UNIVERSIDADE FEDERAL DOS VALES DO JEQUITINHONHA E MUCURI
CURSO DE GRADUAÇÃO EM SISTEMAS DE INFORMAÇÃO
Paulo Henrique Figueiredo Prates

UTILIZAÇÃO DO MODELO CASSIOPEIA PARA MINERAÇÃO DE TEXTOS EM BULAS DE MEDICAMENTOS

Diamantina
2017

Paulo Henrique Figueiredo Prates

UTILIZAÇÃO DO MODELO CASSIOPEIA PARA MINERAÇÃO DE TEXTOS EM BULAS DE MEDICAMENTOS

Trabalho de Conclusão de Curso submetido à Universidade Federal dos Vales do Jequitinhonha e Mucuri para a obtenção do Grau de Bacharel em Sistemas de Informação.

Orientador: Prof. Dr. Marcus Vinícius Carvalho Guelpeli

Diamantina

2017

Paulo Henrique Figueiredo Prates

UTILIZAÇÃO DO MODELO CASSIOPEIA PARA MINERAÇÃO DE TEXTOS EM BULAS DE MEDICAMENTOS

Trabalho de Conclusão de Curso submetido à Universidade Federal dos Vales do Jequitinhonha e Mucuri para a obtenção do Grau de Bacharel em Sistemas de Informação.

Orientador: Prof. Dr. Marcus Vinícius Carvalho Guelpeli

Data de aprovação ____/____/____.

Banca Examinadora:

_____
Prof. Dr. Marcus Vinícius Carvalho Guelpeli

_____
Profa. Dra. Geruza de Fátima Tomé Sabino

_____
Profa. Dra. Maria Lúcia Bento Vilela

Diamatina
2017

À minha família pelo apoio, em especial minha mãe
Sebastiana e meu pai Domingos.
Aos meus amigos pela força que sempre me deram.

AGRADECIMENTOS

Aos meus pais, por acreditarem em mim, sempre me apoiarem e por nunca medirem esforços para que eu chegasse até esta etapa de minha vida.

À minha irmã Edna, pela força que sempre me deu e por estar sempre disposta a me ajudar.

Aos meus amigos, pelo apoio e incentivo que sempre me deram.

À Tatiane, que esteve ao meu lado me ajudando nos momentos mais difíceis, agradeço também pelos momentos de alegria.

Ao meu professor e orientador Marcus Guelpeli, pela paciência, disponibilidade e pelo conhecimento compartilhado. Aprendi muito com você.

A todos os meus professores, que foram tão importantes na minha vida acadêmica.

Às professoras Geruza e Maria lúcia, que aceitaram fazer parte da avaliação deste trabalho.

Agradeço ao pessoal do DTI, em especial ao Douglas, que durante o estágio esteve sempre disposto a compartilhar seus conhecimentos e experiências como analista de sistemas.

A todos que de alguma forma me ajudaram.

RESUMO

O agrupamento de textos é uma técnica de Mineração de Texto (*Text Mining*) que tem o objetivo de organizar textos em grupos que contenham assuntos similares. A frequência das palavras dentro dos textos é um critério de similaridade adotada para se criar esses grupos. Este trabalho tem a proposta de apresentar um estudo sobre a qualidade dos agrupamentos gerados pelo modelo Cassiopeia. Os textos utilizados serão bulas de medicamentos. Para isso, foi utilizado os corpus Contra Indicações, Iterações Medicamentosas, Características Farmacológicas, Uso na Gravidez e um quinto corpus denominado "Tudo" que compreende na união dos demais. Os textos foram processados de forma que possibilitou a análise da relação dos resultados obtidos com a quantidade de palavras que compõe cada corpus. A avaliação dos agrupamentos foi feita com a utilização da métrica interna Coeficiente de *Silhouette* que mede a qualidades de agrupamentos de textos. Os resultados obtidos neste trabalho, possibilita julgar se o modelo Cassiopeia apresenta ou não resultados satisfatórios como agrupador de textos.

Palavras-chave: agrupamento de texto, modelo cassiopeia, bulas de medicamentos, mineração de texto.

ABSTRACT

The Text grouping is a Text Minig technique that aims to organize texts in groups that contain similar subjects. The frequency of words in texts is a criterion of similarity adopted to create such groups. This work has the proposal to present a study about the quality of the groupings generated by the Cassiopeia model. The texts used will be drug leaflet. For this, was used the corpus Contra Indications, Medicinal Iterations, Pharmacological Characteristics, Use in Pregnancy and a fifth corpus that includes the union of the others. The texts were processed in a way that made possible the analysis of the relation of the results obtained with the number of words that compose each corpus. The evaluation of the groupings was done using the internal metric Silhouette coefficient that measures the qualities of groupings of texts. The results obtained in this work makes it possible to judge whether or not the Cassiopeia model presents satisfactory results as a grouping of texts.

Keywords: text grouping, cassiopeia model, medication package, text mining.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE SIGLAS

KD - *Knowledge Discovery*

KDD - *Knowledge Discovery in Databases*

KDT - *Knowledge Discovery in Texts*

RI - Recuperação de Informação

T1 - Teste 1

T2 - Teste 2

T3 - Teste 3

T4 - Teste 4

T5 - Teste 5

SUMÁRIO

1. INTRODUÇÃO

Atualmente, com o crescimento rápido da internet, existe uma quantidade enorme de informações disponível às pessoas. Porém, este volume crescente de informações torna mais difícil a tarefa de assimilação da informação (WIVES, 1999). Devido ao avanço em dispositivos de armazenamento de dados, qualquer desktop ou computador laptop pode armazenar enormes quantidades de dados. Assim, acumular informação é fácil, mas encontrar informação relevante sobre uma busca pode ser difícil. Construir estruturas de dados para facilitar a recuperação de informações relevantes torna-se problemático à medida que os tamanhos das coleções continuam escalar. Cerca de oitenta por cento das informações existentes é na forma de textos, especialmente na internet, por exemplo notícias, livros eletrônicos, documentos de pesquisa, biblioteca digital, páginas web, e-mails e assim por diante. Assim, as pessoas precisam de métodos rápidos e eficientes para encontrar informações (QI E JIANFENG, 2013).

Segundo Levy (2005), o problema de se lidar com muita informação é que se perde um tempo que poderia ser mais bem empregado pensando, refletindo ou raciocinando. A sobrecarga de informação pode gerar vários problemas, dentre eles, o problema que está relacionado à localização da informação relevante e o que está relacionado com a extração de conhecimento presente nas informações relevantes encontradas (WIVES, 1999).

Pode-se definir Mineração de Texto como sendo o processo de extrair padrões ou conhecimento, interessantes e não triviais, a partir de documentos textuais (LOH, 2001). As aplicações da mineração de texto são inúmeras. Qualquer domínio que utilize intensivamente textos poderá beneficiar-se destes sistemas, tal como áreas jurídicas e policiais, os cartórios e órgão de registros, empresas em geral, etc. (LOH, 2001).

Existem várias abordagens ou técnicas de mineração de texto, dentre elas está a mineração de texto por Agrupamento, que é a principal técnica tratada no presente trabalho. Segundo Wives (2004), a tarefa de agrupar objetos, também conhecida por *clustering*, não é recente. O conceito de aglomerado (*cluster*) é tão antigo quanto as bibliotecas. Muitos anos antes da criação dos primeiros computadores, as pessoas já realizavam este processo manualmente, pois agrupar elementos similares facilita a localização de informações. Muitos algoritmos de agrupamento de objetos já foram implementados, estudados e aplicados em diversas áreas do conhecimento, tais como a psiquiatria (com o objetivo de redefinir categorias de diagnóstico existentes), a arqueologia (para investigar os relacionamentos entre os vários tipos de artefatos) e a Genética (especialmente após a criação do projeto GENOMA Humano).

Atualmente, as áreas de marketing e de economia têm despertado grande interesse pelas técnicas de agrupamento, com o objetivo de obter conhecimento sobre padrões de consumo.

A mineração de texto tornou-se, nos últimos anos, um novo tema de pesquisa promissor, especialmente em agrupamento de texto. Nas primeiras pesquisas com a mineração de texto, o agrupamento de texto foi utilizado como uma maneira eficiente de encontrar textos com conteúdo semelhantes. Recentemente, o agrupamento de texto é utilizado na definição de exibição de textos e organização dos resultados de ferramentas de pesquisa de conteúdo na internet, como por exemplo, os motores de busca (QI e JIANFENG, 2013).

O presente trabalho tem como objetivo avaliar a qualidade dos agrupamentos realizados pelo modelo Cassiopeia, utilizando bulas de medicamentos. Essas bulas foram coletadas no site www.medicinanet.com.br e formam um *corpus* que é um conjunto de dados textuais coletados criteriosamente para ser objeto da pesquisa. As bulas estão armazenadas em formato .TXT e divididas em Apresentação, Bula Completa, Indicações, Contraindicações, Posologia, Uso Na Gravidez, Interações Medicamentosas e Reações Adversas. Como resultado do processamento, as bulas de medicamentos estarão agrupadas por algum critério de similaridade que posteriormente terão os grupos avaliados por meio da métrica interna Coeficiente de *Silhouette*[1].

1.1. MOTIVAÇÃO

Atualmente, com o avanço da tecnologia, a área de mineração de textos é bastante promissora, porém ainda é pouco estudada em relação a mineração de dados e outras áreas tecnológicas. Com o crescimento rápido da internet, é indispensável que haja ferramentas que dêem apoio aos usuários na manipulação das informações disponíveis em forma textual. A mineração de texto será de grande importância para o avanço da área de recuperação de informação (RI).

A mineração de texto é um processo da descoberta de conhecimento em bases textuais. Dessa maneira, espera-se que com a análise dos resultados deste trabalho, possa ser mostrado que o modelo Cassiopeia apresenta um bom desempenho como agrupador de textos, assim, confirmando que esse modelo pode ser uma boa opção de ferramenta para ser utilizada em trabalhos futuros que envolvam a descoberta de conhecimento em textos.

[1] Coeficiente de *Silhouette* é uma métrica de avaliação de agrupamentos de textos.

1.2. PROBLEMA

Foi demonstrado no trabalho de Guelpelli (2012) que o modelo Cassiopeia apresenta bons resultados nos domínios jornalístico, jurídico e médico. Poderá o modelo Cassiopeia apresentar resultados satisfatórios como agrupador, utilizando bulas de medicamentos, ou seja, no domínio farmacêutico?

1.3. HIPÓTESE

O modelo Cassiopeia apresentará bons resultados no domíno farmacêutico.

1.4. CONTRIBUIÇÃO

- Avaliação do modelo Cassiopeia como agrupador de textos.
- Processos para a realização de futuros trabalhos que envolvam a Descoberta de Conhecimento em Bulas de Medicamentos.

1.5. METODOLOGIA DE PESQUISA

A metodologia adotada para realização deste trabalho compreende a leitura bibliográfica e métodos quantitativos. Foi utilizado métricas internas para a avaliação dos agrupamentos gerados sobre os *corpus* Contra Indicações, Interações Medicamentosas, Características Farmacológicas, Uso na Gravidez e Tudo[2]. Essa avaliação é relacionada com a quantidade de palavras em cada *corpus.* Foi o utilizado o software *Get FineCount*, versão2.6.1924 para realizar a contagem das palavras.

1.6. ESTRUTURA PROPOSTA

Capítulo 2: Fundamentação Teórica

Neste capítulo serão mostrados os principais conceitos que fundamentam o presente trabalho.

[2] Tudo é um corpus composto por textos que contêm as informações de contra indicações, interações medicamentosas, características farmacológicas e uso na gravidez juntas.

Capítulo 3: Metodologia

Este capítulo têm como objetivo, descrever a metodologia utilizada para a realização desse trabalho.

Capítulo 4: Resultados

No capítulo 4 serão mostrados os resultados da métrica interna ou não supervisionada (Coeficiente de Silhouette), relacionando com a quantidade de palavra dos *corpus*.

Capítulo 5: Discussão dos Resultados

Neste capítulo, será apresentada a análise dos resultados obtidos nos experimentos.

Capítulo 6: Conclusão

No capítulo 6 serão apresentados as limitações e trabalhos futuros.

2. FUNDAMENTAÇÃO TEÓRICA

Neste capítulo serão mostrados os principais conceitos que fundamentam o presente trabalho. A mineração de texto é um processo da descoberta de conhecimento, sendo assim, serão apresentados também conceitos sobre Descoberta de Conhecimento Apoiada por Computador (*Knowledge Discovrery - KD*). Será apresentado, inicialmente, um conceito amplo de Conhecimento. Na seção 2.2 serão mostrados alguns conceitos do KD. Serão mostrados os tipos de descoberta de conhecimento: *Conhecimento em Dados Estruturados – KDD (Knowledge Discovery in Database)* e *Descoberta de Conhecimento em Dados não Estruturados – KDT (Knowledge Discovery From Text)*. Em Descoberta De Conhecimento Em Dados Não Estruturados, serão descritos métodos para a descoberta de conhecimento que envolvem pré-processamento, processamento ou mineração de texto e pós-processamento. Em Descoberta de Conhecimento por Agrupamento, será apresentado o que é um agrupamento de informações textuais e as fases do KDT, serão mostradas também as métricas externas *Recall*, *Precision* e *F-Measure* e as métricas internas *Coesão, Acoplamento* e *Coeficiente de Silhouette*. Por fim, será mostrada uma visão geral do Modelo Cassiopeia, que será utilizado para fazer os agrupamentos de textos.

