivos) |
|---------------------|---|
| KH CODER | Exportação de Arquivo (TXT) |
| NLTK | Não instalado e nem testado |
| PyPLN | Não instalado e nem testado |
| SOBEK <i>MINING</i> | Tela, Imagem JPG, Grafo em XML |
| TAGCROWD | Tela e Saídas em arquivo no formato HTML e PDF |
| TEXTALYZER | Tela |
| WORDCOUNTER | Tela |
| FERRAMENTA IDEAL | Tela / Exportação de Arquivos |

Fonte: Elaborado pela autora

Quadro 14: Comparativo entre as ferramentas selecionadas – Critério #8

| FERRAMENTAS | Critério #8 Análise de texto no Idioma Português |
|---------------------|---|
| KH CODER | Sim |
| NLTK | Não instalado e nem testado |
| PyPLN | Não instalado e nem testado |
| SOBEK <i>MINING</i> | Sim |
| TAGCROWD | Sim |
| TEXTALYZER | Sim |
| WORDCOUNTER | Sim |
| FERRAMENTA IDEAL | Sim |

Fonte: Elaborado pela autora

Quadro 15: Comparativo entre as ferramentas seleccionadas – Critério #9

| FERRAMENTAS | Critério #9 Permite lista de StopWords |
|---------------------|---|
| KH CODER | Não |
| NLTK | Não instalado e nem testado |
| PyPLN | Não instalado e nem testado |
| SOBEK <i>MINING</i> | Sim |
| TAGCROWD | Sim |
| TEXTALYZER | Sim |
| WORDCOUNTER | Não |
| FERRAMENTA IDEAL | Sim |

Fonte: Elaborado pela autora

Quadro 16: Comparativo entre as ferramentas seleccionadas – Critério #10

| FERRAMENTAS | Critério #10 Permite lista de <i>stopwords</i> em Português |
|---------------------|--|
| KH CODER | Não |
| NLTK | Não instalado e nem testado |
| PyPLN | Não instalado e nem testado |
| SOBEK <i>MINING</i> | Sim |
| TAGCROWD | Sim |
| TEXTALYZER | Não |
| WORDCOUNTER | Não |
| FERRAMENTA IDEAL | Sim |

Fonte: Elaborado pela autora

Quadro 17: Comparativo entre as ferramentas selecionadas – Critério #11

| FERRAMENTAS | Critério #11 Entrada da lista de <i>stopwords</i> (Manual / Importação de Arquivos) |
|---------------------|--|
| KH CODER | Não permite lista |
| NLTK | Não instalado e nem testado |
| PyPLN | Não instalado e nem testado |
| SOBEK <i>MINING</i> | Importação de Arquivos |
| TAGCROWD | Manual |
| TEXTALYZER | Manual |
| WORDCOUNTER | Não permite lista |
| FERRAMENTA IDEAL | Importação de Arquivo |

Fonte: Elaborado pela autora

Quadro 18: Comparativo entre as ferramentas selecionadas – Critério #12

| FERRAMENTAS | Critério #12 Usa <i>stemming</i> ou lematização |
|---------------------|--|
| KH CODER | Sim |
| NLTK | Não instalado e nem testado |
| PyPLN | Não instalado e nem testado |
| SOBEK <i>MINING</i> | Não |
| TAGCROWD | Não |
| TEXTALYZER | Não |
| WORDCOUNTER | Não |
| FERRAMENTA IDEAL | Não |

Fonte: Elaborado pela autora

Quadro 19: Comparativo entre as ferramentas selecionadas – Critério #13

| FERRAMENTAS | Critério #13 Considera palavras acentuadas |
|---------------------|---|
| KH CODER | Não |
| NLTK | Não instalado e nem testado |
| PyPLN | Não instalado e nem testado |
| SOBEK <i>MINING</i> | Sim |
| TAGCROWD | Sim |
| TEXTALYZER | Sim |
| WORDCOUNTER | Sim |
| FERRAMENTA IDEAL | Sim |

Fonte: Elaborado pela autora

Quadro 20: Comparativo entre as ferramentas selecionadas – Critério #14

| FERRAMENTAS | Critério #14 Exibe o número da Frequencia das Palavras |
|---------------------|---|
| KH CODER | Sim |
| NLTK | Não instalado e nem testado |
| PyPLN | Não instalado e nem testado |
| SOBEK <i>MINING</i> | Sim |
| TAGCROWD | Sim |
| TEXTALYZER | Sim |
| WORDCOUNTER | Sim |
| FERRAMENTA IDEAL | Sim |

Fonte: Elaborado pela autora

Quadro 21: Comparativo entre as ferramentas selecionadas – Critério #15

| FERRAMENTAS | Critério #15 Visualização Gráfica do Resultado |
|---------------------|---|
| KH CODER | Sim |
| NLTK | Não instalado e nem testado |
| PyPLN | Não instalado e nem testado |
| SOBEK <i>MINING</i> | Sim |
| TAGCROWD | Sim |
| TEXTALYZER | Não |
| WORDCOUNTER | Não |
| FERRAMENTA IDEAL | Sim |

Fonte: Elaborado pela autora

Quadro 22: Comparativo entre as ferramentas selecionadas – Critério #16

| FERRAMENTAS | Critério #16 Visualização Gráfica dos relacionamentos entre palavras |
|---------------------|---|
| KH CODER | Sim |
| NLTK | Não instalado e nem testado |
| PyPLN | Não instalado e nem testado |
| SOBEK <i>MINING</i> | Sim |
| TAGCROWD | Não |
| TEXTALYZER | Não |
| WORDCOUNTER | Não |
| FERRAMENTA IDEAL | Sim |

Fonte: Elaborado pela autora

Quadro 23: Comparativo entre as ferramentas selecionadas – Critério #17

| FERRAMENTAS | Critério #17 Existe Integração com outras ferramentas |
|---------------------|--|
| KH CODER | Não |
| NLTK | Não instalado e nem testado |
| PyPLN | Não instalado e nem testado |
| SOBEK <i>MINING</i> | Não |
| TAGCROWD | Não |
| TEXTALYZER | Não |
| WORDCOUNTER | Não |
| FERRAMENTA IDEAL | Sim |

Fonte: Elaborado pela autora

Quadro 24: Comparativo entre as ferramentas selecionadas – Critério #18

| FERRAMENTAS | Critério #18 Conhecimento prévio de programação computacional |
|---------------------|--|
| KH CODER | Não |
| NLTK | Sim |
| PyPLN | Sim |
| SOBEK <i>MINING</i> | Não |
| TAGCROWD | Não |
| TEXTALYZER | Não |
| WORDCOUNTER | Não |
| FERRAMENTA IDEAL | Não |

Fonte: Elaborado pela autora

4.2 Processamento das Mensagens Textuais por Categoria

A ferramenta SOBEK *Mining* foi selecionada para processar as mensagens textuais do Canal de Atendimento. Apesar de ter sido desenvolvido com fins pedagógicos, ela foi considerada viável para processar as mensagens textuais do Canal de Atendimento.

Como já visto anteriormente, o SOBEK contém uma interface gráfica para apoio ao usuário. Ele extrai as informações de um texto e as apresenta como um grafo, permitindo representar o texto de forma gráfica e auxiliando o usuário a visualizar os principais termos do texto e suas relações. Epstein e Reategui (2015, p. 5) dizem que “a simplicidade para realizar mineração de um texto é uma das características mais importantes do *software*”. Em seu trabalho, (EPSTEIN, REATEGUI, 2015, p. 5 a 7) explicam o funcionamento da ferramenta informando que o grafo resultante da mineração de texto apresenta os termos considerados mais relevantes de acordo com a frequência de ocorrência deles, apresentando as conexões entre os termos. Estas conexões representam relações existentes no texto e podem indicar efeito de causa e consequência, relações temporais ou mesmo termos relacionados pelo seu significado. Outra importante característica presente na representação das informações extraídas do texto é a capacidade de mostrar ao usuário o número de ocorrências de cada termo e as frases do texto onde este termo aparece. Esta característica tem por objetivo permitir que o usuário compreenda melhor as informações referentes a termos específicos, tais como frases relacionadas a esse termo, seu significado, qual seu papel no texto e quão relevante é frente aos demais termos apresentados.

O algoritmo utilizado no SOBEK pode ser dividido em três estágios:

1º estágio: consiste em identificar os termos mais relevantes do texto e organizá-los. Para o minerador de texto SOBEK, um termo será tão importante quanto o número de vezes que ele aparecer no texto.

O primeiro passo para a identificação dos termos relevantes do texto é separar ele em palavras. O SOBEK utiliza espaços em branco, pontuação e sinais de marcação para isso. As palavras são então mapeadas em termos, que podem ser termos simples (termos que contém apenas uma palavra) ou termos compostos (termos que contém duas ou mais palavras). Esse mapeamento de palavras em termos é realizado através de um processo estatístico, que verifica a frequência com que cada palavra é encontrada no texto. Quando um conjunto de palavras aparece constantemente em sequência, é possível que a ideia associada ao conjunto de palavras não possa ser identificada por uma série de termos simples, sendo assim, criado

um termo composto. Durante este processo, o minerador utiliza a lista de *stopwords*. Após identificar todos os termos possíveis de serem extraídos do texto, um processo de *stemming* é utilizado para remover redundância e termos com mesmo significado conjugados em tempos verbais diferentes ou no plural.

2º estágio: consiste na identificação dos relacionamentos entre os termos. Um relacionamento entre termos é formado quando eles estão próximos um ao outro no texto. Isso pode representar diversos tipos de informação sobre estes dois termos, tais como relação de causa e consequência, uma relação temporal ou mesmo que os termos têm significados relativos um ao outro. A análise do texto relaciona dois termos quando estes estão distantes não mais que 5 termos um do outro e quando não há um ponto final entre eles. Para reduzir o número de relações que um termo pode ter e apresentar apenas as relações mais significativas, um número máximo de 7 relações é permitido para cada termo. Estes valores de 5 termos e 7 relações são pré-estabelecidos pelo minerador SOBEK, não permitindo a alteração pelo usuário. Esses valores foram encontrados com base em avaliações de usuários e análise de dados, onde uma distância maior do que 5 termos, costuma apresentar relações de pouca importância entre termos e mais de 7 relações para cada termo produz grafos com um número demasiadamente alto de relações, podendo até mesmo produzir grafos onde todos os termos se relacionam entre si.

3º estágio: consiste na construção do grafo utilizando as informações extraídas do texto. Neste grafo, os termos são representados como vértices (nodos) e as relações entre eles são representadas como conexões (arestas). Para melhorar a visualização do grafo, cada nodo tem um tamanho diferente, baseado na frequência do termo que ele representa, quanto maior for o nodo, maior a frequência de ocorrência desse termo no texto. A frequência de um nodo também influenciará no seu número de ligações, tornando os nodos com maior frequência centrais no grafo, sendo estes mais conectados e destacados dos demais.

Tendo uma ideia de como funciona a ferramenta SOBEK, ela foi utilizada para fazer a mineração das mensagens textuais da seção “Qual a sua Dúvida?”. Ela foi configurada de uma forma geral para extração de 15 termos e sem especificar frequência mínima das palavras. Para melhor aproveitamento e entendimento dos grafos gerados pela ferramenta, foram analisadas somente as categorias que possuíam mais de 400 mensagens cadastradas, conforme Quadro 25.

Quadro 25: As categorias e suas respectivas quantidades de mensagens

| Categorias | Quantidade de mensagens |
|---------------------------------|--------------------------------|
| SECAF | 24.480 |
| Titulação / Qualificação | 2.976 |
| Normas e Legislação | 2.091 |
| Pagamento | 1.703 |
| SECAF - Greve | 1.519 |
| Ressarcimento de Saúde | 1.342 |
| Saúde | 1.244 |
| Outros | 817 |
| Cadastro Pessoal Funcional | 817 |
| Sistemas de RH | 650 |
| Aposentadoria | 530 |
| Avaliação de Desempenho | 505 |
| Políticas e Diretrizes de RH | 493 |
| Promoção / Progressão Funcional | 432 |

Fonte: Elaborado pela autora

Durante a mineração das mensagens, dois problemas foram encontrados que merecem ser relatados:

- a) A conversão de alguns arquivos: o IBGEANDO gera relatórios contendo as mensagens da seção “Qual a sua dúvida?” no formato “.xls” (extensão de arquivo do programa Excel da Microsoft Office) e a ferramenta SOBEK processa arquivos texto (que é uma sequência de linhas e geralmente contém pouca formatação) no formato “.txt”. Alguns problemas foram identificados na conversão dos arquivos do formato “.xls” para o formato “.txt”. Inicialmente, ao visualizar o arquivo convertido no formato “.txt”, não parecia haver problemas, o arquivo texto havia sido gerado com o tipo de codificação de caracteres ANSI. No entanto, ao carregar alguns arquivos na ferramenta SOBEK, o texto aparecia com caracteres especiais ao invés da acentuação correta das palavras em português. A solução encontrada foi realizar a conversão do arquivo no formato “.txt” com a codificação de caracteres UTF-8.
- b) O tamanho do arquivo convertido para ser processado na ferramenta: todas as categorias analisadas possuem menos de 3.000 mensagens, com exceção da categoria SECAF, que possui em torno de 8 vezes mais mensagens. Como a categoria SECAF se refere a um sistema que controla o acesso e a frequência dos servidores no trabalho, gerando débitos de horas não trabalhadas a serem descontados diretamente na folha de pagamento, esta acaba sendo a categoria com mais dúvidas e problemas levantados pelos servidores. A quantidade de mensagens

influencia diretamente no tamanho do arquivo gerado no formato “.txt”. Enquanto o arquivo da categoria SECAF possui 10MB de tamanho, todas as outras categorias possuem arquivos de tamanho entre 150kb a 1MB. Ao carregar estes arquivos na ferramenta SOBEK *Mining*, o tempo de processamento para gerar o grafo da maioria das categorias levou de segundos a nove minutos no máximo. Já para a categoria SECAF, só para importar o arquivo “.txt” na ferramenta, levou 45 minutos e, depois de várias tentativas de processamento, com duração de mais de 24 horas, foi descartada a possibilidade de gerar o grafo para esta categoria na ferramenta SOBEK na versão com interface gráfica. No entanto, em contato direto com os pesquisadores que desenvolveram a ferramenta, eles informaram que existe uma versão da ferramenta para uso em terminais de comando, capaz de processar uma quantidade muito maior de dados que a versão com interface. Mas esta versão não foi obtida para ser testada. Sendo assim, a categoria SECAF não foi analisada neste trabalho.

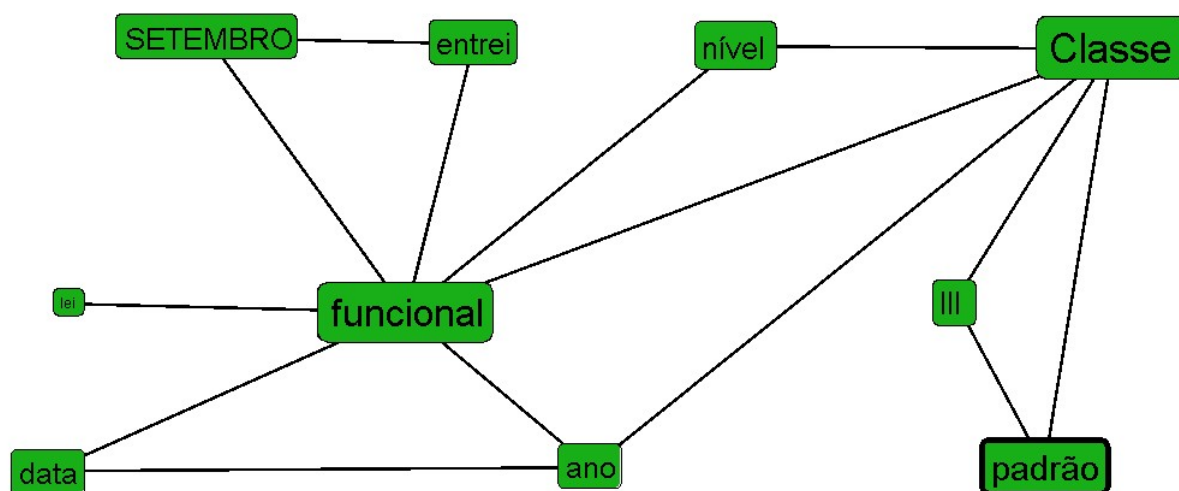
Para fazer o processamento na ferramenta, além das *stopwords* comuns da língua portuguesa, foram inseridas *stopwords* adicionais gerais no contexto das mensagens do IBGE, como por exemplo: IBGE, gostaria, saber, boa, tarde, bom, dia, atentamente, obrigado, Sr, Sra, servidor, caso, servidores, Fundação, gostaria de saber. Para algumas categorias específicas, ainda foi necessário adicionar outras *stopwords*. Isto foi necessário para obter um grafo mais coerente e informativo.

