



MINERAÇÃO DE TEXTOS PARA ORGANIZAÇÃO DE DOCUMENTOS
EM CENTRAIS DE ATENDIMENTO

Maria Luiza Castro Passini

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia Civil, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do Título de Mestre em Engenharia Civil.

Orientador: Nelson Francisco Favilla Ebecken

Rio de Janeiro

Maio de 2012

MINERAÇÃO DE TEXTOS PARA ORGANIZAÇÃO DE DOCUMENTOS
EM CENTRAIS DE ATENDIMENTO

Maria Luiza Castro Passini

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA CIVIL.

Examinada por:

Prof. Nelson Francisco Favilla Ebecken, D.Sc.

Prof. Estevam Rafael Hruschka Júnior, D.Sc.

Prof^a. Beatriz de Souza Leite Pires de Lima, D.Sc.

Prof. Elton Fernandes, Ph.D.

RIO DE JANEIRO, RJ – BRASIL
MAIO DE 2012

Passini, Maria Luiza Castro

Mineração de Textos para Organização de Documentos em Centrais de Atendimento / Maria Luiza Castro Passini. – Rio de Janeiro: UFRJ/COPPE, 2012

XV, 105 p.: il.; 29,7cm.

Orientador: Nelson Francisco Favilla Ebecken

Dissertação (mestrado) – UFRJ/ COPPE/ Programa de Engenharia Civil, 2012

Referências Bibliográficas: p. 95-104.

1. Mineração de Textos. 2. Organização de Textos. 3. Central de Atendimento. I. Ebecken, Nelson Francisco Favilla. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Civil. III. Título.

Dedicatória

Dedico este trabalho, especialmente, ao meu marido Edmundo, maior incentivador dos meus estudos.

Aos meus queridos filhos Eduardo e Henrique que me despertaram o interesse pela área acadêmica, principalmente, pela pesquisa.

Aos meus pais, Marcos e Célia, e irmãos, Marco Aurélio e Antônio, que sempre torceram pelas minhas conquistas profissionais.

À memória de minha querida avó, Mãe Iza, que acompanhou de perto os meus caminhos e sempre demonstrou todo seu carinho e atenção.

Agradecimentos

Agradeço, primeiramente, ao meu orientador, Prof. Nelson Francisco Favilla Ebecken, pela confiança, orientação objetiva, parceria e incentivo na realização deste trabalho. Obrigada pelos seus ensinamentos!

Ao meu querido marido Edmundo, agradeço por toda paciência, amor, compreensão e incentivo permanente. Obrigada por estar sempre ao meu lado!

Agradeço também aos meus familiares, pelo apoio irrestrito, carinho ao cuidar dos meus filhos e compreensão nos momentos em que estive ausente para estudar e pesquisar.

Ao meu querido sogrinho, José Passini, pela revisão do texto deste trabalho, e sugestões que foram importantes para o aperfeiçoamento do mesmo.

À minha amiga Gabi, pela sua amizade e por seu incentivo para que eu iniciasse este curso de mestrado, que desde a graduação tem me acompanhado nas atividades acadêmicas e profissionais.

Aos meus amigos da UFRJ, pelas contribuições e apoio, em especial, à Katiusca Briones pela pesquisa que realizamos e foi aplicada nesse trabalho.

Aos meus amigos e professores da UFJF, Custódio, Alessandreia e Kele, que sempre estiveram ao meu lado em apresentações, salas de aula e trabalhos, trocando experiências e ensinamentos.

Ao Gerente de Tecnologia de Informação, Leonardo Bottino pela confiança e por autorizar o acesso à base de conhecimento, essencial, para realização desse trabalho.

Aos profissionais da Central de Atendimento, pelo esclarecimento de dúvidas e análise dos resultados.

Aos Professores Estevam Rafael Hruschka Júnior, Beatriz de Souza Leite Pires de Lima e Elton Fernandes por aceitarem participar da Banca Examinadora dessa defesa.

Ao CNPq pelo apoio financeiro.

A todos, muito obrigada!

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

MINERAÇÃO DE TEXTOS PARA ORGANIZAÇÃO DE DOCUMENTOS EM CENTRAIS DE ATENDIMENTO

Maria Luiza Castro Passini

Maio/2012

Orientador: Nelson Francisco Favilla Ebecken

Programa: Engenharia Civil

Este trabalho apresenta uma estratégia de Mineração de Textos para organização de conteúdos de documentos e examina o seu desempenho em um estudo de caso de uma Central de Atendimento que presta serviço a uma empresa de petróleo. Em uma Central de Atendimento, a organização e a facilidade na localização de informações são fundamentais para proporcionar maior qualidade e eficiência durante um atendimento. Os procedimentos, normalmente em formato texto, são armazenados em uma Base de Conhecimento, e atualizados manualmente, de acordo com a demanda dos serviços prestados. As técnicas e tarefas de Mineração de Textos foram aplicadas para organizar os documentos da língua portuguesa, a partir de uma base de textos reais. Também são utilizadas duas abordagens recentes: o modelo de tópicos, que representa uma coleção de documentos através de tópicos probabilísticos contendo as palavras-chaves e o mecanismo de seleção imune-supressor, que envolve a seleção dos documentos mais representativos. Os resultados mostram que as palavras mais representativas do modelo de tópicos podem ser comparadas ao melhor resultado obtido pela rotulação automática dos grupos. Além disso, a seleção desses documentos apresenta resultados satisfatórios em grandes bases de documentos.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

TEXT MINING TO ORGANIZATION OF DOCUMENTS
IN CONTACT CENTERS

Maria Luiza Castro Passini

May/2012

Advisor: Nelson Francisco Favilla Ebecken

Department: Civil Engineering

This work presents a Text Mining strategy for the organization of documents contents and examines its performance in a case study of a Contact Center, which provides service to an oil company. In a Contact Center, the organization and facility to find information are essential to provide higher quality and efficiency during a call. The procedures, usually in text format, are stored in a knowledge base and are manually updated, according to the demand of services. The techniques and tasks of Text Mining were applied to organize documents in the portuguese language, from actual textual databases. It also utilizes two recent approaches: the model of topic that represents a collection of documents through probabilistic topics containing the keywords and immune-suppressive mechanism, which involves the selection of the most representative documents. The results show that the most significant words of the type of topic model can be compared to the best result obtained by automatic labeling clusters. Moreover, by choosing these documents achieved satisfactory results with large databases.

SUMÁRIO

1. INTRODUÇÃO	1
1.1. DEFINIÇÃO DO ESCOPO	2
1.2. JUSTIFICATIVA.....	3
1.3. OBJETIVO GERAL	3
1.4. OBJETIVO ESPECÍFICO	4
1.5. CONTRIBUIÇÃO DA PESQUISA.....	4
1.6. ORGANIZAÇÃO DO DOCUMENTO	5
2. TRABALHOS RELACIONADOS	6
2.1. COMENTÁRIO.....	9
3. MINERAÇÃO DE TEXTOS	10
3.1. INTRODUÇÃO.....	10
3.2. PRÉ-PROCESSAMENTO	11
3.2.1. Coleta dos documentos	12
3.2.2. Preparação dos dados	12
3.2.3. Seleção dos dados.....	13
3.2.4. Case folding	13
3.2.5. Stop words	13
3.2.6. Radicalização da Palavra	14
3.2.7. Uso de um Thesaurus	15
3.2.8. Representação de Documentos	15
3.2.9. Modelo LDA.....	18
3.3. PROCESSAMENTO	21
3.4. CLUSTERIZAÇÃO	21
3.4.1. Agrupamento de Partição Total	23
3.4.2. Agrupamento Hierárquico.....	24
3.5. MÉTRICAS DE AVALIAÇÃO DE RESULTADOS DE CLUSTERIZAÇÃO	26
3.6. CLASSIFICAÇÃO	27
3.6.1. Classificação Bayesiana.....	28
3.6.2. Máquinas de Vetor de Suporte	29
3.7. MÉTRICA DE AVALIAÇÃO DE RESULTADOS DE CLASSIFICAÇÃO	30
3.8. PÓS-PROCESSAMENTO.....	31
3.9. COMENTÁRIO.....	32

4. SELEÇÃO DE DADOS DE GRANDES COLEÇÕES DE DOCUMENTOS	33
4.1. INSPIRAÇÃO SISTEMA IMUNE ARTIFICIAL	34
4.2. ALGORITMO SUPRESSOR DE TEXTOS	35
4.3. COMENTÁRIO.....	38
5. FERRAMENTA POLYANALYST.....	39
5.1. Análise dos textos	40
5.2. CLUSTERIZAÇÃO DE TEXTOS	41
5.3. CLASSIFICAÇÃO LINEAR.....	42
5.4. COMENTÁRIO.....	43
6. CENTRAL DE ATENDIMENTO	44
6.1. ESTUDO DE CASO	47
6.2. COMENTÁRIO.....	53
7. METODOLOGIA.....	54
7.1. CLASSIFICAÇÃO DA PESQUISA	54
7.2. RECURSOS UTILIZADOS.....	55
7.3. POPULAÇÃO E AMOSTRA.....	56
7.4. TÉCNICAS DE COLETA DE DADOS	56
7.5. ETAPAS DA MINERAÇÃO DE TEXTOS	57
7.5.1. Base de Documentos	58
7.5.2. Padronização da Coleção.....	59
7.5.3. Preparação dos Documentos	61
7.5.4. Análise Exploratória Inicial.....	62
7.5.5. Geração de Stopt Words, Dicionário, Thesaurus.....	65
7.5.6. Seleção de Atributos	67
7.5.7. RSLP Stemmer	69
7.5.8. Modelo Espaço Vetorial (VSM):.....	71
7.5.9. Alocação Latente <i>Dirichlet</i> (LDA).....	72
7.5.10. Algoritmo Supressor de Textos.....	73
7.6. PROCESSAMENTO	74
7.6.1. Clusterização.....	74
7.6.2. Classificação	77
7.7. PÓS – PROCESSAMENTO	78
7.7.1. Clusterização.....	78
7.7.2. Classificação	79
7.8. COMENTÁRIO.....	80

8. RESULTADOS E ANÁLISES	81
8.1. TAREFA DE AGRUPAMENTO	82
8.2. TAREFA DE CLASSIFICAÇÃO.....	87
8.3. COMENTÁRIO.....	91
9. CONCLUSÃO.....	92
10. REFERÊNCIA.....	95
11. GLOSSÁRIO DE TERMOS.....	105

LISTA DE FIGURAS

Figura 1: Etapas do Processo de Mineração de Textos (Rezende <i>et. al.</i> , 2003)	11
Figura 2: Representação Documento-Termo (LOPES, 2009)	16
Figura 3: Um exemplo de quatro (de 300) tópicos extraídos	18
Figura 4: Comparativo entre os métodos LSA e LDA (STEYVERS et al., 2006)	20
Figura 5: Objetivo do Agrupamento de Informações Textuais (Wives, 1999)	22
Figura 6:Resultado de um agrupamento por partição total e disjunta (Wives, 1999)	23
Figura 7: Agrupamento Hierárquico Aglomerativo (Wives, 1999)	24
Figura 8: Agrupamento Hierárquico Global (Wives, 1999)	25
Figura 9: Exemplo de conjunto não linearmente separável (Lorena e Carvalho, 2003)	29
Figura 10 : Modelo Básico do Funcionamento das Centrais de Atendimento.....	46
Figura 11: Evolução dos Serviços de Comunicação	47
Figura 12: Fluxo de Atendimento	48
Figura 13: Visão Geral da Metodologia aplicada sobre a base de documentos	58
Figura 14: Amostra dos Documentos Representativos	59
Figura 15: Importação de Arquivos no mesmo formato.....	60
Figura 16: Projeto Padronização da Coleção de Documentos	61
Figura 17: Visualização parcial da Base de Documentos Pré-Processada	62
Figura 18: Projeto para análise exploratória geral dos documentos	63
Figura 19: Projeto para análise exploratória dos documentos referente.....	63
Figura 20: Visualização das Palavras Relevantes sem o uso de stop words e dicionários	64
Figura 21: Gráfico Snake para visualização dos termos com maior frequência.....	65
Figura 22: Geração Lista de Stop Words	66
Figura 23: Exemplo do uso do uso de thesaurus a partir dos sinônimos.....	67
Figura 24: Seleção de Atributos.....	68
Figura 25: Relacionamento entre os termos a partir da métrica de Suporte	68
Figura 26: Relacionamento entre os termos a partir da métrica de Tensão	69
Figura 27: Parâmetros do arquivo de configuração do RSLP Stemmer	70
Figura 28: Aplicação do RSLP Stemmer sobre a base de documentos da lingua portuguesa.....	71
Figura 29: Amostra da Matriz Modelo Espaço Vetorial baseado na frequência.....	72

Figura 30: Fluxograma do Algoritmo Supressor de Textos em uma Coleção de Documentos	73
Figura 31: Bases utilizadas nos experimentos da tarefa de Clusterização Hierárquica	74
Figura 32: Configuração do Nó Clusterização de Textos	75
Figura 33: Projeto para Tarefa de Clusterização dos Documentos	76
Figura 34: Bases utilizadas nos experimentos da Tarefa de Classificação dos Documentos	77
Figura 35: Projeto para Tarefa de Classificação dos Documentos	78
Figura 36: Gráfico de visualização da proximidade entre os nove grupos formados a partir do Algoritmo de Clusterização Hierárquica.....	84
Figura 37: Gráfico de visualização da correlação entre cluster e grupos de serviços	85
Figura 38: Percentual de Erro obtido pelo classificador SVM.....	89
Figura 39: Percentual de Erro obtido pelo classificador Bayes	90

LISTA DE TABELAS

Tabela 1: Resultados de Clusterização e Rotulação (NUNES <i>et al.</i> ; 2008).....	21
Tabela 2: Matriz de confusão para duas classes (Rezende, 2003)	30
Tabela 3: Consolidado Indicadores dos Serviços da Central de Atendimento - Abril de 2011	51
Tabela 4: Parâmetros para executar o <i>LDA Gibbs Sampler</i>	72
Tabela 5: Características das Bases de Documentos para a tarefa de Clusterização.....	81
Tabela 6: Consolidado Estatístico da Tarefa de Clusterização	82
Tabela 7: Relação entre o grupo de serviço e o cluster mais representativo.....	85
Tabela 8: Características das Bases de Documentos para a tarefa de Classificação	88
Tabela 9: Resultado dos Algoritmos <i>Naïve Bayes</i> e SVM.....	88
Tabela 10: Tempo de Processamento obtido pelos Classificadores <i>Naïve Bayes</i> e SVM.....	90

LISTA DE QUADROS

Quadro 1: Consolidada com Agentes Multi SKILL	50
Quadro 2: Descrição dos Grupos gerado pelo Algoritmo de Clusterização Hierárquica a partir da base Pré-Processada	83
Quadro 3: Descrição dos Grupos gerado pelo Algoritmo de Clusterização Hierárquica	83
Quadro 4: Descrição dos Grupos gerado pelo Algoritmo de Clusterização Hierárquica a partir da base SeleSupText utilizando a relevância dos termos	83
Quadro 5: Descrição dos Grupos gerado pelo Algoritmo de Clusterização Hierárquica a partir da base SeleSupText utilizando a frequência dos termos	83
Quadro 6: Palavras-Chaves dos Grupos Finais selecionados pelo	86
Quadro 7: Tópicos extraídos da base SeleSup Text utilizando o	86

LISTA DE SÍMBOLOS

AMS: Assistência Multidisciplinar de Saúde
BOW: Saco de Palavras (do inglês, bag of words)
CC: Centro de Chamadas
CCC: Centro de Contato com Consumidores
COFIP: Centro de Operações Financeiras da Petrobras
CTI: Integração Computador Telefone
DAC: Distribuidor Automático de Chamadas
IS: Seleção de Instâncias
LDA: Alocação Latente de *Dirichlet*
ME: Mesa Especializada
MT: Mineração de Textos
PA: *PolyAnalyst*
PABX: Troca Automática de Ramais Privados
PA: Posto Avançado
SIA: Sistemas Imunológicos Artificiais
SAP: Sistema, Aplicações e Produto
SI: Segurança da Informação
SMS: Segurança, Meio Ambiente e Saúde
SVM: Máquinas de Vetor de Suporte
TELECOM: Telecomunicações
TEP: Dicionário Eletrônico para o Português
URA: Unidade de Resposta Audível
VoIP: Voz sobre Protocolo de Internet
VSM: Modelo de Espaço Vetorial
WEB: Rede de Alcance Mundial

1. INTRODUÇÃO

Na Era da Informação, onde a riqueza nasce de idéias inovadoras e do uso inteligente da informação, grande parte da informação eletrônica está disponível em base de dados conhecidas como bases não-estruturadas, ou seja, banco de dados textuais.

O avanço das tecnologias para aquisição e armazenamento de dados tem permitido que o volume de informação gerado em formato digital aumente de forma significativa nas organizações. Cerca de 80% dos dados das empresas se encontram em formatos textuais tais como: cartas de cliente, e-mails, contratos, documentação técnica, notícia, artigos e páginas na internet (TAN, 1999).

As empresas do Século XXI tentam explorar o seu conhecimento para se diferenciar e obter vantagens competitivas sobre os concorrentes (BERNARD e TICHKIEWITCH, 2008). O conhecimento passa a ser considerado o grande diferencial competitivo das organizações. Nesse contexto, as empresas devem ser capazes de obter rapidamente (antes de seus concorrentes) a maior quantidade possível de informações relevantes. Para tanto, devem ser utilizadas técnicas de coleta de informação capazes de selecionar somente as informações pertinentes às empresas.

Essas informações devem ser distribuídas ou repassadas a pessoas específicas da empresa, que expressam sua necessidade e que sabem dar utilidade a ela, transformando-a em ação, produção e competitividade. Para (RIZZI, 2000) quanto melhor o acesso, a disponibilidade e a qualidade da informação, maior é a chance de acerto no processo de tomada de decisão e maior a competitividade das organizações capazes de utilizá-la.

Diante da imensa quantidade de informação disponível em formato textual, a Mineração de Texto engloba técnicas e ferramentas inteligentes, a fim de satisfazer objetivos estratégicos e atender às necessidades do negócio das empresas.

A Mineração de Textos também conhecida como Descoberta de Conhecimento Textual (FELDMAN e DAGAN, 1995) ou Mineração de Dados Textuais (HEARST, 1998), refere-se ao processo de extrair informações úteis e inovadoras em documentos no formato textual não-estruturado através da identificação de conhecimento e exploração de padrões.

Dentre as principais técnicas de Mineração de Textos, as tarefas de Clusterização e Classificação recebem especial atenção na literatura. A Clusterização consiste em agrupar automaticamente os documentos em grupos de acordo com a similaridade. A Classificação, também chamada Categorização de Textos, é utilizada para classificar um conjunto de documentos em uma ou mais categorias ou classes existentes.

No decorrer do trabalho serão apresentados e avaliados alguns modelos e técnicas utilizados na área de recuperação de informações: o método Alocação Latente de *Dirichlet* (LDA), para alocar os termos mais representativos em tópicos (BLEI *et al*, 2003); a nova versão do RSLP *Stemmer* (COELHO, 2007), para radicalização das palavras na língua portuguesa e uma recente proposta para selecionar os documentos mais representativos inspirado no mecanismo supressor de FIGUEREDO *et al*. (2012).

Este trabalho tem como objetivo explorar o processo de descoberta de conhecimento e utilizar as técnicas de Mineração de Textos em uma base de dados textuais reais de uma Central de Atendimento que presta serviço para uma empresa de petróleo.

A Central de Atendimento, também conhecida como *Contact Center*, utiliza a base de dados textuais, que contém todos os documentos necessários para dar suporte ao atendimento. Diante da multiplicidade de ferramentas e especialidade de cada serviço, esses documentos são atualizados, periodicamente, e disponibilizados para consulta via web, através do Portal de Atendimento.

Partindo da premissa de que esses documentos devem ser facilmente localizados e identificados de acordo com o serviço especializado, pretende-se apresentar uma nova forma de organização desses documentos.

1.1. DEFINIÇÃO DO ESCOPO

As Centrais de Atendimento dispõem de um grande volume de informação não estruturada disponível para consulta através do Portal de Atendimento. De acordo com a necessidade do serviço prestado, esses documentos são atualizados periodicamente.

O crescimento periódico desses documentos dificulta sua análise manual, controle, identificação, localização e acesso. Torna-se necessário a organização automática dos documentos disponíveis por serviço para dar suporte ao atendimento quando necessária.

O uso de palavras-chaves também auxilia o processo de busca em uma coleção. Desta forma identifica-se a necessidade de identificar as palavras mais representativas de um grupo de documentos similares, para facilitar essa consulta.

Além disso, trabalhar com grandes bases de documentos implica em trabalhar com matrizes esparsas e de alta dimensionalidade, que exigem um alto tempo de processamento. Desta forma, faz-se necessário a reduzir o número de documentos para viabilizar o uso das técnicas de Clusterização e Classificação.

1.2. JUSTIFICATIVA

A cada dia que passa, as Centrais de Atendimento estão se tornando parte integrante para a maioria das organizações. Essas Centrais dispõem de uma ampla gama de serviços de atendimento.

De acordo com a demanda de um serviço, identifica-se a necessidade de criar meios que possam prover de forma rápida e eficiente a organização dos documentos e facilitar a consulta dos mesmos. No mercado altamente competitivo, qualidade e produtividade são fatores fundamentais para garantir o sucesso.

A extração de conhecimento a partir desses documentos pode ser usada como vantagem competitiva e suporte à tomada de decisão de uma Central de Atendimento.

1.3. OBJETIVO GERAL

Aplicação de técnicas e tarefas de Mineração de Textos para extrair conhecimento útil e possibilitar uma nova forma de organização, a partir de uma base de documentos reais.

1.4. OBJETIVO ESPECÍFICO

Aplicar técnicas de Mineração de Textos em um conjunto de dados textuais nas Centrais de Atendimento de forma a:

- Efetuar o estudo de caso com a utilização da metodologia proposta;
- Explorar a base de documentos sem conhecimento prévio de um especialista de domínio;
- Obter atributos que sejam candidatos a termos do domínio de conhecimento da coleção, selecionando palavras mais significativas na coleção;
- Avaliar a organização de informações de acordo com o conteúdo dos serviços;
- Comparar os grupos encontrados com as ilhas especializadas no atendimento;
- Avaliar a lista de termos representativos (descritores) de cada grupo de serviços;
- Avaliar os algoritmos de Clusterização e LDA para otimização do processo;
- Avaliar a seleção dos documentos mais representativos
- Avaliar o desempenho dos classificadores *Naïve Bayes* e Máquinas de Vetor de Suporte (SVM).

1.5. CONTRIBUIÇÃO DA PESQUISA

Visando alcançar o seu objetivo principal, a partir de uma base de documentos reais, este trabalho pretende proporcionar aos gestores de uma Central de Atendimento, um novo método para organizar automaticamente os documentos, extrair novas palavras-chaves que podem ser utilizadas como consulta no Portal de Atendimento e acompanhar a atualização periódica desses documentos, considerados essenciais para oferecer um atendimento de qualidade.

Com o crescente volume de informações, foi apresentada uma nova abordagem para selecionar instâncias (SI) a partir de bases de dados textuais, envolvendo duas áreas distintas: a Mineração de Textos e o Sistema Imunológico Artificial.

O algoritmo foi especialmente desenvolvido para problemas de classificação de dados estruturados (FIGUEREDO et al., 2012), e nesse trabalho foi avaliado em problemas relacionados à tarefa de clusterização. Uma característica inovadora do algoritmo SeleSupText é selecionar os documentos mais representativos em uma grande coleção de documentos. Esse algoritmo demonstrou resultados satisfatórios em documentos da língua portuguesa e com o crescente volume de informações, torna-se essencial extrair informação útil de forma precisa em um menor tempo computacional.

Por fim, a apresentação de uma nova forma de representação de documentos por meio de tópicos utilizando o modelo LDA (*Latent Dirichlet Allocation*), que de forma rápida e eficiente apresenta a ocorrência de cada palavra em cada tópico de acordo com o conteúdo.

1.6. ORGANIZAÇÃO DO DOCUMENTO

Nos capítulos seguintes estão discorridos os temas que permitem fundamentar e testar a aplicação da Mineração de Textos e estão estruturados da seguinte forma:

O Capítulo 2 apresenta os principais trabalhos relacionados ao uso da Mineração, em documentos da língua portuguesa e para extração do conhecimento em Centrais de Atendimento.

No Capítulo 3 é apresentada uma revisão de literatura referente aos temas da Mineração de Textos, com ênfase nas tarefas abordadas: Modelo de Tópicos, Clusterização e Classificação. A nova abordagem para seleção de dados de grande coleções de documentos é apresentada no Capítulo 4.

O capítulo 5 aborda a ferramenta *PolyAnalyst* utilizada no desenvolvimento desse trabalho. O capítulo 6 apresenta uma visão geral sobre o funcionamento de uma Central de Atendimento e as características da base de documentos, utilizada como Estudo de Caso.

A metodologia de pesquisa proposta para trabalhar com grandes volumes de documentos é detalhada no Capítulo 7. No Capítulo 8, são apresentados os resultados e análises dos experimentos. Por fim, o Capítulo 9 apresenta as conclusões do trabalho e sugestões para pesquisas posteriores.

2. TRABALHOS RELACIONADOS

A literatura mostra que a utilização das técnicas e tarefas de Mineração de Textos atrai o interesse dos pesquisadores diante da enorme quantidade de informação em formato textual, armazenadas pelas organizações.

O uso de uma metodologia para aplicar os processos da Mineração de Textos é apresentado em (YANG, *et. al.*, 2009; NUNES, *et. al.*, 2008; NOGUEIRA, *et. al.*, 2008) Esses trabalhos demonstram a importância de cada uma das etapas, desde o pré-processamento até a utilização do conhecimento.

Os trabalhos relacionados referem-se à aplicação das técnicas de clusterização e classificação em documentos, especialmente, na língua portuguesa.

O estudo do uso da Mineração de Textos para extração e organização não supervisionada de conhecimento, permitindo-se a construção de hierarquias de tópicos de forma automática é realizado por (REZENDE *et al*, 2011; MARCACINI, 2011). Os resultados mostraram que a organização automática de resultados de busca em grupos de temas específicos facilita a exploração das páginas retornadas por uma máquina de busca.

Com a crescente disponibilidade de documentos na WEB, BASTOS (2006) descreve o desenvolvimento e implementação de um sistema que utiliza técnicas de mineração de textos sobre documentos disponíveis na WEB para língua portuguesa. A pesquisa foi considerada satisfatória a partir dos testes realizados sobre documentos em *sites* da WEB.

A aplicabilidade e eficiência do uso de clusterização para textos em língua portuguesa também podem ser encontrados em (LOPES, 2004). Os resultados indicaram que a precisão obtida pelo processo de clusterização está relacionada à qualidade dos dados, ou seja, a falta de atributos em comum nos dados dificulta a identificação de semelhanças entre os documentos.

Para complementar a pesquisa, (OLIVEIRA, 2009) ressalta a importância da combinação das análises semântica e estatística, quando se trata de Mineração de Textos. Em seu trabalho, a autora sugere a utilização de mais de uma técnica num mesmo contexto para apresentar resultados mais relevantes.

Alguns trabalhos evidenciam a aplicabilidade da extração do conhecimento em Centrais de Atendimento.

A partir do uso de algoritmos e ferramentas de mineração (LOPES, 2009) identifica grupos, suas respectivas palavras-chave e as respectivas correlações entre termos representativos. A aplicação da Mineração de Textos, em um problema real, permitiu oferecer conhecimentos sobre a opinião de consumidores de acordo com o espaço geográfico e grupos sociais.

A partir das reclamações dos clientes armazenadas em uma base de dados contendo os registros de ligações, CAPUTO *et al.* (2006) apresenta os benefícios da Mineração de Textos para compreender o perfil dos clientes da empresa e identificar os clientes em potencial para oferta de novos produtos e serviços.

No mesmo contexto, (SCHIESSL, 2007) analisa um Serviço de Atendimento ao Consumidor que centraliza em forma textual, os questionamentos, as reclamações, os elogios e as sugestões, verbais ou escritas dos clientes.

A análise dos registros de atendimento aos usuários de Tecnologia da Informação de uma grande organização é realizada por (CORREA, 2007). Estes registros de solicitação de atendimento são criados, na maioria dos casos, pelos técnicos do *Help-desk* com apoio de uma ferramenta de workflow que gera as bases de dados. O resultado final é apresentado utilizando-se a técnica de mineração de dados denominada regras de associação. Estas regras irão apresentar a relação entre fatos que a princípio não apresentavam relação entre si, revelando conhecimento novo.

O uso da ferramenta *PolyAnalyst* para analisar os registros de um *Call Center* é apresentado em (HVALSHAGEN, 2002; FROELICH *et al.*, 2004). Em um domínio desestruturado, sugere-se um estudo exploratório dos dados. A ferramenta é utilizada, em ambos trabalhos, para estruturar os documentos e desenvolver suporte para várias tomadas de decisões. Além disso, demonstra o potencial da Mineração de Textos para extrair palavras-chave, analisar correlações e utilizar a taxonomia para compreensão dos resultados finais.

Uma importante contribuição na área de redução de dados estruturados (data mining) pode ser encontrada em (CANO *et al.*, 2003). Neste trabalho, os autores apresentam uma revisão dos principais algoritmos de seleção de instância. Além disso, compara o desempenho dos métodos clássicos de seleção de instância com as principais estratégias evolutivas: o algoritmo de gerações genética GGA (GOLDBERG, 1989; HOLLAND, 1975), o algoritmo de estado estacionário genético (ASG) (WHITLEY, 1989), e o algoritmo genético CHC de busca adaptativo (CANO *et al.*, 2003.).

Os autores concluíram que dentre os algoritmos avaliados, o CHC (*Cross generational elitist selection, heterogeneous recombination and Cataclysmic mutation*) é mais robusto e oferece maior percentual de redução sem *overfitting*, apesar do elevado tempo de execução. Além disso, os algoritmos clássicos e evolutivos são afetados quando o tamanho do conjunto de dados aumenta.

2.1. COMENTÁRIO

Direcionou-se a pesquisa para o uso da Mineração de Textos em uma coleção de documentos da língua portuguesa e sua aplicação em bases reais, utilizadas pelas Centrais de Atendimento.

Nota-se que a etapa de pré-processamento demanda especial atenção para garantir a qualidade dos dados. A atualização da lista de *stop words*, a seleção de atributos e a complexidade dos problemas relacionados à alta dimensionalidade dos dados são evidenciados em grande parte dos trabalhos.