2.1. CONHECIMENTO

Na busca do saber, uma pessoa pode adquirir informações empiricamente, aprendendo, sem compreender o nexo causal que dá origem ao fenômeno. Pode ter um conhecimento por experiência como, por exemplo, o modo de dirigir um automóvel sem que tenha compreensão do processo mecânico que sua ação desencadeia. O conhecimento humano inicia-se na primeira infância quando a criança, por imitação, repete os gestos, as expressões faciais e as palavras dos adultos com quem convive. (WERNECK, 2006).

Segundo Mizzaro (1996), os estudos mais importantes e atuais realizados a fim de definir conhecimento e de compreender e explicar seu processo de aquisição e raciocínio estão nas áreas de sociologia, psicologia e cognição. Nessas áreas, o conhecimento é compreendido como sendo a forma com que a pessoa percebe o mundo. Nesse caso, o conhecimento corresponderia à experiência adquirida pelo ser, constituindo-se de respostas figurativas ou motoras aos estímulos sensoriais (ABEL, 2001 *apud* WIVES, 2004). Assim, cada pessoa possui a sua versão do mundo real, mantida internamente (WIVES, 2004).

Uma pessoa está em constante interação com o meio, por isso o seu conhecimento muda com o passar do tempo. Deste modo, o conhecimento de uma pessoa em determinado momento é denominado estado de conhecimento (MIZZARO, 1996 *apud* WIVES, 2004).

Há anos a Inteligência Artificial busca compreender o conhecimento e o processo de inteligência, desenvolvendo formalismos ou mecanismos de representação, armazenamento e utilização de conhecimento através de uma linguagem artificial, tentando aproximar essa representação o melhor possível da forma natural real. O conhecimento aqui descrito trata-se de uma abstração do conhecimento humano real e pode não possuir o mesmo significado de "Conhecimento" que os filósofos tanto buscam definir em termos epistemológicos. Apesar de não corresponder exatamente à forma original ou natural, o conhecimento, no contexto computacional, tem sido extremamente útil para alguns dos processos de raciocínio e inteligência desenvolvidos na área (WIVES, 2004).

2.2. DESCOBERTA DE CONHECIMENTO

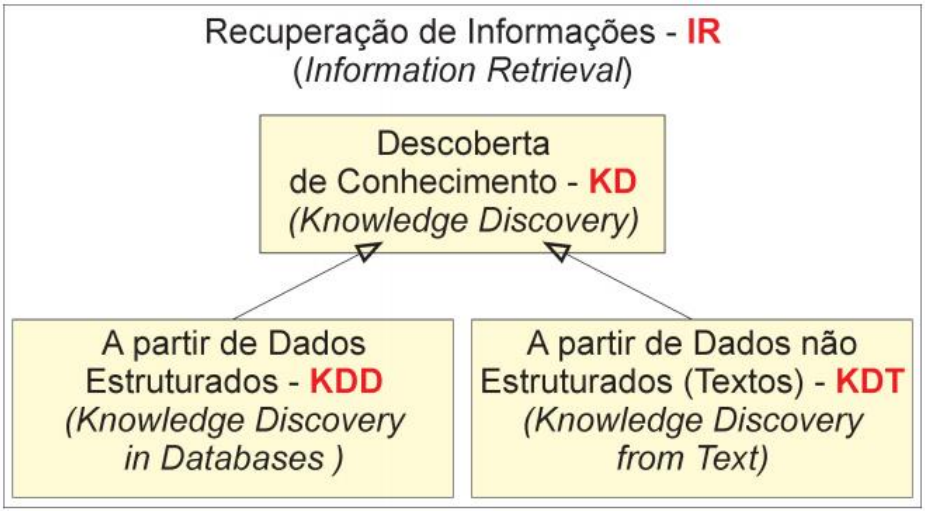
Descobrir conhecimento significa identificar, receber informações relevantes e poder computá-las e agregá-las ao conhecimento prévio, mudando o estado de conhecimento atual, a fim de que determinada situação ou problema possa ser resolvido (WIVES, 2004).

Sabe-se que o volume de informação disponível é muito grande. Neste sentido, mecanismos automáticos de processamento tendem a tornar o processo de descoberta de conhecimento mais eficiente. Logo, faz-se necessário automatizar este processo, principalmente através da utilização de software e computadores. Neste contexto, surge a Descoberta de Conhecimento Apoiada por Computador, que é um processo de análise de dados ou informações, cujo principal objetivo é fazer com que as pessoas possam adquirir novos conhecimentos a partir da manipulação de grandes quantidades de dados (MORAIS e AMBRÓSIO, 2007).

São muitas as discussões em torno das definições das técnicas de extração automática e informações relevantes em grandes massas de dados, nas quais conceitos e termos se misturam, por vezes sendo utilizados como sinônimos: prospecção de conhecimento, descoberta de conhecimento em bases de dados, mineração de dados, descoberta de conhecimento em textos, mineração de textos (RAMOS e BRÄSCHER, 2009). De forma abrangente, utiliza-se o termo "Descoberta de Conhecimento", passando-se a qualificá-lo a partir do conteúdo a ser analisado: se este foi previamente organizado e estruturado, denomina-se *Descoberta de Conhecimento em Dados Estruturados – KDD (Knowledge Discovery in*

Database) ou se encontra disperso em documentos textuais dos mais diversos formatos e tamanhos chama-se *Descoberta de Conhecimento em Dados não Estruturados – KDT (Knowledge Discovery From Text)* (Figura 1).

Figura 1 - Tipos de Descoberta de Conhecimento (MORAIS e AMBRÓSIO, 2007)



2.3. DESCOBERTA DE CONHECIMENTO EM DADOS NÃO ESTRUTURADOS

A evolução da área de Recuperação de Informação teve como consequência o surgimento da área de descoberta de conhecimento em textos. Pode-se então definir descoberta de conhecimento em textos ou *Text Mining* como sendo o processo de extrair padrões ou conhecimento, interessante e não-triviais, a partir de documentos textuais (TAN, 99 *apud* LOH, 2001).

Grandes repositórios textuais contêm informações adormecidas, camufladas, até que o minerador as encontre e as transforme em informações preciosas para uma organização ou para outras finalidades (RAMOS e BRÄSCHER, 2009).

A descoberta de conhecimento em dados textuais pode ser vista como uma extensão da descoberta de conhecimento em dados estruturados. Como a forma mais natural de armazenamento de informação é em texto, acredita-se que a mineração de texto tem um potencial comercial superior ao da mineração de dados. Um estudo recente indicou que 80% das informações estão contidas em documentos de texto. No entanto, a mineração de texto é também uma tarefa muito mais complexa, pois precisa lidar com dados de texto intrinsecamente desestruturados e difusos (AKILAN, 2015).

O KDT, ao invés de deixar que o usuário mesmo procure em um texto, o que lhe interessa, se preocupa em encontrar informações dentro dos textos e trata-las de forma a apresentar ao usuário algum tipo de conhecimento útil e novo. Mesmo que tal conhecimento não seja a resposta direta às indagações do usuário, ele deve contribuir para satisfazer suas necessidades de informação(LOH, 2001).

Segundo Morais e Ambrósio (2007), a análise de dados em formato não estruturado pode ser considerada uma atividade mais complexa, se comparada à análise de dados estruturados, justamente pelo fato dos dados possuírem a características de não estruturação.

O processo de KDT é centrado no processo de Mineração de Textos, que é um campo multidisciplinar, que envolve recuperação de informação, análise textuais, extração de informação, clusterização, categorização, visualização, tecnologia de base de dados, e mineração de dados (MORAIS e AMBRÓSIO, 2007).

Segundo Castro, Mattos, e Simões (2012), o processo de KDT é dividido em três etapas: preparação dos dados (pré-processamento), mineração de texto (processamento) e análise e validação dos resultados (pós-processamento).

2.3.1. PREPARAÇÃO DOS DADOS (PRÉ-PROCESSAMENTO)

O processo de descoberta de conhecimento em textos se inicia na etapa onde se dá a preparação, ou seja, a forma como a seleção dos dados pode representar de maneira significativa o conteúdo dos textos.

Aproximadamente 2% das palavras de um Corpus são usadas para análise textual. Outros elementos como, por exemplo, as *stopword*s, espaços em branco, cabeçalhos e rodapés, não são necessários para análise de padrões e agrupamento de documentos. Dessa forma, é necessário um pré-processamento antes da análise textual e extração de conhecimento. A mineração de texto normalmente envolve um processo de estruturação do texto de entrada (AGNIHOTRI, VERMA, e TRIPATHI, 2014).

Para que a Mineração de Texto obtenha sucesso, é necessário que a preparação dos textos seja eficiente, utilizando-se das seguintes atividades: recuperação da informação, análise de dados e transformação dos dados.

2.3.1.1. RECUPERAÇÃO DE INFORMAÇÃO

É uma etapa da preparação de dados que se apresenta como uma espécie de filtro atuante sobre uma coleção de documentos, com o intuito de retorno do problema a solucionar. Podem-se utilizar o modelo lógico que se vale de um conjunto de termos e índice, que representam o documento e, para efetuar a recuperação da informação, a consulta se dá pela combinação desses termos índice com os operadores lógicos, *and*, *or* ou *not* (FELDMAN, 2007).

2.3.1.2. ANÁLISE DE DADOS

Busca facilitar que as palavras sejam identificadas de acordo com similaridades de significados. São empregados alguns tipos de processos para que a análise dos dados tenha um retorno satisfatório como, por exemplo: a remoção de *stopwords*, que segundo Espina e Rino (2001) visa identificar as palavras que não tenham significado relevante ao texto, como artigos e preposições, e que possam ser removidos sem prejuízo ao conteúdo textual; o *stemming*, onde cada palavra é isolada e na sequencia, é feita uma redução a uma possível raiz, ou *stem*; e o *dicionário de thesaurus*, que pode ser definido como um vocabulário composto de termos, relacionados a uma língua natural e que são usados para representar o conteúdo de documentos de forma condensada (PLUSKWA, 1999).

2.3.1.3. TRANSFORMAÇÃO DOS TEXTOS

A transformação de textos indica um tipo de representação conhecida como *bag of word*, com potencial para conversão em tabelas. Tais tabelas, contudo, não possuem tamanho reduzido, ao contrário, cada termo que não fora descartado na análise é convertido em um atributo (REZENDE, 2003).

2.3.2. MINERAÇÃO DE TEXTO (PROCESSAMENTO)

A mineração de textos está relacionada à busca de informações específi-cas em documentos, à análise qualitativa e quantitativa de grandes volumes de textos, e à melhor compreensão de textos disponíveis em documentos. Essa fase tem como principal objetivo encontrar informações relevantes no texto analisado, observando-se que o grau de relevância da inforamação está diretamente ligado a necessidade do usuário. Para definir essa relevância, deve-se saber primeiramente quais são os objetivos, pois pode-se dizer que informação relevante é aquela que vai ao encontro às necessidades do usuário em determinado momento (WIVES, 2001).

De forma geral, as etapas do processo de mineração de textos são as seguintes: seleção de documentos, definição do tipo de abordagem dos dados (análise semântica ou estatística), preparação dos dados, indexação e normalização, cálculo da relevância dos termos e seleção dos termos (MORAIS e AMBRÓSIO, 2007).

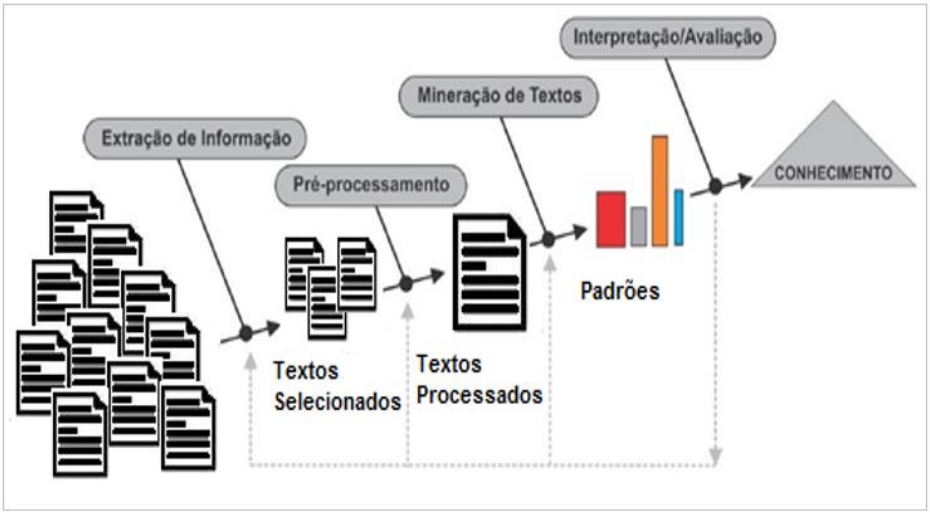
2.3.3. PÓS-PROCESSAMENTO

Fase do KDT em que ocorre a validação das descobertas que foram efetuadas pelo processamento de dados, representando também a fase onde são disponibilizados os dados encontrados em formato visual. São utilizadas medidas para avaliar os dados obtidos, bem como ferramentas de visualização, que se valem do conhecimento de especialistas como forma de avaliação dos resultados.