O resultado da mineração das mensagens textuais por Categoria na ferramenta SOBEK com as percepções desta pesquisadora em relação ao grafo gerado pela ferramenta é apresentado nos subcapítulos a seguir.

4.2.1 Categoria: Promoção / Progressão Funcional

Para esta categoria foi necessário acrescentar as seguintes *stopwords* adicionais: Promoção, Progressão, cargo. A Figura 17 e a Tabela 1 mostram respectivamente, o grafo gerado e a tabela de frequência das palavras nesta categoria.

Figura 17: Grafo - Categoria Promoção / Progressão Funcional – 15 termos



Fonte: SOBEK Mining

Tabela 1: Frequência de Palavras – Categoria Promoção / Progressão Funcional – 15 termos

| Palavras | Frequência | Percentual em relação às mensagens totais do período analisado |
|-----------|------------|--|
| classe | 237 | 54,86% |
| Funcional | 218 | 50,46% |
| Padrão | 138 | 31,94% |
| nível | 108 | 25,00% |
| III | 101 | 23,38% |
| Setembro | 95 | 21,99% |
| ano | 92 | 21,30% |
| Entrei | 88 | 20,37% |
| data | 87 | 20,14% |
| lei | 79 | 18,29% |

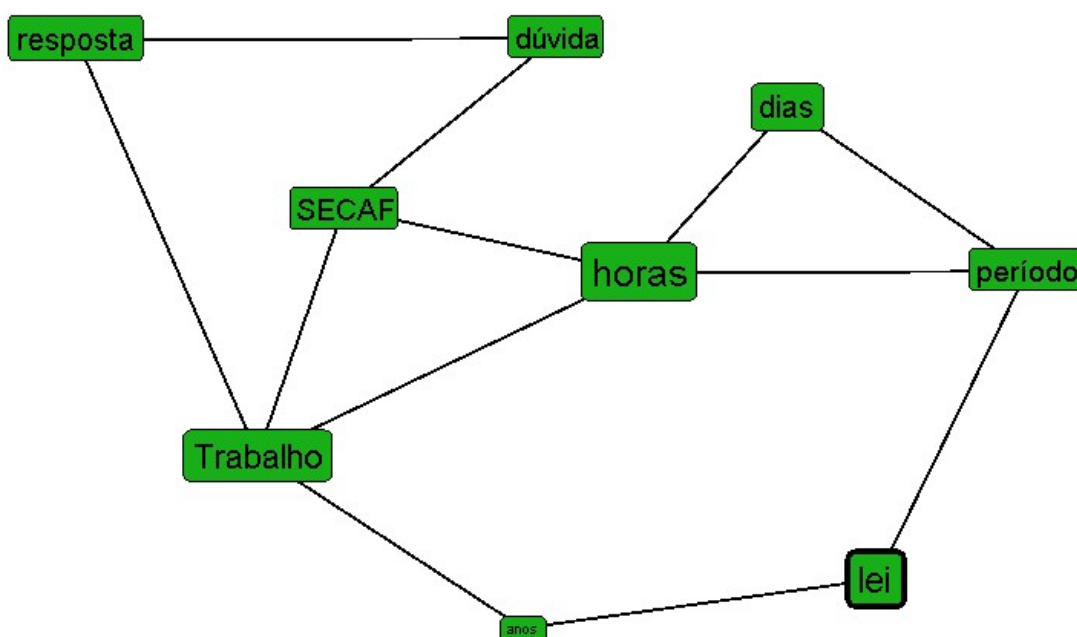
Fonte: Elaborado pela autora

Com o grafo resultante (Figura 17), identifica-se claramente que a maior dúvida dos servidores está na tabela de Cargos e Salários do IBGE quando o servidor está no Padrão III de uma determinada Classe. Outra questão é que o mês de setembro foi a entrada de novos servidores concursados no IBGE, e estes parecem ser os que têm mais dúvidas, principalmente em relação a leis e datas para promoção e progressão.

4.2.2 Categoria: Política e Diretrizes

Para esta categoria foi necessário acrescentar as seguintes *stopwords* adicionais: ART, CRH. A Figura 18 e a Tabela 2 mostram respectivamente, o grafo gerado e a tabela de frequência das palavras nesta categoria.

Figura 18: Grafo - Categoria Política e Diretrizes – 15 termos



Fonte: SOBEK Mining

Tabela 2: Frequência de Palavras – Categoria Política e Diretrizes – 15 termos

| Palavras | Frequência | Percentual em relação às mensagens totais do período analisado |
|----------|------------|--|
| horas | 224 | 45,44% |
| lei | 162 | 32,86% |
| trabalho | 148 | 30,02% |
| dias | 113 | 22,92% |
| resposta | 100 | 20,28% |
| período | 92 | 18,66% |
| SECAF | 89 | 18,05% |
| dúvida | 82 | 16,63% |
| anos | 80 | 16,23% |

Fonte: Elaborado pela autora

Analisando de uma forma geral o grafo da Figura 18, verifica-se que a preocupação dos servidores é em relação a horas de trabalho e a legislação que legitima o assunto. Existe muita dúvida em relação ao sistema SECAF, que é o sistema de controle de acesso e frequência dos servidores no trabalho. Cabe ressaltar que apesar de existir uma categoria específica para o SECAF, este mesmo assunto se repete aqui. Isso acontece porque esta categoria é de responsabilidade do Coordenador de RH, então parece que as dúvidas que não são sanadas na categoria específica, se repetem aqui para um posicionamento de uma autoridade do IBGE ou para que o próprio Coordenador agilize a solução dos problemas e conheça as dúvidas relatadas pelos servidores.

4.2.3 Categoria: Avaliação de Desempenho

Para esta categoria não houve a necessidade de se adicionar novas *stopwords*. A Figura 19 e a Tabela 3 mostram respectivamente, o grafo gerado e a tabela de frequência das palavras nesta categoria.

Figura 19: Grafo - Categoria Avaliação de Desempenho – 15 termos

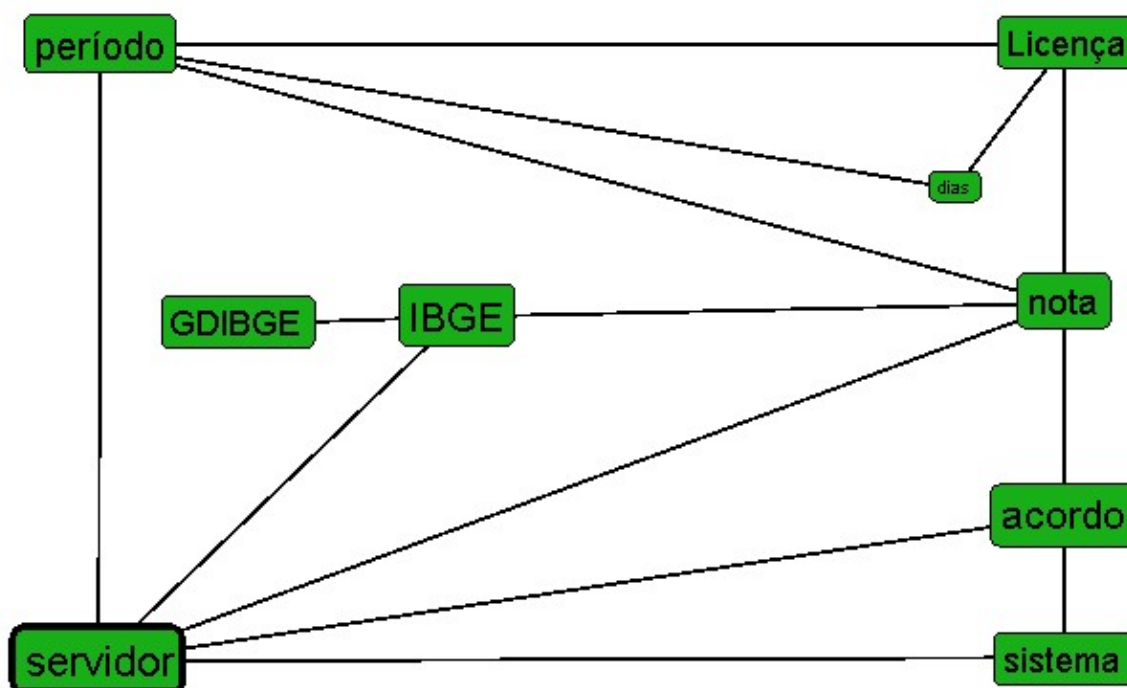


Tabela 3: Frequência de Palavras – Categoria Avaliação de Desempenho – 15 termos

| 126 | Frequência | Percentual em relação às mensagens totais do período analisado |
|----------|------------|--|
| servidor | 180 | 35,64% |
| IBGE | 166 | 32,87% |
| acordo | 156 | 30,89% |
| período | 150 | 29,70% |
| nota | 121 | 23,96% |
| licença | 103 | 20,40% |
| sistema | 101 | 20,00% |
| GDIBGE | 92 | 18,22% |
| dias | 91 | 18,02% |

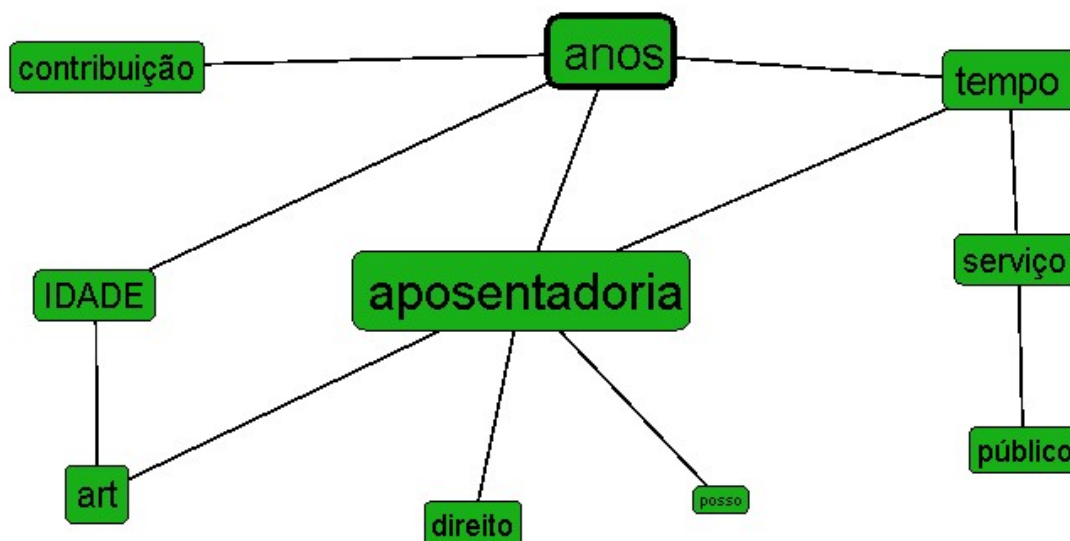
Fonte: Elaborado pela autora

Apesar da palavra IBGE ser uma *stopword* para todas as categorias, ela é exibida no grafo (Figura 19). Acredita-se que isso ocorreu devido às palavras estarem muito vinculadas à sigla IBGE, e isso faz sentido uma vez que a categoria avaliação de desempenho é específica da instituição IBGE, tendo uma gratificação chamada GDIBGE que se divide numa pontuação específica do servidor e outra pontuação da instituição. Analisando o grafo, nota-se que as maiores dúvidas parecem se referir ao acordo de desempenho do servidor no sistema, aos períodos de lançamentos das notas e o que as notas e períodos de acordo interferem ou não nas licenças.

4.2.4 Categoria: Aposentadoria

Para esta categoria houve a necessidade de se adicionar novas *stopwords* adicionais: resposta, aposentar, aposentadoria. A Figura 20 e a Tabela 4 mostram respectivamente, o grafo gerado e a tabela de frequência das palavras nesta categoria.

Figura 20: Grafo - Categoria Aposentadoria – 15 termos



Fonte: SOBEM Mining

Tabela 4: Frequência de Palavras – Categoria Aposentadoria – 15 termos

| Palavras | Frequência | Percentual em relação às mensagens totais do período analisado |
|---------------|------------|--|
| aposentadoria | 588 | 110,94% |
| anos | 476 | 89,81% |
| tempo | 299 | 56,42% |
| art | 226 | 42,64% |
| serviço | 174 | 32,83% |
| idade | 163 | 30,75% |
| contribuição | 159 | 30,00% |
| direito | 120 | 22,64% |
| público | 119 | 22,45% |
| posso | 108 | 20,38% |

Fonte: Elaborado pela autora

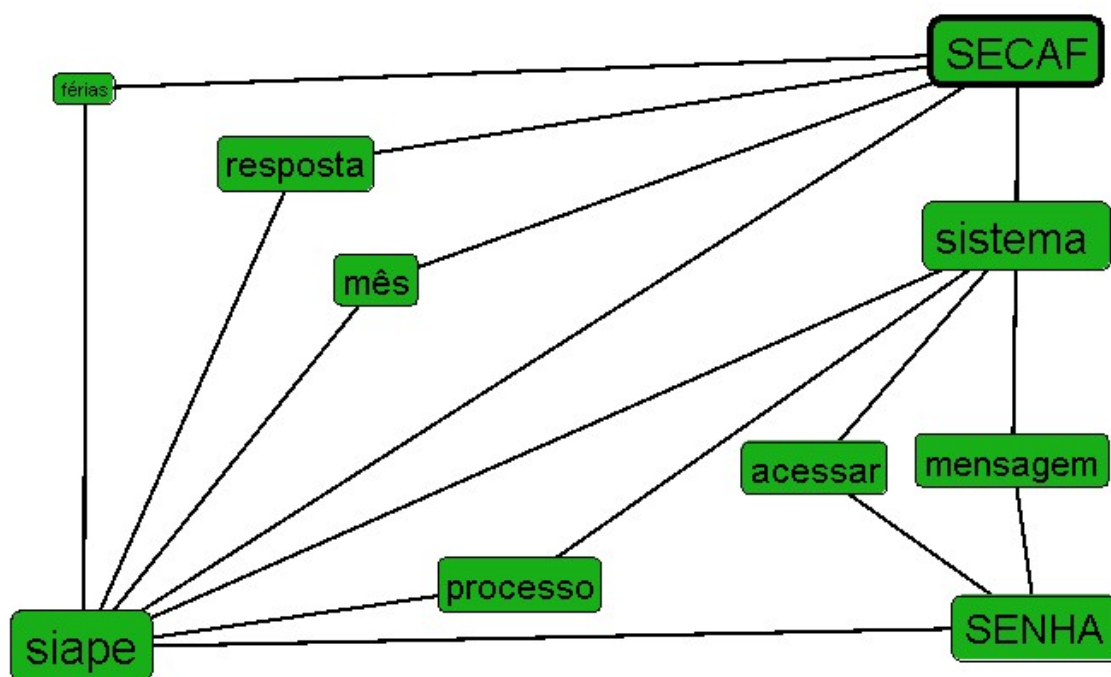
Nesta categoria acontece a mesma situação que na categoria Avaliação de Desempenho. Apesar da palavra aposentadoria ser uma *stopword*, para esta categoria, ela é exibida no grafo (Figura 20). Acredita-se que isso ocorreu devido às palavras estarem muito vinculadas à palavra Aposentadoria que é o assunto central desta categoria. Observa-se que a maior dúvida está em relação aos anos para se aposentar, seja em relação ao tempo de serviço

ou de contribuição ou da idade do servidor. Como aparece o termo “art” que significa artigo de uma lei ou norma, pressupõe-se a preocupação dos servidores em relação a qual artigo os mesmos podem se aposentar. O verbo “posso” e a palavra “direito” indicam que os servidores procuram saber sobre a sua situação particular de quando podem ou tem direito a se aposentar.