3. MINERAÇÃO DE TEXTOS

3.1. INTRODUÇÃO

A partir dos anos 80, com o aumento exponencial do volume de informações armazenadas pelas empresas, surgiu a necessidade de extração, de maneira automatizada, de informações relevantes e de padrões de comportamento em grandes massas de dados (LOPES, 2009).

A Mineração de Textos ou *Text Mining*, também conhecida como Descoberta de Conhecimento em Textos (*Knowledge Discovery in Text - KDT*) surgiu com a finalidade de tratar os dados e as informações não-estruturadas considerando o alto nível de complexidade envolvida neste tipo de representação de informação.

Em um contexto o qual grande parte da informação corporativa é registrada em linguagem natural, a Mineração de Textos (MT) surge como poderosa ferramenta para gestão do conhecimento (EBECKEN *et al.*, 2003).

A Mineração de Textos é considerada uma especialização do processo de mineração de dados, sendo a principal diferença o tipo de dados com que se deseja trabalhar, ou seja, dados estruturados ou dados não-estruturados (REZENDE *et al.*, 2003). Assim, o grande desafio é obter alguma estrutura que represente os textos e então, a partir dessa, extrair conhecimento (WEISS *et al.*, 2005).

O processo de Mineração de Textos é representado por cinco grandes etapas: identificação do problema, pré-processamento, extração de padrões, pós-processamento e utilização do conhecimento (REZENDE *et al.*, 2003).

A Figura 1 apresenta as etapas que formam um ciclo no qual, ao final do processo de MT, obtém-se o conhecimento acerca dos dados analisados.



Figura 1: Etapas do Processo de Mineração de Textos (Rezende *et. al.*, 2003)

Neste trabalho, foram consideradas essenciais as cinco etapas, desde a identificação do problema até a avaliação do conhecimento extraído, para que fosse possível aplicar a teoria em uma coleção de dados textual real.

3.2. PRÉ-PROCESSAMENTO

Essa etapa possui uma relevância fundamental no processo de descoberta de conhecimento. As atividades de obtenção e limpeza dos dados normalmente consomem mais da metade do tempo dedicado ao processo como um todo. Porém, o tratamento inicial dos dados confere maior consistência a eles e pode evitar a obtenção de resultados distorcidos.

Para a obtenção dos conceitos (NASUKAWA e NAGANO, 2001) faz-se necessário a aplicação de uma série de técnicas que permitam eliminar dentro de cada texto as redundâncias e/ou variações morfológicas. O objetivo dessa etapa consiste em transformar o conjunto de documentos em uma base mais limpa, onde o trabalho de representação de documentos, o respectivo processamento dos dados e a consequente interpretação destes, possam ser feitas de maneira mais rápida e eficiente.

A seguir são descritos alguns conceitos e técnicas normalmente utilizados em processos de preparação de dados para a efetiva mineração de dados textuais:

3.2.1. Coleta dos documentos

A primeira tarefa a ser realizada é a coleta dos documentos realmente relevantes para compor a base de textos do trabalho. O principal objetivo é garantir um material de qualidade para aquisição de conhecimento.

Entretanto, o volume de documentos disponíveis para o estudo pode representar um problema dependendo da técnica que se deseja trabalhar.

3.2.2. Preparação dos dados

O manuseio de arquivos texto apresenta alguns desafios. O primeiro a ser citado envolve o próprio formato dos textos com nenhuma ou pouca estruturação, o que dificulta a utilização imediata de várias técnicas de mineração de dados conhecidas.

Um outro desafio diz respeito ao tamanho dos arquivos em formato texto, comumente da ordem de milhares de palavras ou termos. Além disso, muitas dessas palavras são repetidas, expressam o mesmo significado ou são de significado irrelevante. As situações mencionadas acima, assim como outras encontradas quando se lida com dados textuais, devem ser trabalhadas e resolvidas para viabilizar o uso de arquivos texto em primeira instância e, em segunda, aumentar a eficiência de atividades executadas *a posteriori*.

A preparação dos textos é a primeira etapa do processo de descoberta de conhecimento em textos. Essa etapa envolve a seleção dos documentos que constituirão os dados de interesse, ou seja, toda a informação que não refletir nenhuma idéia considerada importante poderá ser desprezada.

Além de promover uma redução dimensional, esta etapa tenta identificar similaridades em função da morfologia ou do significado dos termos (LOPES, 2004).

3.2.3. Seleção dos dados

Selecionar dados significa identificar e segregar o que realmente será relevante para o processo de extração de padrões, como, por exemplo, a escolha das bases de dados e seus respectivos registros e atributos.

Assim, o primeiro passo é determinar a fonte primária de dados a ser utilizada. Diante da definição do problema a ser resolvido, escolhem-se os bancos de dados que serão alvos da pesquisa. Em muitos casos, as fontes de dados poderão estar em formatos diferentes, como planilhas eletrônicas, bancos de dados até mesmo em *Data Warehouse*. Para iniciar a etapa de mineração de dados, todas as fontes deverão estar reunidas em uma única base de dados ou arquivo. A maior diversidade das fontes de dados e dos formatos destes dados implicará em um maior esforço para reunir todas as fontes de dados.

Após a unificação dos dados em uma base única, selecionam-se os registros dos arquivos. Em nosso Estudo de Caso, cada registro de um arquivo representa um procedimento. Supondo a análise dos dados de uma Base de Conhecimento, por exemplo, essa seria o conjunto de todos os procedimentos utilizados para oferecer um atendimento de acordo com a solicitação.

3.2.4. Case folding

É o procedimento de converter todos os caracteres de um documento para um único tamanho, maiúsculo ou minúsculo de forma a conferir maior agilidade na análise dos dados através do processo de indexação (CAPUTO *et al.*, 2006).

3.2.5. Stop words

No processo de análise dos dados identificam-se palavras com baixa frequência, que podem ser removidas, pois nada acrescentam à representatividade da coleção ou que sozinhas nada significam, tais como preposições, pronomes, artigos e advérbios.

De maneira análoga, existe uma relação entre a frequência das palavras e sua importância para o entendimento do contexto das informações. As palavras com frequência muito elevada na base de dados analisada podem ser descartadas, pois não agregam valor ao entendimento do(s) texto(s) analisado(s). (MANNING e SCHÜTZE, 1999)

Tanto os casos de frequência excessiva, como os de frequência muito baixa, podem ser considerados como casos de *Stop Words*, que precisam desta forma ser eliminados no processo de análise de dados.

A eliminação de *Stop Words* reduz significativamente a quantidade de termos, diminuindo o custo computacional das próximas etapas (MANNING *et al.*, 2008).

3.2.6. Radicalização da Palavra

Stemming é uma técnica de redução de termos a um radical comum, a partir da análise das características gramaticais dos elementos, como grau, número, gênero e desinência. Tem o objetivo de retirar os sufixos e prefixos das palavras, e encontrar a sua forma primitiva. Assim, as palavras no plural ou derivadas são reduzidas a um radical único, à sua raiz, simplificando a representação dos termos envolvidos no documento. Isso implica numa única entrada nos índices, aumentando o desempenho do processo.

Os dois erros que costumam ocorrer durante o processo de *stemming* podem ser divididos em dois grupos: *Overstemming* ocorre quando é removido não só o sufixo, mas também uma parte do radical, e *Understemming* quando um sufixo não é removido, ou é apenas reduzido parcialmente. Isto geralmente causa uma falha na confluência de palavras relacionadas, causando a não recuperação de documentos que seriam relevantes. Um desafio corrente aqui é configurar os parâmetros dos algoritmos que executam essa tarefa a fim de que essas distorções sejam evitadas.

Dentre os algoritmos de *Stemming* desenvolvidos para a língua inglesa, mais comumente utilizados em aplicações de Mineração de Textos, destacam-se: *Stemmer S*, Porter (PORTER, 1980) e LOVINS (LOVINS, 1968). Porém, como os algoritmos de *stemming* são dependentes das línguas para os quais foram escritos, há necessidade de se utilizar um *stemmer*, especialmente, projetado para o processamento de palavras escritas em português.

Destacam-se três algoritmos propostos para a língua portuguesa: a versão para português do algoritmo de PORTER (2005), o Removedor de Sufixo da língua Portuguesa (RSLP), proposto por ORENGO e HUYCK (2001) e o algoritmo STEMBR, proposto por ALVARES et al., (2005). Estudos compararam o desempenho desses três algoritmos e consideraram o algoritmo RSLP mais eficiente, por cometer um número menor de erros de *overstemming* e de *understemming* (ORENGO, HUYCK, 2001). A nova implementação do RSLP, proposta por COELHO (2007), foi utilizada nesse trabalho.

3.2.7. Uso de um Thesaurus

Um *Thesaurus* pode ser definido como um dicionário controlado que representa hierarquias, abreviações, sinônimos, acrônimos, ortografias alternativas e relacionamentos associativos entre termos, com o intuito maior de apoiar os usuários na recuperação das informações requeridas (LOPES, 2004).

Assim como ocorre na abordagem da técnica de *Stemming*, no *Thesaurus* vários termos são mapeados para um termo conceito único, que expressa a idéia geral dos elementos. Em outras palavras, pode-se dizer que diferentes usuários costumam definir a mesma consulta através de termos distintos.

Como o conhecimento sobre o domínio dos documentos é fundamental para a elaboração de um *Thesaurus*, optou-se por analisar a lista de sinônimos criada por especialistas da Central de Atendimento como *thesaurus*.

Neste trabalho, utilizou-se o conjunto de sinônimos (*synset*) do dicionário eletrônico de sinônimos para o português do Brasil (doravante, PB), denominado TeP 2.0 desenvolvido por MAZIERO *et al.* (2008).

3.2.8. Representação de Documentos

Uma vez selecionados os termos mais representativos da coleção textual, deve-se buscar a estruturação dos documentos, de maneira a torná-los processáveis pelas tarefas de Mineração de Dados.

O modelo mais utilizado para representação de dados textuais é o modelo espaço-vetorial, proposto originalmente por SALTON (1975), no qual cada documento é um vetor em um espaço multidimensional, e cada dimensão é um termo da coleção (FELDMAN e SANGER, 2006).

Este tipo de representação é também conhecido como BOW (*Bag of Words*). A bag-of-words é uma tabela documento-termo, como ilustrado na Figura 2.

Termo/ Documento	Term 1	Term 2	Term 3	Term 4	Term 5	...	Term n-1	Term n
Doc 1	0	0	1	1	0	...	0	0
Doc 2	0	1	0	0	1	...	0	1
Doc 3	0	1	0	0	0	...	1	1
Doc 4	1	0	1	0	1	...	1	1
Doc 5	1	0	1	0	1	...	0	1
Doc 6	1	1	1	1	0	...	1	0
Doc 7	1	1	0	1	1	...	0	0
...
Doc n-1	1	0	0	1	0	...	1	1
Doc n	0	1	0	0	1	...	0	0

Figura 2: Representação Documento-Termo (LOPES, 2009)

Seguindo esta linha de raciocínio, termos que possuem alta frequência têm importância maior em um documento. Por sua vez, se esses termos aparecem em uma grande quantidade de documentos, terão sua importância diminuída. Desta forma, uma importante tarefa para a caracterização da coleção de documentos examinados é a análise da frequência de palavras na coleção. Este tipo de análise permite verificar a importância de cada palavra em relação aos textos analisados, bem como permite a distinção dos documentos entre si.

De forma a diferenciar os termos em função de sua importância relativa, são atribuídos pesos aos termos. Os valores dos elementos de um vetor - representação de cada documento através das respectivas frequências de cada termo - são calculados como sendo a combinação das estatísticas $TF(t,d)$ e $DF(t)$. Sendo $TF(t,d)$ o número de vezes que um termo “t” ocorre em um documento “d”, enquanto $DF(t)$ é o número de documentos em que tal termo aparece pelo menos uma vez.

Uma maneira usual de se atribuir estes pesos adequadamente é utilizar o cálculo da Frequência Inversa de um Documento, que é calculada pela Equação 1:

$$\text{IDF} = \log \left(\frac{|D|}{DF(t)} \right) \quad \text{Equação 1}$$

Onde: |D| é o número total de documentos.

Pela simples observação da fórmula, pode-se verificar que, quando a IDF de uma palavra ou termo é baixa, este termo ocorre em muitos documentos. Conseqüentemente IDF tem seu valor máximo quando o termo aparece em apenas um único documento.

A estatística TF-IDF combina as duas anteriores:

$$\text{TF-IDF}(t) = \text{TF}(t,d) \cdot \text{IDF}(t) \quad \text{Equação 2}$$

Desta forma de atribuição de peso decorre que: se um termo T_i ocorre frequentemente no documento, este é um termo importante para a identificação deste documento. De outra forma, palavras que ocorrem em muitos documentos são pouco importantes para indexação. Em suma, pode-se concluir que a IDF serve como uma função de ajuste, de modulação da frequência de termos relevantes em um documento.

A representação por meio da tabela documento-termo será empregada em duas técnicas de extração de conhecimento a ser discutidas nesse trabalho: mecanismo de supressão SeleSup (FIGUEREDO *et al.*, 2012) e modelo LDA (BLEI *et al.* (2003).

Deve-se ressaltar que essa etapa de Pré-Processamento pode ser redefinida e então repetida após as próximas etapas, uma vez que a descoberta de alguns padrões pode levar a estabelecer melhorias a serem empregadas sobre a tabela documento-termo, como: ponderar a importância de cada termo ou até mesmo refinar a seleção dos termos (REZENDE *et al.* 2003)

Por fim, para calcular a similaridade entre os documentos, primeiramente são atribuídos pesos que especificam o tamanho e a direção de seu vetor de representação. Para o cálculo da similaridade propriamente dita é calculado o ângulo cosseno formado entre os vetores de dois documentos.

Assim, um valor de cosseno igual a zero indica que os vetores são ortogonais e, portanto, não têm qualquer similaridade, ou seja, os documentos não compartilham nenhum termo em comum. Por outro lado, se o valor da proximidade for próximo de 1, os documentos compartilham termos e são similares.

Outros métodos relevantes para verificação de similaridade são: SWC (*Shared Word Count*), que se baseia na quantidade de palavras comuns entre documentos (WITTEN, 2004), e a Distância de Jaccard (KONCHADY, 2006), que representa o grau de união entre dois documentos.

3.2.9. Modelo LDA

Seguindo a abordagem vetorial para a representação de textos em corpora, recentes propostas de melhorias ao modelo original têm obtido sucesso nas últimas décadas, com a aplicação de sofisticados métodos estatísticos (HOFMANN, 1999; BLEI *et al.*, 2003; GRIFFITHS *et al.*, 2004). Dentre essas propostas, destaca-se o modelo probabilístico LDA de BLEI *et al.* (2003), cujo acrônimo significa *Latent Dirichlet Allocation*, que representa os documentos como uma mistura de tópicos latentes. Cada tópico contém uma lista de palavras mais representativas conforme Figura 3.

Topic 247		Topic 5		Topic 43		Topic 56	
word	prob.	word	prob.	word	prob.	word	prob.
DRUGS	.069	RED	.202	MIND	.081	DOCTOR	.074
DRUG	.060	BLUE	.099	THOUGHT	.066	DR	.063
MEDICINE	.027	GREEN	.096	REMEMBER	.064	PATIENT	.061
EFFECTS	.026	YELLOW	.073	MEMORY	.037	HOSPITAL	.049
BODY	.023	WHITE	.048	THINKING	.030	CARE	.046
MEDICINES	.019	COLOR	.048	PROFESSOR	.028	MEDICAL	.042
PAIN	.016	BRIGHT	.030	FELT	.025	NURSE	.031
PERSON	.016	COLORS	.029	REMEMBERED	.022	PATIENTS	.029
MARIJUANA	.014	ORANGE	.027	THOUGHTS	.020	DOCTORS	.028
LABEL	.012	BROWN	.027	FORGOTTEN	.020	HEALTH	.025
ALCOHOL	.012	PINK	.017	MOMENT	.020	MEDICINE	.017
DANGEROUS	.011	LOOK	.017	THINK	.019	NURSING	.017
ABUSE	.009	BLACK	.016	THING	.016	DENTAL	.015
EFFECT	.009	PURPLE	.015	WONDER	.014	NURSES	.013
KNOWN	.008	CROSS	.011	FORGET	.012	PHYSICIAN	.012
PILLS	.008	COLORED	.009	RECALL	.012	HOSPITALS	.011

Figura 3: Um exemplo de quatro (de 300) tópicos extraídos do corpus TASA (STEYVERS *et al.*, 2006)

O algoritmo LDA, calcula a probabilidade de determinada palavra w_i ocorrer em determinado tópico z_k , denominada $P(w_i|z_k)$, com base numa distribuição *Dirichlet*, tratando a probabilidade de um documento d_j ocorrer em determinado tópico z_k , denominada $P(z_k|d_j)$, como variáveis aleatórias. O principal objetivo é maximizar a seguinte probabilidade pela Equação 3 :

$$p(w|\theta, \alpha, \beta) = \int \sum_z p(w|z, \theta) p(z, \theta) p(\theta, \alpha) p(\phi, \beta) d\theta \quad \text{Equação 3}$$

Onde w é o número de palavras do documento, θ e ϕ são parâmetros sobre os tópicos e as palavras, respectivamente, $p(\theta|\alpha)$ e $p(\phi|\beta)$ são distribuições *Dirichlet* parametrizadas pelos hiperparâmetros α e β . Segundo os autores ZISSERMAN *et al.* (2005) a integral descrita acima é insolucionável, sendo calculada utilizando o algoritmo *Gibbs Sampling* (GRIFFITHS e STEYVERS, 2004). Esse algoritmo foi utilizado no presente trabalho para identificar os tópicos considerando as palavras mais frequentes em cada tópico.

Com isso, calculam-se todas as probabilidades com base em hiperparâmetros, que depois podem ser manipulados numa etapa posterior para se atingir a taxa de acerto desejada. Essa flexibilidade evita que um número excessivo de classes semanticamente similares, porém distintas, e uma taxa de acurácia inferior. (ZISSERMAN *et al.*, 2005).

A Figura 4 apresenta um comparativo entre a Análise Semântica Latente (LSA), que utiliza a decomposição em valores singulares (*Singular Value Decomposition*), e o modelo de tópicos, que utiliza inferência estatística para reduzir a dimensionalidade da matriz gerada pelo Modelo Espaço Vetorial.

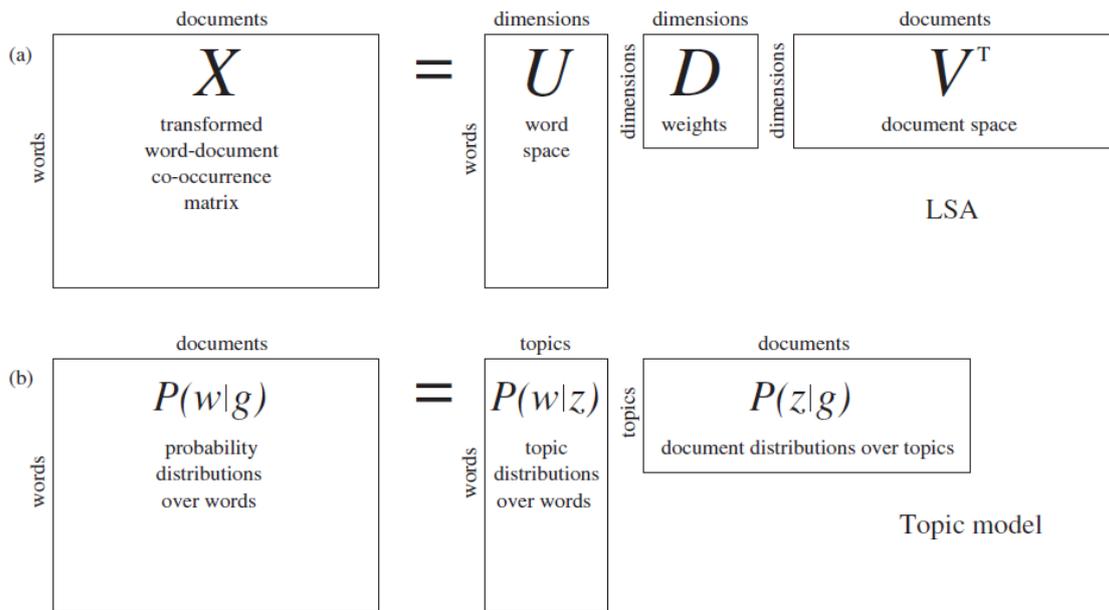


Figura 4: Comparativo entre os métodos LSA e LDA (STEYVERS et al., 2006)

No modelo LDA, a probabilidade de distribuição das palavras em cada tópico é modelada como um conjunto de tópicos probabilísticos. Assim, o modelo de tópicos fornece uma nova forma de representação de um documento para reduzir a dimensionalidade das matrizes geradas na etapa de pré-processamento.

A literatura apresenta alguns métodos com bons resultados para definir as palavras representativas em cada grupo, mas dependentes dos algoritmos de clusterização.

A tarefa de identificar as palavras mais discriminativas em cada grupo, também conhecida como rotulagem de agrupamentos, pode ser vista como um problema de seleção de atributos (REZENDE *et al.*, 2003).

A solução proposta por POPESCU e UNGAR (2000) utiliza as frequências de cada palavra em cada grupo, tomando-se decisões em relação a cada palavra do dendograma (da raiz para as folhas). Entretanto, os autores ressaltam que a rotulagem utilizando as palavras mais frequentes nos grupos, acabam utilizando muitas palavras que são praticamente nulas de poder descritivo. Nesse método é utilizado o estimador X^2 , que não é adequado para dados esparsos.

Na rotulagem pressupõe-se a existência de uma hierarquia de grupos de documento (Secção 3.4.2). A tarefa é atribuir um bom descritor para cada nó de grupo na hierarquia. Os descritores de grupos mais comuns são rótulos concisos, ou listas

de termos e frases (REZENDE *et al.*, 2003). A Tabela 1 apresenta alguns resultados das tarefas de clusterização e rotulação.

Tabela 1: Resultados de Clusterização e Rotulação (NUNES *et al.*; 2008)

Cluster 1	Cluster 2	Cluster 3	Cluster 4 ...
<i>network, learn, neural</i>	<i>intellig, research, system</i>	<i>confer, intern, juli</i>	<i>semant, web, retriev, onto</i>
Doc 2 of class 409866	Doc 1 of class 409866	Doc 87 of class 110	Doc 17 of class 110
Doc 4 of class 409866	Doc 3 of class 409866	Doc 90 of class 110	Doc 11 of class 110
Doc 8 of class 409866	Doc 5 of class 409866	Doc 92 of class 110	Doc 15 of class 110
Doc 9 of class 409866	Doc 6 of class 409866	Doc 93 of class 110	Doc 16 of class 110
...

Ao final dos experimentos da tarefa de agrupamento, as palavras extraídas em cada tópico pelo modelo LDA serão comparadas aos descritores dos grupos obtidos pelo melhor resultado na tarefa de Clusterização.

3.3. PROCESSAMENTO

Essa é a principal etapa do processo de mineração de textos, também denominada Extração de Padrões. Nela ocorre a busca efetiva por conhecimentos inovadores e úteis a partir dos dados textuais. A aplicação dos algoritmos, fundamentados em técnicas que procuram, segundo determinados paradigmas, visa explorar os dados de forma a produzir modelos de conhecimento.

A seguir, a descrição das duas principais tarefas utilizadas na etapa de Processamento: Clusterização e Classificação, e seus respectivos algoritmos.

3.4. CLUSTERIZAÇÃO

O objetivo do agrupamento de informações textuais é separar uma série de documentos dispostos de forma desorganizada em um conjunto de grupos que contenham documentos de assuntos similares (Figura 5).

Para que isto seja feito, parte-se do princípio da *Hipótese de Agrupamento (Cluster Hypothesis)*, levantado por van RIJSBERGEN (1979). Este princípio diz que objetos semelhantes e relevantes a um mesmo assunto tendem a permanecer em um mesmo grupo (*cluster*), pois possuem atributos em comum.

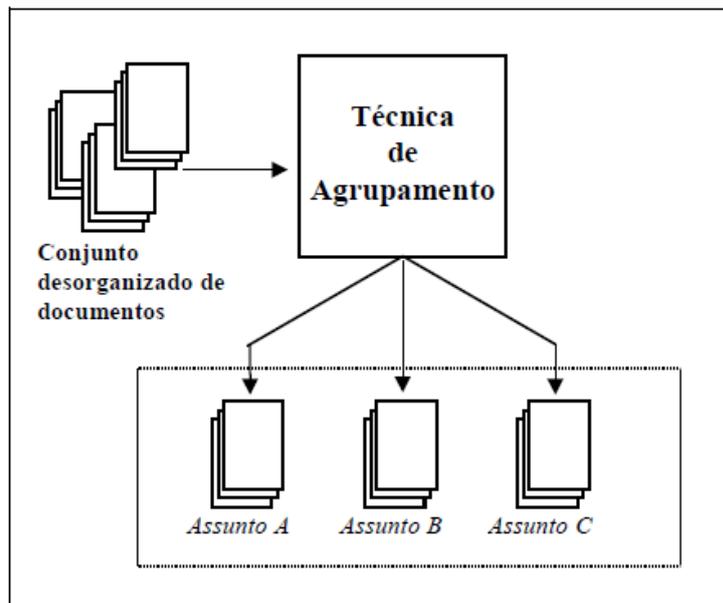


Figura 5: Objetivo do Agrupamento de Informações Textuais (Wives, 1999)

Um das principais características da tarefa de Clusterização de documentos é possibilitar a rápida localização de documentos em uma dada coleção. Quando uma coleção já está devidamente separada por grupos, encontrar um pequeno grupo de documentos relevantes torna-se uma tarefa fácil.

Assim, os agrupamentos podem ser organizados de duas formas (CUTTING, 1992): o agrupamento por partição e o agrupamento hierárquico. Eles dizem respeito à forma na qual os grupos são constituídos. No primeiro tipo de *agrupamento*, denominado *por partição*, os objetos são distribuídos em classes distintas, não havendo relação direta entre as classes. Este tipo de *agrupamento* é denominado *agrupamento de partição total (flat partition)* e os documentos são separados exaustivamente e colocados em grupos totalmente diferentes.

No segundo tipo, denominado *partição hierárquica (hierarchic partition)*, o processo de identificação de grupos é geralmente realimentado recursivamente, utilizando tanto objetos quanto grupos já identificados previamente como entrada para o processamento. Deste modo, constrói-se uma hierarquia de grupos de objetos, estilo uma árvore.

3.4.1. Agrupamento de Partição Total

A técnica consiste em agrupar os documentos em um número predeterminado de *aglomerados (clusters)* distintos. Os documentos são agrupados de forma tal que todos os elementos de um mesmo *aglomerado* possuem um grau mínimo de semelhança, que é indicado pelo número de características em comum que possuem. O algoritmo *K-means* (MACQUEEN, 1967) e suas variantes representam os algoritmos de agrupamento de partição total.

O algoritmo K-Means é uma técnica clássica de Clusterização e considerado o representante mais conhecido para agrupamento particional e muito utilizado em coleções textuais (STEINBACH *et al.*, 2000) É amplamente utilizado para este fim, além de ser considerado um dos métodos mais simples. Esse é um algoritmo de aprendizagem não supervisionada (e seu processo de separação) começa com a definição do usuário quanto ao número inicial *k* de grupos.

Na Figura 6 os documentos são representados pelos pequenos círculos. Os *aglomerados*, que aglomeram os documentos, são representados pelos grandes círculos. Como pode ser visto, os *aglomerados* não possuem ligações entre si, sendo totalmente isolados. Assim, os documentos são totalmente separados.

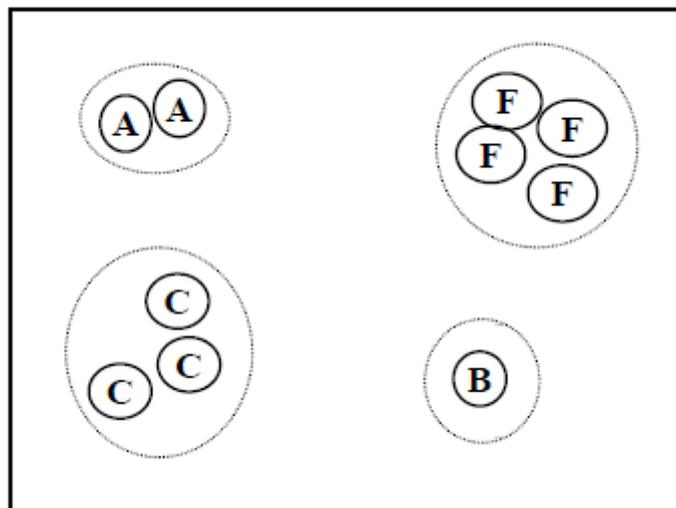


Figura 6:Resultado de um agrupamento por partição total e disjunta (Wives, 1999)

3.4.2. Agrupamento Hierárquico

As técnicas de Agrupamento Hierárquico produzem uma hierarquia de partições com uma simples partição incluindo todos os documentos num extremo, e grupos unitários, cada um composto de um documento individual no outro extremo.

A árvore descrevendo a hierarquia de grupos é chamada de dendrograma. Cada *cluster* ao longo da hierarquia é visto como uma combinação de dois *clusters*, a partir do próximo nível mais alto ou mais baixo, dependendo se a abordagem é aglomerativa ou divisiva, respectivamente.

Na partição estratégica aglomerativa, os elementos são agrupados em pares de maior similaridade. Deste modo, uma hierarquia de *aglomerados* é construída, conforme a Figura 7.