Segundo Rezende(2003), as medidas são de avaliação básicas adotadas pela mineração de texto, ou seja, são métricas adotadas na recuperação de informação. Tem como base a relevância, que consiste em verificar o nível de resposta satisfatória ao usuário, ou seja, averigua se a informação produzida pelo processo de mineração de texto atende as necessidades de informação iniciais.

A visualização de informação atualmente se dá por meio de sistemas gráficos com telas coloridas, aplicações gráficas e interativas, sendo perfeitamente comum esse tipo de representação, visto que a informação disponibilizada em formato visual torna-se bem mais compreensível. A figura 2 apresenta uma visão geral das etapas do KDT.

Figura 2 - Etapas do KDT (NOGUEIRA, 2014)



2.4. DESCOBERTA DE CONHECIMENTO POR AGRUPAMENTO DE TEXTOS

A técnica de descoberta de conhecimento por agrupamento de textos consiste em agrupar textos que tenham características comuns, ou seja, que tenham maior similaridade entre si. Essa técnica facilita o entendimento e identificação de classes potenciais para descoberta de algum conhecimento útil (MORAIS e AMBRÓSIO, 2007).

2.4.1. AGRUPAMENTO

Um cluster é um grupo de objetos de dados que são semelhantes um ao outro dentro do mesmo cluster e são diferentes dos objetos de outros clusters. A clusterização é uma técnica de aprendizagem que se tornou uma importante área de estudo no campo da mineração de dados (AALAM e SIDDIQUI, 2016).

A clusterização tem sido usada em vários campos, como aprendizagem de máquinas, reconhecimento de padrões e análise bioinformática. No cotidiano, clusterização tem suas vantagens em marketing para identificar o comportamento de compra dos clientes, para categorizar pacientes e doenças de acordo com seus sintomas, no gerenciamento de bibliotecas para categorizar os livros com base em seus títulos, autores, custo, dentre outros e em muitas outras áreas onde a categorização dos dados pode ser necessária com base em critério (AALAM e SIDDIQUI, 2016).

Segundo Morais e Ambrósio (2007), a clusterização auxilia o processo de descoberta de conhecimento, facilitando a identificação de padrões (características comuns dos elementos) nas classes. Esta técnica pode ser utilizada para estruturar e sintetizar o conhecimento, quando este é incompleto ou quando há muitos atributos a serem considerados. Também pode ser utilizada para facilitar o entendimento e identificação de classes potenciais para descoberta de algum conhecimento útil.

A tarefa de agrupar objetos, também conhecida por *clustering*, não é recente. O conceito de aglomerado (*cluster*) é tão antigo quanto as bibliotecas. Muitos anos antes da criação dos primeiros computadores, as pessoas já realizavam este processo manualmente, pois agrupar elementos similares facilita a localização de informações (WIVES, 1999).

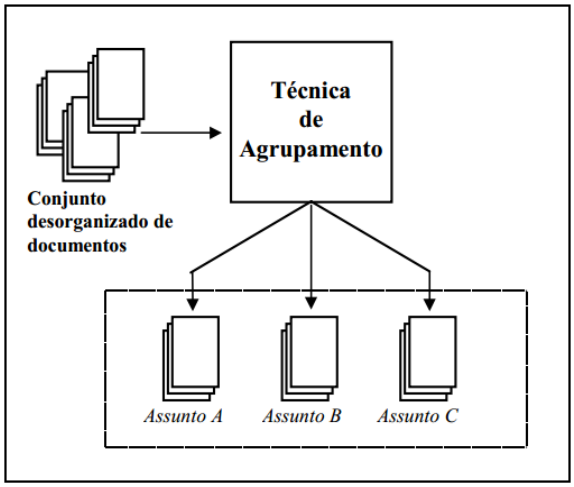
O processo de agrupar significa colocar os elementos (objetos) de uma base de dados (conjunto), de tal maneira, que os grupos formados representem uma configuração, na qual cada elemento tenha maior similaridade com qualquer outro, do mesmo grupo (BERKHIN, 2002 apud GUELPELI, 2012).

Segundo Kow (1997 *apud* WIVES, 1999), um aglomerado é um grupo de objetos similares, geralmente uma classe, que possui um título mais genérico capaz de representar todos os elementos nela contidos.

2.4.2. AGRUPAMENTO DE INFORMAÇÕES TEXTUAIS

Segundo Wives (1999) 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 3).

Figura 3 - Objetivo do agrupamento de informações textuais (WIVES, 1999)



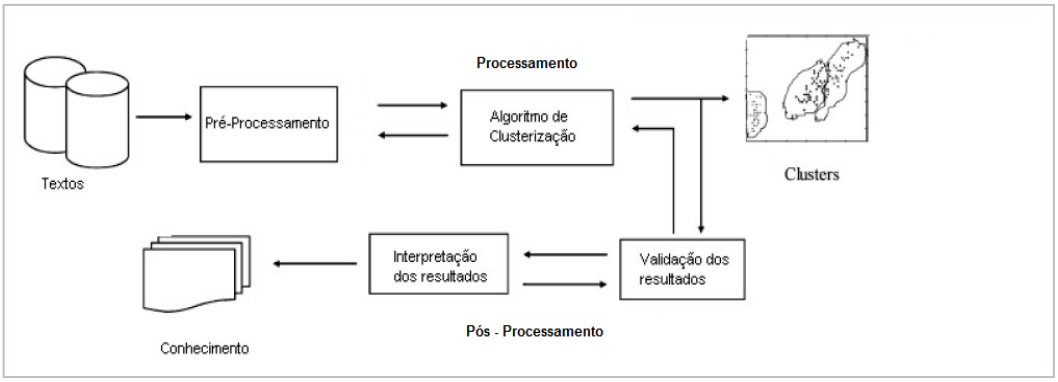
Segundo Guelpeli (2012), o agrupamento de texto é um processo totalmente automático que reparte uma coleção em grupos de textos de conteúdos similares, cujo objetivo é ter maior conhecimento sobre esses textos e suas relações. Assim, este processo consegue reunir uma coleção de padrões desconhecidos (não classificados) em agrupamentos que possuam algum significado.

Conforme Guelpeli (2012), o problema de agrupamento de texto pode ser definido como: dada uma base de texto *T*, devem-se agrupar os elementos de *T* de maneira que os textos mais similares sejam colocados no mesmo grupo, e os menos similares, em grupos distintos. Sendo assim, dado um conjunto com *n* textos *T={T1, T2,..., Tn}*, obtém-se um conjunto de *k* agrupamentos *G={G1, G2,..., Gk}*, cujos textos de um determinado agrupamento *Gi* são similares entre si, mas não são similares aos textos contidos em um conjunto *Gj* qualquer, onde $i \neq j$.

2.4.3. FASES DO AGRUPAMENTO DE TEXTO

Conforme mostrado na figura 4, o agrupamento de texto é dividido em três fases: pré-processamento, processamento e pós-processamento.

Figura 4 - Fases do Agrupamento de Texto (NOGUEIRA, 2014)



Para Goldschmidt e Passos (2005), a etapa de pré-processamento consome 60% de todo o processo, e é uma etapa vital, tanto para a economia de tempo como para o bom funcionamento das seguintes. Preparar os textos para o processo computacional é uma atividade difícil e trabalhosa, que não é nova (GUELPELI, 2012).

Guelpeli (2012) considera a fase de pré-processamento como a parte mais crítica, pois determina a boa qualidade dos agrupadores textuais. Existem, nessa fase, técnicas usadas para reduzir os atributos, como por exemplo, a retirada das *stopwords*, a radicalização de termos com *stemming*, disponível nas línguas portuguesa e inglesa.

Segundo Guelpeli (2012), após a etapa de pré-processamento, os agrupadores de texto precisam de uma estruturação de documentos, para torná-los processáveis pelos algoritmos de agrupamento. Segundo Feldman e Sanger (2006, apud GUELPELI 2012), o modelo mais usual para representação de dados textuais é o modelo espaço-vetorial, no qual cada documento é um vetor em um espaço multidimensional, e cada dimensão é um termo da coleção, obtendo, assim, uma matriz de documento-termo.

Na etapa de processamento, algoritmos de agrupamento de textos são utilizados. Estes algoritmos são um conjunto de técnicas e processos capazes de indicar um conhecimento significativo e inovador em textos (MORAIS e AMBRÓSIO, 2007).

A etapa de pós-processamento é onde ocorre a análise dos resultados, particularmente do processo de mineração de textos.

De acordo com Wives (2002), esta análise pode ser realizada com base em técnicas de uma área conhecida como bibliometria, que é uma sub-área da biblioteconomia encarregada de estudar e aplicar métodos matemáticos e estatísticos em documentos e outras formas de comunicação.

Na última fase, os grupos gerados são interpretados e avaliados. Segundo Guelpeli (2012), na fase de pós-processamento, são usadas as medidas de validação de agrupamentos. Essas medidas são usadas na avaliação de agrupamentos e podem ser distribuídas em duas grandes categorias: externas ou supervisionadas e as internas ou não supervisionadas (HALKIDI, 2001 apud GUELPELI 2012).

2.4.4. MÉTRICAS EXTERNAS OU SUPERVISIONADAS

Segundo Guelpeli (2012), para as métricas externas ou supervisionadas, os resultados dos agrupamentos são avaliados por uma estrutura de classes pré-definidas, que refletem a opinião de um especialista. Para esse tipo de avaliação, são usadas medidas como: Precisão, *Recall*, e como medida harmônica destas duas, o *F-Measure*.

2.4.4.1. RECALL

Segundo Guelpeli (2012) *Recall* mede a proporção de objetos corretamente alocados a um agrupamento, em relação ao total de objetos da classe associada a este agrupamento, onde n(A) é o número de elementos do subconjunto A de acertos[3] e n(D) é o número de elementos do subconjunto D de falsos negativos[4] e n(A ∪D) é o número total de elementos da classe correspondente (Equação 1).

Recall(R): Equação 1:

$$R = \frac{n(A)}{n(A \cup D)}$$

[3] Acertos são elementos que foram corretamente alocados em um grupo.
[4] Falsos Negativos são elementos que deveriam ter sido alocados a um grupo e que foram alocados a outros.

2.4.4.2. PRECISION

O *Precision*, mede a proporção de objetos corretamente alocados a um agrupamento, em relação ao total de objetos deste agrupamento. Onde n(A) é o número de elementos do subconjunto de A de acertos e n(B) é o número de elementos do subconjunto B de falsos positivos e n(A ∪ B) é o número total de elementos do grupo (Equação 2). (GUELPELI, 2012).

Precision(P): Equação 2:

$$P = \frac{n(A)}{n(A \cup B)}$$

2.4.4.3. F-MENSURE

O *F-Measure* é a medida harmônica entre o *Precision* e o *Recall* que, no *F-Measure*, assume valores que estão no intervalo de [0,1]. O valor zero indica que nenhum objeto foi agrupado corretamente, o valor um, que todos os objetos estão contidos corretamente agrupados. Assim, um agrupamento ideal deve retornar um valor igual a 1 (Equação 3). (GUELPELI, 2012).

F-Measure: Equação 3:

$$2 * \frac{Precision(P) * Recall(R)}{Precision(P) + Recall(R)}$$

2.4.5. MÉTRICAS INTERNAS

Nas métricas internas ou não supervisionadas, utilizam-se apenas informações contidas nos grupos gerados para realizar a avaliação dos resultados, ou seja, não se utilizam informações externas. As medidas mais usadas para este fim são Coesão, Acoplamento e Coeficiente de *Silhouette* (GUELPELI, 2012).

2.4.5.1. COESÃO

A *Coesão* mede a similaridade entre os elementos do mesmo agrupamento. Quanto maior a similaridade entre eles, maior a coesão deste agrupamento. Onde *Sim*(*Pi, Pj*) é o cálculo da similaridade entre os textos *i* e *j* pertencentes ao agrupamento *P*, *n* é o número de textos no agrupamento *P*, e *Pi* e Pj são membros do agrupamento *P* (Equação 4). (GUELPELI, 2012).