4.2.5 Categoria: Sistemas de RH

Para esta categoria houve a necessidade de se adicionar novas *stopwords* adicionais: fazer. A Figura 21 e a Tabela 4 mostram respectivamente, o grafo gerado e a tabela de frequência das palavras nesta categoria.

Figura 21: Grafo - Categoria Sistemas de RH – 15 termos



Fonte: SOBEK Mining

Tabela 5: Frequência de Palavras – Categoria Sistemas de RH – 15 termos

| Palavras | Frequência | Percentual em relação às mensagens totais do período analisado |
|-----------------|-------------------|---|
| siape | 259 | 39,85% |
| sistema | 221 | 34,00% |
| SECAF | 210 | 32,31% |
| senha | 183 | 28,15% |
| mensagem | 111 | 17,08% |
| fazer | 109 | 16,77% |
| processo | 108 | 16,62% |
| resposta | 100 | 15,38% |
| acessar | 99 | 15,23% |
| mês | 97 | 14,92% |
| férias | 95 | 14,62% |

Fonte: Elaborado pela autora

Analisando este grafo (Figura 21), nota-se que a palavra “siape” é uma das mais frequentes, isso pode significar a matrícula do servidor que se chama siape ou o sistema governamental que o IBGE usa que se chama SIAPE, assim, nesta categoria os servidores podem estar relatando os seus problemas com o sistema SIAPE ou informando a sua matrícula para se identificar. Sistema também é uma palavra muito frequente, o que era de se esperar por ser uma categoria que trata dos sistemas de RH, mas, cabe ressaltar, que aqui também aparece o SECAF como palavra frequente, o que reforça que existem muitas dúvidas ou problemas relacionados a este sistema. O SECAF está muito relacionado com as palavras siape, férias, mês e resposta, o que pode indicar que os servidores querem resposta, e que pode haver problemas de integração entre o sistema SIAPE e o SECAF em relação ao evento férias. Outra indicação é que parece haver problemas de acesso e de mensagens de senha nos sistemas.

Uma observação importante de se fazer é que para avaliar as mensagens na ferramenta SOBEK, existe uma infinidade de possibilidades de configuração para a geração do grafo, por exemplo, no caso da Categoria de “Sistemas de RH”, o grafo gerado com 15 termos (Figura 21) não mostra os nomes dos sistemas usados no IBGE, somente a palavra sistema e o SECAF. Então, pode-se pensar que aumentando a quantidade de termos a serem apresentados, os nomes dos sistemas mais citados pelos servidores nas mensagens devem ser exibidos no grafo. Assim, foi alterada a configuração da ferramenta para obter 30 termos de forma a avaliar como o grafo se apresenta. A Figura 22 e a Tabela 6 mostram o grafo e a tabela de frequência gerados para 30 termos. Importante notar que, na extração de 15 termos a palavra “fazer” teve uma frequência alta, mas ficou solta no grafo, sem ter nenhuma relação com outra palavra e, por isso, foi excluída do grafo. No entanto, na extração de 30 termos ela aparece associada com a palavra “devo”, assim foi mantida a sua exibição no grafo. A vinculação dos verbos “devo” com “fazer”, parece indicar que os servidores precisam de orientação para usar os sistemas.

Analisando o grafo com 30 termos (Figura 22), é possível identificar os sistemas com mais dúvidas ou mais problemas, indicando inclusive a funcionalidade. Os sistemas identificados que aparecem com mais frequência nas mensagens são:

1º SIAPE com 39,85% (aqui pode haver algum erro uma vez que siape também pode ser a matrícula do servidor, mas poderia ser pensado também numa forma de remover a palavra siape que indicasse matrícula e não sistema para ser mais específico na determinação da frequência)

2º SECAF com 32,31%

3º SDA com 12,15%

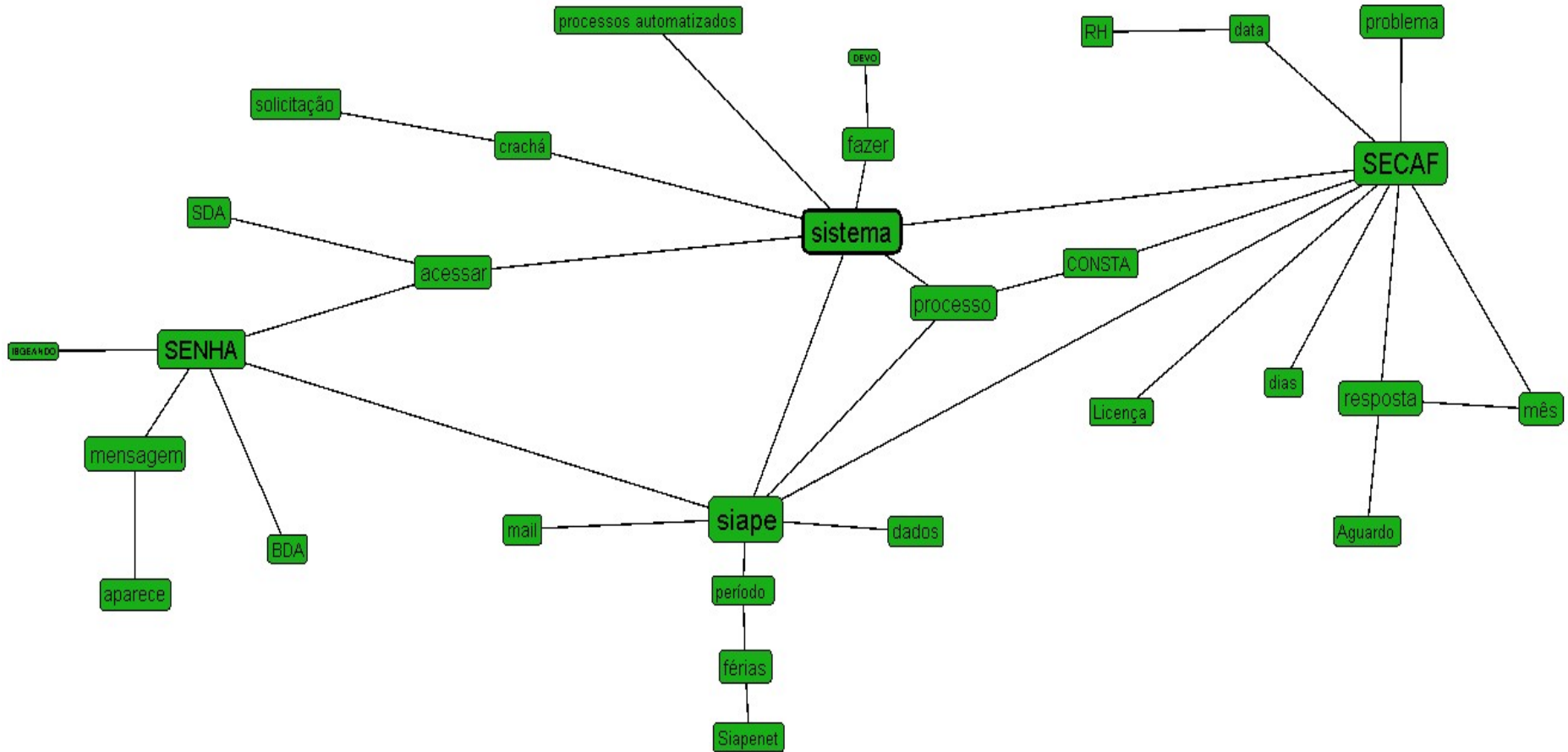
4º Crachá com 10,31%

5º Siapenet com 10,15%

6º Processos Automatizados com 9,23%

7º BDA e IBGEANDO com 8,77%

Figura 22: Grafo - Categoria Sistemas de RH – 30 termos



Fonte: SOBEM Mining

Tabela 6: Frequência de Palavras – Categoria Sistemas de RH – 30 termos
(continua)

| Palavras | Frequência | Percentual em relação às mensagens totais do período analisado |
|-----------------|-------------------|---|
| siape | 259 | 39,85% |
| sistema | 221 | 34,00% |
| SECAF | 210 | 32,31% |
| senha | 183 | 28,15% |
| mensagem | 111 | 17,08% |
| Fazer | 109 | 16,77% |
| processo | 108 | 16,62% |
| resposta | 100 | 15,38% |
| acessar | 99 | 15,23% |
| mês | 97 | 14,92% |
| férias | 95 | 14,62% |
| Problema | 93 | 14,31% |
| Mail | 80 | 12,31% |
| Dados | 80 | 12,31% |
| SDA | 79 | 12,15% |
| RH | 72 | 11,08% |
| Solicitação | 71 | 10,92% |
| Aparece | 71 | 10,92% |
| Consta | 69 | 10,62% |
| Data | 67 | 10,31% |
| Crachá | 67 | 10,31% |
| Siapenet | 66 | 10,15% |

Tabela 6: Frequência de Palavras – Categoria Sistemas de RH – 30 termos
(conclusão)

| Palavras | Frequência | Percentual em relação às mensagens totais do período analisado |
|-------------------------|-------------------|---|
| Licença | 66 | 10,15% |
| Aguardo | 66 | 10,15% |
| Dias | 65 | 10,00% |
| Processos automatizados | 60 | 9,23% |
| Período | 60 | 9,23% |
| BDA | 60 | 9,23% |
| IBGEANDO | 57 | 8,77% |
| Devo | 57 | 8,77% |

Fonte: Elaborado pela autora

O IBGEANDO apesar de ser uma *intranet*, para acessá-lo você deve entrar com uma senha que é a mesma senha do sistema SECAF, e no grafo mostra exatamente esta vinculação do IBGEANDO com a palavra senha.

O BDA era o sistema administrativo interno do IBGE que foi substituído pelo sistema SDA. Esta substituição ocorreu próximo ao período de corte das mensagens do IBGEANDO para esta pesquisa, por isso, aparece claramente que os servidores falam sobre o acesso do SDA enquanto o BDA aparece associado com a palavra senha. O grafo parece indicar que há problemas de senha e de acesso aos sistemas.

Os processos automatizados são citados nas mensagens, mas no grafo não é possível identificar entre as palavras mais frequentes com 30 termos, qual seria o problema.

Já o sistema SIAPENET parece ter problemas de integração em relação ao período de férias com o SIAPE. E, para o sistema de Crachá parece que o problema é a solicitação do crachá.

O sistema SECAF, campeão disparado nas mensagens dos servidores, parece ter realmente problemas, inclusive sendo vinculado à palavra “problema” e a vinculação das palavras “aguardo” e “resposta” pode indicar que os servidores não estão satisfeitos com as

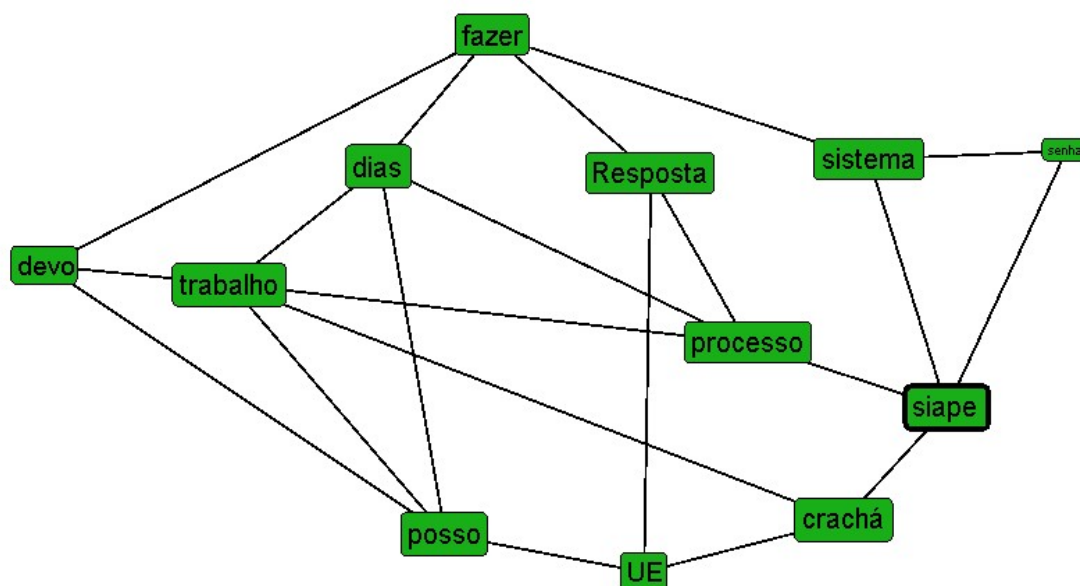
respostas ou não recebem respostas aos seus questionamentos. O SECAF também está associado com datas, dias e mês, além de haver algum problema ao evento “licença”.

A extração desta categoria em 30 termos, serve para mostrar que se pode ter uma análise mais detalhada das mensagens postadas. Também pode-se definir períodos menores de postagens de mensagens para avaliar o que nesse curto período de tempo estava sendo falado ou reportado pelos servidores. Isso mostra que, dependendo da análise desejada, deve-se alterar os parâmetros de configuração da ferramenta.

4.2.6 Categoria: Outros

Para esta categoria não houve a necessidade de se adicionar novas *stopwords*. A Figura 23 e a Tabela 7 mostram respectivamente, o grafo gerado e a tabela de frequência das palavras nesta categoria.

Figura 23: Grafo - Categoria Outros – 15 termos



Fonte: SOBEK Mining

Tabela 7: Frequência de Palavras – Categoria Outros – 15 termos

| Palavras | Frequência | Percentual em relação às mensagens totais do período analisado |
|-----------------|-------------------|---|
| siape | 94 | 11,51% |
| trabalho | 93 | 11,38% |
| dias | 92 | 11,26% |
| fazer | 87 | 10,65% |
| sistema | 86 | 10,53% |
| crachá | 84 | 10,28% |
| processo | 83 | 10,16% |
| resposta | 77 | 9,42% |
| posso | 74 | 9,06% |
| devo | 63 | 7,71% |
| U.E. | 62 | 7,59% |
| senha | 61 | 7,47% |

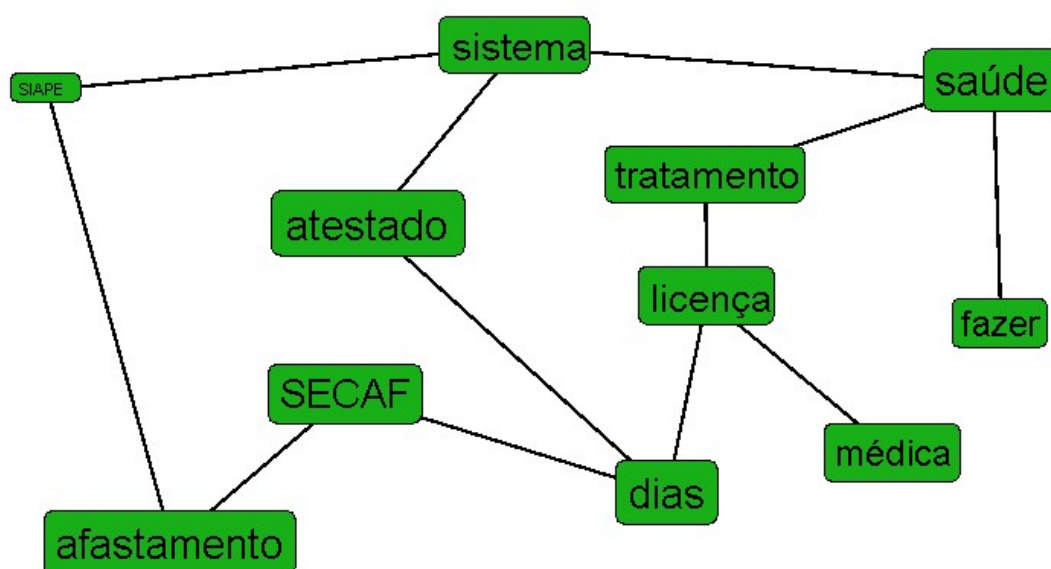
Fonte: Elaborado pela autora

O grafo gerado tem várias linhas que se cruzam, dificultando um pouco a visualização, não foi possível melhorar esta visualização apesar de diversas tentativas. A conclusão que se pode chegar é que esta “confusão” de conexões se deve também à miscelânea de assuntos tratados nesta categoria, como a própria denominação diz “Outros”. Esta categoria foi criada visando dar maior liberdade aos servidores para postar mensagens que não se enquadram nas demais categorias definidas, como por exemplo, postar mensagens que não necessariamente seriam da área de Recursos Humanos. Mas o que se nota no Grafo, é que aqui parece que se postam mensagens duplicadas, que já foram reportadas nas demais categorias, tendo em vista que as palavras mais frequentes são as mesmas que aparecem nas demais categorias.