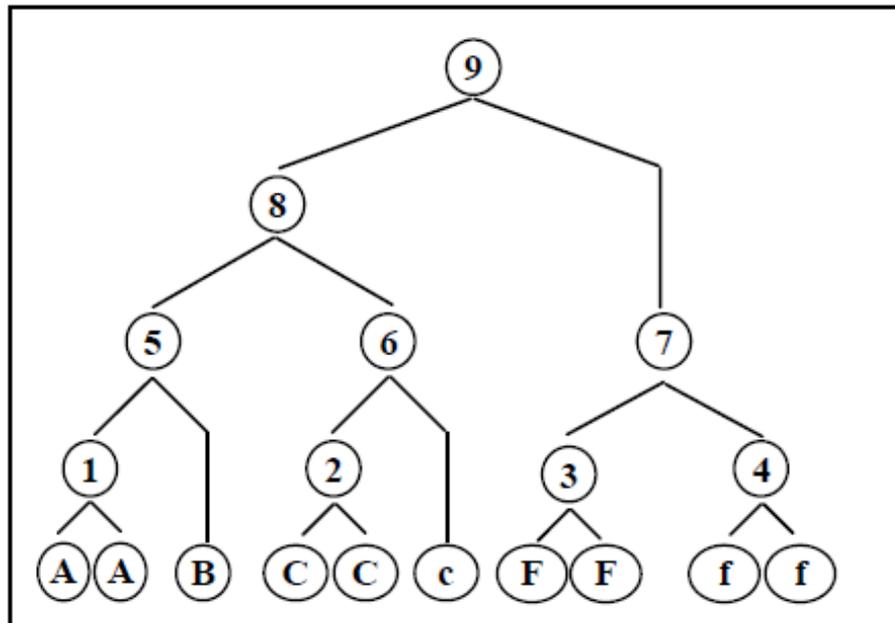


Figura 7: Agrupamento Hierárquico Aglomerativo (Wives, 1999)

A maior parte dos pacotes estatísticos utilizam métodos aglomerativos, dentre os quais os mais conhecidos são: *single linkage*, *complete linkage*, *average linkage*, método de *Ward*, método do Centróide. Uma breve descrição de cada um dos métodos pode ser encontrada em (LOPES, 2004; REZENDE *et.al.*, 2011).

No agrupamento global, os objetos são agrupados de forma similar à forma utilizada no agrupamento de partição total, ou seja, todos os documentos de um grupo são identificados (não somente os pares) como pode ser visualizado na Figura 8.

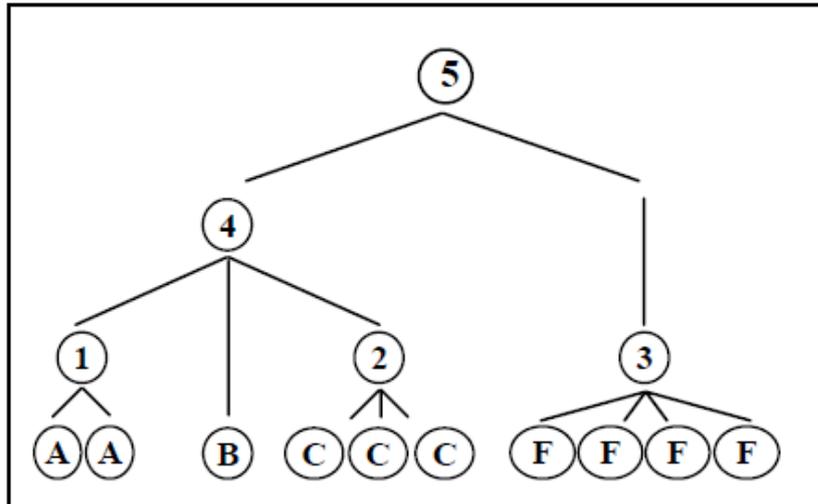


Figura 8: Agrupamento Hierárquico Global (Wives, 1999)

O processo é feito e os *aglomerados* entre os grupos resultantes são identificados. O processo é aplicado recursivamente até que um único grupo seja identificado. Porém, a técnica de análise global de elementos consome muitos recursos computacionais. A maioria dos trabalhos relacionados com agrupamento hierárquico na literatura referenciam as estratégias aglomerativas.

Em ambos os casos é produzida uma árvore, onde as folhas desta árvore representam os elementos individuais e os nós intermediários correspondem a grupos formados pelo agrupamento (*merge*) de seus grupos filhos.

Já que este tipo de agrupamento produz uma árvore, sua grande vantagem diz respeito à facilidade de localização da informação, pois o usuário pode ir navegando pelos ramos de informação mais relevantes à sua necessidade. Isso porque as informações estão agrupadas por assunto, e estes estão interligados de acordo com seus relacionamentos, constituindo uma hierarquia de assuntos. No caso, começa-se pelo nó pai (topo), e decide-se qual dos dois lados é mais similar ao que se procura. A análise é aplicada recursivamente, tomando o rumo do sub-ramo mais adequado até que se chegue ao elemento desejado.

Um dos grandes problemas nessa tarefa de agrupamento diz respeito à atribuição de rótulos (*labels*) aos nós da árvore que vai sendo construída. Se um rótulo incorreto é atribuído ao nó, o usuário não tem condições de compreender (não tem uma idéia correta do assunto) do que se trata o sub-ramo, e não consegue navegar corretamente.

Em geral, acredita-se que os métodos de agrupamentos hierárquicos produzem grupos de melhor qualidade que aqueles produzidos particionais (*K-means* ou uma de suas variantes), o qual tem uma complexidade linear ao número de documentos, mas produz *clusters* de menor qualidade (STEINBACH *et al.*, 2000).

3.5. MÉTRICAS DE AVALIAÇÃO DE RESULTADOS DE CLUSTERIZAÇÃO

Para poder comparar uma técnica de agrupamento com outra, a fim de determinar a melhor técnica, é necessário identificar qual delas oferece melhor resultado e desempenho.

As métricas de avaliação de desempenho foram adotadas da área de Recuperação de Informação e são baseadas na noção de relevância: se um documento atender à necessidade de informação do usuário, ele é considerado relevante à solicitação do usuário (REZENDE, 2003).

As medidas de avaliação derivada da matriz de confusão mais comumente utilizadas em tarefas de agrupamento são as seguintes:

A precisão é uma medida que indica se todos os documentos recuperados tratam realmente do assunto solicitado, ou seja, avalia o quanto o modelo acerta e é definida pela Equação 4.

$$\text{Precisão} = \frac{\text{número de itens relevantes recuperados}}{\text{número total de itens recuperados}}, \quad \text{Equação 4}$$

A abrangência ou revocação (*recall*) mede a quantidade de itens relevantes dentre os existentes na base de dados, que foram recuperados (Wiv,2000). Essa medida avalia o quanto o modelo contabiliza e é definida pela Equação 5:

$$\text{Recall} = \frac{\text{número de itens relevantes recuperados}}{\text{número de itens relevantes na coleção}} \quad \text{Equação 5}$$

A medida F-Measure(F1) é a média harmônica entre precisão e *recall* apresentada pela Equação 6.

$$\text{Medida F} = \frac{2 * \text{Recall} * \text{Precisão}}{\text{Recall} + \text{Precisão}}, \quad \text{Equação 6}$$

3.6. CLASSIFICAÇÃO

A classificação pode ser descrita como um processo de identificação dos principais tópicos de um documento e a sua posterior associação automática a uma ou mais categorias pré-definidas.

O processo de classificação é realizado em dois passos (MOTTA, 2004). O primeiro, conhecido como treinamento ou aprendizado, caracteriza-se pela construção de um modelo que descreve um conjunto predeterminado de classes de dados. Essa construção é feita analisando as amostras de uma base de dados, onde as amostras são descritas por atributos e cada uma delas pertence a uma classe predefinida, identificada por um dos atributos, chamado atributo rótulo da classe ou, simplesmente, classe. O conjunto de amostras usadas neste passo é o conjunto de treinamento, dados de treinamento ou amostras de treinamento.

As formas mais comuns de representar o conhecimento (ou padrões) aprendido na fase de treinamento são regras de classificação, árvores de decisão ou formulações matemáticas. Este conhecimento pode ser usado para prever as classes de amostras desconhecidas futuras, bem como pode permitir um melhor entendimento dos conteúdos da base de documentos.

No segundo passo, o modelo construído é testado, isto é, o modelo é usado para classificação de um novo conjunto de amostras, independentes daquelas usadas no treinamento, chamado conjunto de teste, dados de teste ou amostras de teste. Como este conjunto também possui as classes conhecidas, após a classificação, pode-se calcular o percentual de acertos, comparando as classes preditas pelo modelo com as classes esperadas (ou conhecidas). Este percentual é conhecido como acurácia ou precisão do modelo para o conjunto de teste em questão. Se a acurácia for considerada aceitável, o modelo pode ser usado na classificação de amostras desconhecidas futuras, ou seja, amostras cuja classe não é conhecida.

A classificação de textos é uma tarefa que pode ser aplicada em diferentes áreas, tais como na filtragem de textos, na organização de documentos e na indexação automática para sistemas de recuperação de informação.

A seguir são listados os métodos utilizados nesse trabalho para classificação de textos.

3.6.1. Classificador Naïve Bayes

O classificador *Naïve Bayes* é baseado no Teorema de Bayes e é um dos classificadores mais usados em categorização de textos (McCALLUM e NIGAM, 1998). É um algoritmo para o aprendizado indutivo com abordagem probabilística.

Baseado na probabilidade condicional de determinadas palavras aparecerem em um documento o qual pertence a uma determinada categoria, essa técnica calcula a probabilidade *a posteriori* de um novo documento pertencer a categorias diferentes e atribuir a esse documento a categoria cuja probabilidade *a posteriori* é a mais alta (LEWIS e RINGUETTE, 1994).

O classificador também conhecido como classificador Bayesiano Ingênuo (*Naïve Bayesian Classifier*) se baseia na suposição de que os atributos dos documentos são condicionalmente independentes, ou seja, o classificador assume que existe independência entre as palavras de um texto.

Assim, a fórmula utilizada pelo classificador *Naïve Bayes* é dada por (MITCHELL, 1997):

$$V_{NB} = \operatorname{argmax}_{v_j \in V} P(v_j) \prod_{i=1}^n P(a_i | v_j), v_j \in V \quad \text{Equação 7}$$

Onde:

V_{NB} é a categoria atribuída ao documento;

v_j é cada um dos possíveis categorias pertencentes a V ;

$P(v_j)$ é a probabilidade *a posteriori* da ocorrência de cada hipótese;

$P(a_i | v_j)$ é a probabilidade da ocorrência de cada evidência a_i dada à ocorrência de uma hipótese (categoria) v_j .

Os classificadores bayesianos têm como principal característica produzir resultados rapidamente, mesmo quando aplicados a grandes volumes de dados. Para (HAN e KAMBER, 2006), seus resultados são comparáveis em desempenho aos obtidos pelos classificadores árvores de decisão, redes neurais e vizinhos mais próximos.

3.6.2. Máquinas de Vetor de Suporte

As Máquinas de Vetores Suporte (*Support Vector Machines* - SVMs) constituem uma técnica embasada na Teoria de Aprendizado Estatístico (VAPNIK, 1995) que vem recebendo grande atenção nos últimos anos (HEARST et al., 1998; CRISTIANINI e SHAWE-TAYLOR, 2000).

Um conceito chave das SVMs - Máquinas de Vetor Suporte - é a implementação do mapeamento não-linear dos dados de entrada para um espaço característico de alta dimensão, onde um hiperplano ótimo é construído para separar os dados linearmente em duas classes.

Quando os dados de treinamento são separáveis, o hiperplano ótimo no espaço característico apresenta a máxima margem de separação, minimizando-se assim o erro sobre os dados de treinamento e teste. Já no caso destes dados serem linearmente inseparáveis, é necessária a aplicação de uma função *kernel* com o objetivo de aumentar a dimensão destes dados tornando-os separáveis (HAN e KAMBER, 2006).

A Figura 9 apresenta o exemplo de conjuntos não linearmente separáveis, no qual o espaço é mapeado em função de uma terceira dimensão e os dados podem ser divididos em classes com um hiperplano linear.

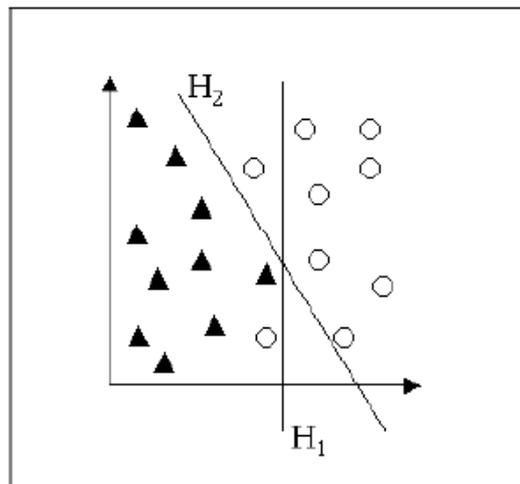


Figura 9: Exemplo de conjunto não linearmente separável (Lorena e Carvalho, 2003)

Utilizando o princípio de Minimização do Risco Estrutural (SMOLA and SCHÖLKOPF, 2002), as SVMs conseguem uma boa capacidade de generalização, mesmo nos casos em que o conjunto de treinamento não seja muito representativo.

Isto é muito importante para diminuir o tempo de treinamento de aprendizagem, e para novos padrões que determinado sistema apresente.

De forma resumida, algumas das principais características das SVMs que tornam seu uso atrativo são, segundo (SMOLA *et al.*, 1999): boa capacidade de generalização, robustez em grandes dimensões, convexidade da função objetivo e a teoria bem definida dentro da Matemática e da Estatística.

Entre as características citadas, essa técnica se caracteriza por alcançar altas taxas de precisão e ser robusta diante de matrizes de alta dimensionalidade geradas a partir de bases textuais, conforme será verificado nos experimentos realizados.

Os resultados da aplicação desta técnica são comparáveis aos obtidos por outros algoritmos de aprendizado, como as Redes Neurais Artificiais (RNAs) (HAYKIN, 2001), e em algumas tarefas tem se mostrado superiores, tal como na detecção de faces em imagens (HEARST *et al.*, 1998), na categorização de textos (HEARST *et al.*, 1999) e em aplicações em Bioinformática (ZIEN *et al.*, 2000).

3.7. MÉTRICA DE AVALIAÇÃO DE RESULTADOS DE CLASSIFICAÇÃO

A matriz de confusão é utilizada para apresentar os resultados obtidos durante uma classificação. Uma matriz de confusão para duas classes (C+ e C-) é apresentada na Tabela 2, onde os dois erros possíveis são denominados falso positivo (FP) e falso negativo (FN). Esses erros foram avaliados após realizada a tarefa de Clusterização pelo algoritmo de Classificação Naïve Bayes.

Tabela 2: Matriz de confusão para duas classes (Rezende, 2003)

Classes	Predita C ⁺	Predita C ⁻
Verdadeira C ⁺	Verdadeiros Positivos (TP)	Falsos Negativos (FN)
Verdadeira C ⁻	Falsos Positivos (FP)	Verdadeiros Negativos (TN)

A medida de avaliação derivada da matriz de confusão utilizadas nesse trabalho para tarefa de classificação é acurácia definida pela Equação 8:

$$\text{Precisão Total} = \frac{(TP + TN)}{(TP + TN + FP + FN)}, \quad \text{Equação 8}$$

Também foi avaliado o desempenho dos classificadores *Naïve Bayes* e SVM em relação ao tempo de processamento.

3.8. PÓS-PROCESSAMENTO

Essa etapa envolve a apresentação, a análise e a interpretação dos resultados, a fim de validar as descobertas efetuadas pela etapa de processamento dos dados. Nela o especialista em mineração de textos e o especialista no domínio da aplicação podem, a partir da avaliação dos resultados alcançados, definir novas alternativas de investigação dos dados.

Essas técnicas eventualmente se utilizam de gráficos em duas e em três dimensões para a apresentação dos resultados, a fim de facilitar ou até mesmo viabilizar a sua compreensão e interpretação.

Assim, o pós-processamento dos dados consiste da fase de validação das descobertas efetuadas pela etapa de processamento dos dados e da visualização dos resultados encontrados. Métricas de avaliação de resultados, ferramentas de visualização, e conhecimento de especialistas ajudam a consolidar os resultados.

3.9. COMENTÁRIO

Comentou-se brevemente as principais técnicas relacionadas ao processo de Mineração de Textos que serão utilizadas nesse trabalho. Cada uma das técnicas na etapa de Pré-Processamento será utilizada na metodologia. Serão utilizadas as duas tarefas principais na etapa de processamento, sendo utilizado o algoritmo de clusterização hierárquica para Clusterização e os algoritmos *Naïve Bayes* e Máquinas de Vetor de Suporte para Classificação.

Foram apresentados o modelo de tópicos LDA para como nova abordagem para representação de documentos e o uso do método *RSLP Stemmer* para radilicalização da língua portuguesa. Ambos métodos serão avaliados utilizando a base de documentos da Central de Atendimento.

4. SELEÇÃO DE DADOS DE GRANDES COLEÇÕES DE DOCUMENTOS

Para lidar com grandes coleções de documentos é necessário utilizar técnicas de amostragem ou de redução de dados para agilizar o processamento. No presente trabalho, optou-se por experimentar a teoria descrita na Seção 4.2, inspirada no Sistema Imune-Artificial.

Comparada a outros paradigmas de inteligência computacional, a área do conhecimento envolvendo Sistemas Imunológicos Artificiais (SIAs) é ainda muito nova e pouco explorada.

Os Sistemas Imunológicos Artificiais (SIA), que surgiram a partir de tentativas de modelar e aplicar princípios imunológicos no desenvolvimento de novas ferramentas computacionais, já vêm sendo utilizados em diversas áreas, como reconhecimento de padrões, detecção de falhas e anomalias, segurança computacional, otimização, controle, robótica, *scheduling*, análise de dados, aprendizagem de máquina, dentre outras, como pode ser encontrado em (DASGUPTA, 1998 a,b; BÄCK *et al.*, 2000 a,b; TIMMIS, 2000 e DE CASTRO, 2001).

Constituído por componentes e mecanismos distintos, porém que atuam de forma conjunta e notavelmente eficaz, o sistema imunológico proporciona ao corpo humano resistência às enfermidades. Os anticorpos, por exemplo, são gerados por células denominadas linfócitos em resposta aos antígenos (agentes infecciosos), e sua presença em um indivíduo reflete as infecções às quais o mesmo já foi exposto.

Os linfócitos são capazes de desenvolver uma memória imunológica, ou seja, reconhecer o mesmo estímulo antigênico caso ele entre novamente em contato com o organismo, evitando assim o restabelecimento da doença. Portanto, mecanismos de aprendizagem e memória dão ao sistema imunológico a capacidade de extrair informações dos agentes infecciosos e disponibilizá-las para uso futuro em casos de novas infecções pelos mesmos agentes ou agentes similares.

A fim de extrair informações relevantes em um grande volume de documentos, é proposta uma nova abordagem capaz de determinar amostras mais significativas de um conjunto de documentos para que seja possível obter uma classificação de sucesso.

Há casos em que o tamanho do conjunto de treinamento é muito grande, havendo necessidade de aprender com as amostras mais relevantes, obtendo um maior desempenho computacional.

A abordagem desenvolvida para seleção de dados de treinamento por (FIGUEREDO *et al.*, 2012) foi inspirada pelo mecanismo de supressão dos sistemas imunológicos artificiais. De acordo com os resultados das experiências realizadas pelos autores, o mecanismo de supressão é eficiente, capaz de reduzir uma grande quantidade de dados do conjunto de treinamento, sem perdas significativas nas medidas de avaliação e desempenho. Além disso, proporciona um ganho no tempo de processamento de todas as bases estudadas.

No caso de classificação de documentos, acredita-se que, escolhendo os mais representativos, pode-se conseguir meios mais eficientes para treinamento dos classificadores.

Nesse trabalho, será avaliada a eficiência do mecanismo de supressão também na tarefa de Clusterização. O objetivo será avaliar e comparar os resultados entre a base de textos original e a base de textos suprimida. Pressupõe-se que a aplicação da nova técnica permitirá otimizar as tarefas de agrupamento e reconhecimento de palavras-chaves em um volume de dados textuais cada vez mais crescentes.

4.1. INSPIRAÇÃO SISTEMA IMUNE ARTIFICIAL

O sistema imunológico apresenta diversas características que lhe possibilitam detectar e combater um agente invasor no organismo de forma otimizada. Um dos mecanismos, que lhe permitem ser capaz de defender o organismo contra todos os patógenos existentes sem desperdício de energia e encontrando rapidamente os melhores agentes de defesa específicos, é a auto-regulação.

De acordo com o mecanismo de auto-regulação dos SIs, as células imunes clones que não são mais necessárias para o organismo, ou seja, aquelas que são auto-imunes, não recebem estímulos de sobrevivência e morrem. Esta característica de auto-regulação impede o desperdício de energia dentro do organismo e faz com que seja mantido apenas o repertório de linfócitos realmente necessários para defesa contra patógenos.

Basicamente, o algoritmo a ser implementado, utiliza este mecanismo onde somente os anticorpos mais representativos são mantidos. Ao invés de trabalhar como um sistema que gera e evolui clones de células B até que os anticorpos reconheçam o conjunto de treinamento, como feito em (WATKINS et al., 2003; WATKINS e TIMMIS, 2002; WATKINS e TIMMIS, 2004; WATKINS *et al.*,2004; WATKINS e BOGGESS, 2002), esta nova abordagem propõe que os dados de treinamento por si só sejam o repertório de anticorpos do sistema como feito em (FIGUEREDO *et al.*, 2008).

O conceito de supressão é empregado no conjunto de treinamento da coleção de documentos para eliminar anticorpos muito similares. Em mineração de textos, o conjunto de treinamento possui um grande volume de documentos e demandam alto custo computacional. Conforme dito anteriormente, a escolha de documentos mais representativos pode obter um melhor desempenho computacional.

4.2. ALGORITMO SUPRESSOR DE TEXTOS

Na tarefa de Classificação, o conceito de supressão é empregado no conjunto de treinamento para eliminar os documentos mais similares e manter os realmente representativos para uma determinada classe.

O algoritmo foi implementado em basicamente três fases: inicialização, supressão e determinação do novo conjunto de treinamento.

Na fase de inicialização, é necessário realizar a indexação dos documentos presentes na coleção utilizando-se da extração dos termos relevantes. Essa extração foi automatizada pela ferramenta *PolyAnalyst*, que permitiu identificar e selecionar os termos mais relevantes de cada documento, sem a intervenção direta do documentarista.

Para atender ao número de classes, que varia de uma coleção para outra, foi utilizada a função *hash* para criação dos índices. Através da função *hash* foi possível gerar uma indexação perfeita, ou seja, uma identificação da relação registro-classe, definindo índices para cada uma das classes da coleção.

Na fase de supressão, o mecanismo divide o conjunto de treinamento em dois subconjuntos. O primeiro subconjunto corresponde a 90% e representa os anticorpos (WBCs). O segundo subconjunto, de 10%, representa os antígenos que irão selecionar os anticorpos com maior similaridade e realizar a supressão.

Ambos os anticorpos e antígenos foram representados como vetores contendo os termos mais relevantes dos documentos obtidos através do Modelo Espaço Vetorial (Secção 7.5.8). Cada vetor foi normalizado para pertencer à mesma escala de valores sendo mapeados para o intervalo [0,1].

Os anticorpos com mais afinidade são determinados pela distância do cosseno. Essa medida é comumente utilizada para calcular a semelhança entre os pares de documentos do corpus, como sugerido por YANG e WILBUR (1996) e é definida pela Equação 9:

$$Sim(W_{WBC}, W_P) = \frac{\sum W_{WBC} \times W_P}{\sqrt{\sum (W_{WBC})^2 \times \sum (W_P)^2}} \quad \text{Equação 9}$$

Onde W_{WBC} e W_P representam os termos do documento do anticorpo (WBC) e do antígeno (P), respectivamente.

O algoritmo busca identificar o melhor subconjunto de anticorpos para reconhecer os antígenos, ou seja, o novo conjunto de treinamento deve ser capaz de identificar novas presenças de antígenos. Os anticorpos mais similares aos antígenos pertencentes a uma mesma classe são eliminados e os mais representativos permanecem.

Finalmente, os anticorpos sobreviventes são representados por uma medida de avaliação e selecionados para fazer parte do novo conjunto de treinamento reduzido.

O pseudocódigo desta técnica é apresentado no Algoritmo 1.

// Fase de Inicialização

Indexar automaticamente os termos mais relevantes;

Gerar o Modelo de Espaço Vetorial

Dividir o Conjunto de Treinamento Original em anticorpos (90%) e antígenos (10%)

Normalizar os dados em [0, 1]

Inicializar a avaliação dos anticorpos igual a 0;

// Fase de Supressão

Para cada antígeno faça

 Encontrar o anticorpo mais similar ao antígeno;

 Se a classe do anticorpo for igual à classe do antígeno então

 //O Anticorpo Similar foi capaz de reconhecer o antígeno//

 Incrementar a avaliação do Anticorpo Similar;

 Fim-Se;

Fim-Para;

// Fase de Identificação do novo Conjunto de Treinamento

Eliminar aqueles anticorpos com avaliação igual a 0;

Gerar o conjunto de anticorpos sobreviventes como um novo conjunto de treinamento reduzido;

Algoritmo 1: Algoritmo de Supressão de Textos

Na tarefa de Clusterização, o mecanismo foi aplicado sobre toda a coleção de documentos, conforme apresentado na Secção 7.6.1.

4.3.COMENTÁRIO

Inspirado no mecanismo de auto-regulação do sistema imunológico, a estratégia de inteligência computacional adotada permite reduzir, significativamente, o esforço computacional envolvido na construção de um modelo eficiente, a partir de grandes coleções de documentos. Com o crescente volume de informações, torna-se essencial utilizar técnicas que permitam extrair o conhecimento de forma ágil e eficiente.

5. FERRAMENTA POLYANALYST

PolyAnalyst (PA) é um programa desenvolvido pela empresa *Megaputer Intelligence* Inc. (<http://www.megaputer.com/polyanalyst.php>), tendo sua primeira versão lançada em 1997. PA é um programa de mineração de dados (estruturados e não estruturados) com variados propósitos, visando a facilitar as pesquisas e a tomada de decisões a partir da análise dos dados. Contém várias ferramentas para exploração dos dados que possibilitam ligá-los, separá-los, analisá-los e sumará-los e que estão disponíveis para que o usuário possa compor seus projetos de análise de forma personalizada, de acordo com as tarefas que pretende realizar.

De forma resumida, algumas tarefas a serem realizadas com o apoio do programa são listadas a seguir: importar bases de dados de diferentes origens; mesclar diferentes bases de dados; adicionar, alterar ou remover colunas ou registros; filtrar, eliminar ou tratar valores ausentes ou com ruído; excluir dados duplicados; visualizar estatísticas básicas a respeito dos dados; realizar tarefas como a descoberta de associações, classificação, clusterização e predição; analisar dados textuais e encontrar palavras-chave; visualizar os dados ou o resultado das análises de forma personalizada; exportar bases de dados; gerar relatórios com os resultados das análises realizadas, entre outras. Neste trabalho serão apresentados os resultados da aplicação de alguns destes recursos oferecidos pela ferramenta.

O *PolyAnalyst* é um pacote de *software* intuitivo com um ciclo de aprendizagem curto. Ele pode ser compreendido e operado em um nível mais simples ou em nível avançado. No nível mais simples ele requer somente conhecimentos básicos em computação, em bancos de dados e em estatística. Já no nível avançado são requeridos conhecimentos em lógica de programação, conceitos básicos de Inteligência Artificial e noções avançadas de estatística.

O software utiliza a arquitetura cliente/servidor, que permite uma administração eficiente dos recursos disponíveis. Tipicamente são montadas várias estações clientes, que são atendidas por um único servidor. Estas estações podem ser configuradas individualmente para dar suporte a tarefas específicas. Por exemplo, algumas podem ser configuradas para as tarefas de análise de dados propriamente ditas, enquanto outras serem voltadas para usuários de negócio, que basicamente terão interesse nos relatórios das análises realizadas.

5.1. ANÁLISE DOS TEXTOS

Esse nó trata das tarefas de processamento dos dados textuais, dando suporte classificação e clusterização dos documentos, bem como a extração de frases, palavras-chave ou entidades contidas nos textos.

O programa permite que sejam usados um ou mais dicionários no apoio a tarefas de análise de dados textuais. Ele oferece um dicionário padrão para isso, denominado WordNet, um dicionário popular disponível na Web, que é um modelo abrangente e que normalmente auxilia de forma satisfatória nas tarefas de análise de textos de temas gerais, e que permite que os usuários cadastrem outros dicionários no sistema para a mesma facilidade.

O *PolyAnalyst* incorpora a técnica de *stemming* para o processamento de dados textuais na língua inglesa. Ele utiliza um algoritmo próprio desenvolvido para esse fim, baseado em algoritmos conhecidos e disponíveis no mercado. Ele utiliza também um dicionário base específico para o algoritmo de *stemming*, na língua inglesa, contendo diversos termos e relacionamento entre eles. Esse dicionário base é uma forma modificada do *WordNet*.

No cálculo de relevância dos termos nos documentos é utilizado um algoritmo para a contagem da pontuação desses termos, que é uma variação do algoritmo *Vector Space Relevance*. O cálculo da relevância leva em consideração o número de documentos da base, o tamanho de cada documento e a frequência das palavras.

Utiliza-se o logaritmo de probabilidade binomial para encontrar os documentos com uma maior concentração de palavras-chave (palavras estão mais próximas umas das outras). Dessa forma se essas palavras-chave ocorrem, frequentemente, em um documento, essas palavras são relevantes para o documento.

A ferramenta permite gerar o modelo de espaço vetorial a partir das medidas booleana, frequência ou relevância das palavras-chaves. Neste trabalho, serão comparados os resultados obtidos pelas medidas frequência e relevância para geração do modelo de espaço vetorial.

A seguir, são apresentadas as principais tarefas de Mineração de Textos disponíveis na ferramenta: Clusterização dos Textos (*Text Clustering*) e Classificação Linear (*Linear Classification*).

5.2. CLUSTERIZAÇÃO DE TEXTOS

Esse nó é utilizado para a clusterização de documentos. Para a geração dos grupos, ele utiliza uma variação do algoritmo *Suffix Tree Clustering* (ZAMIR, 1999).

Esse algoritmo apresenta algumas qualidades. Entre elas, a velocidade de processamento, que é próxima de um processamento linear, ou seja, o tempo é proporcional ao número de registros. Ele apresenta resultados de fácil interpretação por parte dos usuários. Também possui mecanismos de identificação das frases.

As frases apresentam a vantagem de ter um poder descritivo mais alto que os termos isolados. Daí, elas se prestam melhor para descrever o conteúdo dos grupos para os usuários, e de uma maneira mais concisa.

O processo envolve dois passos principais: no primeiro, o algoritmo faz uma busca por registros que compartilham frases; no segundo, ele agrupa os documentos a partir da frequência da ocorrência dessas frases.