Coesão(C): Equação 4:

$$\frac{\sum_{i>j} Sim(P_i, P_j)}{\frac{n(n-1)}{2}}$$

2.4.5.2. ACOPLAMENTO

O *Acoplamento* mede a similaridade média de todos os pares de elementos, sendo que um elemento pertence a um agrupamento e o outro não pertence a esse mesmo agrupamento (KUNZ e BLACK, 1995 apud Guelpeli 2012). Onde *C* é o centroide de determinado agrupamento, presente em *P*, *Sim* (*Ci, Cj*) é o cálculo da similaridade do texto *i* pertencente ao agrupamento *P* e o texto *j* não pertence a *P*, *Ci* centroide do agrupamento *P* e *Cj* é centroide do agrupamento *Pi* e *na* é o número de agrupamentos presentes em *P* (Equação 5).

Acoplamento(A): Equação 5:

$$\frac{\sum_{i>j} Sim(C_i, C_j)}{\frac{n_a(n_a - 1)}{2}}$$

2.4.5.3. COEFICIENTE DE *SILHOUETTE*

O *Coeficiente Silhouette* baseia-se na ideia de quanto um objeto é similar aos demais membros do seu grupo, e de quanto este mesmo objeto é distante dos de outro grupo. Assim, essa medida combina as medidas de coesão e acoplamento. Onde *a* (*i*) é a distância

média entre o *i-ésimo* elemento do grupo e os outros do mesmo grupo. O *b*(*i*) é o valor mínimo de distância entre o *i-ésimo* elemento do grupo e qualquer outro grupo, que não contém o elemento, e *max* é a maior distância entre *a* (*i*) e *b*(*i*) (Equação 6). (GUELPELI, 2012).

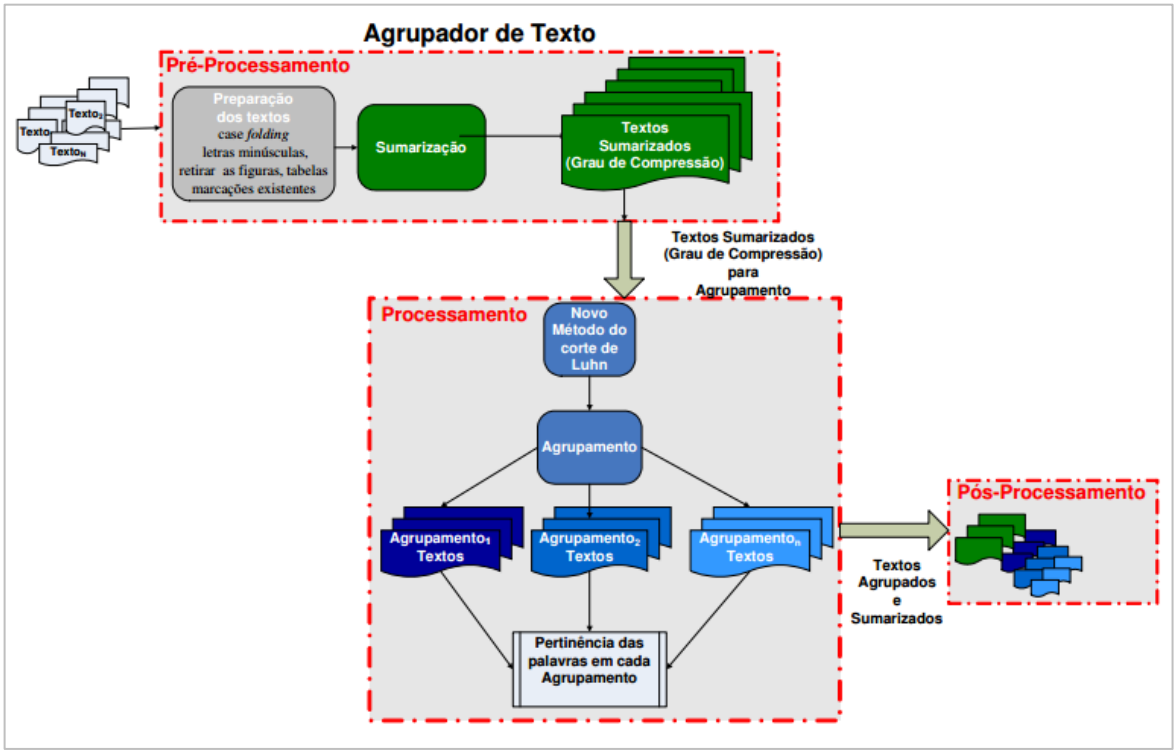
Coeficiente de Silhouette(S): Equação 6:

$$S = \frac{b(i) - a(i)}{max(a(i), b(i))}$$

2.5. MODELO CASSIOPEIA

O Modelo Cassiopeia é um agrupador de texto proposto por Guelpeli (2012). Esse modelo é composto por três macro etapas (pré-processamento, processamento e pós-processamento) conforme figura 5.

Figura 5 - Modelo Cassiopeia (GUELPELI. 2012)



O processo começa com a entrada de textos, que passam pela etapa de pré-processamento, na qual são preparados para o processo computacional, utilizando-se a técnica case *folding* que coloca todas as letras em minúsculas, além de outros cuidados, como descarte de todas as figuras, tabelas e marcações existentes. Nesta etapa, é usado o processo de sumarização, cuja finalidade é diminuir o número de palavras, viabilizando o processamento (GUELPELI, 2012).

Terminada a etapa de pré-processamento, começa a de processamento, que usa o processo de agrupamento de textos hierárquicos e um algoritmo para juntar os textos com similaridade. À medida que novos textos são agrupados ocorre o reagrupamento, podendo surgir agrupamentos, sub-agrupamentos ou até mesmo a fusão destes (GUELPELI, 2012).

A etapa de pós-processamento, na qual cada um dos agrupamentos ou sub-agrupamentos terá, por similaridade, um conjunto de textos-fonte com os sumários correspondentes, que têm alto grau de informatividade e contêm as ideias principais dos textos-fonte, característica da sumarização.

2.5.1. MODELO CASSIOPEIA – PRÉ-PROCESSAMENTO

A etapa de pré-processamento é a que consome mais tempo de toda a mineração de texto, e é essencial, tanto para a economia de tempo como para o bom funcionamento das etapas seguintes da recuperação de informação, principalmente a de processamento, que depende, fundamentalmente, da quantidade e da qualidade das palavras mantidas depois desta etapa (GUELPELI, 2012).

No pré-processamento ocorre a limpeza dos textos, a preparação para o processo computacional, mas a principal preocupação é a redução do número de palavras, não apenas para viabilizar a questão computacional, mas também para obter a informatividade das palavras mantidas, ou seja, proporcionar um ganho qualitativo e quantitativo para o processamento (GUELPELI, 2012).

2.5.2. MODELO CASSIOPEIA – PROCESSAMENTO

O agrupamento de textos, por similaridade, é usado na etapa de processamento, e acontece quando não se conhecem os elementos do domínio disponível, procurando-se, assim,

separar, automaticamente, os elementos em agrupamentos por algum critério de afinidade ou similaridade (GUELPELI, 2012).

2.5.2.1. IDENTIFICAÇÃO DOS ATRIBUTOS

O modelo Cassiopeia identifica as características das palavras no documento, utilizando a frequência relativa, que define a importância de um termo, de acordo com a frequência com que é encontrado no documento. Quanto mais um termo aparecer em um documento, mais importante é, para aquele documento (Equação 7). (GUELPELI, 2012).

Equação 7:

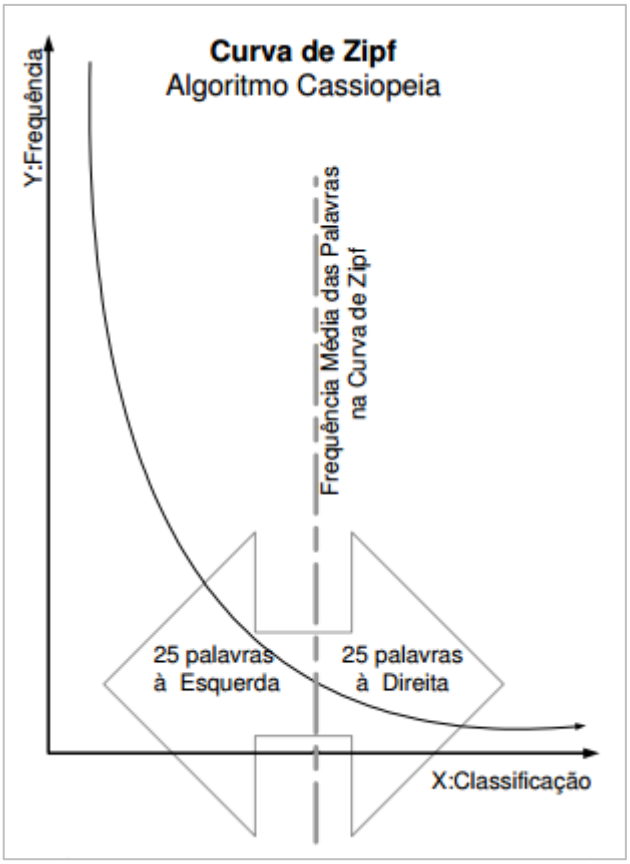
$$F_r X = \frac{F_{abs} X}{N}$$

Onde Frx é igual à frequência relativa de x, FabsX é igual à frequência absoluta de X, ou seja, a quantidade de vezes que X, a palavra aparece no documento e N é igual ao número total de palavras no documento. Considerado um espaço-vetorial, cada palavra representa uma dimensão (existem tantas dimensões quantas palavras diferentes no documento) (GUELPELI, 2012).

2.5.2.2. SELEÇÃO DOS ATRIBUTOS

Tendo como base os pesos das palavras, obtidos na frequência relativa, é calculada a média sobre o total de palavras no documento. Nessa etapa, o modelo usa a *truncagem*, ou seja, um tamanho máximo de 50 posições para os vetores de palavras, realizando um corte que representa a frequência média das palavras obtidas com os cálculos e, em seguida, realiza a organização dos vetores de palavras (Figura 4) (GUELPELI, 2012).

Figura 6 - Seleção dos Atributos no modelo Cassiopeia (GUELPELI, 2012)



O modelo Cassiopeia divide esse vetor de 50 palavras, ordenadas de forma decrescente, com 25 posições à direita e 25 à esquerda da frequência média, calculada para fazer a ordenação do vetor (GUELPELI, 2012).

Exemplificando o passo a passo, como ocorre o novo método para definição do corte de Luhn, proposto no modelo Cassiopeia, ou seja, a seleção dos atributos (GUELPELI, 2012):

1. calcular a frequência relativa: quantas vezes cada palavra aparece no documento, dividido pelo número total de palavras do documento;

2. ordenar as palavras em ordem decrescente de frequência (da maior para a menor);

3. achar a frequência média das palavras, somando as frequências relativas e dividindo pelo número total de palavras do documento;

4. encontrar a primeira palavra cuja frequência mais próxima à média;

5. marcar esta palavra e escolher, incluindo-a, mais as 24 anteriores (esquerda);

6. marcar esta palavra e escolher as 25 posteriores (direita);

7. montar o vetor em ordem decrescente com as 50 palavras escolhidas.

2.5.2.3. USO DO MÉTODO HIERÁRQUICO AGLOMERATIVO E DO ALGORÍTMO CLIQUES

O modelo Cassiopeia utiliza o método hierárquico, para organizar seus textos em agrupamentos, que são particionados, sucessivamente, produzindo uma representação hierárquica, tipo que facilita a visualização dos agrupamentos a cada ciclo de processamento, bem como o grau de similaridade obtido entre eles com uso do algoritmo Cliques. É um método que, de início, não requer definições de número de agrupamentos (GUELPELI, 2012).