4.2.7 Categoria: Saúde

Para esta categoria não houve a necessidade de se adicionar novas *stopwords*. A Figura 24 e a Tabela 8 mostram respectivamente, o grafo gerado e a tabela de frequência das palavras nesta categoria.

Figura 24: Grafo - Categoria Saúde – 15 termos



Fonte: SOBEK Mining

Tabela 8: Frequência de Palavras – Categoria Saúde – 15 termos
(continua)

| Palavras | Frequência | Percentual em relação às mensagens totais do período analisado |
|-------------|------------|--|
| atestado | 547 | 43,97% |
| afastamento | 498 | 40,03% |
| dias | 472 | 37,94% |
| saúde | 445 | 35,77% |

Tabela 8: Frequência de Palavras – Categoria Saúde – 15 termos
(conclusão)

| Palavras | Frequência | Percentual em relação às mensagens totais do período analisado |
|------------|------------|--|
| SECAF | 441 | 35,45% |
| sistema | 380 | 30,55% |
| licença | 354 | 28,46% |
| tratamento | 334 | 26,85% |
| médica | 315 | 25,32% |
| fazer | 242 | 19,45% |
| siape | 226 | 18,17% |

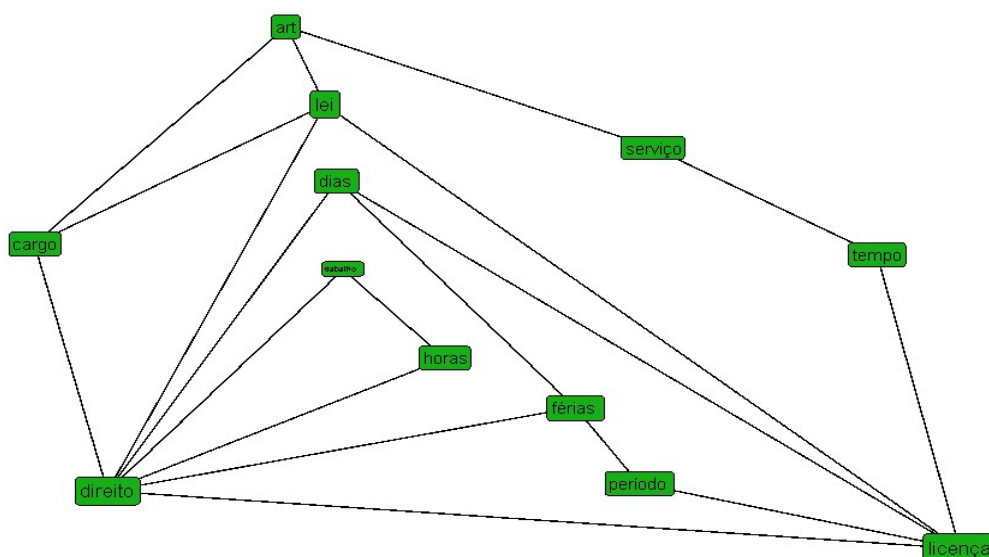
Fonte: Elaborado pela autora

A palavra mais frequente desta categoria é “atestado” e está vinculado às palavras “sistema” e “dias”. O IBGE possui um sistema próprio de saúde onde os servidores anexam o seu atestado médico digitalizado para análise da equipe de saúde do IBGE para concessão de afastamento ou licença. Então deve haver muitas dúvidas ou problemas relacionados ao lançamento deste atestado no sistema e em relação aos dias de direito de licença médica. Mais uma vez o sistema SECAF aparece com frequência alta, e aqui parece indicar que há problemas de registro de afastamento e de dias relacionados ao atestado médico. Importante frisar que o SECAF faz o controle das horas trabalhadas, então, se não há o afastamento médico, o servidor fica com uma notificação de falta no sistema que se não for compensada será debitado o valor das horas não trabalhadas no seu contracheque. Isto justifica uma frequência alta.

4.2.8 Categoria: Normas e Legislação

Para esta categoria não houve a necessidade de se adicionar novas *stopwords*. A Figura 25 e a Tabela 9 mostram respectivamente, o grafo gerado e a tabela de frequência das palavras nesta categoria.

Figura 25: Grafo - Categoria Normas e Legislação – 15 termos



Fonte: SOBEK Mining

Tabela 9: Frequência de Palavras – Categoria Normas e Legislação – 15 termos

| Palavras | Frequência | Percentual em relação às mensagens totais do período analisado |
|----------|------------|--|
| direito | 661 | 31,61% |
| licença | 617 | 29,51% |
| lei | 589 | 28,17% |
| dias | 450 | 21,52% |
| férias | 449 | 21,47% |
| cargo | 385 | 18,41% |
| art | 362 | 17,31% |
| horas | 360 | 17,22% |
| período | 355 | 16,98% |
| serviço | 353 | 16,88% |
| tempo | 350 | 16,74% |
| trabalho | 346 | 16,55% |

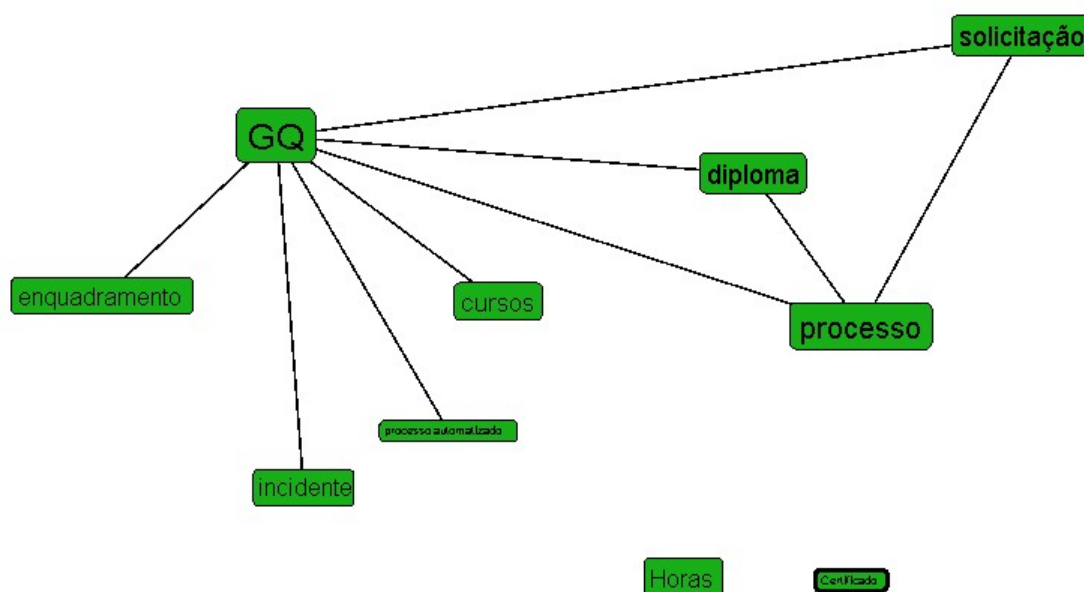
Fonte: Elaborado pela autora

Nesta categoria, as mensagens postadas falam mais sobre os direitos que os servidores têm em relação a seu cargo, a licenças, férias, horas de trabalho e tempo de serviço, levando em consideração às leis que regem estes assuntos.

4.2.9 Categoria: Titulação / Qualificação

Para esta categoria não houve a necessidade de se adicionar novas *stopwords*. A Figura 26 e a Tabela 10 mostram respectivamente, o grafo gerado e a tabela de frequência das palavras nesta categoria.

Figura 26: Grafo - Categoria Titulação/Qualificação – 15 termos



Fonte: SOBEK Mining

Tabela 10: Frequência de Palavras – Categoria Titulação/Qualificação – 15 termos

| Palavras | Frequência | Percentual em relação às mensagens totais do período analisado |
|-----------------------|-------------------|---|
| GQ | 2547 | 85,58% |
| processo | 1626 | 54,64% |
| diploma | 974 | 32,73% |
| solicitação | 966 | 32,46% |
| curios | 826 | 27,76% |
| enquadramento | 795 | 26,71% |
| horas | 721 | 24,23% |
| incidente | 636 | 21,37% |
| processo automatizado | 577 | 19,39% |
| certificado | 568 | 19,09% |

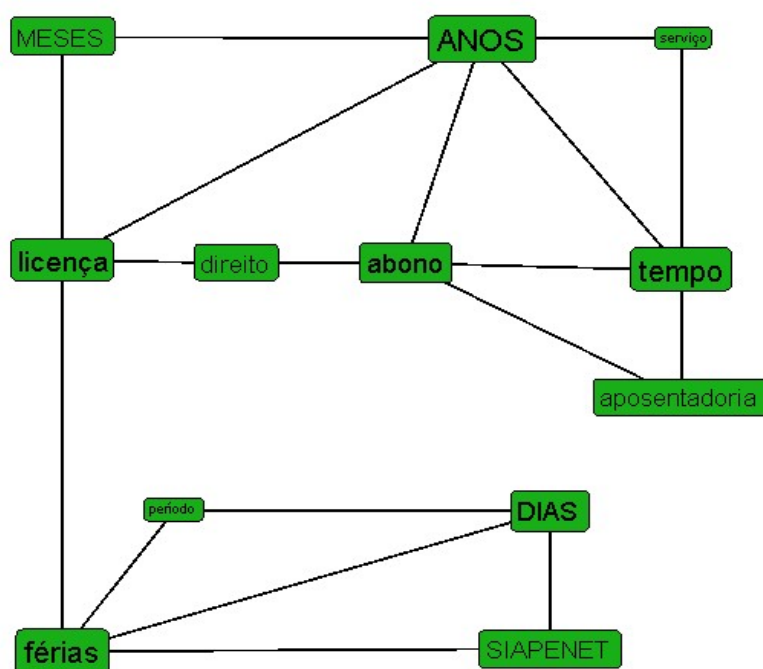
Fonte: Elaborado pela autora

Como mencionado no Quadro 1 que exibe o total de mensagens postadas por categoria, esta categoria teve um aumento enorme de mensagens quando foram implantados os processos automatizados de Gratificação de Qualificação (GQ) e isto pode ser observado na tabela de palavras frequentes desta categoria onde o termo GQ é mencionado mais de 85%. E as dúvidas mais frequentes relacionadas a este termo, são sobre o enquadramento, o incidente gerado nos processos automatizados, os cursos e diplomas aceitos para a qualificação e a solicitação do processo de GQ.

4.2.10 Categoria: Cadastro Pessoal / Funcional

Para esta categoria não houve a necessidade de se adicionar novas *stopwords*. A Figura 27 e a Tabela 11 mostram respectivamente, o grafo gerado e a tabela de frequência das palavras nesta categoria.

Figura 27: Grafo - Categoria Cadastro Pessoal/Funcional – 15 termos



Fonte: SOBEK Mining

Tabela 11: Frequência de Palavras – Categoria Cadastro Pessoal/Funcional – 15 termos

(continua)

| Palavras | Frequência | Percentual em relação às mensagens totais do período analisado |
|---------------|------------|--|
| anos | 301 | 36,84% |
| tempo | 275 | 33,66% |
| férias | 263 | 32,19% |
| licença | 224 | 27,42% |
| abono | 218 | 26,68% |
| dias | 195 | 23,87% |
| aposentadoria | 166 | 20,32% |
| meses | 159 | 19,46% |
| direito | 156 | 19,09% |

Tabela 11: Frequência de Palavras – Categoria Cadastro Pessoal/Funcional – 15 termos (conclusão)

| Palavras | Frequência | Percentual em relação às mensagens totais do período analisado |
|----------|------------|--|
| SIAPENET | 156 | 19,09% |
| período | 154 | 18,85% |
| serviço | 153 | 18,73% |

Fonte: Elaborado pela autora

Nesta categoria pode-se dividir em quatro assuntos mais frequentes:

1º Férias: de acordo com a Grafo, as dúvidas se referem ao período e dias de férias no sistema SIAPENET e o que as férias influenciam ou interferem nas licenças.

2º Licença: dúvidas sobre a duração de licenças em meses e o direito à mesma, assim como a ligação com as férias.

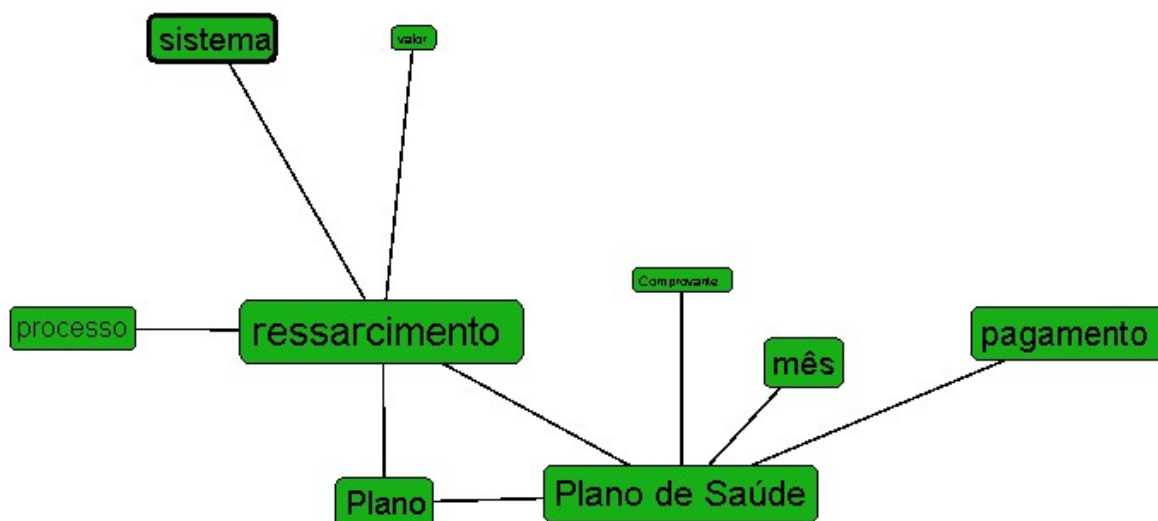
3º Abono: direito ao abono permanência e sua relação com a aposentadoria, assim como o tempo em anos para a concessão.

4º Aposentadoria: tempo de serviço e a ligação com o abono permanência. Provável que neste assunto surja também a questão da contagem de tempo de serviço para solicitar a aposentadoria.

4.2.11 Categoria: Ressarcimento de Saúde

Para esta categoria não houve a necessidade de se adicionar novas *stopwords*. A Figura 28 e a Tabela 12 mostram respectivamente, o grafo gerado e a tabela de frequência das palavras nesta categoria.