É necessário configurar alguns parâmetros matemáticos que manipulam o comportamento do algoritmo, tais como:

- Escolher se o agrupamento a ser feito deverá ser base, exclusivo ou hierárquico (Secção 7.6.1);
- O número máximo de agrupamento máximo de frases individuais que diz respeito ao número máximo de frases individuais e palavras que serão buscadas, no primeiro passo do algoritmo;
- Os números ou percentual mínimo e máximo de registros por grupo;
- O uso ou não de um thesaurus
- O uso ou não de um dicionário e a sua definição.

A ferramenta apresenta, como resultado, a configuração adotada no processamento, um gráfico estatístico com a distribuição dos documentos pelos grupos, uma tabela com os grupos gerados, contendo as suas descrições, a quantidade e a identificação dos documentos presentes, um gráfico mostrando a proximidade dos grupos, entre outros.

5.3. CLASSIFICAÇÃO LINEAR

Esse nó desenvolve um modelo de classificação dependente de um atributo estruturado, através da utilização de uma coluna independente com os textos. Esse modelo é baseado na frequência e na distribuição dos termos no texto. A partir disso, o programa treina um modelo para a classificação automática de textos.

O *PolyAnalyst* apresenta duas abordagens de classificação de textos, baseadas em algoritmos distintos, a saber:

SVM (Support Vector Machine): um algoritmo que requer um processamento mais intensivo em termos computacionais e que tipicamente apresenta uma maior acurácia nos resultados.

Naïve Bayes: um algoritmo de processamento computacional mais rápido, que mais escalável, e que, em geral, obtém resultados menos precisos que o SVM.

As secções 3.9.1 e 3.9.2 apresentam, respectivamente, a fundamentação teórica do funcionamento dos algoritmos *SVM* e *Naïve Bayes*.

Na configuração dos parâmetros para se executar a classificação é necessário definir o atributo a ser utilizado como fonte dos dados, o algoritmo de processamento, o uso ou não de uma *stoplist* e a definição da mesma, o tipo de dado a ser tratado, entre outros.

Durante o processamento, o programa armazena as palavras-chave de forma booleana em uma tabela, indicando se elas aparecem ou não em cada documento, e a frequência com que isso acontece.

Após o processamento, o resultado é apresentado em duas abas: uma contendo informações da configuração adotada e a outra contendo informações da classificação propriamente dita, incluindo uma matriz com as taxas de erro por classe.

5.4. COMENTÁRIO

Este capítulo apresentou uma das ferramentas que apoiam o processo de mineração de textos. A ferramenta *PolyAnalyst* permite realizar todas as etapas da Mineração de Textos, desde o pré-processamento até o pós-processamento.

Diversas outras ferramentas estão disponíveis na literatura, tais como: *LingPipe* que é uma API em Java para mineração em textos distribuída com código-fonte (<http://alias-i.com/lingpipe/>); *Pimiento* que é um ambiente para mineração em textos, baseado em Java (<http://erabaki.ehu.es/jjga/pimiento/>) e *Cortex Intelligence*, que é um sistema de Processamento de Linguagem Natural para mineração de textos aplicada à Inteligência Competitiva (<http://www.cortex-intelligence.com/site/>).

6. CENTRAL DE ATENDIMENTO

Uma definição formal de central de atendimento ou *Call Center* (CC) não existia até os anos 80 HAWKINS *et al.*(2001). No entanto, os consumidores já tinham alguma forma de comunicação com as empresas, por meio de um telefone ou mesmo de correspondências. De forma a atender às informações dos clientes, os atendentes tinham algum acesso a informações, normalmente de forma manual, sobre os produtos e serviços e também sobre os clientes, mas de forma ainda incipiente.

Nos anos 60 e 70, com o advento do computador, as empresas conseguiram melhorar suas formas de atendimento aos consumidores, quase sempre ainda através do telefone. A definição de CC tem mostrado evolução com o decorrer do tempo. Na realidade, observa-se que ela vem incorporando novas formas de comunicação e também se integrando a sistemas mais complexos.

Fica claro que novos conceitos vão sendo agregados ao conceito inicial de um centro para recepção e estabelecimento de chamadas com consumidores. A definição apresentada em *CALL CENTER GUIDE* (2005) mostra bem isso ao dizer que um Centro de Contato com Consumidores (CCC) (antigamente referido como *Call Center*) é uma unidade que permite o contato com consumidores e dos consumidores para a organização, de uma forma efetiva e eficaz. No começo, os CCC eram projetados para processar chamadas telefônicas e atualmente processam todo tipo de mídia, não somente chamadas telefônicas.

Além das referências diretas a CC, alguns conceitos mais recentemente vêm substituindo a denominação de CC como, por exemplo, o *Contact ou Care Center*. ZENONE (2001) mostra essa transição ao dizer que “um *call* ou *contact* ou ainda um *care-center* é formado por um grupo de profissionais atendentes, supervisão e por vezes coordenação, gerência e até diretoria, dependendo de sua força na empresa”.

Segundo Abt (2005), *Call Center*, *Contact Center* ou *telemarketing* são designações para as centrais de atendimento destinadas ao contato com consumidores ou *prospects*, de forma ativa (ligação feita pela empresa para o cliente) ou receptiva (do cliente à empresa), usando telefone ou outros canais de comunicação. O termo mais abrangente é *Contact Center*, que inclui o contato por e-mail, fax, chat e *VoIP* (*voice over IP*), por exemplo.

A possibilidade de interação com usuários através de outras fontes, em especial da *Internet*, criou uma comodidade para os usuários e uma maior complexidade dos sistemas de TI em suporte ao *Call Centers*. *E-mail*, *fax*, conversores de voz em texto e mesmo a utilização de VoIP (*voice over IP*) exigiram maior inteligência dos terminais e maior capacitação dos operadores.

De fato, HOLTGREWE (2005) detectou isso em sua pesquisa. Os e-mails e faxes são utilizados em 89% e 86% dos CCs, respectivamente, representando os maiores complementos às chamadas telefônicas. Ainda da referida pesquisa, 30% utilizam comunicação via *web*, 16% utilizam VoIP e o reconhecimento de voz é a tecnologia menos utilizada com 8%. Também GRAELM (2004) diz que as empresas têm demonstrado maior preocupação em utilizar os *sites* das empresas para oferecer suporte pós-vendas e obter feedback dos clientes.

Com avanço da tecnologia, algumas facilidades, também foram incorporadas e os *Call Center*s tornaram-se mais eficientes, pelo menos tecnologicamente falando (HAWKINS et al., 2001). Entre essas facilidades, a primeira parece ter sido a possibilidade de distribuição automática de chamadas (ACD), que permitia o encaminhamento de chamadas ao serviço específico desejado pelo cliente através de uma interação em que se solicita que os mesmos naveguem por menus pré-definidos, normalmente através de opções via teclado, e não mais sejam encaminhados a um conjunto uniforme de atendentes para posterior(es) transferência(s).

A Figura 10 apresenta uma visão geral do modelo de um *Call Center*. Verifica-se três tipos de chamadas (*emergency*, *service*, *billing*) entrando no *Call Center*. Estas chamadas vão para uma fila e aguardam o primeiro representante do serviço de telefonia disponível. Cada grupo de agentes no modelo representa um ou mais agentes que atendem no mesmo horário. Cada um com um perfil de atendimento. As chamadas de emergência podem entrar através de uma linha direta, ou seja, possuem prioridade na fila.

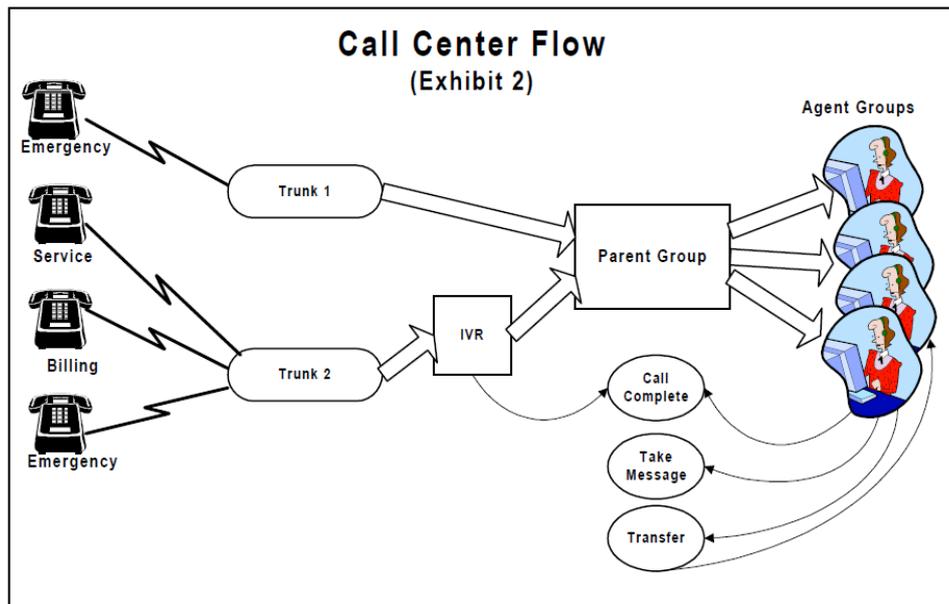


Figura 10 : Modelo Básico do Funcionamento das Centrais de Atendimento (HALL *et al.*, 1998)

A integração computador-telefone (CTI) permitiu que outras facilidades fossem associadas à ACD como a distribuição geográfica de chamadas permitindo um atendimento mais próximo ao consumidor, estatísticas sobre o tempo de retenção das chamadas de atendimento, a taxa de desistências de chamadas ao CC, entre outras, permitiram um melhor dimensionamento dos CCs e conseqüentemente a prestação de um serviço de melhor qualidade aos consumidores.

Além disso, em especial para os serviços das operadoras de telecomunicações (STROUSE, 1999), quando uma chamada é feita, o número do telefone que chama aparece para a atendente que, de antemão, tem as informações dos clientes mostradas em sua tela de computador, facilitando e muito a solução dos problemas ou o correto provimento das informações solicitadas.

A Figura 11 mostra os avanços nas telecomunicações, desde o uso de uma telefonista para realizar uma ligação, passando pela discagem e teclado no aparelho de telefone até o uso de vários tipos de mídia (voz, Web sites, e-mails, chat e vídeo) através de múltiplos canais de comunicação (rede de telefonia, wireless e internet banda larga).



Figura 11: Evolução dos Serviços de Comunicação
(GILBERT *et al.*, 2005)

A evolução dos serviços de telecomunicação impulsionou uma nova geração de serviços inteligentes, capaz de extrair informação útil, sem o mínimo de intervenção humana. A utilização de técnicas inteligentes pode ser usada como vantagem competitiva e suporte à tomada de decisão para aperfeiçoar a operação e os negócios das empresas ou organizações (GILBERT *et al.*, 2005).

Nesse sentido, a Central de Atendimento deve ser vista como fonte contínua de informações que impulsiona cada vez mais a aplicação e utilização de novos serviços de comunicação inteligentes.

6.1. ESTUDO DE CASO

O Centro de Serviços é uma estrutura idealizada pela empresa brasileira de petróleo para centralizar os diversos serviços de atendimento existentes na empresa, unificando as quatro Centrais de Atendimento (Rio de Janeiro, São Paulo, Bahia e Macaé).

O principal site (Central de Atendimento) está localizado na cidade do Rio de Janeiro e outro em Campos dos Goytacazes utilizado como contingência. Ambos os *sites* tem funcionamento em regime 24 x 7 e são monitorados 24 horas por dia, cujas gravações são guardadas por toda a vigência do contrato.

Cada Central de Atendimento possui modelos e definições distintas, além de uma infra-estrutura adequada de forma a viabilizar a manutenção de elevados padrões de atendimento e relacionamento, segundo às exigências de SMS (Segurança, Meio Ambiente e Saúde) e SI (Segurança da Informação) da empresa brasileira de petróleo, além da utilização de tecnologia de ponta para as soluções de VOIP, URA e DAC.

Além disso, a Central disponibiliza um canal de relacionamento estruturado para atendimento às necessidades específicas do negócio do cliente podendo atuar de forma passiva - como principal ponto de contato na recepção, tratamento, resolução e encaminhamento de chamados relacionados ao negócio do cliente - e de forma ativa - realizando chamados com uma abordagem planejada a um público-alvo selecionado para o tratamento de questões de interesse do cliente. A Figura 12 apresenta o fluxo de atendimento atualmente para atender as necessidades específicas do cliente.

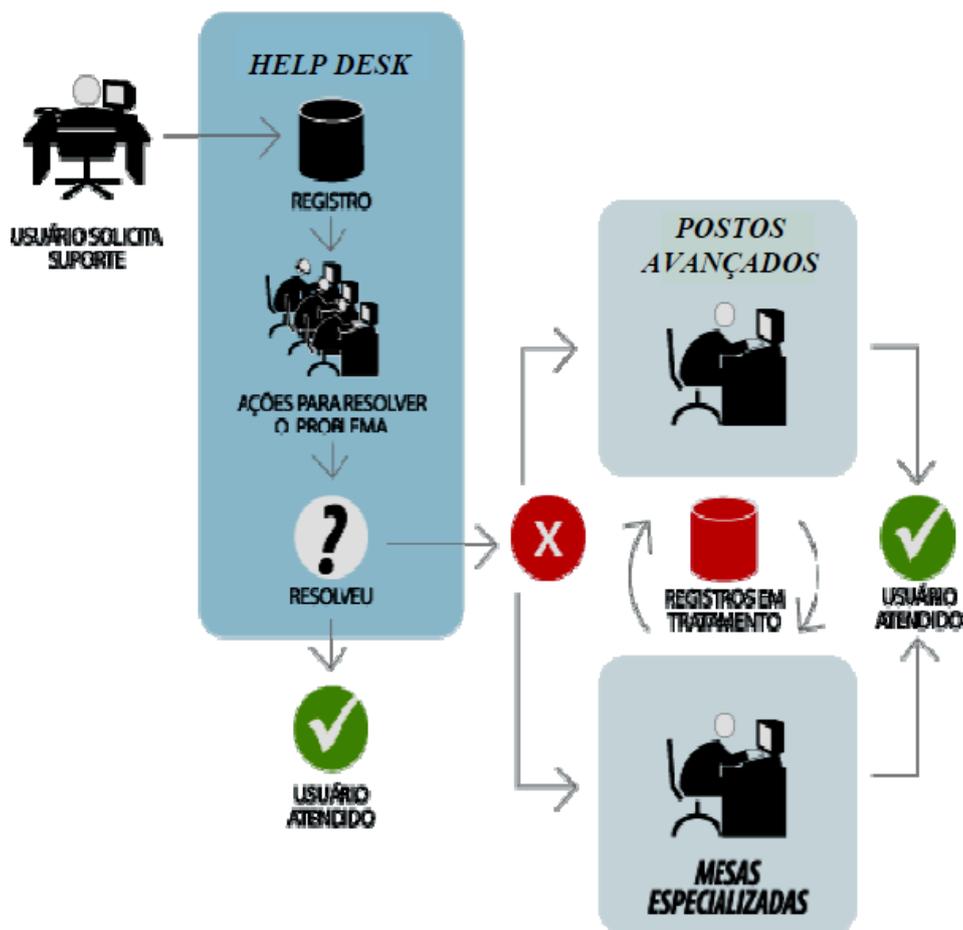


Figura 12: Fluxo de Atendimento
Fonte: Tecnologia da Informação da Empresa de Petróleo

Os principais componentes da estrutura de atendimento ao usuário de TI da empresa de petróleo, apresentados na figura anterior, têm a seguinte descrição básica:

- *Help-Desk* – É denominado atendimento de 1º nível. São equipes de atendimento remoto capaz de gerenciar, coordenar e resolver incidentes no menor tempo possível garantindo que nenhuma solicitação seja perdida, esquecida ou ignorada. É composta de técnicos de informática que recebem treinamento para resolver a maior parte dos atendimentos no primeiro contato. Quando se esgotam as alternativas para solução do atendimento executado pelo *Help-Desk*, o registro é direcionado para a equipe que possui maior capacidade em solucionar aquele registro. A equipe poderá ser uma Mesa especializada (ME) ou Posto avançado (PA). Existem duas estruturas de *Help Desk* na TI um na cidade do Rio de Janeiro e outra em Macaé. Trabalham em regime de 24 x 7 para contemplar todos os clientes do sistema da empresa de petróleo, em especial os trabalham em Plataformas de Exploração & Produção, Navios e Refinarias.

- Posto Avançado (PA) – É denominado atendimento de 2º nível. São equipes de atendimento físico, ou seja, de apoio presencial. São alocados junto ao Cliente. Atendem quando o a solicitação não pode ser resolvida remotamente e não se trata de um caso que possa ser resolvido por uma mesa especializada. Um exemplo típico é um problema de hardware em um equipamento. Existem PAs espalhados por todo o país. Onde houver equipamentos de TI e usuários, teremos pelo menos um PA próximo.

- Mesa Especializada (ME) – É denominada atendimento de 3º nível. São equipes de atendimento especializado de acordo com as características comuns dos serviços e complexidade do atendimento. A TI possui MEs nas áreas de Apoio ao Usuário, Infra-estrutura, SAP R/3 (Sistema de Aplicações e Produtos), Desenvolvimento e Agilidade (sistemas específicos dos clientes das áreas de negócio).

A Central de Atendimento, estudo de caso desse trabalho, apresenta oito principais atendimentos especializados ou grupos de serviços:

Assistência Multidisciplinar de Saúde (AMS): É o plano de saúde da empresa de petróleo.

Informações (PABX): Assuntos relacionados à Companhia, tais como localização de pessoas, informações de funcionários ou empresas terceirizadas da empresa de petróleo.

Telecomunicações (881): Infra-estrutura de telecomunicações, envolvendo consultoria, instalações, manutenção de toda a rede de telefonia e dados e programação das Centrais Telefônicas.

Serviços Gerais: Fornecimento de transporte, alimentação, serviços de infraestrutura, condomínio, manutenção predial e segurança, entre outros.

Centro de Operações Financeira da Petrobras (COFIP): Centraliza as principais atividades transacionais contábeis, financeiras e tributárias da Companhia.

Transporte: Atendimento receptivo sobre transporte aéreo de pessoas, cargas terrestre, marítima e aérea.

Serviços Diferenciados: Informações relacionadas ao setor de recursos humanos, cadastramento de fornecedores de materiais e serviços, verificação de notas de compra, entre outros.

Emergências: Operação, Manutenção e Emergências relacionadas à faixa de Duto, Alta Administração da Companhia, entre outros.

Desta forma, de acordo com as características dos serviços, as ilhas podem ser configuradas para atender mais de uma especialidade, ou seja, as equipes podem ser configuradas como *multi skill*.

O Quadro 1 apresenta um exemplo de configuração de uma ilha consolidada com agentes para atender especialidades de acordo com o perfil de atendimento.

Quadro 1: Consolidada com Agentes Multi SKILL
(Fonte: Central de Atendimento)

Especialidades dos Serviços Gerais	Configuração
AUTO ATENDIMENTO	Ilha Consolidada com Agentes Multi SKIL
BR CADASTRO FORNECEDOR	
GNC ES INFORMAÇÕES	
SERVIÇOS DE PESSOAL	
CAMPANHA BR	
SUPRIMENTOS	
RECADASTRAMENTO	
OUIDORIA	

Verifica-se que os agentes podem atender, por exemplo, tanto a especialidade BR Cadastro quanto Recadastramento nas ilhas especializadas em Serviços Gerais.

Dentre os grandes desafios de uma Central de Atendimento destaca-se garantir a qualidade e a confiabilidade dos serviços prestados através de uma estrutura de processos, indicadores, procedimentos e capacitação. Entende-se por procedimentos, um conjunto de documentos, em forma textual, disponíveis em meio eletrônico para dar suporte ao atendimento quando necessário.

Diante da multiplicidade de sistemas em uso e da variedade de dúvidas e incidentes possíveis, um dos instrumentos de trabalho mais importante das equipes de atendimento é o Portal de Atendimento. Nele há instruções sobre os aplicativos utilizados na empresa que orientem os atendentes durante uma ligação no momento em que os clientes entram em contato com a Central de Atendimento. O atendimento é realizado por meio de telefone, e-mail, intranet, fax ou carta o que retrata um *Contact Center*.

O portal foi inaugurado em agosto de 2011 contendo conteúdos da Tecnologia da Informação e Telecomunicações (TIC). Esse novo conceito de organização dos conteúdos permite a integração em um ambiente único, evita a duplicidade de informações e facilita o acesso e a utilização da intranet como ferramenta de trabalho. Além disso, todos os documentos são padronizados segundo normas internas da empresa brasileira de petróleo.

Atualmente o *site* possui 53 ilhas configuradas para atendimento. A Tabela 3 apresenta a quantidade de documentos por grupo de serviço, a quantidade de ilhas atualmente configuradas para atender o volume de ligações e o tempo médio de atendimento.

Tabela 3: Consolidado Indicadores dos Serviços da Central de Atendimento - Abril de 2011
(Fonte: Central de Atendimento)

	Grupo de Serviço	Documentos	Ilhas	Ligações	TMA
1	CS-AMS	425	9	91.114	0:03:45
2	CS-TELECOM	361	4	17.142	0:02:24
3	CS- EMERGÊNCIA	168	7	2.950	0:01:46
4	CS -COFIP	159	4	8.835	0:03:06
5	DIFERENCIADOS	131	7	4.031	0:03:44
6	CS-PABX	100	5	24.015	0:01:10
7	CS - GERAIS	85	6	9.354	0:01:15
8	CS - TRANSPORTE	78	1	6.639	0:03:23
	TOTAL	1498		164.080	

Em alguns casos, verificou-se junto ao especialista que um mesmo procedimento, contendo as mesmas informações, pode atender a mais de grupo de serviço, como por exemplo, um procedimento disponível para o CS-TELECOM também está disponível para CS-COFIP.

Observa-se que o serviço AMS destaca-se pelo maior volume de ligações (mais de 50% do total de ligações) e número de ilhas configuradas para atender o cliente. Assim como AMS, os serviços COFIP, Diferenciados e Transporte apresentam um maior tempo médio de atendimento em relação aos demais.

No Portal de Atendimento os documentos podem ser consultados através da identificação do grupo de serviço, identificados pela sigla CS, o que facilitou a organização da coleta de documentos.

A partir da coleta dessas informações, verificou-se que o serviço AMS também possui maior número de procedimentos disponíveis para apoiar o atendente durante uma ligação. A partir dessa constatação, ressalta-se a importância de criar uma forma de acompanhamento do volume de documentos disponíveis por grupo de serviço e uma maior facilidade de consulta a esses procedimentos para garantir a eficiência no atendimento.

O presente trabalho tem como objetivo aplicar as técnicas de Mineração de Texto para organizar os documentos de acordo com o grupo de serviço. A partir do conteúdo dos documentos disponíveis no Portal de Atendimento, será possível agrupá-los de acordo com a similaridade e obter palavras-chaves associadas a cada um dos grupos.

6.2.COMENTÁRIO

Um breve resumo sobre as Centrais de Atendimento foi apresentado neste capítulo. O objetivo foi destacar o avanço da tecnologia na área de telecomunicações, e a necessidade de gerenciamento do conhecimento. Utilizam-se cada vez mais de novos meios de comunicação para atender a necessidade de seus clientes, tornando clara a importância de automatizar os processos por meio de poderosas ferramentas como a Mineração de Textos.

7. METODOLOGIA

A metodologia adotada foi norteada pelos objetivos propostos, ou seja, o trabalho consiste na aplicação de técnicas da Mineração de Textos em uma base de dados textuais de uma Central de Atendimento.

A apresentação desta etapa está subdividida em: classificação da pesquisa, população e amostra, técnica de coleta, recursos utilizados e etapas da Mineração de Textos.

7.1. CLASSIFICAÇÃO DA PESQUISA

O presente trabalho pode ser classificado como pesquisa aplicada e descritiva. Segundo MARCONI e LAKATOS (2000) a definição aplicada se justifica por caracterizar o seu interesse prático, em que os resultados sejam utilizados na solução de problemas que ocorrem na realidade. Descritiva porque toca em quatro aspectos: descrição, registro, análises e interpretação do problema, objetivando seu funcionamento no presente.

Quanto aos fins, foi classificada como exploratória. Segundo COLLIS e HUSSEY (2005) uma pesquisa exploratória é realizada sobre um problema ou questão que ainda carece de maiores estudos sobre o assunto, tendo como objetivo procurar idéias ou hipóteses.

Em conformidade, MALHOTRA (2006) afirma que exploratórias são investigações de pesquisa empírica cujo objetivo é a formulação de questões ou de um problema, com tripla finalidade: desenvolver hipóteses; aumentar a familiaridade do pesquisador com um ambiente, fato ou fenômeno para a realização de uma pesquisa futura mais precisa; modificar e clarificar conceitos. Assim abre-se a possibilidade de ampliar a visão sobre o tema e de que novas idéias sejam percebidas, descobertas e testadas em novos estudos.

Desta forma, o escopo é extrair conhecimento da população, no caso a base de documentos, através da aplicação de técnicas de Mineração de Textos reconhecidas em outros trabalhos científicos. Visa ainda apresentar novas formas de acompanhamento e controle dos processos executados diariamente em uma empresa.

7.2. RECURSOS UTILIZADOS

O principal programa utilizado foi a ferramenta de Mineração de Textos, *PolyAnalyst* (PA) descrita na (Secção 3). A ferramenta permitiu realizar uma análise exploratória desde a fase inicial correspondente ao pré-processamento até a visualização dos resultados, ou seja, o pós-processamento.

A ferramenta *PolyAnalyst* foi selecionada para ser utilizada nas etapas de pré-processamento e classificação de textos. É um software desenvolvido pela empresa *Megaputer Intelligence Inc.*, tendo sua primeira versão lançada em 1997. PA é um programa de mineração de dados (estruturados e não estruturados) com variados propósitos, visando facilitar as pesquisas e a tomada de decisões a partir da análise dos dados. Contém várias ferramentas para exploração dos dados, que possibilitam ligá-los, separá-los, analisá-los e sumarizá-los e que estão disponíveis para que o usuário possa compor seus projetos de análise de forma personalizada.

Para etapa de pré-processamento, foram utilizadas as seguintes ferramentas:

TotalDocConverter e *SmartPdfCreator* : padronização dos documentos para o formato txt.

Snowball¹: além da lista de *stopword* criada, manualmente, a partir da análise dos documentos, acrescentou-se a lista amplamente divulgada na literatura.

Lista de sinônimos do Portal Atendimento: a lista de sinônimos, utilizada como *thesaurus*, é criada por analistas da qualidade para auxiliar na consulta aos documentos durante o atendimento.

Base de Sinônimos TEP 2.0²: representa o *Thesaurus* Eletrônico para o Português do Brasil, onde é possível obter todos os *synsets*, (do inglês, *synonym set*) para o português.

RSLP Stemmer³: Removedor de Sufixos da Língua Portuguesa utilizado como técnica de *stemming* sobre os documentos. A partir das instruções para aplicar o RSLP, foi necessário criar um programa escrito em linguagem C.

¹ Disponível em: <http://snowball.tartarus.org/algorithms/portuguese/stemmer.html> (último acesso em 01/03/2012)

² Disponível em: <http://www.nilc.icmc.usp.br/tep2/download.htm> (último acesso em 01/03/2012)

³ Disponível em: http://www.inf.ufrgs.br/~arcoelho/rspl/integrando_rslp.html (último acesso em 01/03/2012)

Programa Matlab: extrair tópicos pelo LDA⁴. O programa executa um exemplo do *LDA Gibbs Sampler* em um pequeno conjunto de documentos para extrair um conjunto de tópicos e apresentar as palavras mais representativas em cada tópico. Também foi necessário criar um algoritmo GeraTxtLDA.m adaptado para importar os documentos da pesquisa.

Para complementar o estudo, foi implementada em C uma nova abordagem para selecionar os documentos mais representativos, inspirado no Sistema Imunológico Artificial (Secção 2.9), sendo considerada uma estratégia para redução da dimensionalidade.

Todos os experimentos realizados foram avaliados a partir do mesmo computador para obter uma melhor avaliação do desempenho. A idéia é utilizar um computador a nível de consumidor, ou seja, um computador de uso geral.

7.3. POPULAÇÃO E AMOSTRA

Dentre os 1507 documentos coletados, 1344 documentos foram considerados válidos para os objetivos da pesquisa, o que representa uma amostra de aproximadamente 89.11% da população. Esses documentos referem-se ao período compreendido entre agosto de 2010 à abril de 2011.

7.4. TÉCNICAS DE COLETA DE DADOS

Uma das primeiras fontes de informação a serem consideradas é a existência de registros na própria organização, sob a forma de documentos, fichas, relatórios ou arquivos em computador. As informações utilizadas nessa pesquisa foram extraídas da base de conhecimento de uma Central de Atendimento prestadora de serviços de tecnologia de informação, que faz parte de um sistema informatizado (Portal de Atendimento) desenvolvido por uma empresa brasileira de petróleo. O uso de documentos disponíveis reduz tempo e custo de pesquisas para avaliação.

⁴ Disponível em: http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm (último acesso em 01/03/2012)

O Portal de Atendimento reúne informações das áreas de Tecnologia da Informação e Telecomunicações (TIC). Por motivos de segurança das informações, a pesquisa teve acesso somente aos documentos utilizados pela Central de Atendimento, por meio de uma chave de segurança autorizado pelo responsável da área de TI.

Ao acessar o Portal de Atendimento, verificou-se que os documentos são organizados por grupo de serviço e identificados pela sigla CS (Central de Serviço). Desta forma foi possível coletar, manualmente, cada documento e organizar de acordo com o serviço.

Segundo Schiessl apud Weiss (2005), se os documentos já estão identificados, a principal tarefa é efetuar a eliminação de ruídos e assegurar que a amostra seja de boa qualidade. Assim, como nos dados não textuais, a intervenção humana pode comprometer a integridade dos dados no processo de coleta, por isso requer extremo cuidado nesta tarefa.

Devido à sua importância, a coleta de documentos, contou com o apoio de analistas da qualidade para sua avaliação e esclarecimentos de dúvidas em relação ao volume de documentos por grupo de serviços.

7.5. ETAPAS DA MINERAÇÃO DE TEXTOS

O processo de Mineração de Textos pode ser resumida em três etapas principais: Pré-Processamento, Processamento e Pós-Processamento.