Nesse método, os agrupamentos são recursivamente criados através do algoritmo Aglomerativo, considerando a medida de similaridade. Sendo assim, no início, os agrupamentos são em número reduzido, com baixo grau de similaridade entre os documentos de cada agrupamento. Os passos do Algoritmo Aglomerativo são (GUELPELI, 2012):

1. Procure pelo par de clusters com a maior semelhança;

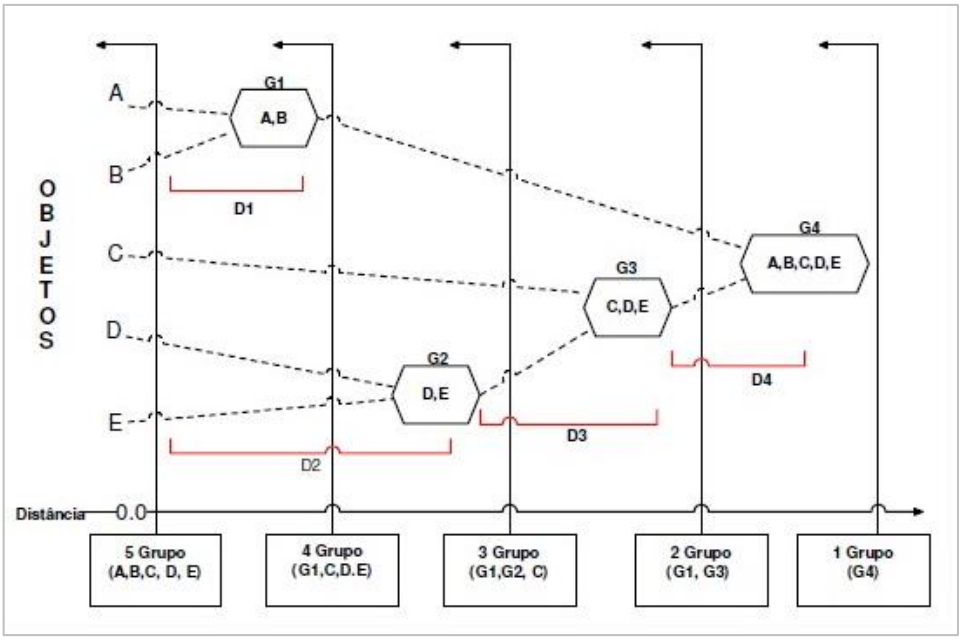
2. Crie um novo cluster que agrupe o par selecionado no passo 1;

3. Decremente em 1 o número de clusters restantes;

4. Volte ao passo 1 até que reste apenas um cluster.

Um exemplo do funcionamento do método hierárquico pode ser visto na Figura 7, com o uso do algoritmo Aglomerativo.

Figura 7 - Dendograma do método hierárquico aglomerativo (GUELPELI, 2012)



No modelo Cassiopeia, os textos são agrupados e adota o algoritmo *Cliques* para garantir a similaridade entre documentos e agrupamentos. O algoritmo *Cliques* baseia-se em teoremas e axiomas conhecidos da teoria dos grafos, e possue alta capacidade dedutiva, tendo, portanto, maior fundamentação teórica (GUELPELI, 2012).

A adaptação para o trabalho de Guelpeli (2012) foi definir o grau de similaridade, ou seja, contabilizar o total de palavras comuns entre os textos nos seus vetores e nos agrupamentos em seus centroides. Na primeira fase, a contabilização das palavras ocorre nos vetores dos textos para criar os agrupamentos. Na fase seguinte, todos os textos já estão em agrupamentos, cada um desses agrupamentos contém um centroide de palavras obtidos na primeira fase, começa então o reagrupamento. O reagrupamento contabiliza o total de palavras comuns entre centroides dos agrupamentos, podem surgir agrupamentos, subagrupamentos ou até mesmo a fusão destes(GUELPELI, 2012).

Passos do Algoritmo Cliques(GUELPELI, 2012):

1. Seleciona 1° elemento e coloca em um novo cluster;
2. Procura o próximo objeto similar;
3. Se o objeto é similar a todos os outros elementos do cluster, este objeto é o agrupado;

4. Voltar ao passo 2, enquanto houver objetos;

5. Para os elementos não alocados, repetir o passo 1.

2.5.3. MODELO CASSIOPÉIA – PÓS-PROCESSAMENTO

No pós-processamento, o modelo fornece uma estrutura hierárquica capaz de apresentar bons resultados para a recuperação de informação (GUELPELI, 2012). Nessa etapa, o modelo terá como saída os textos agrupados por similaridade. Com os textos agrupados no pós-processamento, é possível realizar a recuperação de documentos e, a partir da sua análise, pode-se obter outros similares. Com essa organização estrutural, uma generalização e/ou especificação de documentos pode ser feita, já que a partir do momento da recuperação, parece ser interessante possibilitar a consulta a outros documentos mais específicos ou mais genéricos. Quando o documento for encontrado pela recuperação de informação, a estrutura possibilitará ter o texto-fonte e o seu sumário correlato, ou seja, com alto grau de informatividade (GUELPELI, 2012).

3. METODOLOGIA

Este capítulo tem como objetivo, descrever a metodologia utilizada para a realização desse trabalho. Será apresentando o processo de seleção e coleta das bulas de medicamentos, bem como o processo de limpeza e organização das mesmas e por fim o processamento utilizando o Modelo Cassiopeia.

3.1. CRIAÇÃO DO CORPUS

Corpus é um conjunto de textos selecionados com base em algum critério significativo para um determinado objetivo e *corpora* é o conjunto de *corpus*. Os *corpus* criados para este trabalho são compostos por bulas de medicamentos. As bulas de medicamentos são um conjunto de informações sobre um determinado medicamento que obrigatoriamente os laboratórios farmacêuticos devem acrescentar à embalagem de seus produtos vendidos no varejo. A coleta das bulas de medicamentos foram feitas no site www.medicinanet.com.br.

3.1.1. COLETA DAS BULAS

Existem dois tipos de bulas de medicamentos, bula do profissional e a bula do paciente. Para este trabalho foi utilizado a bula do profissional, uma vez que essas bulas possuem informações mais detalhadas e técnicas, elas também facilitam o trabalho junto a um profissional da área farmacêutica. O site MedicinaNet disponibiliza as bulas de medicamento distribuídas pelas categorias que vão de A até Z, onde a categoria A possui as bulas dos medicamentos que tem seu nome comercial iniciado com a letra ‘a’, a categoria B possui as bulas dos medicamentos que possui o nome comercial iniciado com a letra ‘b’ e assim por diante. Inicialmente foram coletadas as primeiras 363 bulas em ordem alfabética da categoria A, as primeiras 123 bulas em ordem alfabética da categoria B e depois também em ordem alfabética, as 30 primeiras bulas das seguintes categorias. Inicialmente, o objetivo era criar um corpus contendo todas as bulas de medicamentos disponíveis no site MedcinaNet, porém não seria concluído em tempo hábil para a conclusão do trabalho, uma vez que existem mais de 4000 bulas e o processo de download e limpeza das bulas foi feito manualmente neste trabalho. Desse modo, observando trabalhos anteriores, concluiu-se que um corpus com no mínimo 700 bulas seria suficiente para realização dos testes. Por questão metodológica e melhor

organização, definiu-se obter 30 bulas de cada categoria do site e manter as 486 bulas somadas das categorias A e B que já haviam sido coletadas.

3.1.2. ORGANIZAÇÃO E LIMPEZA

Para cada bula coletada foi criado um diretório com o seu nome comercial, estas pastas contém os arquivos referentes a estas bulas. As bulas de medicamentos foram salvas em arquivos de texto (".txt") que é o formato compatível para o processamento computacional. As bulas possuem as seguintes informações: Apresentação, Informações, Indicações, Contraindicações, Uso na Gravidez, Interações Medicamentosas, Reações Adversas e Posologia. Para cada informação da bula foi gerado um arquivo de texto. Os arquivos de texto foram salvos com o nome comercial do medicamento seguido da informação, como por exemplo, o medicamento Advil deu origem aos arquivos "Advil_Apresentação.txt", "Advil_Indicacoes.txt" e assim por diante.

Na limpeza foram retiradas as tabelas, caracteres especiais e imagens. A limpeza foi feita por automatização de tarefas por meio de *Shell Script* no sistema Linux.

3.1.3. SELEÇÃO DAS INFORMAÇÕES

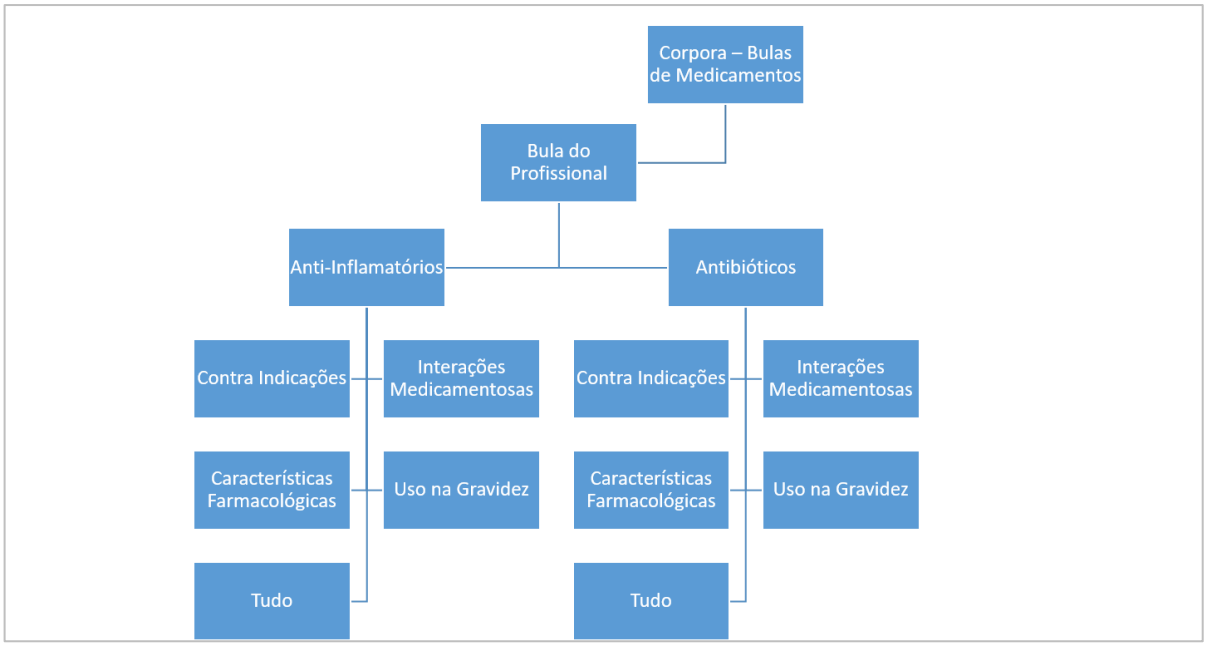
Depois de feita a coleta e manipulação das bulas de medicamento, foi feita uma seleção das informações mais interessantes para o trabalho. Essa seleção foi feita com a orientação de um profissional da área farmacêutica. Disney Oliver Sivieri-Jr, professor Adjunto de Farmacologia, Farmácia Clínica e Primeiros Socorros do Curso de Farmácia da Universidade Federal dos Vales do Jequitinhonha e Mucuri – UFVJM foi quem deu o suporte.

Foi selecionado para este trabalho, os Anti-inflamatórios e Antibióticos. A escolha dessa classe de medicamento é pelo motivo deles serem amplamente utilizados para vários tipos de tratamento e também são bastante utilizados em associação com outras drogas. Existem várias formas farmacêuticas de medicamentos, como por exemplo, comprimidos, cremes, cápsulas, drágeas, injetáveis, dentre outros. Cada forma farmacêutica apresenta características diferentes entre elas, desse modo, todos as bulas selecionadas para os testes estão na forma de comprimido, obtendo assim, testes com medicamentos com características semelhantes quanto ao tempo de absorção, velocidade de ação, via de administração, dentre outras características. Os Anti-inflamatórios selecionados foram: Aceclofenaco, Ácido Mefenamico, Advil, Belfaren,

Bioflac, Codaten, Deltaren, Diclofenaco, Ibufran e Indocid. Os Antibióticos selecionados foram: Amoxilina e Ampicilina.

Depois de selecionados os medicamentos, foi escolhido as informações mais relevantes das bulas de medicamentos para o trabalho. Ainda seguindo as orientações do profissional da área farmacêutica, foi selecionado as informações Contra Indicações, Características Farmacológicas, Interações Medicamentosas e Uso na Gravidez para o processamento. Essas são as informações que apresentam maior relevância a respeito do medicamento. Também foi criado um corpus denominado "Tudo", que é a união de todas as informações selecionadas anteriormente. A figura 8 mostra o Corpora criado.

Figura 8 - Corpora criado no trabalho



3.2. PROCESSAMENTO – CASSIOPEIA

Após a organização e limpeza dos dados (Pré-processamento), inicia-se o processamento com o modelo Cassiopeia (GUELPELI, 2012). O modelo Cassiopeia é um agrupador de texto que utiliza métodos de geração combinando Clusterização e/ou Sumarização e/ou Remoção ou não das *stopwords*. As bulas de medicamentos são textos pequenos, desse modo, não foi necessário utilizar a sumarização, sendo executado somente o método de clusterização com remoção de *stopwords*.

3.2.1. REMOÇÃO DE *STOPWORDS* E CLUSTERIZAÇÃO

As *stopwords* são palavras muito frequentes, que não tem significado relevante ao texto, como artigos e preposições e que podem ser removidas sem prejuízo ao contexto textual. Para se obter bons resultados é importante que se utilize uma boa lista de *stopwords*. Desse modo, foi utilizado uma lista de *stopwords* para o idioma português, disponibilizada no Blog do pesquisador Stanley Loh. Essa lista foi criada com base em análises estatísticas e foi utilizada em vários trabalhos do autor. A lista está no formato ".txt", portanto foi necessário adaptá-la para o formato ".sw" que é o formato compatível com o modelo Cassiopeia. Link da lista de *stopwords*: http://miningtext.blogspot.com.br/2008/11/listas-de-stopwords-stoplist-portugues.html – Acessado em 02/03/2017.