Figura 28: Grafo - Categoria Ressarcimento de Saúde – 15 termos



Fonte: SOBEM Mining

Tabela 12: Frequência de Palavras – Categoria Ressarcimento de Saúde – 15 termos

| Palavras | Frequência | Percentual em relação às mensagens totais do período analisado |
|----------------|------------|--|
| ressarcimento | 999 | 74,44% |
| plano de saúde | 754 | 56,18% |
| pagamento | 616 | 45,90% |
| plano | 517 | 38,52% |
| mês | 507 | 37,78% |
| sistema | 465 | 34,65% |
| processo | 320 | 23,85% |
| valor | 271 | 20,19% |
| comprovante | 269 | 20,04% |

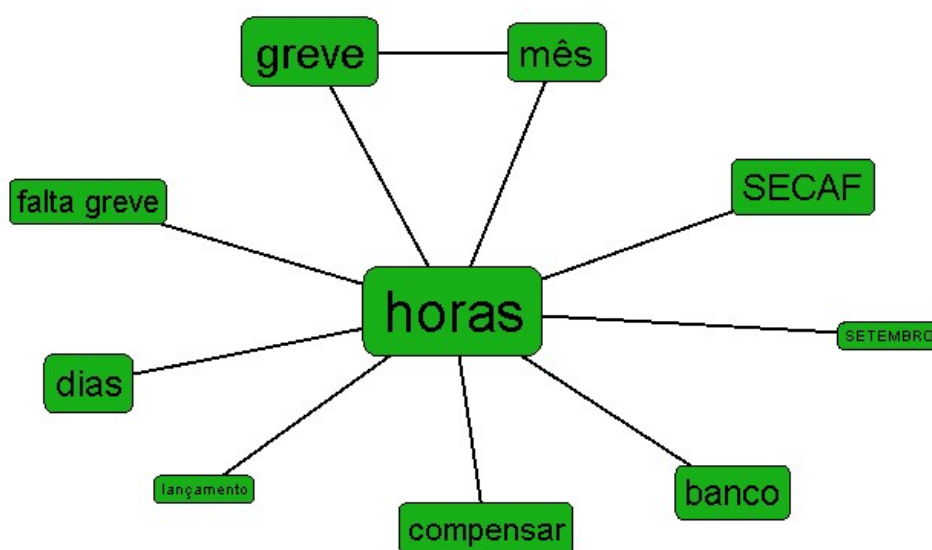
Fonte: Elaborado pela autora

Todo o grafo está vinculado à palavra “ressarcimento”, onde o assunto principal é o pagamento do plano de saúde que deve ter um comprovante para ser anexado ao sistema mensalmente. Este grafo reflete todo o funcionamento do processo do benefício de Assistência à Saúde ao Servidor onde há o ressarcimento mensal de um valor parcial ao servidor que paga um Plano de Saúde.

4.2.12 Categoria: SECAF Greve

Para esta categoria não houve a necessidade de se adicionar novas *stopwords*. A Figura 29 e a Tabela 13 mostram respectivamente, o grafo gerado e a tabela de frequência das palavras nesta categoria.

Figura 29: Grafo - Categoria SECAF Greve – 15 termos



Fonte: SOBEK Mining

Tabela 13: Frequência de Palavras – Categoria SECAF Greve – 15 termos

| Palavras | Frequência | Percentual em relação às mensagens totais do período analisado |
|-------------|------------|--|
| horas | 2480 | 163,27% |
| greve | 1234 | 81,24% |
| mês | 834 | 54,90% |
| SECAF | 736 | 48,45% |
| dias | 728 | 47,93% |
| banco | 637 | 41,94% |
| compensar | 434 | 28,57% |
| Falta Greve | 393 | 25,87% |
| setembro | 334 | 21,99% |
| lançamento | 329 | 21,66% |

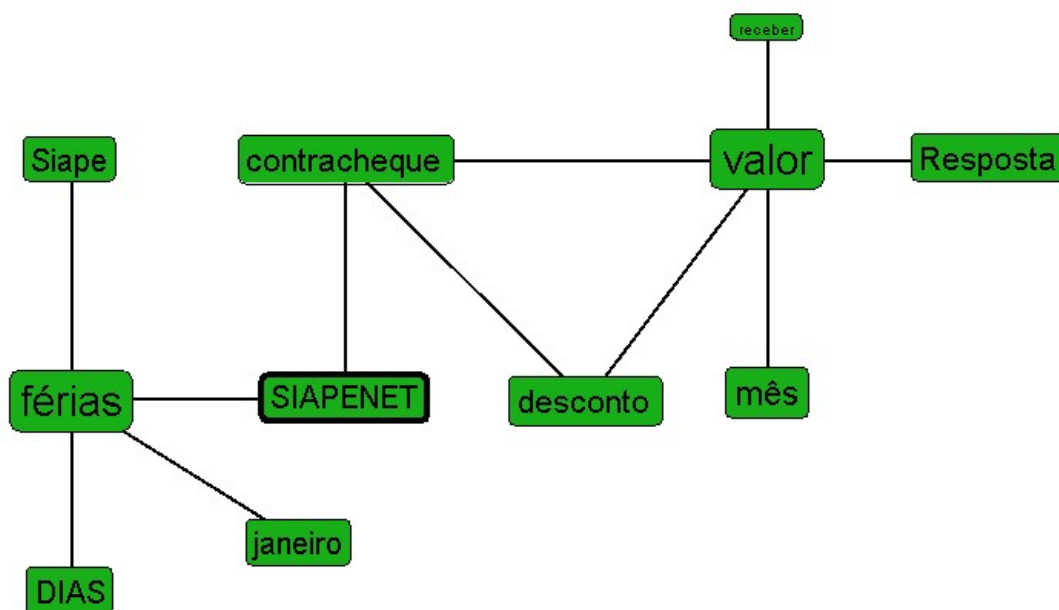
Fonte: Elaborado pela autora

A categoria SECAF Greve foi desmembrada da categoria SECAF para tratar somente das questões do SECAF referente ao assunto greve. Assim, nota-se que todo o grafo está vinculado à palavra “horas”, onde há dúvidas sobre o termo falta greve, banco de horas, compensação das horas e lançamento. Há um indicativo de que algo aconteceu no mês de setembro por ser um termo frequente.

4.2.13 Categoria: Pagamento

Para esta categoria houve a necessidade de se adicionar a seguinte *stopword* adicional: pagamento. A Figura 30 e a Tabela 14 mostram respectivamente, o grafo gerado e a tabela de frequência das palavras nesta categoria.

Figura 30: Grafo - Categoria Pagamento – 15 termos



Fonte: SOBEK Mining

Tabela 14: Frequência de Palavras – Categoria Pagamento – 15 termos

| Palavras | Frequência | Percentual em relação às mensagens totais do período analisado |
|--------------|------------|--|
| férias | 633 | 37,17% |
| valor | 616 | 36,17% |
| mês | 365 | 21,43% |
| desconto | 321 | 18,85% |
| contracheque | 312 | 18,32% |
| resposta | 307 | 18,03% |
| SIAPENET | 285 | 16,74% |
| janeiro | 250 | 14,68% |
| siape | 247 | 14,50% |
| dias | 245 | 14,39% |
| receber | 238 | 13,98% |

Fonte: Elaborado pela autora

Aqui o evento férias é o mais comentado e está relacionado com o mês de janeiro, aos dias e ao lançamento nos sistemas SIAPE e SIAPENET. Além disso, parece haver muitas dúvidas em relação aos descontos no contracheque e seus valores. O surgimento da palavra “resposta” como mais frequente também indica que os servidores esperam respostas, como se eles não estivessem sendo respondidos.

4.3 Validação dos Resultados

No capítulo anterior, os grafos gerados pela ferramenta SOBEK *Mining* foram analisados de forma pragmática de acordo com a percepção da pesquisadora que trabalhou na CRH durante 7 anos, respondendo aos questionamentos dos servidores na seção “Qual a sua Dúvida?” por 5 anos, entre outras atividades. Assim, pode não ter havido uma imparcialidade nesta análise.

Por isso, houve a necessidade da validação dos resultados apresentados pela ferramenta de Mineração de Textos para o grupo de servidores que trabalham na CRH. A proposta foi aplicar um questionário de análise dos grafos resultantes da mineração das mensagens textuais, com o intuito de verificar se os servidores entendem o que está sendo apresentado. O objetivo da aplicação deste questionário foi averiguar a percepção e análise dos servidores da CRH, que já responderam ou ainda respondem aos questionamentos dos servidores na seção “Qual a sua Dúvida?”, em relação ao grafo gerado pela SOBEK *Mining*, verificando a viabilidade de obter insumos para tomada de decisão e a aplicabilidade de uma ferramenta similar na instituição.

A ideia inicial era aplicar o questionário para os responsáveis de cada categoria da seção “Qual a sua dúvida?”, mas teria no máximo 3 respondentes para cada categoria e ficaria difícil elaborar um resultado estatístico final. Assim, para melhor validação e apuração dos resultados apresentados pela ferramenta, foi elaborado um Questionário de Análise do Grafo (Apêndice A) gerado pela ferramenta para apenas uma Categoria, “Sistemas de RH”. A escolha de elaborar o questionário somente para essa categoria teve como principal motivo a possibilidade de utilizar a estatística para análise dos resultados e, por ser essa a categoria mais conhecida por todos os respondentes, pois mesmo que não sejam responsáveis por responder aos questionamentos desta categoria, os mesmos utilizam os sistemas, sejam como gestores ou simples usuários, entendendo o vocabulário e siglas pertencentes a este contexto.

Assim, este questionário pôde ser aplicado a 18 pessoas. Cabe ressaltar que o questionário foi aplicado sem a informação do nome da ferramenta que gerou o grafo.

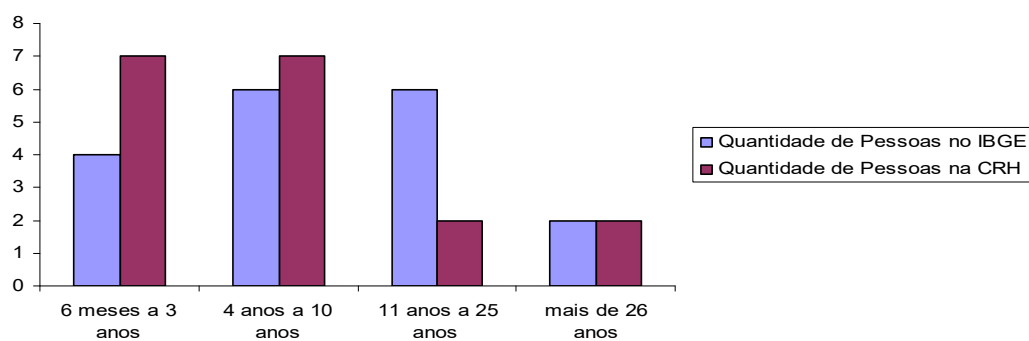
O questionário foi dividido em três partes:

1ª parte – Identificação: sem identificação nominal de quem preenche o questionário, somente as informações de tempo de trabalho no IBGE e na CRH e o período que responde ou respondeu aos questionamentos dos servidores na seção “Qual a sua Dúvida?”. Nesta parte, o intuito é verificar se há alguma diferença ou predominância de resposta de acordo com o tempo de experiência na instituição, no setor específico de Recursos Humanos ou na execução da atividade específica, que é responder as dúvidas do canal de atendimento.

2ª parte – Análise da imagem da categoria “Sistemas de RH”: para melhor entendimento do público-alvo da aplicação deste questionário foi utilizada a nomenclatura “imagem” ao invés de “grafo”, por muitas pessoas desconhecerem o significado da palavra “grafo” e evitar confusão desnecessária ao preenchimento do questionário. Além disso, foi apresentado o grafo com a configuração de 30 termos da categoria de “Sistemas de RH”, tendo em vista que, como explanado no capítulo 6.1.5, foi considerada uma imagem mais clara e de fácil entendimento para pessoas que nunca viram e analisaram uma imagem similar. Nesta parte, os respondentes deveriam responder a três perguntas fechadas (“a”, “b”, “c”) e depois uma pergunta aberta (“d”) para verificar o entendimento e percepção dos mesmos em relação ao grafo. As perguntas fechadas servem para fazer o cálculo estatístico e induzir ao respondente como ele deveria responder à pergunta aberta. O mesmo acontecendo nas perguntas “e” e “f”, para verificar a viabilidade de obter insumos utilizando a imagem gerada pela ferramenta de mineração de textos.

3ª parte – Aplicabilidade da imagem: os questionamentos desta parte servem para avaliar se os respondentes acham útil e/ou aplicável a utilização de uma ferramenta similar ao SOBEK *Mining* para resumir o que os servidores estão falando nas mensagens textuais postadas na seção “Qual a sua Dúvida?”.

O questionário foi aplicado a 18 pessoas nos dias 26 e 27 de abril de 2017. A maioria dos respondentes trabalha no IBGE entre 4 a 25 anos, mas apenas 10 anos trabalhando na CRH, conforme Gráfico 1.

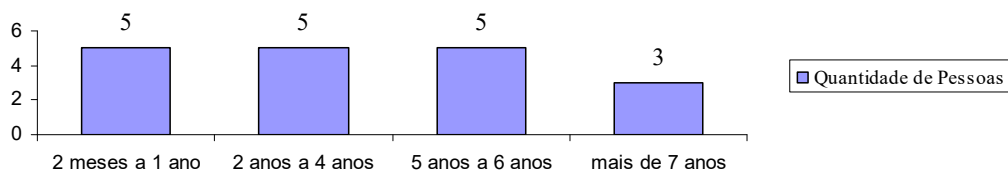
Gráfico 1: Quantidade de pessoas por tempo de Serviço no IBGE e na CRH

Fonte: Elaborado pela autora

Apenas 3 pessoas respondem à seção “Qual a sua Dúvida?” desde o seu surgimento (ano de 2009) e 5 pessoas possuem menos de um ano nesta atividade, conforme Gráfico 2.

Gráfico 2: Quantidade de pessoas respondendo à seção “Qual a sua Dúvida?” por tempo

Quantidade de Pessoas respondendo a seção por tempo

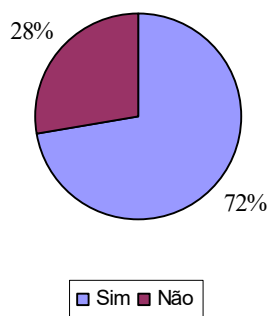


Fonte: Elaborado pela autora

A maioria das pessoas responde atualmente aos questionamentos, enquanto 5 pessoas não realizam mais esta atividade conforme Gráfico 3.

Gráfico 3: Quantidade de pessoas que continuam respondendo à seção “Qual a sua Dúvida?”

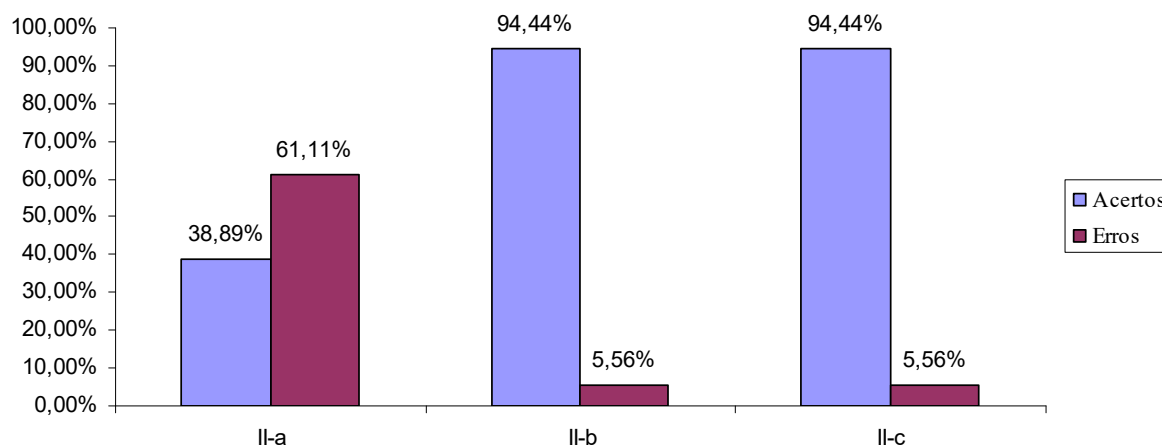
Quantidade de Pessoas que continuam respondendo a seção



Fonte: Elaborado pela autora

O Gráfico 4 mostra a consolidação das respostas referentes às perguntas fechadas da 2ª parte do questionário. Os acertos e erros se referem à percepção das conexões entre os termos.

Gráfico 4: Consolidação das respostas referentes às perguntas fechadas do 2ª parte do questionário



Fonte: Elaborado pela autora

A questão II-a foi a única com o maior percentual de erro, 61,11%. É uma questão de percepção da associação envolvendo quatro termos, talvez este resultado seja proveniente de um entendimento mais complexo pelo número de termos envolvidos e por ter sido a primeira questão a ser respondida. Das pessoas que acertaram, não foi identificado um perfil específico, teve desde os que respondem há muito tempo até os mais novos.