Dentre as etapas, a que exigiu maior atenção foi a etapa de pré-processamento para garantir a qualidade dos dados e aplicar duas recentes abordagens na área de Mineração de Textos: o Algoritmo Supressor de Textos (SeleSupText) e o modelo Alocação Latente de *Dirichlet* (LDA).

Esquemáticamente as etapas se inter-relacionam como descrito na Figura 13:

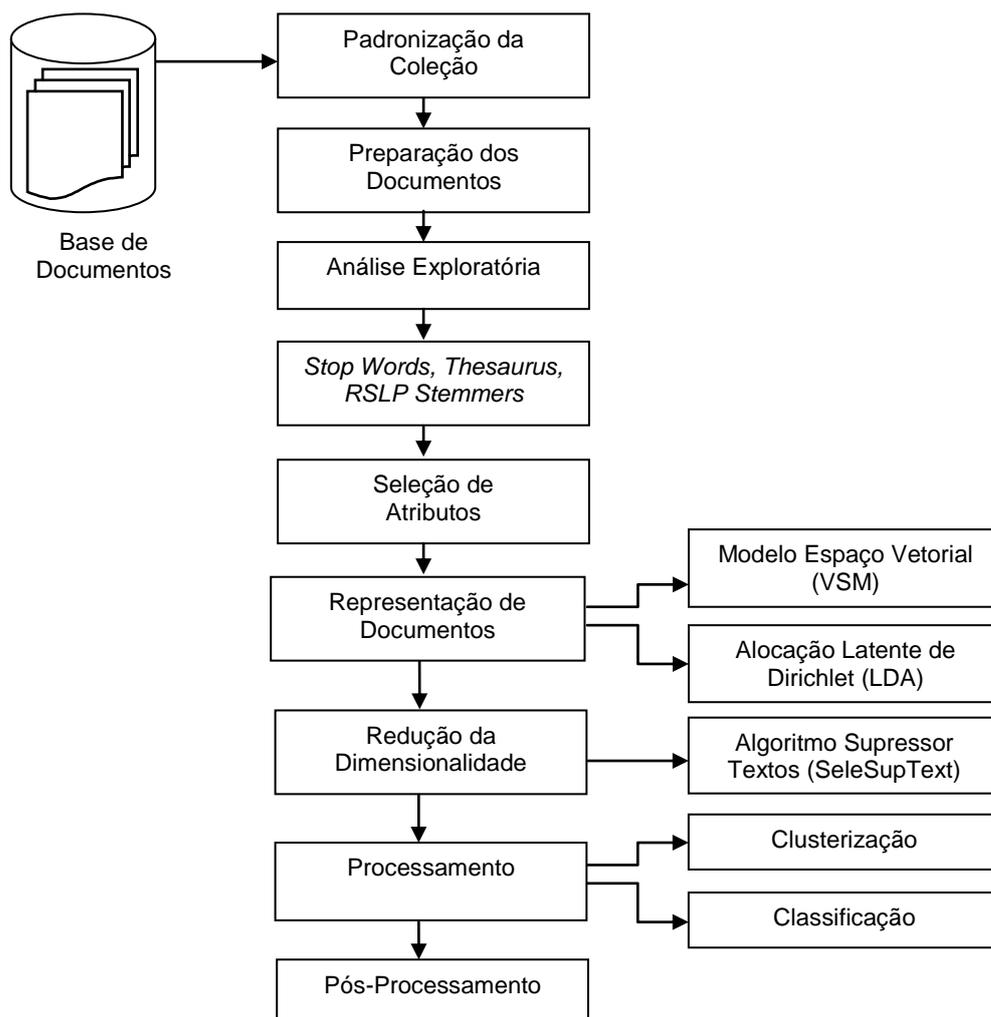


Figura 13: Visão Geral da Metodologia aplicada sobre a base de documentos

O detalhamento de cada uma das etapas apresentadas na metodologia é descrita a seguir. Dentre estas etapas, a etapa de pré-processamento demanda uma atenção especial para garantir a qualidade dos dados analisados e para assegurar maior fidelidade aos resultados obtidos pelos algoritmos de aprendizado.

7.5.1. Base de Documentos

A base de documentos na qual nos referimos é denominada Base de Conhecimento. É formada por documentos que auxiliam a execução de atividades necessárias para oferecer um atendimento de qualidade conforme instruções e

procedimentos internos da Central de Atendimento. Esses documentos representam o objeto de estudo a ser minerado.

A Figura 14 apresenta um conjunto de documentos agrupados de acordo com o tipo de serviço. Nessa pesquisa apenas os documentos e os grupos de serviços, respectivamente, registros e classe, são considerados relevantes para as etapas de processamento.

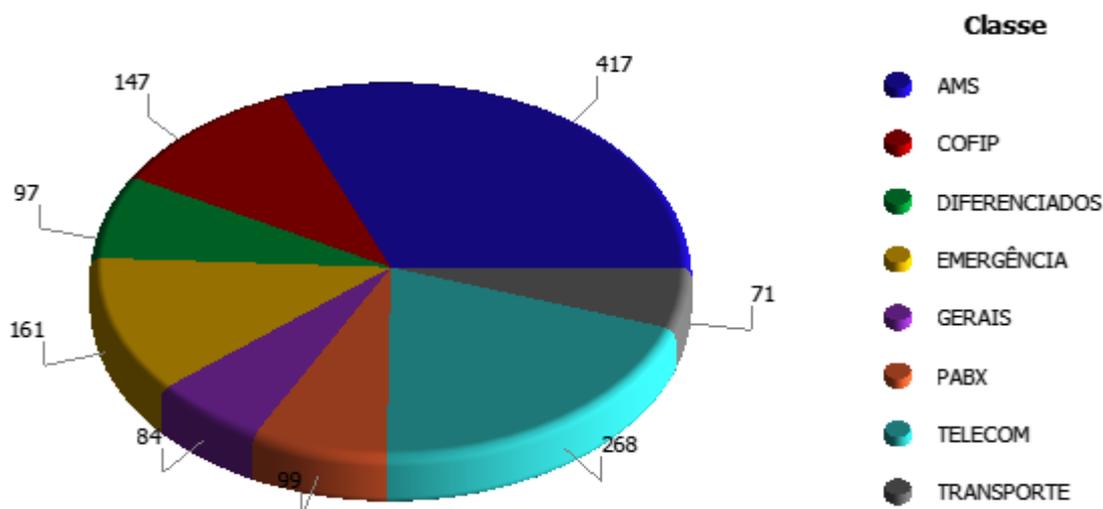


Figura 14: Amostra dos Documentos Representativos

Detalhes sobre cada um dos serviços é apresentada no Estudo de Caso (Secção 4). Trata-se de uma base desbalanceada, com variação do número de registro por classe, o que representa um grande desafio para as tarefas de Mineração de Textos.

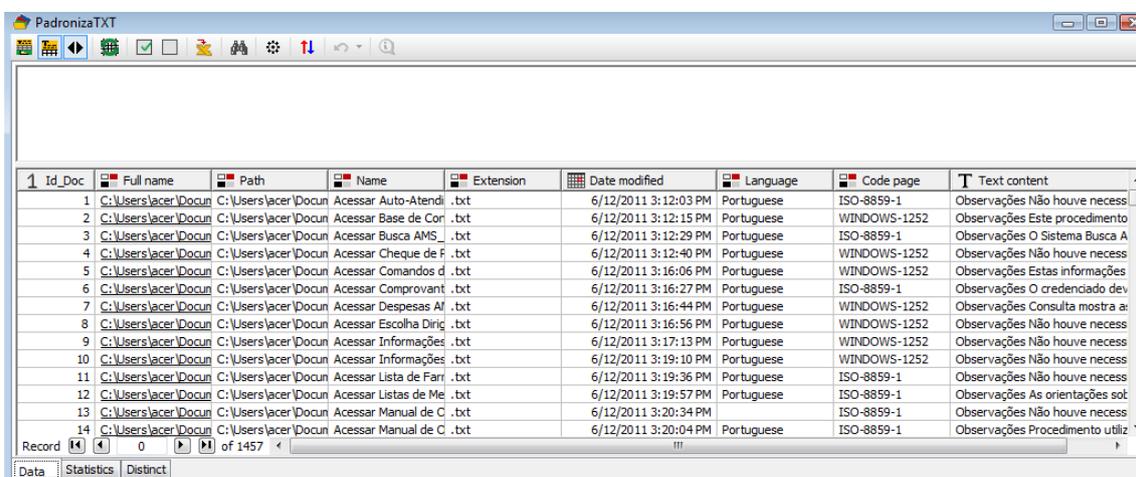
7.5.2. Padronização da Coleção

Nesse passo, analisa-se a coleção com a qual se irá trabalhar, principalmente no que remete a dois aspectos: verificação de representatividade da coleção, identificando se os documentos disponíveis são suficientes e representativos sobre o domínio; e se os documentos não possuem problemas, como caracteres corrompidos (NOGUEIRA, 2008).

Assim, aplica-se uma sequência de passos de forma recorrente, a fim de que a coleção de textos trabalhada apresente os requisitos necessários.

Após a coleta dos dados foi possível observar que existiam vários formatos, onde 79% correspondiam aos documentos do tipo doc e o restante distribuídos entre os formatos pdf, ppt, xls e odt. Neste momento, foi realizada a conversão de 97% dos documentos, colocando-os na forma de texto (txt) e descartando aqueles que não puderem ser convertidos (normalmente, documentos que continham somente figuras).

Após a padronização dos arquivos em um mesmo formato, organizados de acordo com o serviço, foi realizada a importação automática de todos os documentos. A Figura 15 destaca algumas características da coleção, tais como: diretório onde se encontra o arquivo, o nome do arquivo, extensão, o conteúdo, linguagem, entre outros, representando a unificação dos documentos em uma base única, onde cada registro corresponde a um documento.



Id_Doc	Full name	Path	Name	Extension	Date modified	Language	Code page	Text content
1	C:\Users\lacer\Docum	C:\Users\lacer\Docum	Acessar Auto-Atendi	.txt	6/12/2011 3:12:03 PM	Portuguese	ISO-8859-1	Observações Não houve necess
2	C:\Users\lacer\Docum	C:\Users\lacer\Docum	Acessar Base de Con	.txt	6/12/2011 3:12:15 PM	Portuguese	WINDOWS-1252	Observações Este procedimento
3	C:\Users\lacer\Docum	C:\Users\lacer\Docum	Acessar Busca AMS_	.txt	6/12/2011 3:12:29 PM	Portuguese	ISO-8859-1	Observações O Sistema Busca A
4	C:\Users\lacer\Docum	C:\Users\lacer\Docum	Acessar Cheque de F	.txt	6/12/2011 3:12:40 PM	Portuguese	WINDOWS-1252	Observações Não houve necess
5	C:\Users\lacer\Docum	C:\Users\lacer\Docum	Acessar Comandos d	.txt	6/12/2011 3:16:06 PM	Portuguese	WINDOWS-1252	Observações Estas informações
6	C:\Users\lacer\Docum	C:\Users\lacer\Docum	Acessar Comprovant	.txt	6/12/2011 3:16:27 PM	Portuguese	ISO-8859-1	Observações O credenciado dev
7	C:\Users\lacer\Docum	C:\Users\lacer\Docum	Acessar Despesas Af	.txt	6/12/2011 3:16:44 PM	Portuguese	WINDOWS-1252	Observações Consulta mostra a:
8	C:\Users\lacer\Docum	C:\Users\lacer\Docum	Acessar Escolha Dirij	.txt	6/12/2011 3:16:56 PM	Portuguese	WINDOWS-1252	Observações Não houve necess
9	C:\Users\lacer\Docum	C:\Users\lacer\Docum	Acessar Informações	.txt	6/12/2011 3:17:13 PM	Portuguese	WINDOWS-1252	Observações Não houve necess
10	C:\Users\lacer\Docum	C:\Users\lacer\Docum	Acessar Informações	.txt	6/12/2011 3:19:10 PM	Portuguese	WINDOWS-1252	Observações Não houve necess
11	C:\Users\lacer\Docum	C:\Users\lacer\Docum	Acessar Lista de Farr	.txt	6/12/2011 3:19:36 PM	Portuguese	ISO-8859-1	Observações Não houve necess
12	C:\Users\lacer\Docum	C:\Users\lacer\Docum	Acessar Listas de Me	.txt	6/12/2011 3:19:57 PM	Portuguese	ISO-8859-1	Observações As orientações sot
13	C:\Users\lacer\Docum	C:\Users\lacer\Docum	Acessar Manual de C	.txt	6/12/2011 3:20:34 PM	Portuguese	ISO-8859-1	Observações Não houve necess
14	C:\Users\lacer\Docum	C:\Users\lacer\Docum	Acessar Manual de C	.txt	6/12/2011 3:20:04 PM	Portuguese	ISO-8859-1	Observações Procedimento utiliz

Figura 15: Importação de Arquivos no mesmo formato

Nessa etapa, foram identificados manuais das ferramentas utilizadas pelos atendentes como procedimentos com várias páginas. Recomenda-se, nesse caso, dividir o manual por capítulos para facilitar a consulta do atendente, enquanto o cliente aguarda na linha a sua solicitação.

7.5.3. Preparação dos Documentos

Esta etapa envolve o processo de limpeza da base através da remoção de caracteres sem representatividade, que não agregam valor à análise, como por exemplo, acentos, pontuação, cedilhas, números e underlines.

Nessa etapa, utiliza-se a conversão de caracteres para a sua forma minúscula, para conferir maior agilidade no processo de indexação (LOPES, 2009).

O projeto criado para a preparação da coleção pode ser visualizado através da Figura 16, utilizando-se os nós *Replace Terms*, *Extract Terms* e *Derive Node*. Nessa etapa foi identificada a duplicidade de alguns documentos.

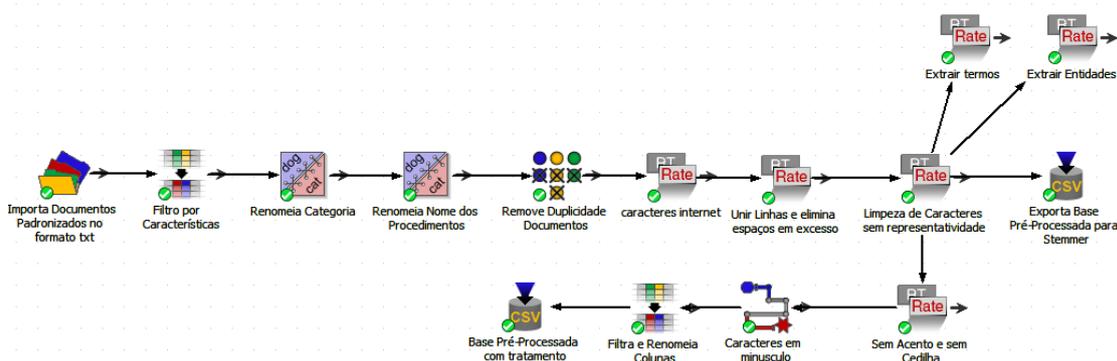


Figura 16: Projeto Padronização da Coleção de Documentos

No projeto, identifica-se a saída de dois tipos de arquivos, sendo um para aplicação do *stemmer* conforme regras do método RSLP e outro para realizar diretamente o processamento.

Já nessa etapa foi possível extrair conhecimento útil a partir desses documentos. Como os procedimentos seguem uma padronização foi possível extrair os Gerentes e os Responsáveis pelos respectivos procedimentos por grupo de serviço. Dessa forma, caso ocorra mudanças na organização referente a essas responsabilidades, como por exemplo, transferência, demissão ou remanejamento, é possível identificá-los de forma a garantir a informação disponível sempre atualizada.

Também foram identificados documentos em duplicidade (Secção 6.1), reduzindo o número de documentos de 1457 para 1344. Foi verificado junto ao especialista que um mesmo procedimento pode atender a mais de um serviço.

Por fim, faz-se uma avaliação subjetiva da coleção disponível com o objetivo de verificar a representatividade da coleção. Este processo de preparação da base de dados foi repetido várias vezes, até atingir um ponto considerado satisfatório.

A Figura 17 apresenta a base pré-processada, conforme projeto de padronização, utilizada como fonte de pesquisas ao longo de todo o trabalho.

1 Id_Doc	Classe	Nome do Procedimento	Procedimento
1	AMS	acessar auto atendimento telefônico	observacoes nao houve necessida
2	AMS	acessar base de conhecimento consultar sams	observacoes este procedimento e
3	AMS	acessar busca ams	observacoes o sistema busca ams
4	AMS	acessar cheque de pagamento de credenciado pitx	observacoes nao houve necessida
5	AMS	acessar comandos do sam	observacoes estas informacoes tar
6	AMS	acessar comprovante de imposto de renda credenciado	observacoes o credenciado deve c
7	AMS	acessar despesas ams empregados ativos	observacoes consulta mostra as de
8	AMS	acessar escolha dirigida e livre escolha	observacoes nao houve necessida
9	AMS	acessar informações sobre o pad na intranet	observacoes nao houve necessida
10	AMS	acessar informações sobre o pasa na intranet	observacoes nao houve necessida
11	AMS	acessar lista de farmácias	observacoes nao houve necessida
12	AMS	acessar listas de medicamentos	observacoes as orientacoes sobre
13	AMS	acessar manual de orientações aos beneficiário	observacoes nao houve necessida
14	AMS	acessar manual de orientações técnicas	observacoes procedimento utilizad
15	AMS	acessar pae programa de assistência especial	observacoes nao houve necessida
16	AMS	acessar portal ams	observacoes somente pela intrane
17	AMS	acessar processo de reembolso	observacoes nao houve necessida

Record 0 of 1344

Figura 17: Visualização parcial da Base de Documentos Pré-Processada

Assim, a coleção de documento é representada pelos seguintes campos: Id_doc: identificador do documento; classe: grupo de serviço no qual pertence o documento; nome do procedimento: nome do arquivo do documento; procedimento: o conteúdo do documento.

7.5.4. Análise Exploratória Inicial

No intuito de obter informações que pudessem estar ocultas no conteúdo dos documentos de cada grupo de serviço, foi realizada uma análise exploratória geral sobre os documentos por grupo (Figura 18) e, posteriormente, uma análise em cada um dos grupos (Figura 19).

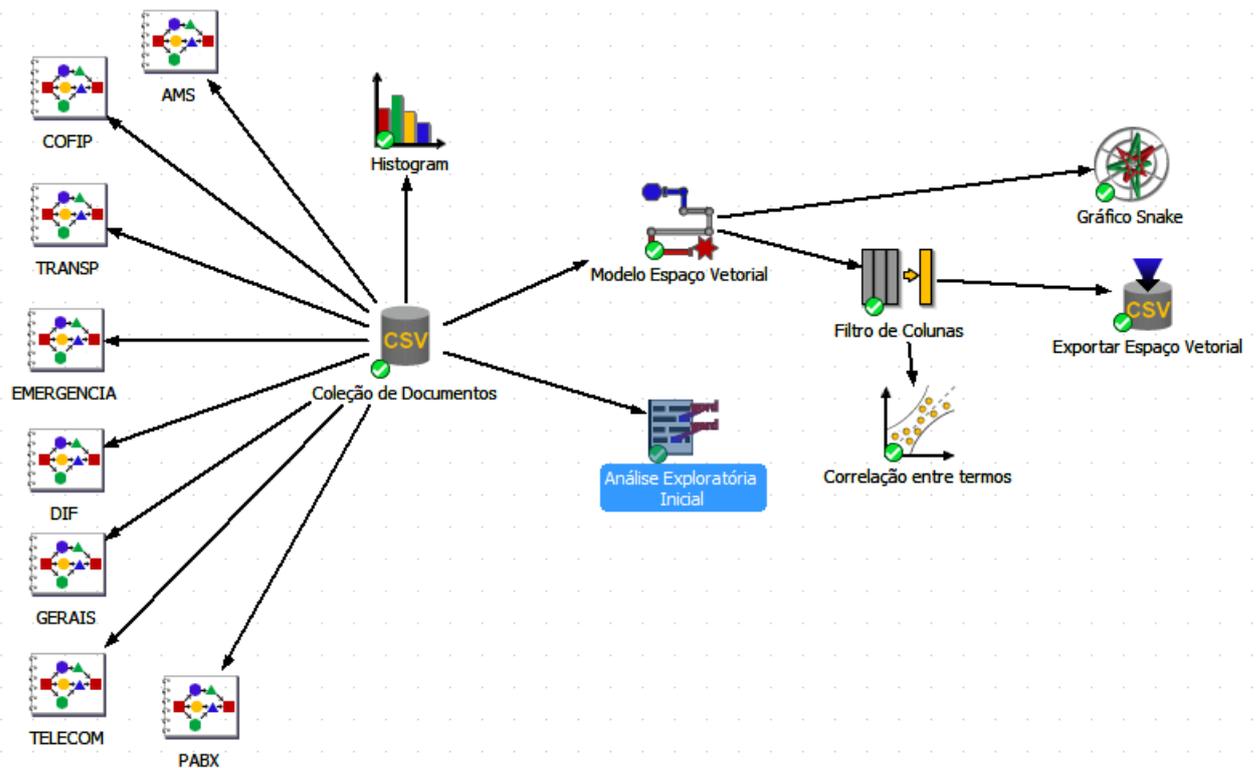


Figura 18: Projeto para análise exploratória geral dos documentos

Os documentos foram analisados estatisticamente utilizando gráficos e as análises de correlação entre os termos.

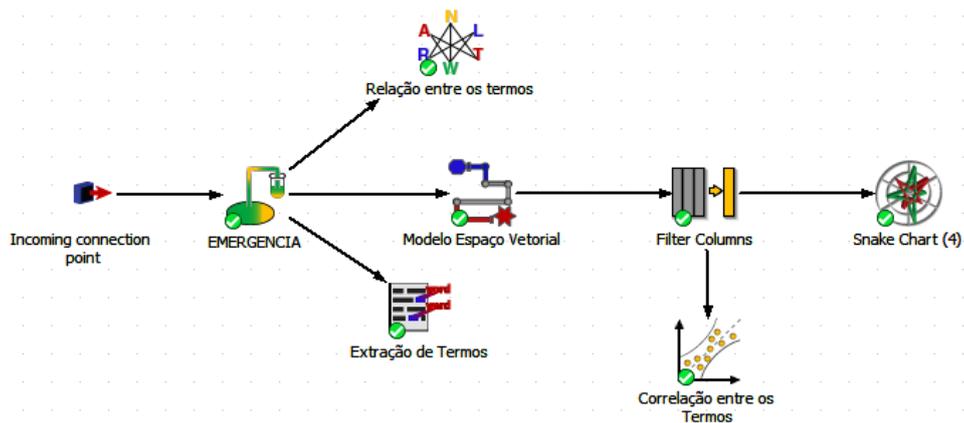


Figura 19: Projeto para análise exploratória dos documentos referente ao grupo de serviço Emergência

Primeiramente, utilizou-se do nó *Keyword Extraction* para obter uma visão geral das palavras identificadas nos documentos. O nó é geralmente usado na fase de pré-processamento para exploração dos textos. É possível verificar informações tais como as palavras-chave mais frequentes.

A Figura 20 apresenta a visualização de uma lista dos registros e conteúdo do texto, onde uma palavra-chave é mencionada. A palavra-chave é destacada para a exploração fácil no contexto. Note que a ferramenta também adiciona uma coluna temporária para esse conjunto de dados chamado Relevância, que contém uma medida que relaciona o quão relevante é um registro. A relevância é calculada numa escala de 0 a 100, considerados os 100 registros mais relevantes.

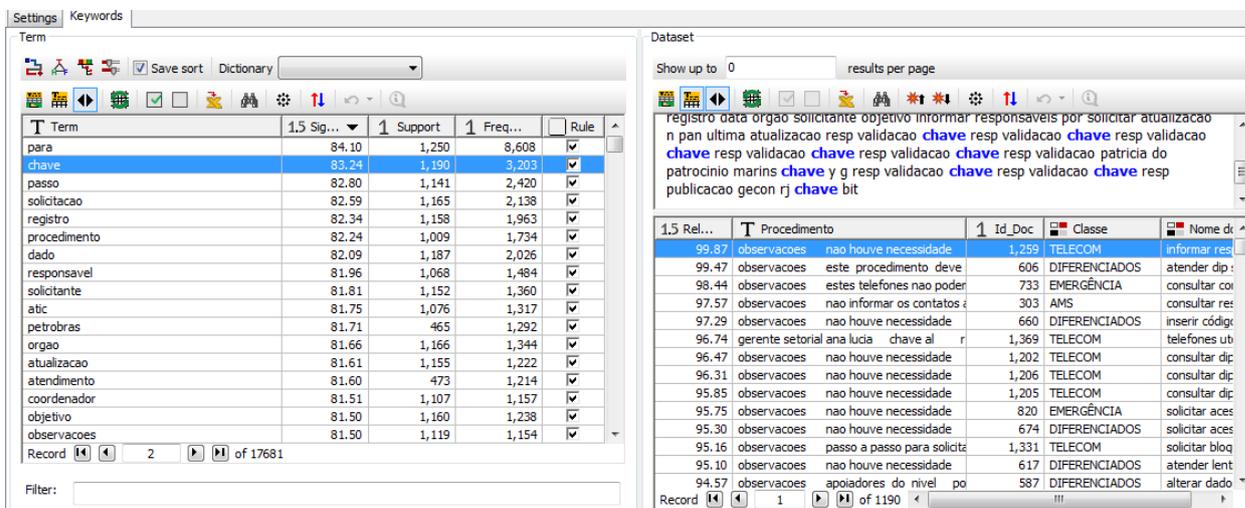


Figura 20: Visualização das Palavras Relevantes sem o uso de stop words e dicionários

Observa-se a geração de um grande número de termos (17.681) que representam a coleção. Consideram-se, aqui, termos como palavras simples (*one-grams*). Todos os termos são gerados considerando cada documento como uma *bag-of-words*, sem considerar informações a respeito do contexto em que se encontra. Segundo (FORMAN, 2003), o número de palavras candidatas a atributos excede o número de documentos em mais de uma ordem de magnitude, gerando matrizes esparsas e de alta dimensionalidade.

Diante do grande volume de termos, foi necessário recorrer ao gráfico *Snake* para visualizar as distribuições de diversas dos termos dos documentos de uma só vez. É possível visualizar os termos que ocorrem com maior frequência por grupo.

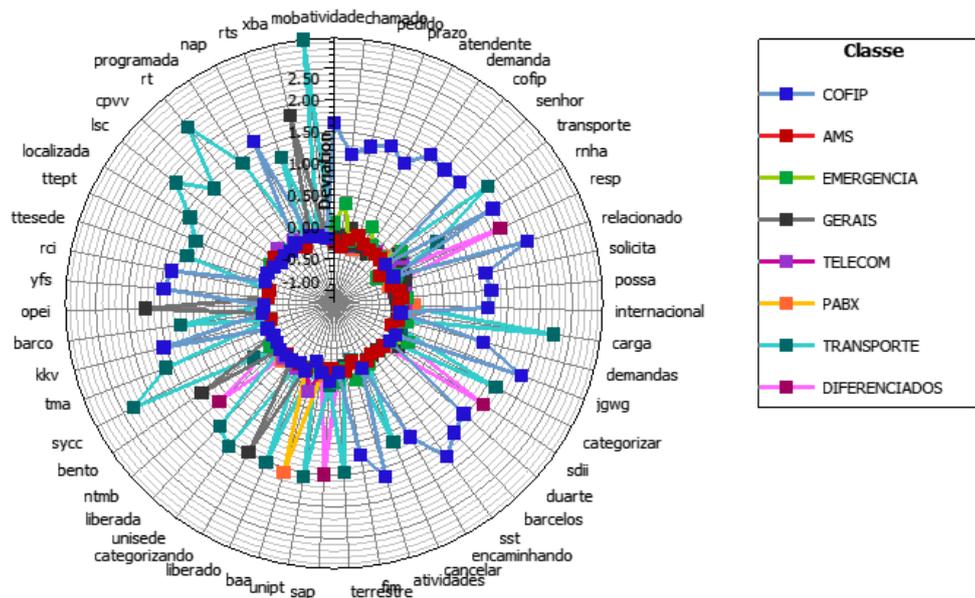


Figura 21: Gráfico Snake para visualização dos termos com maior frequência

A Figura 21 mostra um gráfico *Snake*, onde foram selecionados alguns termos e as oito categorias de serviços. Alguns termos com apenas quatro caracteres foram definidos como *stop words*, pois segundo a especialista de domínio tais termos identificam os analistas da qualidade que atualizaram os procedimentos.

7.5.5. Geração de Stop Words, Dicionário, Thesaurus

Em seguida, foi construída a lista de *stop words* na língua portuguesa, a partir da lista *snowball*, que foi aperfeiçoada à medida que as análises exploratórias foram aplicadas utilizando recursos da ferramenta *PolyAnalyst*, utilizando como por exemplo, a tarefa *Extract Terms*. Ou seja, em um processo de Mineração de Dados típico, os aperfeiçoamentos, inclusive da lista de *Stop Words* são realizadas de maneira interativa. (LOPES, 2009).

Além da exclusão das palavras sob os critérios descritos acima, foram adicionadas à *Stop List*, o Dicionário TEP 2.0 (*Thesaurus Eletrônico para o Português*), que contém 19.888 conjuntos de sinônimos e 44.678 unidades lexicais, tendo a média de 2,5 unidades por conjunto de sinônimos (MAZIERO *et al.*, 2008).