3.2.2. PROCESSAMETO COM E SEM ANTIBIÓTICOS

A estratégia utilizada nesse trabalho é processar separadamente as classes Contra Indicações, Interações Medicamentosas, Características Farmacológicas e Uso na Gravidez. Primeiramente é feito o processamento de dez anti-inflamatórios de uma das classes, depois faz o processamento desses mesmos anti-inflamatórios incluindo mais dois antibióticos da mesma classe, resultando, dessa forma, resultados de antes e depois da inclusão de antibióticos para análise. A avaliação é feita utilizando a métrica interna Coeficiente de *Silhouette*. O objetivo de introduzir antibióticos, é que esse tipo de medicamento possui características diferentes dos anti-inflamatórios, logo, as informações que se encontram em um anti-inflamatório são diferentes das informações encontradas nos antibióticos, ou seja, espera-se que o conjunto de palavras comumente encontrado em bulas dos anti-inflamatórios seja diferente do encontrado em bulas de antibióticos.

No processamento, foi utilizado o método de geração "Clusterizacão + *Stopwords*" no modelo Cassiopeia. Primeiro foram selecionados dez anti-inflamatórios de um corpus, depois, na configuração de *logger*, foi utilizado o formato "HTML" e a opção de Métricas "Coesão/Acoplamento", depois de gerado os resultados com os dez anti-inflamatórios, foram adicionados mais dois antibióticos do mesmo corpus e repetidos os testes com as mesmas configurações. Para todos os testes foi configurado a execução com 100 iterações. A tabela 1 mostra um resumo dos testes.

Tabela 1 - Resumos dos testes no modelo Cassiopeia

TESTE	TIPO	DESCRIÇÃO
Teste 1 - T1	Contra Indicações Sem Antibióticos	Processar 10 bulas de medicamentos de anti-inflamatórios da classe Contra Indicações.
	Contra Indicações Com Antibióticos	Processar as mesmas 10 bulas de medicamentos de anti-inflamatórios da classe Contra Indicações com inclusão de mais 2 bulas de medicamentos de antibióticos da classe Contra Indicações
Teste 2 - T2	Interações Medicamentosas Sem Antibióticos	Processar 10 bulas de medicamentos de anti-inflamatórios da classe Interações Medicamentosas.
	Interações Medicamentosas Com Antibióticos	Processar as mesmas 10 bulas de medicamentos de anti-inflamatórios da classe Interações Medicamentosas com inclusão de mais 2 bulas de medicamentos de antibióticos da classe Interações Medicamentosas.
Teste 3 - T3	Características Farmacológicas Sem Antibióticos	Processar 10 bulas de medicamentos de anti-inflamatórios da classe Características Farmacológicas.
	Características Farmacológicas Com Antibióticos	Processar as mesmas 10 bulas de medicamentos de anti-inflamatórios da classe Características Farmacológicas com inclusão de mais 2 bulas de medicamentos de antibióticos da classe Características Farmacológicas.
Teste 4 - T4	Uso na Gravidez Sem Antibióticos	Processar 10 bulas de medicamentos de anti-inflamatórios da classe Uso na Gravidez.
	Uso na Gravidez Com Antibióticos	Processar as mesmas bulas de medicamentos de anti-inflamatórios da classe Uso Na Gravidez com inclusão de mais 2 bulas de medicamentos de antibióticos da classe Uso na Gravidez.

Teste 5 - T5	Tudo Sem Antibióticos	Processar 10 bulas de medicamentos de anti-inflamatórios da união de todas as classes.
	Tudo Com Antibióticos	Processar as mesmas 10 bulas de medicamentos de anti-inflamatórios da união de todas as classes com inclusão de mais 2 bulas de medicamentos de antibióticos da união de todas as classes.

3.2.3. ESTATÍSTICA DO CORPORA

A estatística dos corpus foi obtida por meio do software *Get FineCount*, versão2.6.1924 de 05/12/2014.

A tabela 2 mostra a estatística no corpus Contra Indicação. Dentre os corpus, esse é o que tem o menor número de palavras, sendo um total de 928 palavras no corpus contendo somente anti-inflamatório e um total de 1015 palavras no corpus contendo anti-inflamatório e antibióticos.

Tabela 2 - Estatística do corpus Contra Indicações

Contra Indicações				
Arquivos	Caracteres	Palavras	Palavras + Números	% de Números
Aceclofenaco_ContraIndicacoes	382	57	57	0
AcidoMefenamico_ContraIndicacoes	553	88	88	0
Advil(ibuprofeno)_ContraIndicacoes	342	62	64	3,13
Belfaren_ContraIndicacoes	296	44	44	0
Bioflac_ContraIndicacoes	607	102	103	0,97
Codaten_ContraIndicacoes	1169	179	179	0
Deltaren_ContraIndicacoes	487	74	74	0
Diclofenaco_ContraIndicacoes	862	135	136	0,74
Ibufran_ContraIndicacoes	186	29	29	0
Indocid_ContraIndicacoes	1003	158	158	0

Amoxilina_ContraIndicacoes	327	50	50	0
Ampicilina_ContraIndicacoes	228	37	37	0
Total Sem Antibióticos	5887	928	932	0,43
Total Com Antibióticos	6442	1015	1019	0,39

A tabela 3 mostra a estatística no corpus Interações Medicamentosas. Dentre os *corpus*, esse é o que tem o segundo maior número de palavras, sendo um total de 3976 palavras no corpus contendo somente anti-inflamatório e um total de 4266 palavras no corpus contendo anti-inflamatório e antibióticos.

Tabela 3 - Estatística do corpus Interações Medicamentosas

Interações Medicamentosas				
Arquivos	Caracteres	Palavras	Palavras + Números	% de Números
Aceclofenaco_InteracoesMedicamentosas	1638	272	273	0,37
AcidoMefenamico_InteracoesMedicamentosas	537	86	86	0
Advil(ibuprofeno)_InteracoesMedicamentosas	1087	158	159	0,63
Belfaren_InteracoesMedicamentosas	75	14	14	0
Bioflac_InteracoesMedicamentosas	1900	283	283	0
Codaten_InteracoesMedicamentosas	5684	877	879	0,23
Deltaren_InteracoesMedicamentosas	2970	455	456	0,22
Diclofenaco_InteracoesMedicamentosas	3477	528	529	0,19
Ibufran_InteracoesMedicamentosas	3202	510	515	0,97
Indocid_InteracoesMedicamentosas	5060	793	799	0,75
Amoxilina_InteracoesMedicamentosas	1187	182	182	0
Ampicilina_InteracoesMedicamentosas	656	108	108	0
Total Sem Antibióticos	25630	3976	3993	0,43
Total Com Antibióticos	27473	4266	4283	0,40

A tabela 4 mostra a estatística no corpus Características Farmacológicas. Dentre os corpus, esse é o que tem o maior número de palavras, sendo um total de 5792 palavras no corpus contendo somente anti-inflamatório e um total de 6618 palavras no corpus contendo anti-inflamatório e antibióticos.

Tabela 4 - Estatística do corpus Características Farmacológicas

Características Farmacológicas				
Arquivos	Caracteres	Palavras	Palavras + Números	% de Números
Aceclofenaco_Caracteristicas_farmacologicas	1848	316	334	5,39
Acido_Mefenamico_Caracteristicas_Famacologicas	1368	236	251	5,98
Advil_Caracteristicas_Famacologicas	1374	230	256	10,16
Belfaren_Caracteristicas_farmacologicas	2254	382	400	4,5
Bioflac_Caracteristicas_Famacologicas	5033	881	937	5,98
Codaten_Caracteristicas_Famacologicas	14734	2477	2643	6,28
Diclofenaco_Sodico_Caracteristicas_Famacologicas	3115	522	529	1,32
Ibufran_Caracteristicas_farmacologicas	1094	183	188	2,66
Indocid_Caracteristicas_farmacologicas	1675	291	310	6,13
Proflam_Caracteristicas_farmacologicas	1655	274	289	5,19
Amoxilina_Caracteristicas_farmacologicas	4670	732	738	0,81
Ampicilina_Caracteristicas_farmacologicas	519	94	100	6
Total Sem Antibióticos	34150	5792	6137	5,62
Total Com Antibióticos	39339	6618	6975	5,12

A tabela 5 mostra a estatística no corpus Uso na Gravidez. Esse corpus Contém 1539 palavras no corpus contendo somente anti-inflamatório e um total de 1711 palavras no corpus contendo anti-inflamatório e antibióticos.

Tabela 5 - Estatística do corpus Uso na Gravidez

Uso Na Gravidez				
Arquivos	Caracteres	Palavras	Palavras + Números	% de Números
Aceclofenaco_UsoNaGravidez	728	132	132	0
AcidoMefenamico_UsoNaGravidez	1135	198	198	0
Advil(ibuprofeno)_UsoNaGravidez	604	106	106	0
Belfaren_UsoNaGravidez	747	139	140	0,71
Bioflac_UsoNaGravidez	1895	325	328	0,91
Codaten_UsoNaGravidez	1155	207	207	0

Deltaren_UsoNaGravidez	647	114	115	0,87
Diclofenaco_UsoNaGravidez	964	167	168	0,6
Ibufran_UsoNaGravidez	746	127	127	0
Indocid_UsoNaGravidez	136	24	24	0
Amoxilina_UsoNaGravidez	631	108	108	0
Ampicilina_UsoNaGravidez	376	64	64	0
Total Sem Antibióticos	8757	1539	1545	0,39
Total Com Antibióticos	9764	1711	1717	0,35

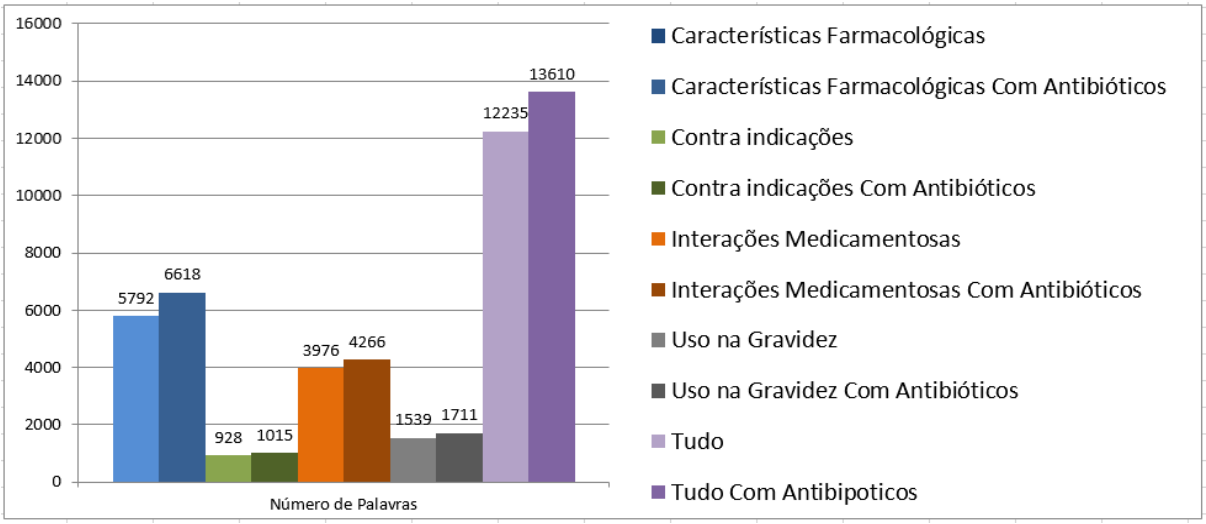
A tabela 6 mostra a estatística no corpus Tudo que é a união de todos os corpus. Esse corpus apresenta um total de 12235 palavras no corpus contendo somente anti-inflamatório e um total de 13610 palavras no corpus contendo anti-inflamatório e antibióticos.

Tabela 6 - Estatística do corpus Tudo

Tudo				
Arquivos	Caracteres	Palavras	Palavras + Números	% de Números
Aceclofenaco_Tudo	4596	777	796	2,39
AcidoMefenamico_Tudo	3593	608	623	2,41
Advil(ibuprofeno)_Tudo	3407	556	585	4,96
Belfaren_Tudoz	3371	579	598	3,18
Bioflac_Tudo	9435	1591	1651	3,63
Codaten_Tudo	22742	3740	3908	4,3
Deltaren_Tudo	5758	917	934	1,82
Diclofenaco_Tudo	8418	1352	1362	0,73
Ibufran_Tudo	5228	849	859	1,16
Indocid_Tudo	7874	1266	1291	1,94
Amoxilina_Tudo	6815	1072	1078	0,56
Ampicilina_Tudo	1779	303	309	1,94
Total Sem Antibióticos	74422	12235	12607	2,95
Total Com Antibióticos	83016	13610	13994	2,74

A figura 9 apresenta a quantidade de palavras dos corpus com e sem antibiótico.