Para responder à questão II-b, bastava entender a visualização da disposição dos termos, quanto maior o nodo maior a frequência, ou o quadro de estatística. Apenas uma pessoa parece não ter entendido esta questão.

Apenas uma pessoa não acertou a questão II-c que é uma questão de percepção da associação envolvendo três termos.

Analisando estas questões, percebe-se que a maioria das pessoas conseguiu entender a imagem. Esse resultado é reforçado nos resultados das perguntas abertas, questão II-d, onde as respostas foram agrupadas em três tipos:

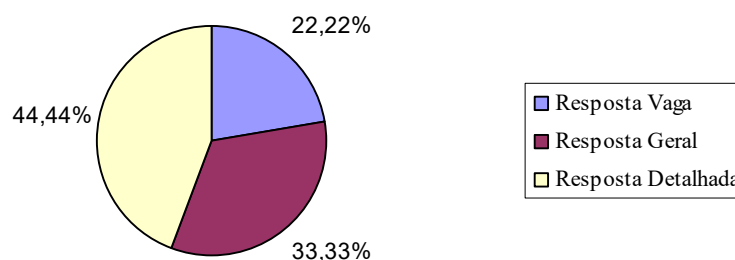
Resposta Vaga: não houve clareza na resposta em relação à imagem.

Resposta Geral: analisou a imagem de forma geral e deu uma resposta resumida.

Resposta Detalhada: detalhou as conexões entre os termos, tentando identificar exatamente o que os servidores estão falando.

Conforme Gráfico 5, 44,44% deu uma resposta detalhada em relação à imagem.

Gráfico 5: Análise das respostas da questão II-d do questionário



Fonte: Elaborado pela autora

Quanto às questões II-e e II-f, que perguntavam se é possível se basear na imagem para elaborar ações que minimizem as dúvidas e reclamações dos servidores, 83,33% responderam que “Sim”, sendo que a grande maioria sugeriu as mesmas ações, como elaborar manuais, divulgar melhor os processos de trabalho e disponibilizar um canal com as perguntas e respostas frequentes. Somente uma pessoa deu uma resposta diferente sugerindo que poderia aumentar as ações de manutenção preventiva nos sistemas, bem como melhoria da

acessibilidade e navegabilidade dos mesmos. Apenas 3 pessoas, ou seja, 16,67%, deram “Não” como resposta, sendo estas 3 pessoas do grupo que respondem ou responderam à seção no período de 2 a 5 anos. Mas somente uma pessoa justificou informando que apenas olhando a imagem não dá para propor ações, que seria necessário ler as mensagens.

Na 3ª parte, o resultado foi que 100% dos respondentes acham aplicável e útil ter uma ferramenta de mineração de textos similar a essa da SOBEK *Mining*. E na justificativa surgiram boas respostas como:

- visualizar os grandes problemas de forma clara e simples, agrupá-los e avaliar sua relevância para aplicar ações de correção ou até mesmo prevenção;
- identificar o maior foco das dúvidas ou problemas e propor ações de melhorias a partir dessa análise;
- interessante ter uma imagem que apresente de forma resumida uma determinada situação;
- indicador útil e claro de questões que estão mais evidentes em determinado universo;
- fácil visualização e entendimento;
- priorizar melhorias e ações corretivas;
- identificar gargalos em processos, diferentes problemas, expectativas, mapeia e identifica demandas;
- auxiliar processo de tomada de decisão gerencial;
- ajudar a incrementar ações objetivas nas questões mais perguntadas;
- facilita dar tratamento geral ao sistema sem olhar caso a caso.

Em relação aos comentários e sugestões, 7 pessoas não deram sugestões e 5 pessoas comentaram sobre a falta de entendimento das palavras menores, ou seja, com menos frequência; a possibilidade de se utilizar mais cores e que as linhas poderiam fornecer alguma informação específica. No entanto, alguns comentários são bem interessantes e merecem ser destacados, como:

- “Analisando a imagem, percebo que para as palavras maiores (aquelas que se repetem muitas vezes) é eficiente para reconhecer rapidamente os problemas maiores, parece-me um teste rápido e eficiente. Contudo, para as palavras menores tais como: aparece, fazer, dados, etc, não nos traz qualquer fonte de

informação relevante, seria necessário uma análise mais próxima para reconhecer o problema”.

- “Essa ferramenta para propor ações no RH em foco na pesquisa, não é muito útil. Mas acho fundamental como meio de diagnóstico especialmente por ser um indicador sistemático quanto às questões que mais se revelam”.
- “Muito interessante e com aplicabilidade no dia a dia da CRH”.
- “Imagem bem didática”.
- “Achei excelente e acredito que pode contribuir muito para a gestão da comunicação interna”.

4.4 Discussão sobre os Resultados

Conforme exposto no capítulo 4.1, existem muitas ferramentas de Mineração de Textos para serem utilizadas, mas nem sempre as mesmas são completas para a necessidade do usuário. Deve haver uma seleção, incluindo testes com uma pequena massa de dados para verificar a viabilidade de utilização da ferramenta à necessidade demandada.

Neste trabalho, mesmo a ferramenta escolhida, *SOBEK Mining*, não atendeu todas as necessidades para a mineração das mensagens. No entanto, como esta ferramenta foi desenvolvida por um grupo de pesquisa da UFRGS, é possível criar uma parceria com este grupo para tentar minimizar os *gaps* encontrados. No último contato realizado com o grupo de desenvolvedores do *SOBEK Mining*, foi informado que as seguintes funcionalidades estavam sendo desenvolvidas:

- A exportação de um arquivo no formato XML com o resultado da saída da mineração, contendo os trechos onde cada termo aparece no texto, as relações e a frequência do termo. Existindo esta funcionalidade, é possível fazer uma integração com outras ferramentas.
- Outra versão do Minerador para conseguir extrair conceitos em textos grandes, utilizando terminal de comando.

Como apresentado no capítulo 4.2, a mineração das mensagens no *SOBEK Mining* deu um panorama geral para cada Categoria da seção “Qual a sua Dúvida?” do IBGEANDO, ressaltando que podem existir diversas formas para analisar uma categoria, seja de uma forma

mais genérica ou específica, para isso é necessário ajustar as configurações da ferramenta de acordo com a análise que se deseja. Em alguns grafos, foi reforçado o funcionamento de um processo, outros evidenciaram os principais problemas de uma área, outros mostraram que a denominação de uma Categoria pode influenciar no tipo do grafo gerado.

Para ter uma visão geral da área de Recursos Humanos do IBGE, de acordo com a mineração das mensagens, foi elaborada a Tabela 15 que mostra os termos mais frequentes que aparecem em mais de uma categoria analisada:

Tabela 15: Termos mais frequentes em mais de uma categoria analisada

| Palavras | Aparece em quantas categorias | Percentual em relação às 13 categorias analisadas | Frequência | Percentual em relação às mensagens totais (40.165) do período analisado |
|-----------------|--------------------------------------|--|-------------------|--|
| horas | 4 | 30,77% | 3785 | 9,42% |
| dias | 8 | 61,54% | 2386 | 5,94% |
| processo | 4 | 30,77% | 2137 | 5,32% |
| meses | 5 | 38,46% | 1962 | 4,88% |
| SECAF | 4 | 30,77% | 1476 | 3,67% |
| férias | 4 | 30,77% | 1440 | 3,59% |
| licença | 5 | 38,46% | 1364 | 3,40% |
| sistema | 5 | 38,46% | 1253 | 3,12% |
| anos | 4 | 30,77% | 949 | 2,36% |
| siape | 4 | 30,77% | 826 | 2,06% |
| período | 5 | 38,46% | 811 | 2,02% |
| resposta | 4 | 38,46% | 584 | 1,45% |

Fonte: Elaborado pela autora

A tabela indica que a área de RH trabalha muito com a questão de prazos e tempo uma vez que as palavras “horas”, “meses”, “dias”, “anos” e “período” são as que aparecem mais frequentemente nas mensagens. Depois vem a palavra “processo” que indica burocracia e trâmites na área, o que normalmente acontece numa instituição governamental. Em seguida vem o sistema SECAF que parece ser o sistema mais falado pelos servidores. Outra palavra que chama a atenção é “férias”, um evento que deveria ser simples e normal e parece apresentar alguns problemas, seguido também do evento “licença”. Sistema é outra palavra que merece atenção pela área, parece haver muitos questionamentos em relação aos sistemas de RH, incluindo o SIAPE. E o aparecimento da palavra “resposta” também pode indicar que os servidores não estejam sendo respondidos satisfatoriamente ou nem tenham uma resposta.

Com esta visão geral, a Coordenação de Recursos Humanos poderia investigar melhor os assuntos mais comentados com o intuito de minimizar os problemas e reclamações dos servidores, assim como avaliar as equipes que tratam das questões de sistemas, por exemplo, investindo e melhorando os procedimentos. E isso é percebido também pelos servidores da CRH que responderam ao questionário aplicado por esta pesquisa. A grande maioria entendeu a visualização do grafo da categoria “Sistemas de RH” e conseguiu extrair insumos para a tomada de decisão e melhoria da comunicação interna, além de propor ações para minimizar os problemas encontrados.

Para os 18 respondentes do questionário, desde os servidores que ingressaram recentemente no IBGE aos mais antigos, não houve nenhuma falta de entendimento quanto aos termos apresentados, que incluíam siglas e nomes específicos de sistemas usados na instituição. No questionário não foi colocada nenhuma legenda para explicar o significado dos termos. Isso mostra que os respondentes pertencem a uma mesma comunidade discursiva, com uma linguagem própria da instituição. Provavelmente, se esse questionário fosse aplicado a pessoas externas, fora do contexto do IBGE, estas não conseguiriam entender os termos e conseqüentemente a relação entre eles.

Analisando os resultados obtidos tanto com o processamento das mensagens, quanto a sua posterior validação com os servidores da CRH, conclui-se que a prova de conceito atingiu o seu propósito, de verificar que o pressuposto afirmado anteriormente tem aplicação num mundo real e é viável tecnicamente, ou seja, que a utilização de uma ferramenta de mineração de textos para processar as mensagens textuais do canal de atendimento serve para obter insumos para a tomada de decisão.

5 CONSIDERAÇÕES FINAIS

Considerando tudo o que foi exposto neste trabalho, o objetivo geral desta pesquisa foi atendido, o de analisar a viabilidade do uso de ferramenta de mineração de textos para processamento automático das mensagens textuais no Canal de Atendimento do Portal IBGEANDO de forma a obter insumos para auxiliar a tomada de decisão e melhorar a comunicação da área de RH.

Dentre as várias ferramentas selecionadas para minerar as mensagens textuais, a ferramenta SOBEK *Mining* foi a escolhida e, apesar de ter sido desenvolvida inicialmente com fins pedagógicos, ela foi considerada viável para processar as mensagens textuais do Canal de Atendimento. Isso é ratificado com a aplicação do questionário aos servidores da CRH, onde a grande maioria entendeu o grafo gerado pela SOBEK e conseguiu extrair insumos e ações para os problemas levantados nas mensagens postadas. Assim, foi possível ter uma avaliação geral da ferramenta de mineração de textos para verificar a viabilidade de obter insumos a partir das mensagens textuais postadas, ou seja, de descobrir o que os servidores estão falando. Cabe ressaltar que este entendimento do grafo gerado pode ser melhorado com o treinamento da equipe e maior utilização da ferramenta.

A Figura 31 ilustra de uma forma geral o entendimento do objetivo da pesquisa utilizando o processamento feito com as mensagens textuais da Categoria Aposentadoria do Canal de Atendimento.

Este trabalho mostrou como as mensagens textuais foram processadas na ferramenta SOBEK *Mining*, identificando os termos e suas relações contidos nas mensagens, que após serem analisadas pelos servidores de RH do IBGE, foram obtidos os insumos para auxiliar na tomada de decisões e definir as ações a serem executadas. A obtenção dos insumos indica que o IBGE funciona como uma comunidade discursiva, pois somente servidores da Fundação conseguem de maneira efetiva dar um sentido para os termos das mensagens e suas relações após o processamento da ferramenta de mineração de textos. Os insumos obtidos a partir destas mensagens podem ser considerados como um processo de consumo de informação que foi produzido num âmbito de relações sociais entre os servidores do IBGE, indicando assim que o IBGE funciona como uma Comunidade Discursiva, produzindo sentido somente para quem integra esta comunidade.

estudo, é que, uma vez que já foi comprovado que as mensagens textuais postadas pelos servidores, após o processo de mineração de textos, geram insumos para a tomada de decisão no âmbito da área de RH, que estes insumos sejam efetivamente utilizados, gerando ações por parte da CRH e dando clareza ao processo decisório, possibilitando um ambiente organizacional mais participativo, onde os servidores se sintam realmente ouvidos.

No IBGE, o processo de tomada de decisão, que se inicia com a busca da informação, já pode ter como subsídios a análise das mensagens textuais, de modo a facilitar a elaboração de alternativas para resolver uma situação, dispondo assim de informação confiável e consistente e, em tempo para que possam tomar decisões eficazes e eficientes.

Além disso, com a adoção de uma ferramenta de mineração de textos pelo IBGE, atende-se ao Programa de Melhoria da Qualidade na Gestão Institucional, baseado em seus princípios:

- Agilidade: uma vez que são identificados os problemas ou questões principais da maioria dos servidores, de forma automática e sistematizada, que geram insumos para dar fluidez aos processos de trabalho;
- Qualidade da Informação: uma vez que, ao utilizar uma ferramenta para minerar as mensagens, garante-se o máximo de precisão da informação;
- Confiabilidade: uma vez que é dada transparência na forma como os insumos para a tomada de decisão são obtidos;
- Relacionamento: fortalecendo continuamente a rede de relacionamentos com os servidores;
- Cordialidade: dando tratamento justo e de acordo com o demandado ou apontado pelos servidores.

Desta forma, é possível estabelecer um processo participativo com os servidores, estimulando o comprometimento cada vez maior dos mesmos com o sucesso do IBGE, além de torná-los mais motivados, produtivos e satisfeitos com a instituição. Além disso, esta ferramenta pode ser utilizada em textos de *blogs* de outras *intranets* ou outros canais de atendimentos dentro do próprio IBGE, como o de relacionamentos com os clientes externos.

Conclui-se que este estudo serve para contribuir para uma gestão da informação mais eficiente dentro do IBGE que acarrete em implantação de políticas, serviços ou sistemas que atendam às necessidades dos servidores e da própria Fundação.

A partir desta pesquisa, também se pode pensar em outras vertentes de estudos dentro do próprio IBGEANDO, como se basear nas respostas dos questionamentos para elaborar automaticamente uma página de Perguntas Frequentes, realizar um novo estudo para determinar a Categorização da seção “Qual a sua Dúvida?”, construir uma ontologia para o Portal do RH, entre outros.

REFERÊNCIAS

ALVARES, Lilian; DE ARAÚJO JÚNIOR, Rogério Henrique. Marcos históricos da ciência da informação: breve cronologia dos pioneiros, das obras clássicas e dos eventos fundamentais. **Transinformação**, Campinas, v. 22, n. 3, p. 195-205, setembro, 2012. Quadrimestral.

BARBOSA, M. L.; SEVERO, C.; REATEGUI, E. B. Mineração de padrões no gênero textual blog. **Revista Novas Tecnologias na Educação**, Porto Alegre, v. 7 n° 3, dezembro, 2009.

BASTOS, V. M. **Ambiente de Descoberta de Conhecimento na Web para a Língua Portuguesa**. Rio de Janeiro: COPPE, 2006. Disponível em: <http://www.coc.ufrj.br/teses/doutorado/inter/2006/Teses/BASTOS_VM_06_t_D_int.pdf>. Acesso em: set. 2016.

BIRD, Steven; KLEIN, Ewan; LOPER, Edward. **Natural Language Processing with Python: analyzing text with the Natural Language Toolkit**. 1ª Edição. Sebastopol: O'Reilly, 2009. 479 p.

BRAGA, Kátia Soares. Aspectos relevantes para seleção de metodologia adequada à pesquisa social em Ciência da Informação. In: **MUELLER**, Suzana P. M. (Org.). **Métodos para a pesquisa em Ciência da Informação**. 1ª Edição. Brasília: Thesaurus, 2007. p. 17-37.