A Figura 22 apresenta listas de *stop words*, dicionários e sinônimos criados para explorar a base e garantir a representatividade dos termos considerados

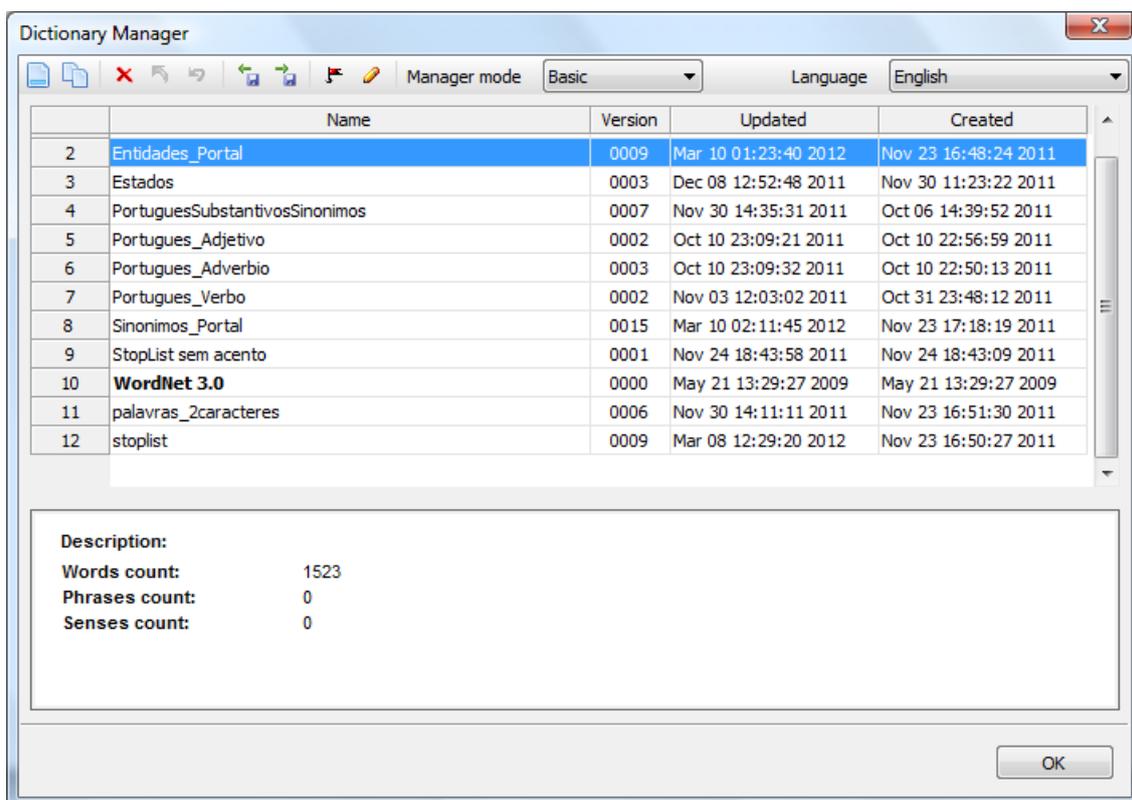


Figura 22: Geração Lista de Stop Words

O uso da lista de sinônimos disponibilizada no Portal de Atendimento também foi utilizada como *thesaurus* e avaliada na etapa de processamento.

A Figura 23 apresenta um exemplo da utilização do thesaurus sobre o nome do procedimento dos documentos. Nesse caso, as palavras auto atendimento, disque cidadão e comperj foram substituídas por uma única palavra, autoatendimento.

1,5 Id_doc	Classe	T Nome Procedimento	T Nome Procedimento_new
1.00	AMS	acessar auto atendimento telefonico	acessar autoatendimento telefonico
689.00	EMERGENCIA	ars disque cidadao	ars autoatendimento cidadao
683.00	EMERGENCIA	acessar aplicacoes da comperj via citrix	acessar aplicacoes da autoatendimento via citrix
694.00	EMERGENCIA	atender aplicacoes da comperj via citrix	atender aplicacoes da autoatendimento via citrix

Figura 23: Exemplo do uso do uso de thesaurus a partir dos sinônimos criado pela própria Central de Atendimento

Optou-se por avaliar o *thesaurus*, da lista de sinônimos do Portal de Atendimento, para evitar a substituição de termos que não correspondem a realidade do Centro de Serviços.

7.5.6. Seleção de Atributos

Como o número de termos permanece muito grande, o uso de métodos eficazes para seleção de atributos torna-se essencial para garantir a validade e da eficiência do processo de extração de conhecimento, na medida em que delimita o domínio a ser tratado pelos algoritmos (NOGUEIRA, 2008).

Diante desta constatação, foram realizados dois critérios de filtro, a partir das medidas estatísticas, significância e a frequência dos termos (TF) descritos a seguir:

1. As palavras com alta significância, consideradas maior ou igual 80%, foram eliminadas devido a presença em grande parte dos documentos.
2. Os termos que frequência menor ou igual a três, foram eliminados como poder ser visualizado na Figura 24, fundamentado em JOACHIMS (1998). De fato, ao final da aplicação destas técnicas o número de termos foi reduzido de 11.329 para 3.835 termos.

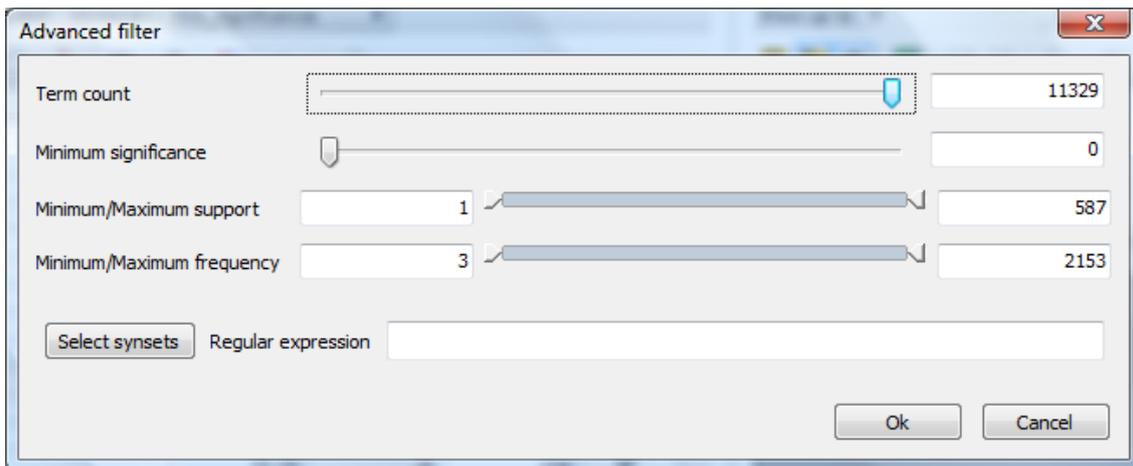


Figura 24: Seleção de Atributos

Após a redução de termos utilizou-se o nó *Link Terms* para gerar uma visualização das associações entre várias palavras-chave. Identificam-se rapidamente as relações entre as palavras de um conjunto de documentos, revelando mais detalhes sobre as idéias originais encontradas no texto.

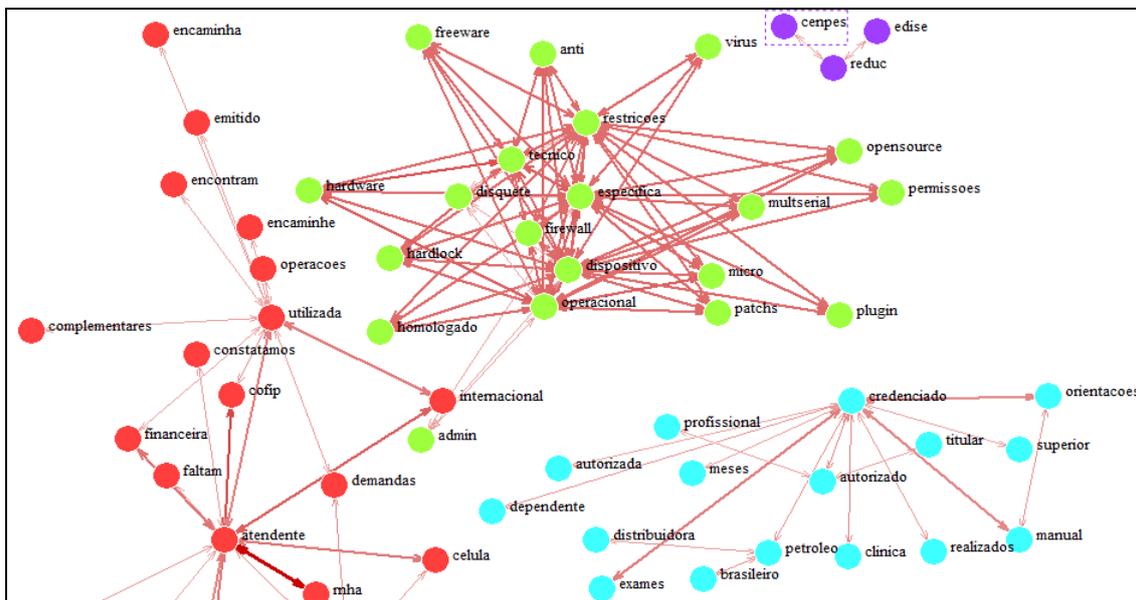


Figura 25: Relacionamento entre os termos a partir da métrica de Suporte

A métrica Suporte da ferramenta *PolyAnalyst* representa a contagem de registros em que duas palavras ocorreram em conjunto, ou seja, determina quão fortemente essas duas palavras estão relacionadas umas com as outras.

Aumentando o suporte mínimo reduz o número de associações encontradas pelo algoritmo, evidenciando, ainda mais, as fortes correlações.

Alguns serviços são facilmente identificados pela associação das palavras, como por exemplo, as palavras credenciado e exames retratam o serviço AMS, cofip e atendente refere-se ao serviço COFIP, e hardware e dispositivo são termos relacionados ao serviço TELECOM.

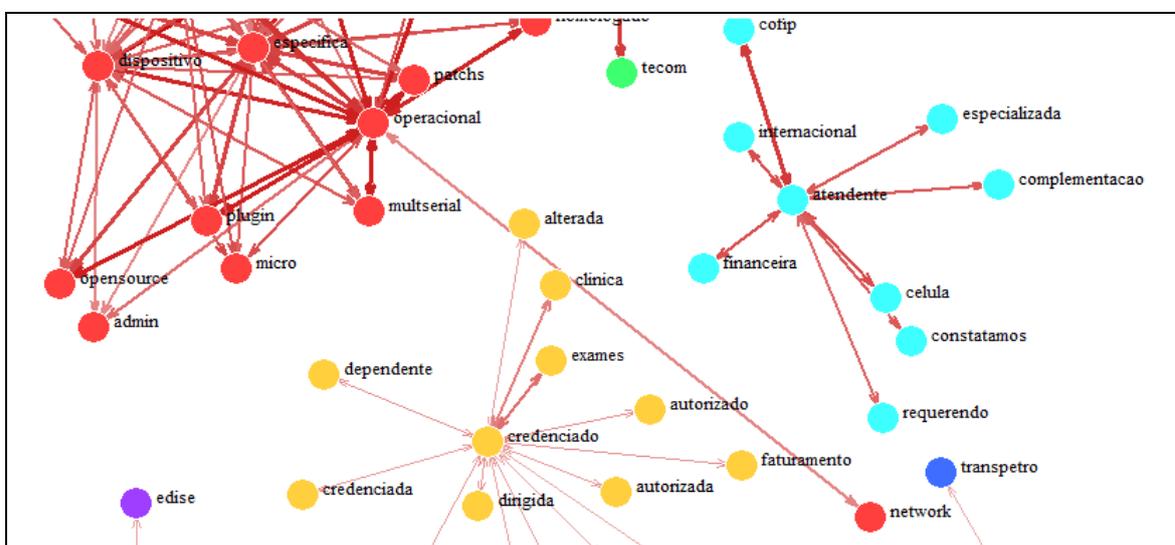


Figura 26: Relacionamento entre os termos a partir da métrica de Tensão

Por outro lado, a métrica tensão é calculada como um logaritmo do valor de probabilidade de uma relação entre dois termos. Uma maior tensão representa uma maior relação entre os termos. Observa-se pela Figura 26, a visualização de palavras que podem ser consideradas tanto *stop words* quanto palavras-chaves.

Utilizando-se essa métrica é possível destacar ainda mais a associação entre as palavras apresentadas como exemplo na métrica suporte.

7.5.7. RSLP Stemmer

Por fim, aplicou-se o RSLP *Stemmer* para avaliar e comparar os resultados na coleção de documentos. Primeiramente, para aplicar o *Stemmer* foi necessário preparar a base original mantendo acentos e nomes próprios iniciados por letras maiúsculas conforme apresentado na etapa de padronização da coleção.

Desta forma, sempre que uma palavra a ser processada pelo *stemmer* inicia por letras maiúsculas, é verificada sua presença no dicionário de nomes próprios. Caso a palavra seja encontrada neste dicionário, o algoritmo passa diretamente à função de remoção de acentos, ou seja, não é aplicado qualquer passo de redução sobre esta palavra. Caso contrário, a palavra é processada normalmente.

Com o intuito de evitar o *stemming* de nomes próprios, que foi uma das dificuldades apontadas por ORENGO e HUYCK (2001), adicionou-se um dicionário denominado Entidades Nomeadas, à nova implementação do RSLP.

Em seguida, foi incorporado ao dicionário criado por COELHO (2007) o dicionário denominado Entidades Portal, contendo os nomes próprios da base de textos da pesquisa e por fim, o arquivo de configuração do RSLP Stemmer foi atualizado conforme a seguir:

```
STEPS_FILE=steprules.txt
DO_STEMMING=YES
REPLACE_ISO_CHARS=YES
USE_STEM_DICTIONARY=YES
USE_NAMED_ENTITIES=YES
NAMED_ENTITIES_FILE=entidades_portal.txt
STEM_DICT_MAX_SIZE=512
NAMED_ENTITIES_DICT_MAX_SIZE=50
```

Figura 27: Parâmetros do arquivo de configuração do RSLP Stemmer

A Figura 28 apresenta um exemplo de documentos com o uso do RSLP Stemmer. Observa-se que as preposições, os artigos e os nomes próprio, em destaque, não sofreram o processo de *stemmer*. Foi possível aplicar o *stemmer* em todos os 1.457 registros. As palavras com acento foram também atualizadas.

observação: não houve necessidade de passar a senha para acessar o atendimento telefônico regional, pois o endereço www.portalpetrobras.com.br, acesso central de serviços, informe no campo o serviço desejado, aut atende no link indicado, outra abrangência selecionada, compartilhe clique em busca geral, acesso ao link, senha de atendimento telefônico, clique em acesso direto, informe a chave e a senha da solicitação registrada, data da organização, solicitação, infratratativa, arquivo objeto, procedimento orientado, com acesso ao serviço de atendimento telefônico, última atualização, responsável: Maria Luiza Castro Passini, chave, senha, coordenador, supervisor, gerente, aprovador.

1.5 Id_Doc	Serviço	Nome do P...	Documento_Pré-Processado
1.00	AMS	acessar auto atendi	observação: não houve necess... pass a pass pa
2.00	AMS	acessar base de con	observação: est proced e utiliz na region sp e b
3.00	AMS	acessar busca ams	observação: o sistem busc am e utiliz par encor
4.00	AMS	acessar cheque de p	observação: não houve necess... pass a pass pa
5.00	AMS	acessar comandos de	observação: est inform também pod ser verific
6.00	AMS	acessar comprovant	observação: o credenci dev cri sua propr senh
7.00	AMS	acessar despesas an	observação: consult mostr as desp debit no ult
8.00	AMS	acessar escolha dir	observação: não houve necess... pass a pass pa
9.00	AMS	acessar informações	observação: não houve necess... pass a pass pa
10.00	AMS	acessar informações	observação: não houve necess... pass a pass pa
11.00	AMS	acessar lista de farm	observação: não houve necess... pass a pass pa
12.00	AMS	acessar listas de med	observação: as orient sobr as patolog assist de
13.00	AMS	acessar manual de oi	observação: não houve necess... pass a pass pa
14.00	AMS	acessar manual de oi	observação: proced utiliz soment na region es
15.00	AMS	acessar pae program	observação: não houve necess... pass a pass pa
16.00	AMS	acessar portal ams	observação: soment pel intranet... cas o usu se

Record 1 of 1457

Figura 28: Aplicação do RSLP *Stemmer* sobre a base de documentos da língua portuguesa

Além disso, o número de termos original foi reduzido de 19.590 para 12.466 termos, ou seja, 36% dos termos foram eliminados, demonstrando a eficiência do processo de stemming para remoção de sufixos, inspirado em PORTER (1980).

Com a aplicação do RSLP *Stemmer*, espera-se melhorar o processamento das tarefas de Mineração de Textos, tanto na clusterização quanto na classificação. Serão avaliados os resultados sobre a base de documentos com e sem *stemming*. Além disso, comprovou-se a facilidade de integração entre a nova implementação do RSLP *Stemmer* ao código criado para importação da coleção de documentos.

7.5.8. Modelo Espaço Vetorial (VSM):

A partir da redução do número de termos foi gerado o Modelo Espaço Vetorial, ou Vector Space Model (VSM), ou seja, essa é a etapa de transformação dos documentos.

A Figura 29 mostra uma parte da matriz VSM gerada, pois é gerada uma matriz com alta dimensionalidade, conforme apresentado na literatura, onde cada linha representa o documento e cada coluna um termo ou palavra. Foram avaliados dois modelos de espaço vetorial, um deles gerados a partir da relevância e outro a partir da frequência das palavras.

1	Id_Doc	Classe	Nome do P...	T	Procedimento	1	credenciado	1	petroleo	1	atendente	1	engenharia	1	tecnico	1	comperj	1	compartilhado	1	ramais	
1	AMS	acessar auto atendi	observacoes	nao houve r		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	AMS	acessar base de conl	observacoes	este proced		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	AMS	acessar busca ams	observacoes	o sistema bu		2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	AMS	acessar cheque de p	observacoes	nao houve r		3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	AMS	acessar comandos d	observacoes	estas inform		8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	AMS	acessar comprovant	observacoes	o credenciai		4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	AMS	acessar despesas an	observacoes	consulta mo		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	AMS	acessar escolha dir	observacoes	nao houve r		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	AMS	acessar informaçõe	observacoes	nao houve r		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	AMS	acessar informaçõe	observacoes	nao houve r		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	AMS	acessar lista de farm	observacoes	nao houve r		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	AMS	acessar listas de me	observacoes	as orientac		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	AMS	acessar manual de o	observacoes	nao houve r		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	AMS	acessar manual de o	observacoes	procedimen		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	AMS	acessar pae prograr	observacoes	nao houve r		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	AMS	acessar portal ams	observacoes	somente pe		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	AMS	acessar processo de	observacoes	nao houve r		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	AMS	acessar programaçã	observacoes	nao houve r		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	AMS	acessar programaçã	observacoes	nao houve r		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	AMS	acessar programaçã	observacoes	nao houve r		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figura 29: Amostra da Matriz Modelo Espaço Vetorial baseado na frequência

Segundo documentação da ferramenta, no Modelo Espaço Vetorial (VSM), os documentos são tratados como “bags of words”. Para cada palavra, identifica-se a quantidade de vezes que essa ocorre no documento da coleção.

A partir do Modelo Espaço Vetorial (VSM) foi possível obter uma nova forma de representação dos documentos conhecida como Alocação Latente de *Dirichlet* (LDA) apresentada a seguir.

7.5.9. Alocação Latente *Dirichlet* (LDA)

Utiliza-se o modelo LDA como uma das mais recentes formas de representação dos documentos, cujo objetivo é extrair da base de documentos os tópicos contendo os principais termos.

A partir da matriz e dos termos gerados pelo VSM foi criado um arquivo organizado em três colunas, onde cada linha contém o índice do documento, o índice da palavra e a contagem da palavra, ou seja, o arquivo apresenta a quantidade de vezes que a palavra aparece em cada documento. Foram ajustados os tópicos e iterações conforme Tabela 4.

Tabela 4: Paramêtros para executar o *LDA Gibbs Sampler*

Número de Tópicos	T= 10; T =15;
Número de termos por tópico	S= 10; S=15
Iterações	N=100; N=1000

Ressalta-se a importância de definir o número de iterações de acordo com a matriz gerada pelo modelo espaço vetorial para não prejudicar o desempenho do método.

7.5.10. Algoritmo Supressor de Textos

Diante da enorme quantidade de documentos armazenados, cada vez mais as técnicas de redução dos dados estão sendo exploradas. Inspirado no algoritmo supressor para dados desenvolvido por FIGUEREDO *et al.* (2012), o algoritmo Supressor de Textos foi adaptado para Mineração de Textos. Conforme observado na etapa de pré-processamento trabalha-se com matrizes com alta dimensionalidade e demandam um maior tempo de processamento. Um dos grandes desafios dessa nova abordagem para redução de dados é selecionar os documentos mais representativos para realizar pesquisas mais eficientes.

A base pré-processada sem o uso de *stemming* constituiu a coleção de documentos para os experimentos realizados. Assim como utilizado no modelo LDA, também foram avaliados os espaços vetoriais gerados a partir da relevância e frequência das palavras. As matrizes dos espaços vetoriais são as entradas para o algoritmo de supressão. A Figura 30 apresenta o fluxograma do Algoritmo Supressor de Textos aplicado em uma coleção de documentos. Cada uma das etapas discriminadas corresponde às funções necessárias para realizar a supressão dos documentos.

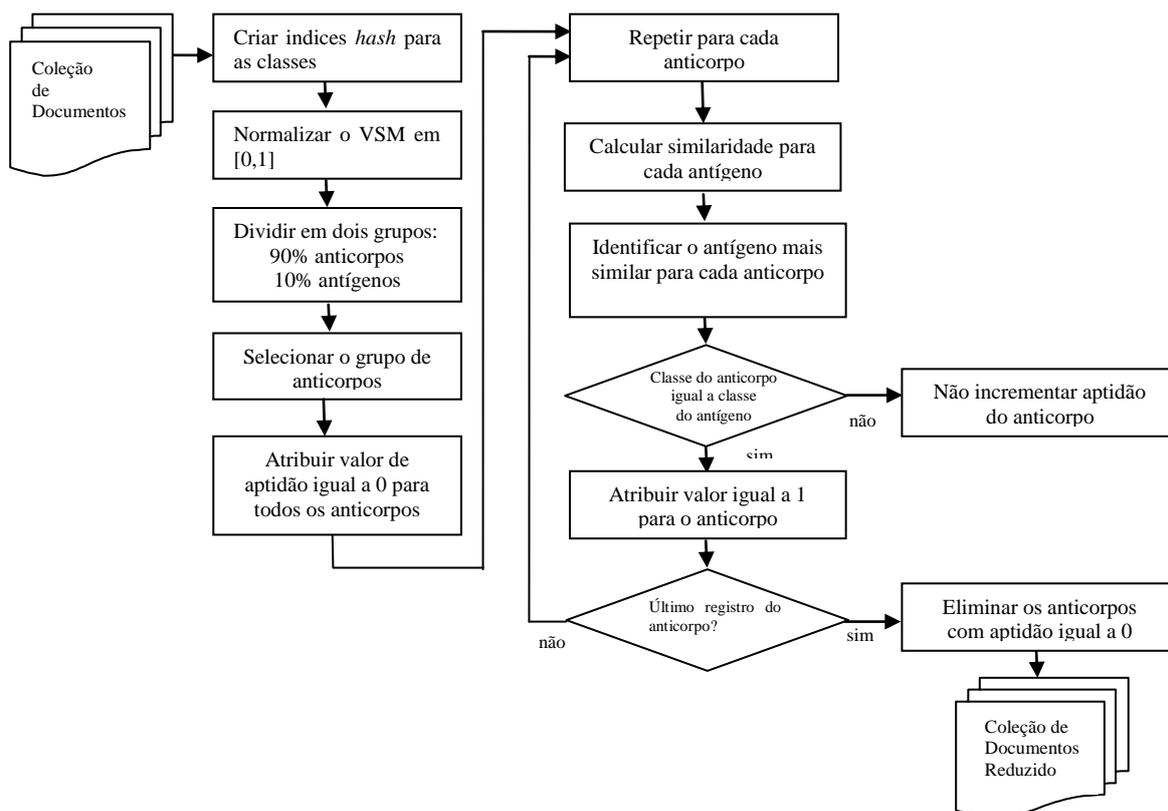


Figura 30: Fluxograma do Algoritmo Supressor de Textos em uma Coleção de Documentos

No presente trabalho, o Algoritmo Supressor de Textos será avaliado nas principais tarefas de Mineração de Textos. Entretanto, na Clusterização o algoritmo é aplicado diretamente sobre a coleção de documentos, e na Classificação, somente o conjunto de treinamento será submetido ao processo de supressão.

Acredita-se que a escolha de documentos mais representativos pode obter um melhor desempenho computacional sem perder a acurácia.

7.6. PROCESSAMENTO

7.6.1. Clusterização

A tarefa de Clusterização Hierárquica foi realizada sobre três tipos de bases conforme apresentado na Figura 31:

- 1- Base de documentos pré-processada;
- 2- Base de documentos com RSLP *Stemmer*;
- 3- Base Suprimida pelo SeleSupText.

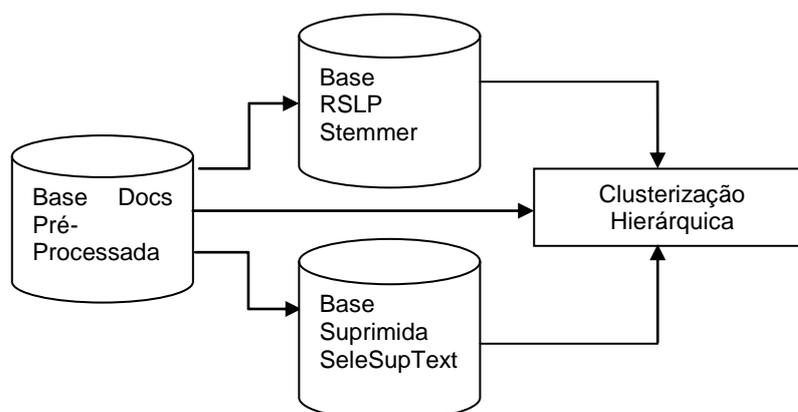


Figura 31: Bases utilizadas nos experimentos da tarefa de Clusterização Hierárquica

O agrupamento hierárquico utiliza o algoritmo *Suffix Tree Clustering* (ZAMIR, 1999), no qual os grupos que contêm registros compartilhados (em percentual) com outros *clusters* são parcialmente ou totalmente representados como filhos destes *clusters*, em uma hierarquia.

São disponibilizadas três tarefas de agrupamento *Base Clustering*, *Exclusive Clustering* e *Hierarchical Clustering*, sendo escolhida a última por ser mais utilizada na literatura.

Dado que a construção eficiente de um agrupamento é feita através de “tentativa e erro”, o *PolyAnalyst* permite variadas combinações de parâmetros descritas a seguir.

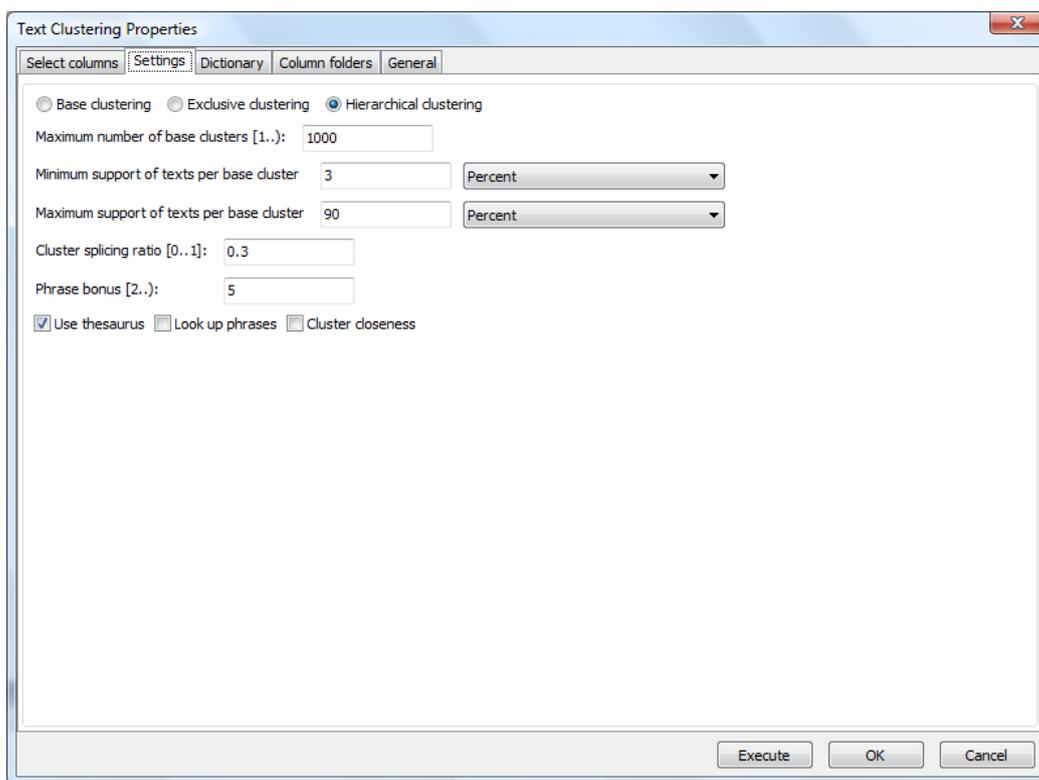


Figura 32: Configuração do Nó Clusterização de Textos

Cluster Method: o método *Base Clustering* desenvolve um conjunto unidimensional de grupos não excludentes através do algoritmo de agrupamento em árvore. O método *Exclusive Clustering* é uma extensão do método Base, pois é realizada uma segunda checagem que irá atribuir cada um dos documentos ao grupo de maior relevância, tornando os grupos excludentes. Por fim, o método *Hierarchical Clustering* explora mais o algoritmo de agrupamento em árvore com grupos não excludentes, através de uma verificação do percentual de documentos dentro de cada grupo que formariam outro grupo.;

Maximum number of base clusters: refere-se ao número máximo de frases ou palavras (base) que são encontrados na primeira etapa do algoritmo de agrupamento em árvore;

Minimal percent of texts per base cluster: refere-se à quantidade mínima de documentos que devem conter as frases ou palavras (base) encontradas na primeira etapa do algoritmo. Ao aumentar esse parâmetro, menos bases serão utilizadas por não atingirem o percentual mínimo de documentos que as contenham;

Maximal percent of texts per base cluster: refere-se à quantidade máxima de documentos que devem conter as frases ou palavras (base) encontradas na primeira etapa do algoritmo. Ao diminuir esse parâmetro, as bases que foram encontradas num percentual de documentos maior do que ele serão descartadas;

Cluster splicing ratio: refere-se a uma taxa utilizada para unir grupos com forte semelhança. Calcula a relação entre a interseção entre as duas frases (a contagem dos registros onde ocorram ambas as frases) e a união entre as duas frases (o conjunto de registros que contenham uma ou outra frase), em seguida, compara isso a um limite mínimo definido pelo usuário. Se a razão calculada é maior do que o limite, as duas frases são fundidas em um único cluster;

Phrase bonus: refere-se a um coeficiente de ponderação da importância de frases do conteúdo no conjunto de documentos, que impacta em quanto mais importantes são as frases, como identificadoras de um grupo, em relação às palavras. O aumento deste parâmetro influenciará o algoritmo para analisar frases como melhor identificadoras do que as palavras individuais.