Figura 9 - Gráfico da quantidade de palavras dos corpus



4. RESULTADOS

No capítulo 4 serão mostrados os resultados da métrica interna ou não supervisionada (Coeficiente de *Silhouette*) explicada na seção 2.4.5.3. relacionando com a quantidade de palavra dos *corpus*. Os valores de Coeficiente de *Silhouette* situam-se na faixa de 0 a 1, sendo que os melhores agrupamentos obtêm valores próximos a 1.

4.1. NÚMERO DE PALAVRAS

A tabela 7 demonstra a quantidade de palavras dos corpus antes e depois da inclusão de antibióticos.

"Características Farmacológicas" apresenta a quantidade de 5792 palavras e "Características Farmacológicas com Antibióticos" apresenta 6618 palavras, com um aumento de 0,14% após a inclusão de antibióticos. "Contra Indicações" contém 928 palavras, já "Contra Indicações Com Antibióticos" possui 1015 palavras, com um aumento de 0,09% após a inclusão de antibióticos. "Interações Medicamentosas" possui 3976 palavras e "Interações Medicamentosas Com Antibióticos" possui 4266 palavras, com um aumento de 0,07% após a inclusão de antibióticos. "Uso na Gravidez" têm 1539 palavras e "Uso na Gravidez Com Antibióticos" a quantidade de palavras é de 1711, com um aumento de 0,11% após a inclusão de antibióticos. "Tudo" possui 12235 palavras e "Tudo Com Antibióticos" possui 13610 palavras, com um aumento de 0,11% após a inclusão de antibióticos.

Tabela 7 - Quantidade de palavras dos corpus

CORPUS	QUANTIDADE DE PALAVAS	VARIAÇÃO (%)
Características Farmacológicas	5792	0,14
Características Farmacológicas Com Antibióticos	6618	
Contra Indicações	928	0,09
Contra Indicações Com Antibióticos	1015	
Interações Medicamentosas	3976	0,07

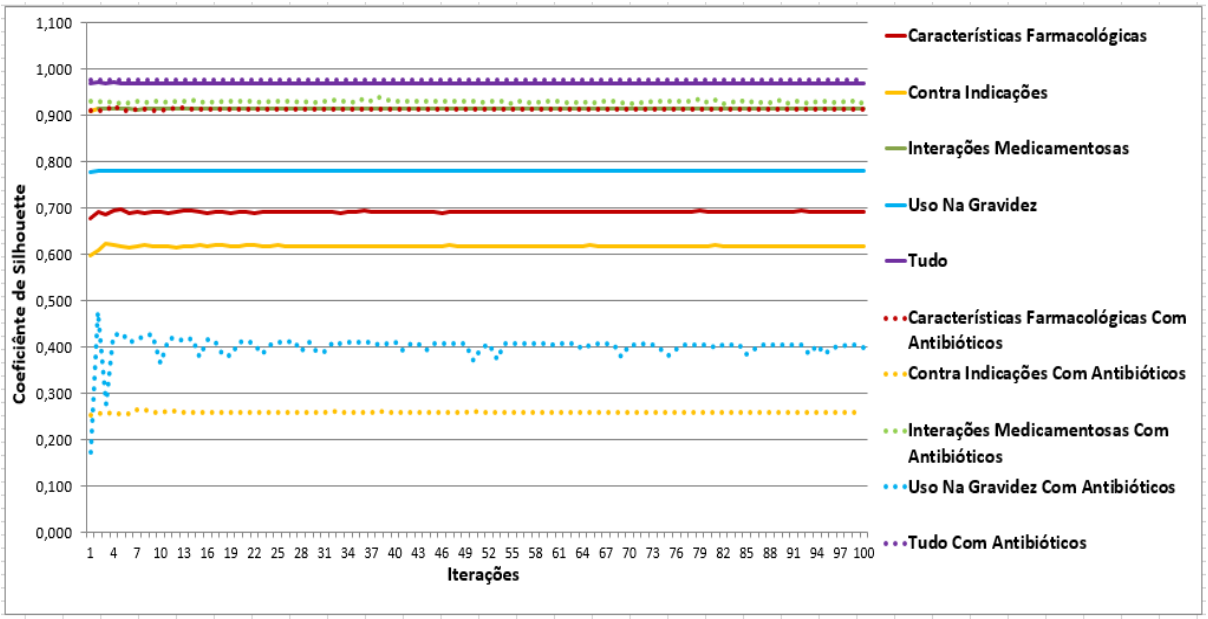
Interações Medicamentosas Com Antibióticos	4266	
Uso na Gravidez	1539	0,11
Uso na Gravidez Com Antibióticos	1711	
Tudo	12235	0,11
Tudo Com Antibióticos	13610	

4.2. COEFICIENTE DE SILHOUETTE

A figura 10 mostra o gráfico com os valores das médias acumuladas de Coeficiente de Silhouette dos corpus com e sem inclusão de antibióticos. As linhas pontilhadas representam os *corpus* que contêm a inclusão de antibióticos, as outras linhas correspondem aos *corpus* que não contêm a inclusão de antibióticos. Os valores do eixo x demonstram as iterações feitas no processamento no Modelo Cassiopeia. O eixo y apresenta os valores das médias acumuladas da métrica interna.

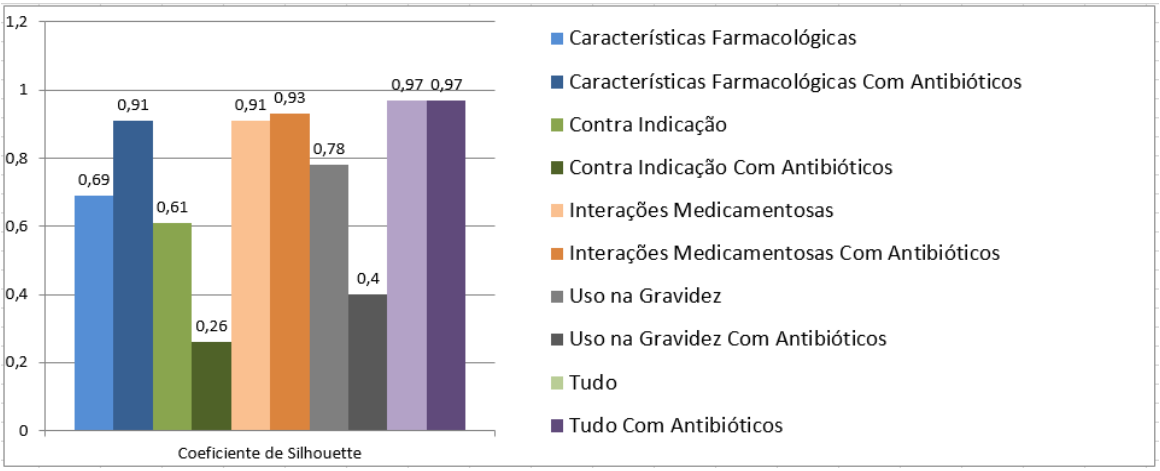
"Características Farmacológicas" apresenta valores médio de 0,69, já "Características Farmacológicas com Antibióticos" apresenta valores médio de 0,91. "Contra Indicações" tem valores médio de 0,61 e "Contra Indicações Com Antibióticos" os valores médio são 0,26. "Interações Medicamentosas" possui valores médio de 0,91 e "Interações Medicamentosas Com Antibióticos" possuem os valores médio de 0,93. "Uso na Gravidez" têm valores médio de 0,78 e "Uso na Gravidez Com Antibióticos" os valores médio são 0,40. "Tudo" e "Tudo Com Antibióticos" apresentam os mesmos valores médio de 0,97.

Figura 10 - Gráfico dos valores médios acumulados de Coeficiente de Silhouette



A figura 11 apresenta os valores médio do Coeficiente de *Silhouette* com e sem antibióticos em cada *corpus* obtidos das 100 execuções. Pelo gráfico percebe-se que os *corpus* Características Farmacológicas e Interações Medicamentosas tiveram valores maiores do Coeficiente de *Silhouette* após a inclusão de antibióticos. Os *corpus* Contra Indicação e Uso na Gravidez tiveram valores menores, enquanto que o corpus Tudo teve valores iguais antes e depois da inclusão de antibióticos.

Figura 11 - Gráfico dos valores médio do Coeficiente de *Silhouete*



4.3. RELAÇÂO ENTRE QUANTIDADE DE PALAVRAS E COEFICIENTE DE SILHOUETTE

A tabela 8 apresenta a quantidade de palavras e os valores do Coeficiente de *Silhouette* em cada *corpus*. Observa-se nos resultados obtidos que quando há um maior número de palavras nos textos processados, os valores do Coeficiente de *Silhouette* são melhores, ou seja, mais próximos do valor 1. Conforme na seção 2.4.5.3, o Coeficiente de *Silhouete* baseia-se na ideia de quanto um texto é similar aos demais do seu grupo, e de quanto esse mesmo texto é distante dos de outros grupos. Desse modo, observa-se que quanto maior são os textos, mais palavras similares entre eles são encontradas, melhorando assim, os valores do Coeficiente de *Silhouete*. Por outro lado, quanto menor são os textos, menos palavras em comum são identificadas e consequentemente são gerados grupos com palavras menos similares, assim obtendo valores do Coeficiente de *Silhouette* menores.

"Características Farmacológicas" e "Interações Medicamentosas", desconsiderando "Tudo" que é a união de todos os textos, são os que contêm o maior número de palavras, um total de 5792 e 3976 respectivamente, esses corpus tiveram aumento no Coeficiente de *Silhouette* após a inclusão de antibióticos. Como os textos desses dois tipos de informação da bula possuem maior número de palavras, o grau de similaridade entre as palavras ou a quantidade de palavras em comum, tendem a aumentar quando se inclui mais textos no processamento, assim há uma melhora dos valores da métrica interna.

Os *corpus* "Contra Indicações" e "Uso na Gravidez" que contém menor número de palavras em relação aos demais, com um total de 928 e 1539 respectivamente, tiveram uma diminuição do Coeficiente de *Silhouette* após a inclusão de antibióticos. Como os textos desses dois tipos de informação da bula possuem poucas palavras, obtém-se grupos com palavras pouco similares em relação a grupos com textos maiores.

"Tudo", por ser a união de todos os textos, possui o maior número de palavras, um total de 12235 palavras. A média acumulada do Coeficiente de *Silhouette* manteve a mesma com a inclusão de antibióticos. Os resultados da métrica interna foi o melhor obtido nos testes, já que os textos processados contém um maior número de palavras em relação aos demais, e dessa forma o resultado são grupos que contém grande similaridade entre eles.

Tabela 8 - Relação quantidade de palavras por Coeficiente de *Silhouette*

CORPUS	COEFICIENTE DE SILHOUETTE	QUANTIDADE DE PALAVAS
Contra Indicações	0,61	928
Contra Indicações Com Antibióticos	0,26	1015
Características Farmacológicas	0,69	5792
Características Farmacológicas Com Antibióticos	0,91	6618
Interações Medicamentosas	0,91	3976
Interações Medicamentosas Com Antibióticos	0,93	4266
Uso na Gravidez	0,78	1539
Uso na Gravidez Com Antibióticos	0,4	1711
Tudo	0,97	12235
Tudo Com Antibióticos	0,97	13610

5. DISCUSSÃO DOS RESULTADOS

Este capítulo tem como objetivo discutir os resultados obtidos nos testes feitos no modelo Cassiopeia e apresentados no capítulo 4.

5.1. VISÃO GERAL

Conforme a seção 3.2.2, foram executados um total de cinco testes, T1, T2, T3, T4 e T5. Cada teste é composto por duas execuções, a primeira corresponde o processamento de somente bulas de medicamentos de anti-inflamatórios e a segunda é o processamento das bulas de medicamentos de anti-inflamatórios mais a inclusão de antibióticos. Cada teste gera dois valores de Coeficiente de *Silhouette*, ou seja, um valor antes e um valor depois da inclusão de bulas de antibióticos.

No geral, foram obtidos bons resultados, uma vez que, dentre os 10 valores de Coeficiente de *Silhouete* que foram obtidos como resultado, apenas dois ficaram abaixo de 0,5. No teste T1 foi obtido o valor de 0,26 e no teste T4 foi obtido o valor de 0,4. Os demais resultados tiveram 0,61 para o menor valor e 0,97 para o maior.