BRITO, Edeleon Marcelo Nunes de. **Mineração de Textos: Detecção automática de sentimentos em comentários nas mídias sociais**. 2016. 53 f. Dissertação (Mestrado) - Curso de Mestrado Profissional em Sistemas de Informação e Gestão do Conhecimento, Programa de Pós-graduação Stricto Sensu, Universidade Fundação Mineira de Educação e Cultura, Belo Horizonte, 2016. Disponível em: <<http://www.fumec.br/revistas/sigc/article/view/3737>>. Acesso em: out. 2016.

BUCHANAN, Leigh; O'CONNELL, Andrew. Uma breve história da tomada de decisão. **Harvard Business Review**, jan. 2006. Mensal. Disponível em : <http://tupi.fisica.ufmg.br/michel/docs/Artigos_e_textos/MPE_e_empresa_familiar/ISO_Pequenas_empresas.pdf> Acesso em: out. 2016. Não paginado.

CAPRA, Fritjof. **A teia da vida: uma nova compreensão científica dos sistemas vivos**. 6ª. ed. São Paulo: Cultrix, 2001. 256 p.

CÂNDIDO, Gesinaldo Ataíde; et al. A interpretação organizacional em empresas de tecnologia da informação e comunicação (TIC): um estudo na Incubadora Tecnológica de Campina Grande – PB. **Ciência da Informação**, Brasília, DF, v. 40, n. 2, p. 192-206, mai/ago. 2011. Quadrimestral. Disponível em: <<http://revista.ibict.br/ciinf/article/view/1310>>. Acesso em: out. 2016.

CAPURRO, Rafael. Epistemology and Information Science. Royal Institute of Technology Library, Stockholm, 1985, Report TRITALIB- 6023. Disponível em: <<http://www.capurro.de/trita.htm>>. Acesso em: set. 2016.

CARDOSO, A. M. P. Retomando possibilidades conceituais: uma contribuição à sistematização do campo da informação social. **Revista da Escola de Ciência da Informação da UFMG**, Belo Horizonte, MG, v. 23, n. 2, p. 107-114, jul./dez. 1994. Semestral.

CASSARRO, Antonio Carlos. **Sistemas de informações para tomada de decisões**. 4ª ed. São Paulo: Cengage, 2010. 136 p.

CASSETTARI, Rafael-roeck-borges et al. Comparação da Lei de Zipf em conteúdos textuais e discursos orais. **El Profesional de La Información**, [s.l.], v. 24, n. 2, p.157-167, 11 mar. 2015. Disponível em : <<https://recyt.fecyt.es/index.php/EPI/article/view/34736>> Acesso em: mai. 2017.

CHOO, Chun Wei. **A organização do conhecimento**: como as organizações usam a informação para criar significado, construir conhecimento e tomar decisões. 3ª ed. São Paulo: Senac São Paulo, 2003. 416 p. Tradução: Eliana Rocha.

CORRÊA, Ângela M. C. Jorge; SFERRA, Heloisa Helena. Conceitos e aplicações de data mining. **Revista de Ciência & Tecnologia**, Santa Maria, RS, v. 11, n. 22, p.19-34, jul. 2003.

DAMASCENO, Fábio Rafael. **Mineração textual em teleconsultorias**: aprimoramento da educação permanente de equipes da saúde da família no projeto telessaúde-RS. 2015. 131 f. Tese (Doutorado) - Curso de Informática na Educação, Programa de Pós Graduação, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, 2015.

DANTAS, Edmundo Brandão. A importância da pesquisa para a tomada de decisões. **Biblioteca Online de Ciências da Comunicação**, Brasília, DF, 2013. Disponível em: <<http://www.bocc.ubi.pt/pag/dantas-edmundo-2013-importancia-pesquisa-tomada-decisoes.pdf15>>. Acesso em: nov. 2016.

EPSTEIN, Daniel; REATEGUI, Eliseo . Uso de mineração de textos no apoio à compreensão textual. **RENOTE. Revista Novas Tecnologias na Educação**, Porto Alegre, RS, v. 13, n. 1, p. 200-210, jul. 2015.

FAIRCLOUGH, Norman. **Language and power**. 2ª. ed. Reino Unido, Longman: Routledge, 2001. 240 p.

FELDMAN, Ronen; SANGER, James. **Text Mining Handbook: advanced approaches in analyzing unstructured data**. 2ª. ed. New York: Cambridge University Press NY, 2006. 424 p.

FERNANDES, Geni Chaves. Quatro visões no campo da Ciência da Informação. Rio de Janeiro, 2006. 42p. Trabalho apresentado à banca avaliadora do concurso para Professor Adjunto 1 DE do Departamento de Ciência da Informação da Universidade Federal de Santa Catarina como requisito parcial para aprovação.

FRANCIS, Louise A. Text Mining Handbook. **Casualty Actuarial Society Forum: Spring**, Arlington, Virginia, p.1-61, set. 2010. Trimestral. Disponível em: <https://www.casact.org/pubs/forum/10spforum/Francis_Flynn.pdf>. Acesso em: set. 2016.

FUJITA, Mariângela Spotti Lopes. A importância teórica e prática da indexação na fundamentação científica da organização e representação do conhecimento. In: DODEBEI, Vera; GUIMARÃES, José Augusto Chaves (Org.). **Complexidade e organização do conhecimento: desafios de nosso século**. 1ª ed. Rio de Janeiro: ISKO-Brasil; Marília: FUNDEPE, 2013, v. 1, p. 147-159.

GONZÁLEZ DE GOMEZ, María Nelida. O caráter seletivo das ações de informação. **Informare**, Rio de Janeiro, v. 5, n. 2, p. 7-31, 1999. Disponível em: <<http://repositorio.ibict.br/bitstream/123456789/126/1/GomezInformare1999.pdf>>. Acesso em: out. 2016.

GUEDES, Vânia Lisbôa da Silveira. Estudo de um critério para indexação automática derivativa de textos científicos e tecnológicos. **Ciência da Informação**, Brasília, DF, v. 23, n. 3, p.318-326, set. 1994. Trimestral.

GUEDES, Vânia LS; BORSCHIVER, Suzana. Bibliometria: uma ferramenta estatística para a gestão da informação e do conhecimento, em sistemas de informação, de comunicação e de avaliação científica e tecnológica. **Encontro Nacional de Ciência da Informação**, Salvador, BA, v. 6, p. 1-18, 2005. Semestral.

GUPTA, Vishal; LEHAL, Gurpreet S. A Survey of Text Mining Techniques and Applications. **Journal of Emerging Technologies in Web Intelligence**, v. 1, n. 1, 2009. Disponível em:
<<http://www.academypublisher.com/jetwi/vol01/no1/jetwi01016076.pdf>> Acesso em: set. 2016

HEMSLEY, Bronwyn; PALMER, Stuart. Two Studies on Twitter Networks and Tweet Content in Relation to Amyotrophic Lateral Sclerosis (ALS): Conversation, Information, and ‘Diary of a Daily Life’. In: **Digital Health Innovation for Consumers, Clinicians, Connectivity and Community**: Selected Papers from the 24th Australian National Health Informatics Conference (HIC 2016). p. 41-47, 2016. IOS Press.

HEARST, Marti, **What is Text Mining**, outubro, 2003. Disponível em:
<<http://people.ischool.berkeley.edu/~hearst/text-mining.html>>. Acesso em: mai. 2017
Não paginado.

HJØRLAND, B. Birger; ALBRECHTSEN, Hanne. Toward a new horizon in information science: domain analysis. **Journal of the American Society for Information Science**, v. 46, n. 6, p. 400-425, 1995.

HJØRLAND, Birger. **Information seeking and subject representation**: an activity-theoretical approach to Information Science. New York: Greenwood Press, 1997. 213 p.

HJØRLAND, Birger. Domain analysis in information science: Eleven approaches – traditional as well as innovative. **Journal of Documentation**, v.58, n.4, p. 422-462, 2002a.

HJØRLAND, Birger. “Epistemology and the socio-cognitive perspective in information science”, **Journal of the American Society for Information Science and Technology**, v. 53 n. 4, p. 257-270. 2002b.

IBGE. Política de Comunicação Integrada. Rio de Janeiro: IBGE, 2012. Disponível em:
<http://w3.presidencia.ibge.gov.br/comites_comissoes_grupos/interna/pdf/cpc/politica_comunicacao.pdf>. Acesso em: mai. 2014.

IBGE. Portal IBGEANDO, Rio de Janeiro, IBGE, 2008.
<<http://w3.ibgeando.ibge.gov.br>> Acesso em: mai. 2014.

IBGE. Programa de Melhoria da Qualidade na Gestão Institucional 2008-2011, Rio de Janeiro, IBGE, ago. 2008. Disponível em: <<http://portal.de.ibge.gov.br/web/cps1/programa-de-melhoria>>. Acesso em: jul. 2014.

IBGE. Projeto Portal do Servidor. Rio de Janeiro: IBGE, DE/CRH, mai. 2008.

JAIN, V. K.; KUMAR, S. An Effective Approach to Track Levels of Influenza-A (H1n1) Pandemic in India Using Twitter. **Procedia Computer Science**, v. 70, p. 801–807, 2015. ISSN 1877-0509. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S1877050915032846>>. Acesso em: set. 2016

KLEMMANN, Miriam; REATEGUI, Eliseo; RAPKIEWICZ, Clevi. Análise de ferramentas de mineração de textos para apoio a produção textual. In: **Anais do Simpósio Brasileiro de Informática na Educação**. 2011. Porto Alegre, RS.

LUHN, H. P. The automatic creation of literature abstracts. **IBM Journal of Research and Development**, v.2, n.2, p.159–165, 1958

LIMA, Fanny do Couto Ribeiro. Mineração de Textos para a extração automática de sintagmas nominais: tendências e estratégias para recuperação da informação jurídica. In **Anais [recurso eletrônico] / XV Encontro Nacional de Pesquisa em Ciência da Informação: além das nuvens, expandindo as fronteiras da Ciência da Informação**, 27-31 de outubro em Belo Horizonte, MG. 2014. p. 955 a 974. Disponível em: <<http://enancib2014.eci.ufmg.br/documentos/anais/anais-gt2>> Acesso em: set. 2016

LOH, Stanley. et al. Formalizando e Explorando Conhecimento Tácito com a tecnologia de Text Mining para Inteligência. In: INTERNATIONAL SYMPOSIUM ON KNOWLEDGE MANAGEMENT/DOCUMENT MANAGEMENT, ISKM/DM. Curitiba, PR, 2001. Disponível em: <https://www.researchgate.net/publication/242508948_FORMALIZANDO_E_EXPLORANDO_CONHECIMENTO_TACITO_COM_A_TECNOLOGIA_DE_TEXT_MINING_PARA_INTELIGENCIA> Acesso em: nov. 2016. Não paginado.

LOUSADA, M.; VALENTIM, M. L. P. Informação orgânica como insumo do processo decisório empresarial. In: VALENTIM, M. L. P. (Org.) Gestão da informação e do conhecimento no âmbito da Ciência da Informação. São Paulo: Polis: Cultura Acadêmica, 2008. 268p Perspectivas em Ciência da Informação, v.16, n.1, p.147-164, jan./mar. 2011 Disponível em: <<http://www.scielo.br/pdf/pci/v16n1/a09v16n1.pdf>>. Acesso em: out. 2016.

MACEDO, Alexandra Lorandi. **Rede de conceitos**: uma ferramenta para contribuir com a prática pedagógica no acompanhamento da produção textual coletiva. 2009. 166 f. Tese (Doutorado) - Curso de Doutorado em Informática na Educação, Programa de Pós-graduação em Informática na Educação, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, 2010. Disponível em: <<http://www.lume.ufrgs.br/handle/10183/19030>>. Acesso em: set. 2016

MARAVILHAS-LOPES, Sérgio Paulo. A Web 2.0 como ferramenta de análise de tendências e monitorização do ambiente externo e sua relação com a cultura de convergência dos media. **Perspectivas em Ciência da Informação**, v. 18, n. 1, p. 126-137, jan./mar. 2013. Disponível em: <<http://portaldeperiodicos.eci.ufmg.br/index.php/pci/article/view/1658>>. Acesso em: jun. 2014.

MARCONI, M. D. A.; LAKATOS, E. M. **Técnicas de pesquisa**: planejamento e execução de pesquisas, amostragens e técnicas de pesquisas, elaboração, análise e interpretação de dados. 3ª ed. São Paulo: Atlas, 1996.

MARCUSCHI, Luiz Antônio. Gêneros Textuais: Definição e Funcionalidade. 2002a. Disponível em: <https://disciplinas.stoa.usp.br/pluginfile.php/133018/mod_resource/content/3/Art_Marcuschi_G%C3%AAneros_textuais_defini%C3%A7%C3%B5es_funcionalidade.pdf> Acesso em: out. 2016 Não paginado

MARCUSCHI, Luiz Antônio. Gêneros textuais emergentes e no contexto da tecnologia digital. Texto da Conferência pronunciada na 50ª Reunião do GEL – Grupo de Estudos Linguísticos do Estado de São Paulo, USP, São Paulo, 23-25 de maio de 2002b. Disponível em: <http://www.twiki.faced.ufba.br/twiki/pub/GEC/RefID/marcuschi-generos_textuais_emergentes_no_.....doc> Acesso em: jun.2015.

MINAYO, Maria Cecília de Souza. **Ciência, técnica e arte**: o desafio da Pesquisa Social. In: _____ Pesquisa Social: Teoria, método e criatividade. 30 ed. Petrópolis: Vozes, 2011.p. 9-29 (cap.1).

MOLINA, Rafael Antonangelo; STEINBERGER-ELIAS, Margarethe Born. Criando um corpus sobre desastres climáticos com apoio da ferramenta NLTK. In: 8º Brazilian Symposium in Information and Human Language Technology. Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology, p. 194-198, Cuiabá, MT, Outubro 24-26, 2011.

MORAIS, E. A. M. Contextualização de Documentos em Domínios Representados por Ontologias Utilizando Mineração de Textos. Dissertação de Mestrado do Instituto de Informática da Universidade Federal de Goiás. Goiás, GO, 2007. Disponível em: <http://www.inf.ufg.br/sites/portal.inf.ufg.br/mestrado/files/ds_Edison.pdf>. Acesso em: set. 2016

MORAIS E.A.M.; AMBRÓSIO A.P.L. Mineração de Textos. In: Technical Report – INF_005/07, Goiás, GO, 2007. 29 p. Disponível em: <http://www.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_005-07.pdf> Acesso em: nov. 2016.

NASCIMENTO, D. M.; MARTELETO, R. M. A “informação contruída” nos meandros dos conceitos da teoria social de Pierre Bordieu. **Datagramazero**, Rio de Janeiro, v. 5, n. 5, out. 2004. Disponível em: <http://www.dgz.org.br/out04/F_I_art.htm>. Acesso em: jun. 2015.

NEVES, Ronald Gonçalves das; ZACCARO, Thiago Chagas. **Mineração de texto para análise de interações em redes sociais acadêmicas**. 2016. 57 f. Monografia (Especialização) - Curso de Análise e Gestão de Sistemas de Informação, Pós-graduação Lato Sensu, Instituto Federal de Educação, Ciência e Tecnologia Fluminense, Campos dos Goytacazes, RJ, 2016. Disponível em: <<http://bd.centro.iff.edu.br/xmlui/handle/123456789/1124>>. Acesso em: nov. 2016.

NOBLIA, M. V. The computer-mediated communication, a new way of understanding the language. In: Anais da international conference. Bristol: conference papers. 1998. Disponível em: <<http://nl.ijs.si/janes/wp-content/uploads/2014/09/noblia98.pdf>>. Acesso em: set. 2016. Não paginado.

PINHEIRO, Lena Vânia Ribeiro, LOUREIRO, José Mauro Matheus. Traçados e limites da ciência da informação. **Ciência da Informação**, Brasília – DF, v. 24, n. 1, 1995. Disponível em: <<http://revista.ibict.br/ciinf/article/view/609/611>>. Acesso em: nov. 2016.

REEVES, T.C.; HARMON, S.W. Systematic Evaluation Procedures for interactive Multimedia for Education and Training, In S. Riesman(Ed), *Multimedia Computing: Preparing for the 21 set century*, Idea Group: Harrisburg, PA, pp.472-505, 1998

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

RUSSEL, Matthew A.. **Mineração de dados da Web Social**. Primeira edição - São Paulo: Novatec, 2011.