Foram utilizados o nó Clusterização de Textos, para o método Agrupamento Hierárquico, e o nó Gráfico de Ligações para visualização das relações entre os grupos originais e os grupos criados. O projeto para a tarefa de Clusterização é apresentado na Figura 33.

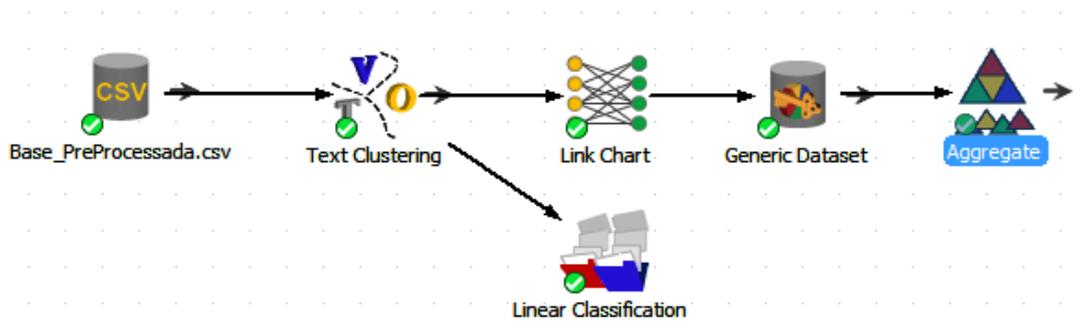


Figura 33: Projeto para Tarefa de Clusterização dos Documentos

Pela descrição dos nós verifica-se que foram avaliados os resultados obtidos com o dicionário TEP, o *thesaurus* e o uso de frases como identificadores de clusters.

7.6.2. Classificação

Como foi possível coletar os dados de acordo o grupo de serviço, ou seja, conhecendo-se a classe a qual um documento pertence, foi possível realizar também a tarefa de Classificação.

Nesse estudo, foi possível avaliar o desempenho dos classificadores *Naïve Bayes* e Máquinas de Vetor de Suporte (SVM), sobre os seguintes conjuntos de treinamento:

1. Conjunto de treinamento da base pré-processada;
2. Conjunto de treinamento da base com *RSLP Stemmer*;
3. Conjunto treinamento suprimido (x%) obtido pelo o mecanismo de supressão;

O conjunto de treinamento refere-se ao conjunto obtido dividindo-se a base de documentos pré-processada original em 70% para conjunto de treinamento e 30% para conjunto de teste.

A partir da base de treinamento correspondente a 70% da base original foram aplicadas as técnicas *RSLP Stemmer* e o Algoritmo de Supressão de Textos, conforme apresentado na Figura 34:

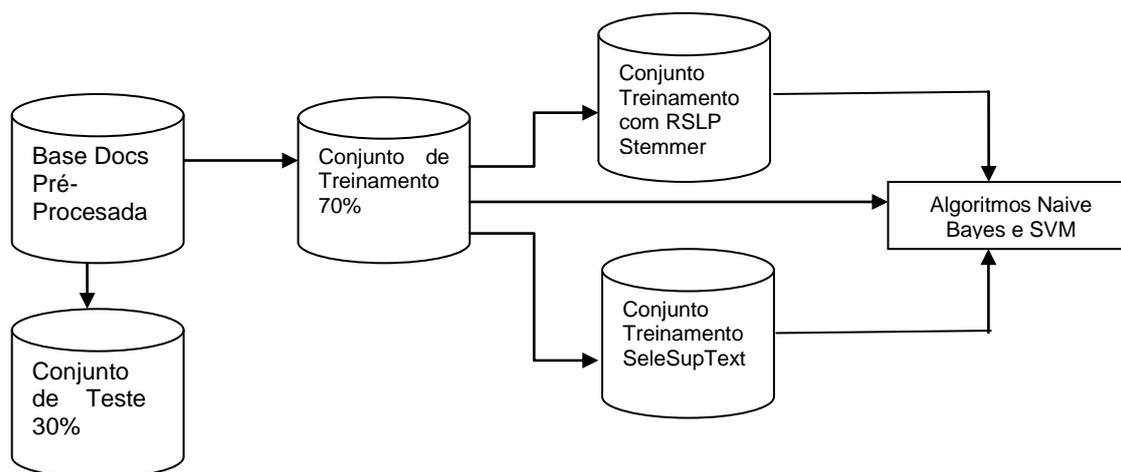


Figura 34: Bases utilizadas nos experimentos da Tarefa de Classificação dos Documentos

A Figura 35 apresenta o projeto criado para realizar a tarefa de classificação para cada um dos conjuntos de treinamento avaliados.

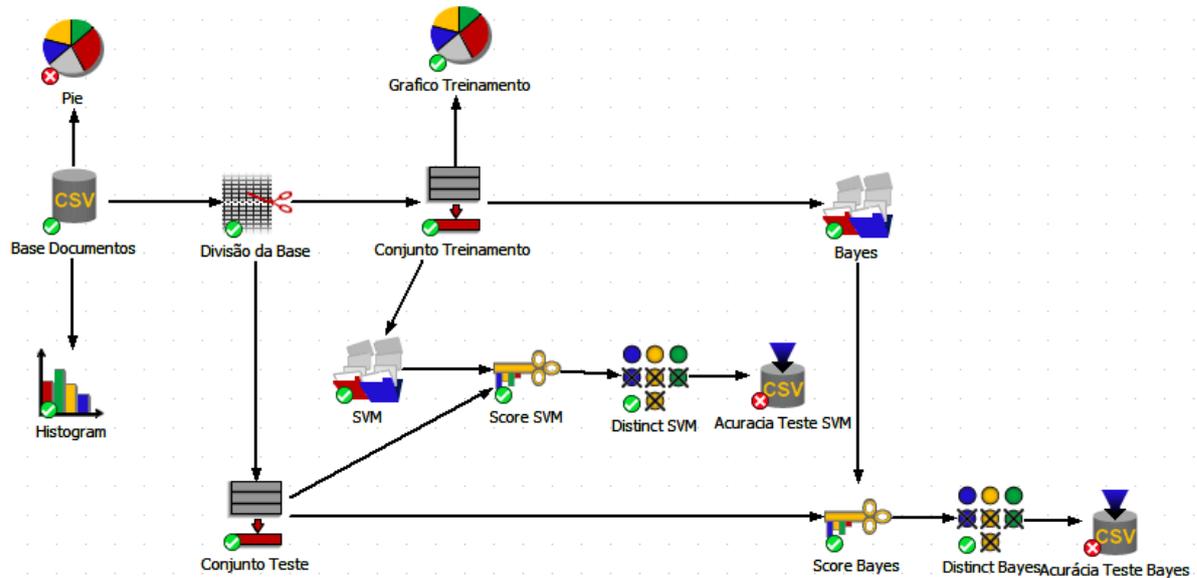


Figura 35: Projeto para Tarefa de Classificação dos Documentos

O classificador é avaliado por seu desempenho em classificar corretamente um conjunto de dados cuja classificação de cada registro não é conhecida. O conjunto de avaliação é chamado conjunto de teste.

O teste do modelo é realizado através da ferramenta *PolyAnalyst*, utilizando-se o nó *Score* que aplica o modelo sobre o conjunto de teste. Calcula-se através da matriz de confusão, o percentual de acertos das classes previstas pelo modelo em relação as classes esperadas (ou conhecidas).

7.7. PÓS – PROCESSAMENTO

7.7.1. Clusterização

As medidas de avaliação consideradas para avaliação dos agrupamentos foram Suporte e Significância a partir do nó *Link Chart*.

É possível visualizar e identificar a relação entre os clusters criados para cada grupo de serviço. As linhas mais grossas representam fortes correlações positivas entre os grupos de serviços e os grupos gerados. Um grupo de serviço pode ter relação com mais de um grupo.

Nesse caso, as correlações mais evidentes também podem ser identificadas por meio de dois indicadores: a significância, medida que representa a importância do cluster para determinado grupo de serviço e o suporte, que representa a quantidade de documentos por cluster.

Para avaliar o resultado da Clusterização será considerada a quantidade de registros do cluster que possuir maior significância para o grupo de serviço.

7.7.2. Classificação

Os critérios considerados para avaliação dos classificadores foram a acurácia das bases de treinamento e de teste utilizando-se a matriz de confusão. Além de ser avaliado o desempenho (tempo de processamento) dos algoritmos *Naïve Bayes* e SVM.

7.8.COMENTÁRIO

Este capítulo apresentou o detalhamento de cada uma das etapas da metodologia proposta para extrair conhecimento do Estudo de Caso. Os recursos utilizados, a partir de ferramentas distintas, foram previamente testados e analisados para garantir a execução dos experimentos .

Vale ressaltar que os documentos da Central de Atendimento são, devidamente, padronizados de acordo com o formato e regras definidas pelo cliente, e redigidos por especialistas de domínio. Dessa forma, foi possível aplicar facilmente as principais técnicas de Mineração de Textos.

8. RESULTADOS E ANÁLISES

Com o objetivo de avaliar as tarefas de Clusterização e classificação em um domínio específico, os experimentos foram realizados a partir da base textual composta por 1.343 documentos da língua portuguesa, provenientes da Base de Conhecimento, de uma Central de Atendimento que presta serviços para a empresa petróleo.

Os experimentos foram conduzidos a partir de três subconjuntos de documentos obtidos através da referida Base de Conhecimento:

Base Pré-Processada: base após etapa de pré-processamento

RSLP *Stemmer*: base após uso do Removedor de Sufixos da Língua Portuguesa;

SeleSupText: base suprimida pelo mecanismo de supressão SeleSupText (**Seleção por Supressão para Textos**).

Nas Seções 7.5.6 e 7.5.10 foram apresentados o critério de seleção de atributos aplicado em cada um desses subconjuntos e o Algoritmo Supressor de Textos para selecionar os dados a partir de grandes coleções de documentos. A Tabela 5 mostra a redução do número de termos e percentual de supressão após a utilização dessas duas técnicas.

Tabela 5: Características das Bases de Documentos para a tarefa de Clusterização

Base de Documentos	Número de Documentos	% Supressão	Quantidade de Termos Inicial	Quantidade de Termos Final	% Redução Termos
Base Pré-Processada	1343	0	15343	5453	64.46
Base com RSLP Stemmer	1343	0	12496	4206	66.34
Base SeleSupText ¹	102	92.40	2786	858	69.20
Base SeleSupText ²	88	93.44	2786	833	70.10

¹ SeleSupText gerado a partir da relevância dos termos

² SeleSupText gerado a partir da frequência dos termos

Observa-se que, a partir da frequência dos termos, a base SeleSup obteve maior percentual de redução, com uma pequena diferença em relação ao número de documentos suprimidos. É possível também notar que a nova abordagem para

redução de dimensionalidade obteve um percentual de redução da base original acima de 90%.

8.1. TAREFA DE AGRUPAMENTO

A Tabela 6 apresenta as características de cada um dos subconjuntos obtidos a partir da base Pré-Processada para tarefa de Clusterização.

Entre as execuções do método de Clusterização Hierárquica, onde foram testadas diversas variações de parâmetros, o melhor resultado obtido para cada uma das bases encontra-se descrito na Tabela 6.

Tabela 6: Consolidado Estatístico da Tarefa de Clusterização

Base de Documentos	Total Clusters	% Documentos Clusterizados	Mínimo de Documentos por Cluster	Máximo de Documentos por Cluster
Base Pré-Processada	14	92	72	857
Base com RSLP Stemmer	23	98	39	920
Base SeleSupText ¹	6	95	11	50
Base SeleSupText ²	9	75	5	25

¹ SeleSupText gerado a partir da relevância dos termos

² SeleSupText gerado a partir da frequência dos termos

Diante dos resultados, as bases suprimidas destacam-se em relação as demais bases, pelo total de grupos relevantes recuperados próximo ao número de grupos reais da coleção de documentos, ou seja, oito grupos de serviços.

Após várias tentativas, as bases pré-processada e com RSLP conseguiram obter um maior percentual de documentos agrupados, somente, a partir de um maior número de grupos formados, o que eleva o custo computacional em grandes bases de documentos.

De uma forma geral, verifica-se que a base SeleSup Text com base na frequência dos termos obteve melhores resultados.

A descrição de cada um dos grupos gerados para cada base é apresentada nos seguintes quadros: Quadro 2, Quadro 3, Quadro 4, Quadro 5 para melhor visualização do resultado encontrado.

Quadro 2: Descrição dos Grupos gerado pelo Algoritmo de Clusterização Hierárquica a partir da base Pré-Processada

+ cpf; faz; som; pabx; site; saida; video; virus; clique; ramais; \$idioma; \$perfil; captura; ddd ddi; entrada; \$arteria; \$lusiada; [857]
 cofip; prazo; uteis; abertura; demandas; atendente; empregado; utilizada; observacao; nota fiscal; \$abastecedor; conta pagar ext [661]
 dip; stic; custo; centro; assunto; estrutural; [229]
 tissweb; credenciado; [187]
 reparar; instalar; designado; facilidade; utilizacao; telefonia ddp; regionais classificacao; designacao organizacao infra; [184]
 apoio; carga; transporte; \$marinheiro; \$programacao; [176]
 titular; \$matricula; dependente; beneficiario; [155]
 sam; medica; vinculado; assistencia; [154]
 edisp; \$botao compartilhado infra; [142]
 gerais; manutencao; [108]
 publico; completo; [73]
 padrao; sinpep; [72]

Quadro 3: Descrição dos Grupos gerado pelo Algoritmo de Clusterização Hierárquica a partir da base com RSLP Stemmer

+ cab; cat; dur; mes; uso; feri; laud; micr; past; acion; descr; emerg; idiom; orad; pagin; plant; recur; corpor; detalh; execut; [920]
 + seg; val; adic; emit; nasc; pend; poss; abert; cofip; email; estej; relat; aguard; client; empreg; encerr; envolv; espec; finan [889]
 ddp; gel; pet; cust; padr; facil; repar; assunt; sinpep; aparelh; desativ; dip stic; disponibil; region classific; organiz supor [665]
 ddd; pit; cart; cobr; pabx; util; canal; vincul; restrit; tic ase; validade; sam digit; rot extern; mudanc faix; press digit; ana [506]
 ob; paci; medic; cirurg; [342]
 tic orient; [223]
 liber; marit; transport; [147]
 edis; edit; reduc; transpetr; [129]
 bot; compartilh; [120]
 dent; peric; odontolog; [95]
 risc; grand; pequen; particip; [88]
 tro; patric; [39]

Quadro 4: Descrição dos Grupos gerado pelo Algoritmo de Clusterização Hierárquica a partir da base SeleSupText utilizando a relevância dos termos

infra; regional; [50]
 cofip; centro; chamado; receber; demandas; atendente; empregado; financeira; dip bandeja; relacionado; situacao resposta script; [36]
 video; virus; \$placa; \$idioma; captura; licenca; versao build \$fabricante; \$licenciamento freeware desenvolvido trial rede \$loca [28]
 reparar; instalar; designado; \$localidade; \$transpetro; designacao organizacao; regionais classificacao; [20]
 fax; saude; assistencia; \$credenciado; beneficiarios; [17]
 carga; transporte; \$programacao; [11]

Quadro 5: Descrição dos Grupos gerado pelo Algoritmo de Clusterização Hierárquica a partir da base SeleSupText utilizando a frequência dos termos

cpf; cnpj; cofip; venda; demandas; proceder; atendente; empregado; recebivel; nota fiscal; prazo uteis; internacional; nacional [25]
 via; instalar; designado; regionais; \$transpetro ddp; disponibilidade; posto \$avancada; reparar classificacao; designacao organi [17]
 rede; banco; \$lacaio; hardware; impressao; [13]
 medica; titular; \$efeituar; vinculado; \$matricula; beneficiario; [12]
 dip; centro; demanda; [11]
 carga; reclamacao; transporte; [9]
 limite; \$odontologia; [7]
 apoio \$marinheiro; [5]
 \$botao compartilhado; [5]

Os documentos são visualizados em hierarquia de tópicos com seus respectivos descritores. Cada tópico contém os documentos relacionados a um mesmo tema.

Verifica-se que os documentos correspondentes aos grupos AMS, COFIP, TELECOM e TRANSPORTE são facilmente identificados pelos seus respectivos descritores em todas as bases. As palavras cofip, centro, chamado, demanda, atendente, entre outras, descrevem, por exemplo, o grupo de serviço COFIP. Esta forma de visualização permite avaliar o quão os termos gerados representam seus respectivos procedimentos.

Em contrapartida, observa-se que alguns serviços não foram corretamente agrupados, como por exemplo, os documentos dos Grupos DIFERENCIADOS, EMERGENCIA e GERAIS. Esse fato demonstra que alguns serviços com menor número de documentos, algumas vezes similares em conteúdo dificultam a tarefa de agrupamento. Vale ressaltar que a base de documentos do referido Estudo de Caso é desbalanceada em relação ao número de documentos por serviço.

A Figura 36 e a Figura 37 apresentam o melhor resultado obtido pela Base SeleSup Text, com base na frequência dos termos, onde é possível visualizar a proximidade entre os clusters, a partir das palavras-chave e os clusters mais representativos identificados pelo Algoritmo de Clusterização Hierárquica. O resultado demonstra que é possível, a partir da seleção de documentos mais representativos, garantir maior confiabilidade do uso do Mecanismo de Supressão SeleSupText.

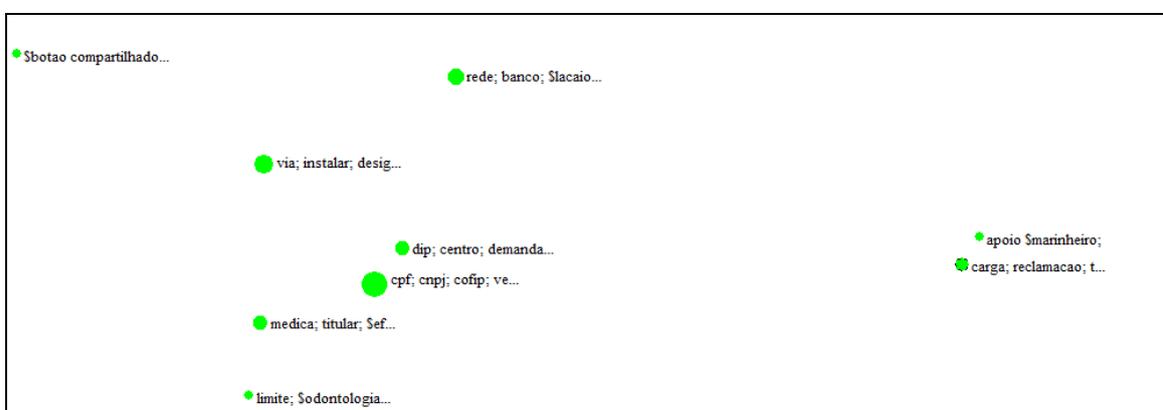


Figura 36: Gráfico de visualização da proximidade entre os nove grupos formados a partir do Algoritmo de Clusterização Hierárquica

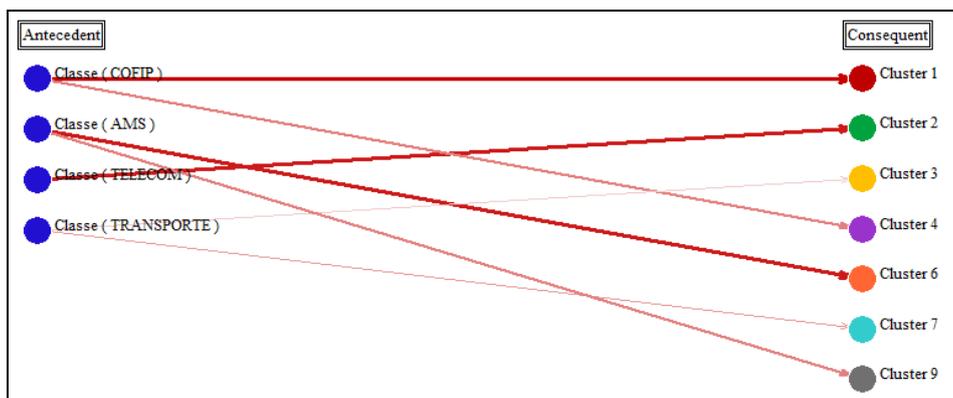


Figura 37: Gráfico de visualização da correlação entre cluster e grupos de serviços

A Figura 37 permite visualizar a relação entre os clusters criados para cada grupo de serviço identificado pelo Algoritmo de Clusterização Hierárquica, onde as linhas mais grossas representam fortes correlações positivas. Verifica-se que um grupo de serviço pode ter relação com mais de um cluster.

A Tabela 7 destaca o cluster que melhor representa o Grupo de Serviço baseado na importância do cluster para o grupo de serviço (significância) e na quantidade de documentos relevantes recuperados (suporte).

Tabela 7: Relação entre o grupo de serviço e o cluster mais representativo

Grupo de Serviço	Cluster	Significância	Suporte
AMS	Cluster 6	11.95	11
AMS	Cluster 9	8.02	7
COFIP	Cluster 1	17.49	12
COFIP	Cluster 4	10.34	7
TELECOM	Cluster 2	11.85	11
TRANSPORTE	Cluster 7	13.77	6
TRANSPORTE	Cluster 3	8.58	4

Assim, os melhores descritores para os grupos de serviços identificados corretamente pelo SeleSupText são apresentados no Quadro 6, o que demonstra a capacidade do SeleSupText em encontrar os melhores descritores para cada um dos grupos.

Os documentos organizados de forma automática oferece à Central de Atendimento uma nova forma de acompanhamento periódico dos procedimentos criados de acordo com a demanda do serviço.

Quadro 6: Palavras-Chaves dos Grupos Finais seleccionados pelo Algoritmo de Clusterização Hierárquica

Cluster	Grupo	Palavra-Chave
Cluster 6 e 9	AMS	medica, titular, beneficiario, vinculado, matricula, limite, odontologia
Cluster 1 e 4	COFIP	cpf, cnpj, cofip, venda, demanda, proceder, atendente, empregado, fiscal, internacional, nacional, dip, centro
Cluster 2	TELECOM	via, instalar, ddp, disponibilidade, organização, reparar, designação, rede, hardware
Cluster 3 e 7	TRANSPORTE	carga, reclamação, transporte, apoio, marinheiro

Adicionalmente, esses descritores ou palavras-chaves obtidos pelo SeleSupText, melhor resultado na tarefa de Clusterização, foram comparados ao modelo LDA (Alocação Latente *Dirichlet*), uma das mais recentes formas de representação dos documento.

O Quadro 7 apresenta dez tópicos extraídos da base SeleSupText, onde as palavras são representadas como pertencentes a um conjunto de tópicos probabilísticos.

Quadro 7: Tópicos extraídos da base SeleSupText utilizando o método LDA com número de iterações igual a 100 (N=100)

TÓPICO_1	TÓPICO_2	TÓPICO_3	TÓPICO_4	TÓPICO_5
exodontia	descricao	credenciado	prazo	server
digito	auditoria	paciente	apoio	chamadas
retalho	conta	guias	emergencia	origina
dente	valor	psicoterapia	transporte	subnet
fabricado	aut	beneficiario	acordo	dns
protese	idade	tiss	maritimo	gateway
pagina	demanda	prorrogacao	laboratorio	broadcast
implante	sala	titular	almojarifado	network
restauracao	criteriou	inclusao	uso	range
parcial	inicia final	sam	nota	rua
TÓPICO_6	TÓPICO_7	TÓPICO_8	TÓPICO_9	TÓPICO_10
medico	chamado	rt	Rnha	telecom
guia	pedido	botao	banco	ponto
pacote	dip	carga	ocorrencia	designacao
posto	estacao	mesa	tecnico	organizacao
avancado	atendente	sunidade	chamado	fazenda
laudo	nacional	status	web	classificacao
tabela	resposta	psv	compartilhado	fax
referente	empregado	agua	ronda	localidade
internacao	cofip	amosconnect	caixa	planejamento
tecle	demandas	servidor	cadastro	disponibilidade

O tempo de processamento do algoritmo criado para gerar a arquivo texto, contendo o número de vezes que uma palavra ocorre em um documento, foi de apenas nove minutos. A partir desse arquivo texto foi possível rodar o modelo LDA para extração dos tópicos.

Identificam-se as palavras nos tópicos 1, 3 e 6 como sendo do Grupo AMS, que também são apresentas como descritores dos clusters 6 e 9. Os tópicos 4 e 8 representam o grupo de serviço TRANSPORTE, os tópicos 5 e 10 representam o grupo TELECOM e por fim, os tópicos 2 e 7 descrevem o serviço COFIP. Assim, como pode ser observado, alguns descritores identificados pela tarefa de clusterização também foram identificados pelo modelo de tópicos.

Desta forma, identifica-se o modelo LDA como um importante método para extrair termos e palavras-chaves dos documentos de forma rápida e eficiente, a partir de uma coleção de documentos. Além disso, pode criar de forma automática uma lista de sinônimos específica para o atendimento de forma a agrupar os documentos a partir dessas palavras.

Conforme verificado, os especialistas da qualidade, responsáveis pelo controle dos procedimentos, podem utilizar a tarefa de Clusterização organizar e acompanhar o volume de documentos disponíveis por grupo de serviço.

Além disso, a nova abordagem para selecionar os documentos mais representativos, utilizando o Mecanismo de Supressão adaptado para Textos, é sem dúvida uma nova forma de estratégica para obter maior eficiência na organização dos documentos diante da enorme quantidade de informação disponível em uma Central de Atendimento.

Em seguida, realizou-se a tarefa de Classificação para avaliar a capacidade de generalização do modelo de treinamento.

8.2. TAREFA DE CLASSIFICAÇÃO

A característica de cada base é apresentada na Tabela 8 sendo que na tarefa de classificação, a nova abordagem proposta para reduzir a dimensionalidade da base de documentos foi aplicada na base de treinamento obtida aleatoriamente a partir de 70% dos documentos da base pré-processada.

Tabela 8: Características das Bases de Documentos para a tarefa de Classificação

Base de Documentos	Número de Documentos	Número de Docs. Treinamento	Número de Docs. Teste	Número de Termos
Base Pré-Processada	1343	940	403	4438
Base com RSLP Stemmer	1343	940	403	4150
Base SeleSupText ¹	474	71	403	3114
Base SeleSupText ²	470	67	403	2607

¹ SeleSupText gerado a partir da relevância dos termos

² SeleSupText gerado a partir da frequência dos termos

Oberseva-se que o número de termos (atributos) excede o número de documentos, principalmente em relação as bases suprimidas, onde o número de documentos da base de treinamento é significativamente inferior a quantidade de termos, o retrata uma matriz de esparsa e de alta dimensionalidade (FORMAN, 2003).

A Tabela 9 apresenta os resultados dos algoritmos *Naïve Bayes* e SVM para cada uma das bases de documentos. Observa-se que as bases suprimidas pelo método SeleSupText obtiveram uma taxa de redução maior que 90% , na qual foi possível reduzir a dimensionalidade da base original através de documentos mais representativos.

Tabela 9: Resultado dos Algoritmos *Naïve Bayes* e SVM

Base de Documentos	Redução (%)	Tempo de Execução (s)	Acurácia <i>Naïve Bayes</i> Treino (%)	Acurácia <i>Naïve Bayes</i> Teste (%)	Acurácia SVM Treino (%)	Acurácia SVM Teste (%)
Base Pré-Processada	0.00	-	91.58	87.84	99.06	89.08
Base com RSLP Stemmer	0.00	-	87.64	84.86	99.07	87.10
Base SeleSupText ¹	92.44	00:15	91.84	74.94	100	72.46
Base SeleSupText ²	92.87	00:12	92.86	66.75	100	63.52

¹ SeleSupText gerado a partir da relevância dos termos

² SeleSupText gerado a partir da frequência dos termos

Além disso, o método para supressão é processado em segundos, o que representa uma característica importante para reduzir o tempo de aprendizagem dependendo da quantidade de informações que deseja-se trabalhar.

Comparando-se os resultados entre os classificadores, verifica-se que as SVMs, devido sua robustez, alcançam resultados superiores ao *Naïve Bayes* no processo de aprendizagem em todas as bases. As bases suprimidas conseguiram atingir um percentual de acurácia similar ao percentual obtido pela base Pré-Processada.

Por outro lado, no teste da acurácia a base suprimida baseada na relevância dos termos, obteve um resultado superior, quando comparado à base suprimida baseada na frequência dos termos. Entretanto, devido alta taxa de supressão maior que 90%, a base suprimida baseada na relevância não atingiu a mesma precisão em relação base pré-processada.

Também foi possível observar que os resultados da base com *RSLP Stemmer* não apresentaram uma melhoria em relação a acurácia, a partir da radicalização dos termos. A base Pré-Processada sem o uso da técnica *stemming* apresentou os melhores resultados em ambos algoritmos. Esse fato demonstra que a técnica desenvolvida depende não somente da língua, mas também do domínio da aplicação.

Com o objetivo de identificar os grupos que apresentaram maior percentual de erro, a Figura 38 e a Figura 39 apresentam os valores obtidos na acurácia teste para cada grupo de Serviço. Tanto nos classificadores SVM quanto Bayes, o serviço TELECOM apresentou o maior percentual de erro em todas bases de documentos.

Os serviços EMERGENCIA e TELECOM apresentaram o maior percentual de erro na base SeleSupText baseado na frequência dos termos, provavelmente, pelo fato de alguns termos serem comuns aos dois serviços.