Observando os resultados, verificou-se que o valores de Coeficiente de *Silhouette*, geralmente são melhores quando o processamento é feito com textos que contêm maior número de palavras. O melhor resultado obtido foi no teste T5, onde o resultado foi 0,97 com um total de 13610 palavras, já o pior resultado foi no teste T1, com o resultado de 0,26 e um total de 1015 palavras, ou seja, o teste T1 foi feito com 7,45% do total de palavras do teste T5.

5.2. ANÁLISE DOS TESTES

Conforme já observado, geralmente, os processamentos com textos que contem maior número de palavras geram melhores resultados, o que não quer dizer que seja sempre verdade. Além da quantidade de palavras, é preciso levar em consideração a qualidade das palavras que compõem os textos, ou seja, o quanto essas palavras influenciam no grau de similaridade dos textos no processo de agrupamento.

Nos testes T1 e T4, observa-se que nem sempre os resultados melhoram quando se aumenta o número de palavras. No Teste T1, foi obtido o valor de 0,61 como resultado do Coeficiente de *Silhouette* antes da inclusão de antibióticos e o resultado de 0,26 após a inclusão.

No Teste T4, foi obtido o valor de 0,78 antes da inclusão de antibióticos e o resultado de 0,40 após a inclusão. Esses resultados mostraram que após a inclusão de antibióticos e consequentemente um aumento do número de palavras, não houve melhora dos resultados, mas sim uma piora, concluindo que as palavras que compunham os textos referente as bulas dos antibióticos tinham pouca qualidade.

Já nos testes T2 e T3, houve melhores resultados após a inclusão de antibióticos. No Teste T2, foi obtido o valor de 0,91 como resultado do Coeficiente de *Silhouette* antes da inclusão de antibióticos e o resultado de 0,93 após a inclusão. No Teste T3, foi obtido o valor de 0,69 antes da inclusão de antibióticos e o resultado de 0,91 após a inclusão. Estes resultados mostram que as palavras referentes aos antibióticos proporcionaram um aumento do grau de similaridade entre os textos.

O teste T5 não houve diferença de valores do Coeficiente de *Silhouette* de antes e depois da inclusão de antibióticos, foi o obtido o valor 0,97 em ambos os casos. Com estes resultados, verifica-se que houve um aumento de palavras após a inclusão de antibióticos, porém essas palavras não foram fortes o suficiente para aumentar o grau de similaridade entre os grupos gerados.

6. CONCLUSÃO

O objetivo deste trabalho foi analisar os agrupamentos textuais gerados na fase de processamento do modelo Cassiopeia utilizando a métrica interna Coeficiente de *Silhouette*.

Os resultados apresentados pelo Modelo Cassiopeia, no geral foram satisfatórios, uma vez que dentre os 10 resultados obtidos no processamento, apenas dois foram ruins. Foram analisados os agrupamentos obtidos, relacionando o Coeficiente de *Silhouette* com a quantidade de palavras dos *corpus*. Os melhores resultados foram obtidos nos testes em que a quantidade de palavras nos textos eram maiores, porém não só a quantidade de palavras proporciona bons resultados, mas também a qualidade delas.

Os testes T2 e T4 tiveram melhores resultados em relação aos testes T1 e T3 que utilizaram corpus que continham poucas palavras. O melhor resultado foi obtido no teste T5 que utilizou um *corpus* que é a união te todos os outros utilizados nos testes T1, T2, T3 e T4 e que consequentemente contêm maior quantidade de palavra em relação aos demais *corpus*.

Conclui-se que modelo Cassiopeia atende ao propósito como agrupador de textos, sendo uma boa opção para processos de descoberta de conhecimento em textos.

Na seção seguinte, serão apresentadas as dificuldades e limitações deste trabalho. Na última seção, serão mostradas algumas possibilidades de trabalhos futuros que poderão aprofundar ainda mais o tema e ampliar o estudo aqui apresentado.

6.1. DIFICULDADES E LIMITAÇÕES

Neste trabalho foram coletadas 758 bulas de medicamentos para a criação do corpora. Um número maior de bulas poderia ser obtido para se ter um conjunto de bulas mais completo, porém não foi possível, uma vez que o processo de download e limpeza das bulas utilizado foi feito manualmente, o que impossibilitaria a conclusão do trabalho em tempo hábil para apresentação.

No pós processamento, a análise dos resultados, além da métrica interna utilizada neste trabalho, poderia também ser usadas as métricas externas. Para isso, é necessário a opinião de um profissional da área farmacêutica. Porém por limitações de prazo, não foi possível concluir os trabalhos junto ao profissional.

6.2. TRABALHOS FUTUROS

Uma possibilidade de trabalho futuro é a criação de um corpus maior, por meio da utilização de ferramentas que automatizam o processo de obtenção dos textos na Internet. É possível também o desenvolvimento dessas ferramentas.

Outro trabalho futuro que poderá ser realizado pelo modelo Cassiopeia, é o de Descoberta de Conhecimento em Bulas de Medicamentos, utilizando as métricas internas e externas, com o apoio de um profissional da área farmacêutica.

REFERÊNCIAS

AALAM, P.; SIDDIQUI, T. **Comparative Study of Data Mining Tools uused for Clustering.** 2016 Internacional Conference on Computing for Stainable Global Development (INDIACom). 2016.

AGNIHOTRI, D.; VERMA, K.; TRIPATHI, P. **Pattern and cluster mining on text data**. 2014 4th International Conference on Communication Systems and Network Technologies, CSNT 2014. Disponível em: <http://doi.org/10.1109/CSNT.2014.92> Acesso em: 08 nov. 2016.

AKILAN, A. **Text Mining: Challenges and Future Directions**. IEEE Sponsored Second International Conference On Eletronics And Communication Systems (Icecs). 2015.

CALVILLO, E. A.; PADILLA, A.; MUNOZ, J.; PONCE, J.; FERNANDEZ, J. T. **Searching research papers using clustering and text mining**. 23rd International Conference on Electronics, Communications and Computing, CONIELECOMP 2013. Disponível em: <http://doi.org/10.1109/CONIELECOMP.2013.6525763> Acesso em: 08 nov. 2016

CHERNYSHOVA, G.; SMORODIN, G.; OVCHINNIKOV, A. **Technique of cluster validity for text mining**. 2016 6th Internacional Conference – Cloud System and Big Data Engineering (Confluence). Disponível em: <http://doi.org/10.1109/CONFLUENCE.2016.7508139> Acesso em: 08 nov. 2016.

CASTRO, F. DOS S.; MATTOS, M. C.; SIMÕES, P. W. T. DE A. **Mineração de textos na saúde por meio da utilização da ferramenta eurekha.** 2012**.**

DELGADO, C. C. N.; DIAS, H. D.; GUELPELI, M. V. C; **Utilização de sumários humanos no modelo Cassiopeia**. Computer on The Beach 2013. 2013.

GALHO, T. S.; MORAES, S. M. W. **Categorização Automática de Documentos de Texto Utilizando Lógica Difusa**. I Workshop de Computação Da Região Sul, 2004. Acessado em: <http://inf.unisul.br/~ines/workcomp/cd/index.html> Acesso em: 01 out. 2016.

GUELPELI, M.V.C. **Cassiopeia: Um modelo de agrupamento de textos baseado em sumarização**. – Tese (doutorado) – Universidade Federal Fluminense. Programa de Pós-graduação em Computação, Niteroi, BR – RJ, Brasil, 2012.

LOH, S.; AMARAL, L. A., WIVES, L. K.; OLIVEIRA, J. P. M. **Descoberta de Conhecimento em Textos Através da Análise de Sequencias Temporais**. Workshop em Algoritmos e Aplicações de Mineração de Dados, WAAMD, II; SBBD, 2006, Florianópolis, ISBN 85-7669-088-8. Florianópolis: Sociedade Brasileira de Computação, 2006. p. 49-56.

LOH, S. **Abordagem Baseada em Conceitos para Descoberta de Conhecimento em Textos** Universidade Federal do Rio Grande do Sul-Instituto de Informática-Curso de Pós-graduação em Ciência da Computação.Tese de Doutorado- UFRGS, 2001.

MARTINS, E. S.; RIBEIRO, M.; LISBOA-FILHO, J.; REINALDO, F.; FREDDO, A.; REIS, L. P. **Uso de Clusterização em Dados Espaciais para Extração de Conhecimentos**. 2016

11th Iberian Conference on Information Systems and Technologies (CISTI). Disponível em: <http://doi.org/10.1109/CISTI.2016.7521626> Acesso em: 10 nov. 2016.

METZ, J.; MONARD, M. C. **Clustering hierarquico: uma metodologia para auxiliar na interpretação dos clusters**. XXV Congresso da Sociedade Brasileira de Computação. 2005.

MORAIS, E. A. M.; AMBRÓSIO, A. P. L. **Mineração de Textos**. Revista de Ciências Exatas. Vol. 3, N°.3, Ano 2008. Acessado em: <http://doi.org/RT-INF_005-07> Acesso em: 01 out. 2016.

NOGUEIRA, T. C. **Mineração de texto em bulas de medicamentos.** XXIV Congresso Brasileiro De Engenharia Biomédica. 2014.

NUNES, M.; BORGES, T. **Técnica para Analisar a Evolução do Perfil do Usuário com base nas suas Publicações**. II Escola Regional de Banco de Dados. Acessado em: <http://paginas.ucpel.tche.br/~loh/pdfs/erbd-2007-marcos.pdf> Acesso em: 01 out. 2016.

QI, C; JIANFENG, L. **A Text Mining Model Based on Improved Density Clustering Algorithm**. Major Program of National Natural Science Foundation of China. 2013.

RAMOS, H. D. S. C.; BRÄSCHER, M. **Aplicação da descoberta de conhecimento em textos para apoio à construção de indicadores infométricos para a área de C&T**. Ci. Inf., Brasília, v. 38, n. 2, p. 56-68, maio/ago. 2009. Disponível em: <http://doi.org/10.1590/S0100-19652009000200005> Acesso em: 01 out. 2016.

SCHRÖDER, S.; TUMMEL, C.; ISENHARDT, I.; JESCHKE, S.; RICHERT, A. **Benchmarking of scientific research clusters by use of text mining algorithms on textual artefacts**. Proceedings - 2016 International Conference on Information Systems Engineering, ICISE 2016, 22–28. Disponível em: <http://doi.org/10.1109/ICISE.2016.9> Acesso em: 08 nov. 2016.

SERAPIÃO, P. R. B.; SUZUKI, K. M. F.; MARQUES, P. M. DE A. **Uso de mineração de texto como ferramenta de avaliação da qualidade informacional em laudos eletrônicos de mamografia**. Radiologia Brasileira, 43(2), 103–107. Disponível em: <http://doi.org/10.1590/S0100-39842010000200010> Acesso em: 08 nov. 2016.

TRAPPEY, C.; WU, H.-Y.; LIU, K.-L.; LIN, F.-T. **Knowledge Discovery of Service Satisfaction Based on Text Analysis of Critical Incident Dialogues and Clustering Methods**. 2013 IEEE 10th International Conference on E-Business Engineering, Disponível em <http://doi.org/10.1109/ICEBE.2013.40> Acesso em: 08 nov. 2016

WANG, Y.; XU, W.; JIANG, H. **Using Text Mining and Clustering to Group Research Proposals for Research Project Selection**. 2015 48th Hawaii International Conference on System Sciences. Disponívem em: <http://doi.org/10.1109/HICSS.2015.153> Acesso em: 08 nov. 2016

WERNECK, V. R. **Sobre o processo de construção do conhecimento: O papel do ensino e da pesquisa**. Ensaio Aval Pol Públ Educ, 14(51), 173–96. Disponível em: <http://doi.org/10.1590/S0104-40362006000200003> Acesso em: 08 nov. 2016.

WIVES, L. K. **Utilizando conceitos como descritores de textos para o processo de identificação de conglomerados (clustering) de documentos**. Universidade Federal do Rio Grande do Sul-Instituto de Informática-Curso de Pós-graduação em Ciência da Computação.Tese de Doutorado-UFRGS, 2004.

WIVES, L. K. **Um estudo sobre agrupamento de documentos textuais em processamento de informações não estruturadas usando técnicas de" clustering."**. Universidade Federal do Rio Grande do Sul-Instituto de Informática-Curso de Pós-graduação em Ciência da Computação. Dissertação submetida à avaliação como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação-UFRGS, 1999.

YANG, C. L.; BENJAMASUTIN, N.; CHEN-BURGER, Y. H. **Mining hidden concepts: Using short text clustering and wikipedia knowledge**. 2014 IEEE 28th International Conference on Advanced Information Networking and Applications Workshops, IEEE WAINA 2014. Disponível em: <http://doi.org/10.1109/WAINA.2014.109> Acesso em: 08 nov. 2016.

ZHANG, N.; WANG, J.; HE, K.; LI, Z. **An Approach of Service Discovery Based on Service Goal Clustering**. 2016 IEEE International Conference on Services Computing (SCC). Disponível em: <http://doi.org/10.1109/SCC.2016.22> Acesso em: 08 nov. 2016.