SCHENKER, Adam. **Graph-theoretic techniques for web content mining**. 2003. 131 f. Unpublished PhD thesis, University of South Florida - Curso de Philosophy, Department Of Computer Science And Engineering, University Of South Florida, Florida, 2003. Disponível em: <<http://scholarcommons.usf.edu/cgi/viewcontent.cgi?article=2466&context=etd>>. Acesso em: nov. 2016.

SCHIESSL, José Marcelo. Descoberta de Conhecimento em Texto aplicada a um sistema de atendimento ao consumidor. Orientador: Profa. Dra. Marisa Bräscher, 2007. Dissertação (Mestrado em Ciência da Informação) – Departamento de Ciência da Informação e Documentação, Universidade de Brasília. RICI: R.Ibero-amer. Ci. Inf., Brasília, v. 4, n. 2, p. 94-110, ago./dez.2011. Disponível em: <periodicos.unb.br/index.php/RICI/article/download/6212/5106> Acesso em: set.2016

SIMON, Herbert. **Administrative Behavior**: a study of decision-making processes in administrative organization. 4. ed. New York: Free Press, 1997. 384 p.

SOUZA, Renato Rocha; COELHO, Flávio Codeço; GONÇÁLVES, Alexandre. O projeto Media Cloud Brasil: uma análise do tratamento de informações em ambientes de big data. 2015. 11 f. Dissertação (Mestrado) - Curso de Mestrado, Base de Dados, Repositório Fgv Teses/diss./papers/etc., Fgv, [s.l.], 2015. Disponível em: <http://bibliotecadigital.fgv.br/dspace/bitstream/handle/10438/15036/O_Projeto_Media_Cloud_Brasil.pdf?sequence=1>. Acesso em: nov. 2016.

SWALES, John M. The concept of discourse community. In: _____. *Genre analysis*. Cambridge: Cambridge University Press, 1990. p.21-32.

TAN, A. Text Mining: The State of the Art and the Challenges. In: WORKSHOP ON KNOWLEDGE DISCOVERY FROM ADVANCED DATABASES, 1999, Beijing. Proceedings... 1999. p. 71-76. Disponível em: <http://www.ntu.edu.sg/home/asahtan/Papers/tm_pakdd99.pdf> Acesso em: set. 2016

TORRES, José Alberto Sousa. Descoberta de Informação Através da Mineração de Texto: Fundamentos e Aplicações. In: RAMOS, Jorge Luis Cavalcanti (Org.). Anais da XII Escola Regional de Computação Bahia Alagoas Sergipe. Juazeiro, Ba: Univasf, 2012. Cap. 12. p. 349-384. Disponível em: <<https://www.researchgate.net/publication/289540770>>. Acesso em: nov. 2016.

XAVIER, Rodolfo Coutinho Moreira; COSTA, Rubenildo Oliveira da. Relações mútuas entre informação e conhecimento: o mesmo conceito? **Ciência da Informação**, Brasília, DF, v. 39, n. 2, p.75-83, maio 2010. Trimestral. Disponível em: <<http://revista.ibict.br/ciinf/article/view/1278/1456>>. Acesso em: nov. 2016.

APÊNDICES

Apêndice A – Questionário de Análise do Grafo da Categoria “Sistemas de RH”

Rio de Janeiro, _____ de _____ de 2017

QUESTIONÁRIO

I- IDENTIFICAÇÃO

Há quanto tempo trabalha no IBGE? _____

Há quanto tempo trabalha ou trabalhou na CRH? _____

Há quanto tempo responde ou respondeu aos questionamentos dos servidores na seção “Qual a sua Dúvida?” _____

Continua respondendo a estes questionamentos?

Sim Não

II- ANÁLISE DA IMAGEM DA CATEGORIA “SISTEMAS DE RH”

Ao utilizar uma ferramenta para obter as palavras mais usadas nas mensagens postadas pelos servidores na categoria “Sistemas de RH” (de Maio/2009 a Maio/2015), foi gerada a imagem reproduzida no ANEXO, onde o tamanho da palavra é proporcional à quantidade de vezes que ela é usada nas mensagens (frequência da palavra) e as linhas entre as palavras representam as ligações ou associações entre elas.

Observando essa imagem, você pode afirmar que:

a) A maioria dos servidores está relatando que há algum problema ou dúvida em relação ao período de férias entre o Siape e Siapenet?

Sim Não

b) O SDA é o sistema que apresenta mais questões dos servidores?

Sim Não

c) Os servidores estão falando que aguardam uma resposta sobre o sistema SECAF?

Sim Não

d) Analisando toda a imagem, o que mais você acredita que os servidores estão falando nas mensagens?

e) Sendo o “Qual a sua Dúvida?” uma seção de dúvidas, reclamações e elogio, e, considerando as respostas acima, você pode se basear nesta imagem para elaborar ações que minimizem as dúvidas e reclamações dos servidores? Exemplo: Elaborar um vídeo explicativo de como acessar os sistemas.

Sim Não

f) Se sim, quais outras ações você poderia sugerir para minimizar as dúvidas dos servidores?

III- APLICABILIDADE DA IMAGEM

Você acha útil e aplicável ter uma ferramenta que gere uma imagem similar à apresentada para resumir o que os servidores mais estão falando nas mensagens?

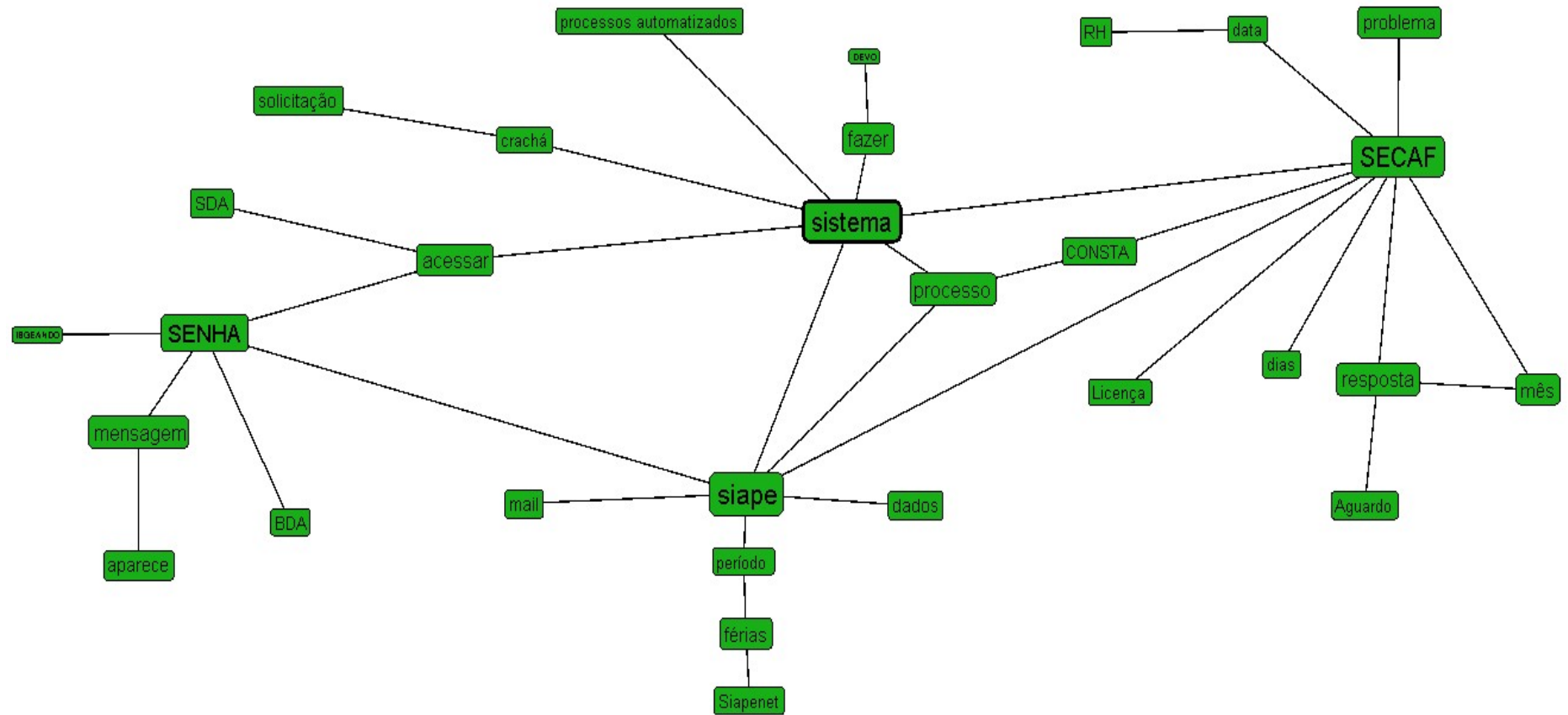
Sim Não

Por que? _____

Comentários / sugestões sobre a imagem gerada pela ferramenta:

ANEXO

IMAGEM – CATEGORIA SISTEMAS DE RH



| Palavras | Frequencia | Percentual em relação às mensagens totais do período analisado |
|-----------------|-------------------|---|
| siape | 259 | 39,85% |
| sistema | 221 | 34,00% |
| SECAF | 210 | 32,31% |
| senha | 183 | 28,15% |
| mensagem | 111 | 17,08% |
| Fazer | 109 | 16,77% |
| processo | 108 | 16,62% |
| resposta | 100 | 15,38% |
| acessar | 99 | 15,23% |
| mês | 97 | 14,92% |
| férias | 95 | 14,62% |
| Problema | 93 | 14,31% |
| Mail | 80 | 12,31% |
| Dados | 80 | 12,31% |
| SDA | 79 | 12,15% |
| RH | 72 | 11,08% |
| Solicitação | 71 | 10,92% |
| Aparece | 71 | 10,92% |
| Consta | 69 | 10,62% |
| Data | 67 | 10,31% |
| Crachá | 67 | 10,31% |

| Palavras | Frequencia | Percentual em relação às mensagens totais do período analisado |
|-------------------------|-------------------|---|
| Siapenet | 66 | 10,15% |
| Licença | 66 | 10,15% |
| Aguardo | 66 | 10,15% |
| Dias | 65 | 10,00% |
| Processos automatizados | 60 | 9,23% |
| Período | 60 | 9,23% |
| BDA | 60 | 9,23% |
| IBGEANDO | 57 | 8,77% |
| Devo | 57 | 8,77% |

Apêndice B - Quadro único com o comparativo das ferramentas de mineração de textos

| FERRAMENTAS | KH CODER | NLTK | PyPLN | SOBEK MINING | TAGCROWD | TEXTALYZER | WORDCOUNTER | PARÂMETROS DA FERRAMENTA IDEAL |
|---|---|--|--|--|--|---|--|--------------------------------|
| CRITÉRIOS | | | | | | | | |
| Instalação (Local / Online) | Local | Local | Local | Local | Online | Online | Online | Local |
| Possui documentação sobre a ferramenta | Sim | Sim | Sim | Sim | Não | Não | Não | Sim |
| Facilidade de Operação | Médio | Difícil | Difícil | Fácil | Fácil | Fácil | Fácil | Fácil |
| Processamento do teste executou satisfatoriamente | Médio | Não instalado e nem testado | Não instalado e nem testado | Sim | Sim | Sim | Sim | Sim |
| Linguagem da Interface | Inglês/Espanhol | Não instalado e nem testado | Não instalado e nem testado | Português | Inglês | Inglês | Inglês | Português |
| Entrada de Dados (Manual / Importação de Arquivos) | Importação de Arquivo (XLS, CSV, TXT) | Não instalado e nem testado | Não instalado e nem testado | Importação de arquivos (DOC, PDF, TXT) | Manual / Importação de Arquivo | Manual / Importação de Arquivo | Manual | Importação de Arquivo |
| Saída de Dados (Tela / Exportação de Arquivos) | Exportação de Arquivo (TXT) | Não instalado e nem testado | Não instalado e nem testado | Tela, Imagem JPG, Grafo em xml | Tela e Saídas em arquivo no formato HTML e PDF | Tela | Tela | Tela / Exportação de Arquivos |
| Análise de texto no Idioma Português | Sim | Não instalado e nem testado | Não instalado e nem testado | Sim | Sim | Sim | Sim | Sim |
| Permite lista de <i>stopwords</i> | Não | Não instalado e nem testado | Não instalado e nem testado | Sim | Sim | Sim | Não | Sim |
| Permite lista de <i>stopwords</i> em Português | Não permite lista | Não instalado e nem testado | Não instalado e nem testado | Sim | Sim | Não | Não permite lista | Sim |
| Entrada lista de <i>stopwords</i> (Manual / Importação de Arquivos) | Não permite lista | Não instalado e nem testado | Não instalado e nem testado | Importação de Arquivos | Manual | Manual | Não permite lista | Importação de Arquivo |
| Usa stemming ou lematização | Sim | Não instalado e nem testado | Não instalado e nem testado | Não | Não | Não | Não | Não |
| Considera palavras acentuadas | Não | Não instalado e nem testado | Não instalado e nem testado | Sim | Sim | Sim | Sim | Sim |
| Exibe o número da Frequencia das Palavras | Sim | Não instalado e nem testado | Não instalado e nem testado | Sim | Sim | Sim | Sim | Sim |
| Visualização Gráfica do Resultado | Sim | Não instalado e nem testado | Não instalado e nem testado | Sim | Sim | Não | Não | Sim |
| Visualização Gráfica dos relacionamentos entre palavras | Sim | Não instalado e nem testado | Não instalado e nem testado | Sim | Não | Não | Não | Sim |
| Existe Integração com outras ferramentas | Não | Não instalado e nem testado | Não instalado e nem testado | Não | Não | Não | Não | Sim |
| Conhecimento prévio de programação computacional | Não | Sim | Sim | Não | Não | Não | Não | Não |
| Comentário Geral | Dispõe de diversas funcionalidades úteis, mas que dependem de um tempo a ser demandado para seu estudo e utilização. Faz a lematização mas de uma forma pouco compreensível para o português. | Foi considerada difícil por necessitar de conhecimentos prévios de programação, não sendo de fácil uso para usuários leigos. Ela não foi instalada e nem testada devido ao tempo a ser demandado para o estudo em questão. | Foi considerada difícil por necessitar de conhecimentos prévios de programação, não sendo de fácil uso para usuários leigos. Ela não foi instalada e nem testada devido ao tempo a ser demandado para o estudo em questão. | Bem fácil de ser utilizada e documentada, mostrando uma rede de conceitos de frequência de palavras clara e objetiva, sem dispêndio de tempo de instalação e entendimento da ferramenta. | Muito fácil de ser utilizada, mas existem algumas limitações como não permitir inserir uma lista de <i>stopwords</i> . | Fácil uso, mas com uma visualização pouco agradável e com algumas limitações, como não aceitar lista de <i>stopword</i> e não exportar o resultado. | Muito simples e fácil de utilizar, no entanto, tem algumas desvantagens como não permitir importar o arquivo com o texto a ser analisado assim como, exportar o resultado, e não permitir inserir uma lista de <i>stopword</i> . | |

ANEXOS

Anexo A – Autorização de acesso aos dados do IBGEANDO


FE.323

TERMO DE CONSENTIMENTO

Autorizo a servidora **Luciana Lopes Monteiro**, siape 1531352, acesso aos dados e informações disponíveis no Portal IBGEANDO, principalmente aos conteúdos postados na seção "Qual a sua Dúvida?", tendo em vista ser essas informações o objeto de estudo do projeto de mestrado "**Mensagens textuais no Canal de Atendimento do Portal IBGEANDO como insumos para a tomada de decisão**", do Curso de Mestrado em Ciência da Informação, da Universidade Federal Fluminense (UFF).

A servidora se comprometerá a manter sigilo sobre todas as informações obtidas e analisadas do IBGEANDO, durante todo o estudo em questão, bem como disponibilizar todos os resultados para que o IBGE possa utilizá-los visando à melhoria na tomada de decisões e comunicação interna.

Rio de Janeiro, 17 de setembro de 2015.



Fernando José de Araújo Abrantes
Diretor-Executivo - IBGE