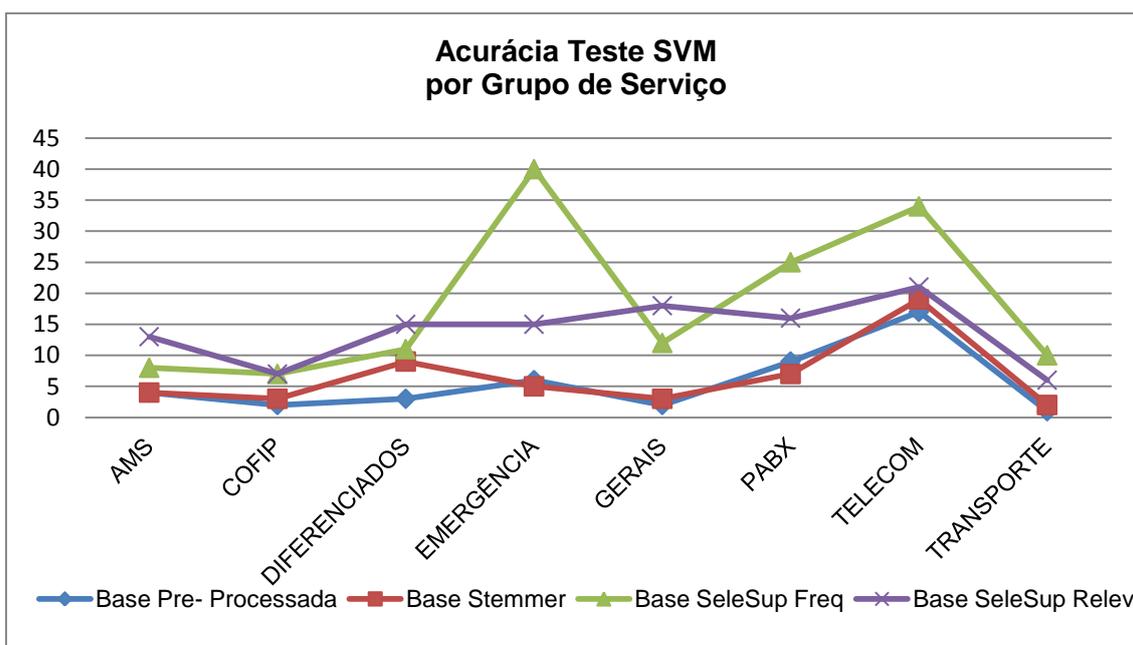


Figura 38: Percentual de Erro obtido pelo classificador SVM

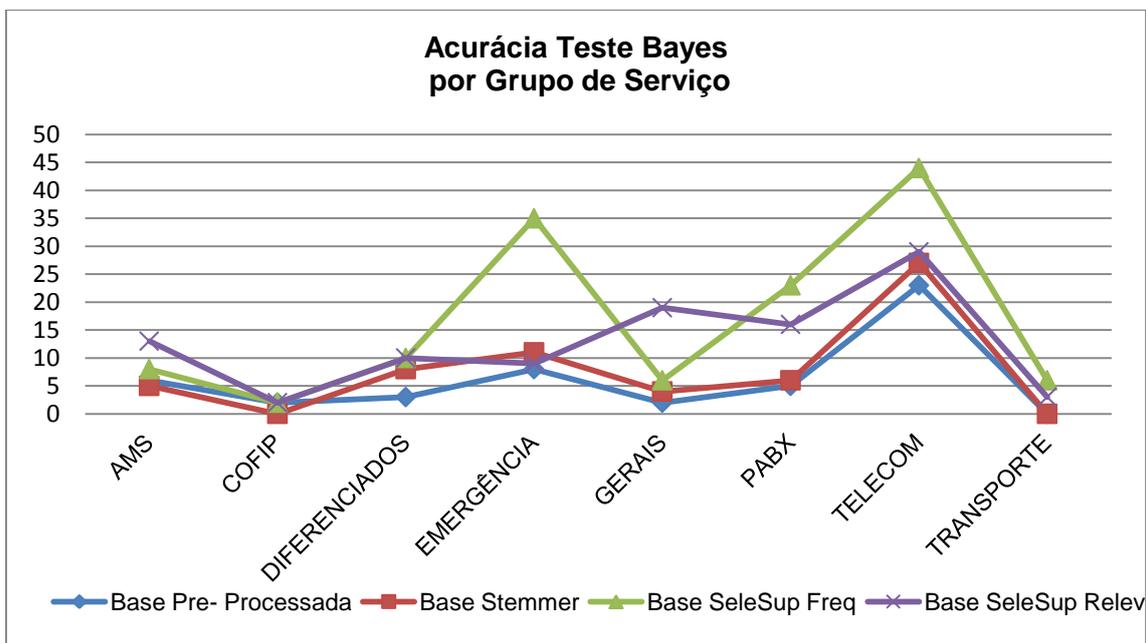


Figura 39: Percentual de Erro obtido pelo classificador Bayes

Em seguida, o desempenho dos classificadores *Naïve Bayes* e SVM foi avaliado com base no tempo de processamento conforme apresentado na Tabela 10.

Tabela 10: Tempo de Processamento obtido pelos Classificadores *Naïve Bayes* e SVM

Base de Documentos	<i>Naïve Bayes</i> Tempo de Processamento	SVM Tempo de Processamento
Base Pré-Processada	00:00:02	00:00:10
Base com RSLP Stemmer	00:00:02	00:00:11
Base SeleSupText ¹	00:00:00	00:00:00
Base SeleSupText ²	00:00:00	00:00:00

¹ SeleSupText gerado a partir da relevância dos termos

² SeleSupText gerado a partir da frequência dos termos

Observa-se que o tempo de processamento do SVM é superior ao *Naïve Bayes* e que as bases suprimidas foram processadas imediatamente. Em grandes coleções de documentos, o desempenho do classificador sobre uma base suprimida contribui de forma significativa para reduzir o custo computacional durante o processamento.

8.3.COMENTÁRIO

O Estudo de Caso evidenciou a utilização das principais técnicas e tarefas de Mineração de Textos, a partir da coleção de documentos reais disponível.

A metodologia foi desenvolvida para construir um modelo eficiente que atenda a escalabilidade de grandes quantidades de documentos

Verificou-se que a ponderação dos termos influencia no processamento da base, ou seja, para a Tarefa de Agrupamento o melhor resultado obtido foi utilizando a frequência dos termos. Por outro lado, na tarefa de Classificação o uso da relevância das palavras obteve o melhor resultado.

Foi avaliado o desempenho do uso da técnica de *stemming* sobre a coleção de documentos da língua portuguesa. Entretanto, os resultados mostraram que a técnica depende também do domínio da aplicação.

Por fim, a lista de sinônimos criada, manualmente, pelos especialistas não apresentou-se eficiente como *thesaurus* . Identificaram-se alguns termos comuns nas listas de sinônimos durante a importação do dicionário. Esse fato revela a necessidade de utilizar a taxonomia como nova forma de extração de termos comuns.

9. CONCLUSÃO

Aplicar as técnicas e as principais tarefas de Mineração de Textos, em uma base de documentos reais, é uma tarefa desafiadora, pois a etapa de pré-processamento consome maior esforço e tempo para garantir a qualidade dos dados e assegurar a fidelidade dos resultados na etapa de processamento.

Os experimentos foram realizados, a partir da coleção de documentos disponível, que trata da realidade atual de uma Central de Atendimento, que presta serviço de atendimento a uma empresa de petróleo. Tais documentos necessitam ser atualizados de acordo com a demanda do serviço e facilmente localizados para garantir a qualidade e produtividade das equipes.

O mais recente método para redução de instâncias, denominado mecanismo de seleção imune-supressor para textos foi comparado aos resultados obtidos por duas bases de documentos: a base pré-processada, ou seja, a base preparada para a extração de padrões, e a base de documentos utilizando o RSLP *Stemmer*, que utiliza a técnica de radicalização da língua portuguesa.

Os resultados mostraram que o algoritmo supressor *SeleSupText* reduz significativamente a quantidade de registros de uma coleção de documentos, chegando a uma taxa de supressão superior a 90% dos documentos. Além disso, deve-se destacar que a aplicação do algoritmo, não exige alto custo computacional, sendo facilmente executado.

Dessa forma, o algoritmo pode ser considerado como poderosa ferramenta para redução de instâncias, permitindo-se construir um modelo que atenda a escalabilidade não somente em grandes coleções de dados, como também, em grandes coleções de documentos.

Ainda no decorrer do trabalho, surgiu o interesse em avaliar o modelo de tópicos denominado Alocação Latente *Dirichlet*, que representa uma coleção de documentos por meio de tópicos, contendo as palavras mais representativas. O desempenho do modelo de tópicos foi comparado ao melhor resultado obtido pelo *SeleSupText*, a partir das palavras-chaves identificadas como rótulos de cada grupo.

Um das contribuições desse trabalho refere-se a organização automática dos documentos em hierarquia de tópicos e seleção dos descritores para os grupos formados. Nos primeiros níveis das hierarquias estão os documentos mais genéricos e, nos níveis mais profundos, estão os documentos específicos. Em cada nível é apresentado o número de documentos disponível.

Assim, a tarefa de agrupamento oferece uma nova forma de acompanhamento dos procedimentos por grupo de serviço, permitindo avaliar a frequência com que esses documentos são criados de acordo com a solicitação do cliente. A informação facilmente localizada por um atendente contribui para a produtividade da equipe, reduzindo o tempo de atendimento para solucionar um problema.

Cada tópico gerado, tanto pela tarefa de agrupamento quanto pelo modelo de alocação *Dirichlet*, contém as palavras mais representativas de forma a facilitar a interpretação dos documentos em cada grupo ou tópico, respectivamente. As novas palavras-chaves identificadas também podem ser incorporadas à lista de sinônimos utilizadas atualmente pelo Portal de Atendimento.

Além disso, a organização da coleção em grupos de documentos similares, permitiu ao especialista validar as palavras-chaves identificadas como descritores de cada grupo encontrado. Vale ressaltar que a avaliação subjetiva de uma especialista do domínio contribuiu para seleção dos termos (atributos), excluindo palavras muito frequentes e sem representatividade.

Outra contribuição importante desse trabalho é utilizar a tarefa de classificação para classificar automaticamente um novo documento, de forma rápida e eficiente. Esse processo é útil, principalmente, nos casos em que a informação necessita ser repassada o mais rápido possível para as equipes, que já se encontram em atendimento.

A partir da aplicação da metodologia desenvolvida, foi possível extrair as entidades responsáveis por cada grupo de serviço. Segundo as normas de qualidade estabelecidas pela empresa de petróleo, todos os documentos devem ser identificados através dos responsáveis. Esse conhecimento inovador permite manter a informação sempre atualizada, quando ocorrer mudança na estrutura organizacional.

Como melhoria no processo atual, identifica-se a necessidade de automatizar a lista de sinônimos do Portal de Atendimento, visto que é realizado, manualmente, por mais de uma analista de qualidade. Como solução recomenda-se desenvolver de forma automática um dicionário de domínio específico para a Central de Atendimento, de forma a evitar que termos sejam comuns a mais de um conceito.

Referente aos documentos coletados, sugere-se que os manuais referentes ao uso das ferramentas de acordo com o serviço, sejam segmentados em capítulos para oferecer maior agilidade durante uma ligação onde o cliente aguarda a resposta.

Embora o desenvolvimento da metodologia tenha cumprido os objetivos propostos, algumas limitações devem ser consideradas. Além de obter o maior número de documentos válidos para a realização dos experimentos, a base é considerada desbalanceada, dificultando a identificação dos documentos similares nos grupos de serviço com menor quantidade de documento. Outra limitação refere-se a aplicação das técnicas e tarefas de Mineração de Textos em documentos da língua portuguesa, cuja a língua é mais complexa gramaticalmente e necessita utilizar dicionários e métodos específicos para a realização dos experimentos.

Em trabalhos futuros, pretende-se comparar o resultado do mecanismo supressor com o o resultado obtido pelo método evolutivo CHC. O método CHC também caracteriza-se pela capacidade em reduzir o número de instâncias e manter o desempenho, apesar do elevado tempo computacional.

A partir da coleção atual de documentos, será realizado um estudo comparativo entre a nova métrica de ponderação dos termos *Flexi-PR* (MODANI *et al.*, 2009) e a tradicional medida *TFxIDF*. Conforme demonstrado as métricas frequência e relevância dos termos influenciaram na qualidade dos resultados das principais tarefas de Mineração de Textos.

O caso estudado será avaliado por períodos mais longos, de forma que, a coleção de documentos possa atingir uma maior quantidade de documentos. Além disso, espera-se aplicar a mesma metodologia em uma Central de Atendimento que presta serviço para o mercado de massa.

Conclui-se que o modelo de tópicos LDA e o mecanismo de supressão SeleSupText, são consideradas importantes contribuições na área de pesquisa envolvendo técnicas de Mineração de Textos.

10. REFERÊNCIA

- ALVARES, R. V., GARCIA, A. C.B., FERRAZ, I., 2005, "STEMBR: A Stemming Algorithm for the Brazilian Portuguese Language". In: *Portuguese Conference on Artificial Intelligence, 12. Proceedings ... Covilhã, Portugal*.
- ASSOCIAÇÃO BRASILEIRA DE TELEMARKEETING (ABT), 2005, Brasil: Pólo de Qualidade em Call Center – casos de excelência no relacionamento com o cliente. São Paulo: ABT.
- BÄCK, T.; FOGEL, D.B. & MICHALEWICZ, Z., 2000a, "Evolutionary Computation 1 Basic Algorithms and Operators". *Institute of Physics Publishing, Bristol, UK*.
- BÄCK, T.; FOGEL, D.B. & MICHALEWICZ, Z., 2000b, "Evolutionary Computation 2 Advanced Algorithms and Operators". *Institute of Physics Publishing, Bristol, UK*.
- BASTOS, V. M., 2006, *Ambiente de Descoberta de Conhecimento na WEB Para a Língua Portuguesa*. Tese de D.Sc., COPPE/UFRJ, Rio de Janeiro, RJ, Brasil.
- BLEI, D. M., NG, A. Y., JORDAN, M. I., 2003, "Latent Dirichlet Allocation". *Journal of Machine Learning Research*, 3, pp.993-1022.
- BERNARD, A., TICHKIEWITCH, S., 2008, *Methods and Tools For Effective Knowledge Life-cycle-management*, Springer, ISBN 978-3-540-78430-2.
- CAPUTO, G.M., BASTOS, V.M., EBECKEN, N.F.F., 2006, "Using Text Mining to Understand the call center customers claims". In: *Data Mining & Information Engineering*, Prague.
- CALL CENTER GUIDE, 2005. Disponível em: <http://www.callcenterguide.com/Definitions/>. Acesso em: 23/03/2011.
- CANO, J. R., HERRERA, F., LOZANO, M., 2003, "Using evolutionary algorithms as instance selection for data reduction". In *KDD: An experimental study. IEEE Transaction on Evolutionary Computation*, 7, 561-575.

- CORREA, U., 2007, *Mineração de Dados de Help Desk usando RATTLE – O Caso Petrobras*. Tese de M.Sc., Faculdade de Economia e Finanças, IBMEC, Rio de Janeiro, RJ, Brasil.
- COELHO, A. R., 2007, *Stemming para Língua Portuguesa: estudo, análise e melhoria do Algoritmo RSLP*. Tese de M.Sc., Universidade Federal do Rio Grande do Sul, UFRGS, Porto Alegre, RS, Brasil.
- CRISTIANINI, N., SHAW-TAYLOR, J., 2000, *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press.
- CUTTING, D.R., KARGER, D.R., PEDERSEN, J.O., TUKEY, J. W., 1992, “Scatter/Gather: A cluster-based approach to browsing large document collections”. *SIGIR '92*, pp. 318-329.
- DASGUPTA, D., 1998a, *Artificial Immune Systems and Their Applications*, Springer-Verlag.
- DASGUPTA, D., 1998b, “An Overview of Artificial Immune Systems and Their Applications”, *Artificial Immune Systems and Their Applications*, Springer-Verlag, pp. 3-21.
- DE CASTRO, L.N., 2001, *Engenharia Imunológica: Desenvolvimento e Aplicação de Ferramentas Computacionais Inspiradas em Sistemas Imunológicos Artificiais*. Tese de D. Sc., Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e Computação, Campinas, SP, Brasil.
- EBECKEN, N. F. F., LOPES, M. C. S., de ARAGÃO COSTA, M. C., 2003, “Mineração de Textos”, In: *Sistemas Inteligentes: Fundamentos e Aplicações*, Ed. Manole Ltda, Cap. 13, pp. 337-370.
- FELDMAN, R., DAGAN, I., 1995, “Knowledge discovery in textual databases (KDT)”. In: *KDD, Montréal, Québec. Menlo Park, CA: AAAI*, pp. 112-117.

- FELDMAN, R., SANGER, J., 2006, *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press.
- FIGUEREDO, P. G., 2008, *Mecanismo Imuno-Supressor para Seleção de Dados de Treinamento em Problemas de Classificação*, Tese de M.Sc., COPPE/UFRJ, Rio de Janeiro, RJ, Brasil.
- FIGUEREDO, G.P., EBECKEN, N.F.F., BARBOSA, H.J.C., AUGUSTO, D.A., 2012, "An Immune-inspired Data Selection Mechanism for Supervised Classification", *Memetic Computing*. Volume 4, Number 2, Pages 135-147.
- FORMAN, G., 2003, "An extensive empirical study of feature selection metrics for text classification", *The Journal of Machine Learning Research*, 3, pp. 1289–1305.
- FROELICH, J., ANANYAN, S., OLSON, D.L., 2004, *The Use of Text Mining to Analyze Public Input*, Megaputer Intelligence and University of Nebraska..
- GILBERT, M.;WILPON, J.G.;STERN, B.; et al., 2005, "Intelligent Virtual Agents for Contact Center Automation", *IEEE Signal Processing Magazine*, vol 22. nº 5, pp. 32-41.
- GOLDBERG, D., 1989, *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, Reading, Mass.
- GRAELM, A. R., 2004. "A Internet e os seus Impactos nas Atividades de *Back-Office*: A Utilização da Internet por Empresas Industriais no Brasil". In: XXVIII Encontro Anual da Associação Nacional de Pós-Graduação em Administração, Curitiba.
- GRIFFITHS, T., STEYVERS, M., TENENBAUM, J. B., 2007, "Topics in semantic representation". *American Psychological Association*. Vol. 114, No. 2, 211–244.
- GRIFFITHS, T. and STEYVERS, M., 2004, "Finding scientific topics". In *Proceedings of the National Academy of Sciences*, 101, 5228-5235.

- HAN, J., KAMBER, M., 2006, *Data Mining: Concepts and Techniques*, 2ª edição, Morgan Kaufmann Publishers.
- HAWKINS, L.; MEIER, T.; NAINIS, S. et al., 2001, The Evolution of the Call Center to Customer Contact Center. Information Technology Support Center, White Paper.
- HAYKIN S., 2001, *Redes Neurais: Princípios e Prática*, 2ª ed., Porto Alegre, RGS, Brasil, Bookman Companhia Editora.
- HALL, B.; ANTON, J. 1998, Optimizing your Call Center through simulation. White Paper. Disponível em: <http://www.erlang.com.br/artigos/Optimizing%20Your%20Call%20Center%20Through%20Simulation.pdf>. Acesso em: 23/03/2011
- HEARST, M. A., SCHOLKOPF, B., DUMAIS, S., et al., 1998. Trends and controversies - support vector machines. *IEEE Intelligent Systems*, 13(4), pp. 18–28.
- HEARST, M., 1999, “Untangling text data mining”. In: *Proceedings of the 37th Annual Meeting of the ACL*, pp. 3–10, College Park, Maryland
- HOFMANN, T., 1999, “Probabilistic latent semantic indexing”. In: *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: ACM Press, pp. 50-57.
- HOLLAND, J.H., 1975, *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor, MI.
- HOLTGREWE, U., 2005, Call Centres in Germany – Preliminary Findings from the Global Call Centers Project. Duisburg/Essen University.
- HVALSHAGEN, M., 2002, *Call Center Data Analysis, Case Study in Text Mining*, Megaputer Intelligence.

- JOACHIMS, T., 1997, Text categorization with support vector machines. Technical report, LS VIII Number 23, University of Dortmund.
- KONCHADY M., 2006, *Mineração de Textos Application Programming*.
- LEWIS, D. D., RINGUETTE, M., 1994, "A Comparison of Two Learning Algorithms for Text Categorization". In: *Symposium on Document Analysis and RI, ISRI, Las Vegas*.
- LOPES, M. C. S., 2004, *Mineração de dados textuais utilizando técnicas de clustering, para o idioma português*. Tese de D.Sc., COPPE/UFRJ, Rio de Janeiro, RJ, Brasil.
- LOPES, R. B., 2009, *Mineração de Textos com Georeferenciamento*. Tese de D.Sc., COPPE/UFRJ, Rio de Janeiro, RJ, Brasil.
- LORENA, A. C.; CARVALHO, A. C. P. L., 2003, *Introdução às Máquinas de Vetores Suporte*. São Carlos, SP.
- LOVINS, J. B., 1968, *Development of a stemming algorithm*. *Mechanical Translation and Computacional Linguistics*, Volume 11, Number 1-2, pp. 22-31.
- MACQUEEN, J. B., "Some methods for classification and analysis of multivariate observations," *In Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, L. M. L. Cam and J. Neyman, Eds., vol. 1. University of California Press, 1967, pp. 281–297.
- MALHOTRA, N. K., 2006, *Pesquisa de Marketing: Uma Orientação Aplicada*. 4. ed. Porto Alegre: Bookman, pp. 720.
- MANNING, C. D., SCHÜTZE, H., 1999, *Foundations of Statistical Natural Language Processing*. MIT Press.
- MANNING, C. D., RAGHAVAN, P., SCHUTZE, H., 2008, *An Introduction to Information Retrieval*. Cambridge University Press.

- MARCACINI, R. M., 2011, *Aprendizado não Supervisionado de Hierarquia de Tópicos a partir de coleções textuais dinâmicas*, Tese de M.Sc., USP, São Carlos, SP, Brasil.
- MARCONI, M. de A.; LAKATOS, E. M., 2000, *Metodologia científica: ciência e conhecimento científico, métodos científicos, teoria, hipóteses e variáveis, metodologia jurídica*. 3. ed. rev. e ampl. São Paulo: Atlas.
- MAZIERO, E.G., PARDO, T.A.S., DI FELIPPO, A., et al., 2008. A Base de Dados Lexical e a Interface Web do TeP 2,0 - Thesaurus Eletrônico para o Português do Brasil. *VI Workshop em Tecnologia da Informação e da Linguagem Humana (TIL)*, pp, 390-392.
- McCALLUM, A. K, NIGAM, K., 1998, "A Comparison of event models for *Naïve Bayes Text Classification*". In: *Proceedings of th 1st AAAI Workshop on Learning for Text Categorization*, pp 41-48, Madison, USA.
- MITCHELL, T. M., 1997. "Bayesian Learning". In C. L. Liu & A. B. Tucker (Eds.), *Machine Learning*, McGraw-Hill, pp. 154-200.
- MODANI, N., RAMAKRISHNAN, G., Godbole, S., 2009. Tunable Feature Weights for Flexible Text Retrieval. In *Proceedings of SIG-KDD 2009*
- MOTTA, C. G. Lopes da, 2004, *Sistema inteligente para avaliação de riscos em vias de transporte terrestre*, Tese de D.Sc., COPPE/UFRJ, Rio de Janeiro, RJ, Brasil.
- NASUKAWA T., NAGANO, T.T., 2001, "Knowledge discovery using robust natural language processing", *Pacific Association for Computational Linguistics (PACLING)*, pp.189-198.
- NOGUEIRA B. M., MOURA M. F., CONRADO M. S., et al., 2008, "Avaliação de método não-supervisionados de seleção de atributos para Mineração de Textos", In: *I Workshop on Web and Text Intelligence*.

- NUNES, M., CABRAL, L., LIMA, R., et al., 2008, "Docs-Clustering: A System for Hierarchical Clustering and Document Labeling", In: I Workshop on Web and Text Intelligence.
- OLIVEIRA, I. M., 2009, *Estudo de uma Metodologia de Mineração de textos Científicos em Língua Portuguesa*. Tese de M.Sc., COPPE/UFRJ, Rio de Janeiro, RJ, Brasil.
- ORENGO, V. M.; HUYCK C., 2001, "A Stemming Algorithm for Portuguese Language". In: *Symposium on String Processing and Information Retrieval*, 8., Proceedings.. Chile.
- POPESCU, A., UNGAR, L., 2000. Automatic labeling of document clusters. Disponível em <http://citeseer.nj.nec.com/popescul100automatic.html>,
- PORTER, M., 1980, An algorithm for suffixing stripping. Program, Volume 14, Number 3, pages 130-137.
- PORTER, M. F., The stemming algorithm. 2005. Disponível em <http://snowball.tartarus.org/algorithms/portuguese/stemmer.html>. Acesso em 11 de maio de 2011
- REZENDE, S. O., PUGLIESI, J. B., MELANDA et al., 2003, "Mineração de dados," In: *Sistemas Inteligentes: Fundamentos e Aplicações*, Editora Manole Ltda., ch. 12, pp. 307–335.
- REZENDE, S. O., MARCACINI, R. M., MOURA, M. F., 2011, "O uso da Mineração de Textos para Extração e Organização Não Supervisionada de Conhecimento". *Revista de Sistemas de Informação da FSMA*, n. 7, pp. 7-21
- RIZZI, C., B., 2000, *Categorização de Textos por Rede Neural Estudo de Caso*. Tese de M.Sc., Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brasil.

- SALTON, G., WANG, A., YANG, C. S., 1975, "A Vector Space Model for Information Rretrieval", *Journal of the American Society for Information Science*, 18(11), pp.613–620.
- SCHIESSL, J., M., 2007, *Descoberta de Conhecimento em Texto aplicada a um Sistema de Atendimento ao Consumidor*. Tese M. Sc. UNB, Brasília, DF, Brasil.
- SMOLA, A. J., BARLETT, P., SCHÖLKOPF, B., et al., 1999. *Advances in Large Margin Classifiers*. MIT Press.
- SMOLA, A. J. and SCHÖLKOPF, B., 2002. *Learning with Kernels*. The MIT Press, Cambridge, MA.
- STEINBACH, M., KARYPIS, G., AND KUMAR, V., 2000, "A comparison of document clustering techniques". In *KDD Workshop on Text mining*. <http://citerserr.nj.nec.com/steinback00comparison.html>
- STEYVERS, M., GRIFFITHS, T., 2006, "Probabilistic topic models". In T. Landauer D. McNamara S. Dennis W. Kintsch (Editors), *Latent semantic analysis: A road to meaning*. Mahwah, NJ: Erlbaum.
- STROUSE, K. G., 1999, *Marketing Telecommunications Services: New Approaches for a Changing Environment*. United States: Artech House.
- TAN, A.-H., 1999, "Text mining: the state of the art and the challenges". In: *KDAD, Beijing, China. PAKDD*, pp. 71-76.
- TIMMIS, J., 2000, *Artificial Immune Systems: A Novel Data Analysis Technique Inspired by the Immune Network Theory*. Tese de D. Sc., University of Whales, Department of Computer Science, Aberystwyth, Ceredigion, Wales.
- van RIJSBERGEN, C. J., 1979, *Information Retrieval*. 2.ed., Butterworths, London.
- WEISS, M. S. Indurkhya, N. Zhang, T. Damerou, F. J., 2005, *Text Mining - Predictive Methodsfor Analyzing Unstructured Information*, 1st ed. Springer Science+Business Media, Inc.

- WHITLEY, D., 1989, "The genitor algorithm and selective pressure: Why rank based allocation of reproductive trials is best". In *Proceedings of 3rd Int. Conf. GAs*, pp 116-121.
- WITTEN , 2004, "Adaptative text mining: inferring structure from sequences", *Journal of Discrete Algorithms*, 2, pp. 137-159.
- WIVES, L. K., 2009, *Um estudo sobre Agrupamento de Documentos Textuais em Processamento de Informações não Estruturadas Usando Técnicas de Clustering* Tese de M.Sc., UFRGS, Porto Alegre, RS, Brasil.
- VAPNIK, V. N., 1995, *The Nature of Statistical Learning Theory*. Springer-Verlag.
- WATKINS, A., TIMMIS, J., 2002, "Artificial immune recognition system AIRS: Revisions and refinements", In: *1st Intl. Conference on Artificial Immune Systems (ICARIS2002)*, (TIMMIS, J., BENTLEY, P., eds.), University of Kent, pp. 173–181.
- WATKINS, A., TIMMIS, J., 2004, "Exploiting parallelism inherent in AIRS, an artificial immune classifier", In: *In Proc. of the 3rd Intl. Conference on Artificial Immune Systems (ICARIS 2004) (NICOSIA, G., CUTELLO, V., BENTLEY, P., et al., eds.)*, (Catania, Italy), pp. 427–438.
- WATKINS, A., TIMMIS, J., BOGGESS, L., 2004, *Artificial Immune Recognition System (AIRS): An Immune-Inspired Supervised Learning Algorithm*, v. 5, pp. 291 – 317. Springer Netherlands.
- WATKINS, A. B., BOGGESS, L. C., 2002, "A new classifier based on resource limited artificial immune systems", In: *Congress on Evolutionary Computation, IEEE World Congress on Computational Intelligence, Honolulu, HI, USA*, pp. 1546–1551.
- YANG, Y.; WILBUR, J., 1996, "Using corpus statistics to remove redundant words in text categorization". *Journal of the American Society for Information Science*, v. 47, n. 5, pp. 357–369.

- YANG, J.M, WU, W.C., LIAO, W.C, et al., 2009, "Trend Analysis of Machine Learning – A Text Mining And Document Clustering Methodology". *In: International Conference on New Trends in Information and Service Science*.
- ZAMIR O., 1999, Clustering Web Documents: A Phrase-Based Method for Grouping Search Engine Results. Ph.D. Dissertation, University of Washington, Seattle, WA, EUA.
- ZENONE, L. C. Customer Relationship Management (CRM) Conceitos e Estratégias: Mudando a estratégia sem comprometer o negócio. São Paulo: Atlas, 2001.
- ZIEN, A., RÄTSCH,G., MIKA,S., et al., 2000, "Engineering support vector machine kernels that recognize translation initiation sites". *Bioinformatics*, 16, pp. 799–807.
- ZISSERMAN, A., SIVIC, J., RUSSEL, B. C., et al., 2005, "Discovering object categories in image collections". *MIT-CSAIL*, 12.

11. GLOSSÁRIO DE TERMOS

CASE FOLDING: todas as palavras da sentença são transformadas em minúsculo

OVERFITTING: ajuste demasiado do classificador aos dados de treinamento. Uma vez que, quanto maior o espaço de exemplos, menos representativo torna-se uma amostra de tamanho fixo, pior tende a ser o desempenho do classificador para os dados de teste.

STEMMING: as palavras do texto são substituídas pelas suas respectivas raízes (*stems*).

STOP WORDS: são palavras muito comuns e, portanto, irrelevantes para o processamento em questão.

STOP LIST: é a lista de palavras muito comuns, ou seja, sem representatividade

SYNSETS: é o conjunto de sinônimos

THESAURUS: é o termo empregado por diferentes especialistas para designar diferentes objetos.