Universidade Federal dos Vales do Jequitinhonha e Mucuri

Thaís Caldoncelli Nogueira

MINERAÇÃO DE TEXTO EM BULAS DE MEDICAMENTOS

Diamantina

2014

Thaís Caldoncelli Nogueira

MINERAÇÃO DE TEXTO 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. Sob orientação do Prof. Dr. Marcus Vinícius Carvalho Guelpeli.

Diamantina

2014

Monografia de projeto final de graduação sob o título "Mineração de Texto em Bulas de Medicamento", defendida por Thaís Caldoncelli Nogueira e aprovada em 22 de julho de 2014, em Diamantina, Minas Gerais.

Banca Examinadora:

Prof. Dr. Marcus Vinícius Guelpeli
Orientador

Profª. MSc. Cláudia Beatriz Berti

Profª. Drª. Geruza Tomé Sabino

Prof. MSc. Eduardo Pelli

Dedico este trabalho a todos que contribuíram direta ou indiretamente em minha formação acadêmica.

AGRADECIMENTOS

Agradeço a todos que contribuíram no decorrer desta jornada, especialmente:

À Deus, a quem devo o dom da vida.

À minha família que sempre me apoiou com palavras de carinho nos momentos de dificuldade.

Ao orientador Prof. Marcus Vinícius Carvalho Guelpeli que teve papel fundamental na elaboração deste trabalho.

Aos meus colegas de sala pelo companheirismo e disponibilidade para me auxiliar em vários momentos.

Aos meus amigos distantes, por terem se mantido presente mesmo com minha ausência.

RESUMO

Atualmente, com o advento da Internet, o volume de informações textuais vem crescendo exponencialmente. Tais informações são encontradas em maiores quantidades em diversas fontes. Através das informações textuais, conhecidas também como não estruturadas, é possível extrair conhecimento útil e implícito que devido ao grande volume são impossíveis de serem processadas por um ser humano. As técnicas para Mineração de Texto (*Text Mining*- TM) estão sendo constantemente desenvolvidas e aprimoradas para tratar dados não estruturados. O TM tem como principal finalidade extrair padrões em dados não estruturados, desta forma ela executa vários processos em várias etapas que transforma ou organiza uma quantidade de documentos em uma estrutura sistematizada, normalmente com algum critério de similaridade. Isso possibilita, posteriormente, a sua utilização de forma eficiente e inteligente. A partir disso, propõe-se no presente trabalho, a utilização de um algoritmo TM sobre uma base de Bulas de Medicamentos (BM) coletados em sites específicos da medicina. Para manipulação dessas bases textuais, o Modelo Cassiopeia foi utilizado empregando seu algoritmo TM de agrupamento textual que tem como principal finalidade gerar agrupamentos, ou seja, *clusters* (grupos) de documentos textuais que apresentam algum tipo de similaridade. O Modelo Cassiopeia possui três fases: pré-processamento, processamento e pós-processamento das bases textuais. Com as BM coletadas, elas passam primeiramente pela fase de limpeza de dados no pré-processamento, logo após, a utilização do algoritmo no processamento e por fim, as análises dos resultados no pós-processamento. Os resultados obtidos neste trabalho mostram valores significativos quanto à similaridade dos documentos dentro de um *cluster* e entre os *clusters*, medidos através das medidas de agrupamento textual, proposto pelo Modelo Cassiopeia.

Palavras-chave: Descoberta de Conhecimento em Textos, Agrupamento Textual, Bulas de Medicamentos, Modelo Cassiopeia.

ABSTRACT

Currently, the volume of textual information has grown exponentially with the advent of the Internet, where the information is found in larger quantities from various sources. Through the textual information, also known as unstructured, it is possible to extract useful and implicit knowledge and that due to the volume are impossible to be processed by a human being. Techniques for Text Mining (TM) are constantly being developed and improved to handle unstructured data. The TM has as main purpose to extract patterns in unstructured, so it runs several processes in several stages, which transforms and organizes a number of documents in a systematic structure, usually with some criterion of similarity data. This makes it possible subsequently to use efficiently and intelligently. From this, it is proposed in this paper, the use of a TM algorithm on Medicine Package (BM) collected at specific sites of medicine. For handling these textual bases, Cassiopeia model was used, employing your TM algorithm textual grouping which has as main purpose to generate clusters of textual documents that have some similarity type. Cassiopeia model has three stages: pre-processing, processing and post-processing of the textual bases. With BM collected, they first pass through the stage of data cleaning in preprocessing, soon after, the use of the algorithm in processing and final analysis of results post-processing. The results of this study show significant values as the similarity of documents within a cluster and between clusters, as measured by the clustering of textual measures proposed by Model Cassiopeia.

Keywords: Knowledge Discovery in Texts, Clustering Textual, Medicine Package, Model Cassiopeia.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AI	***A**rtificial **I**ntelligence*
BM	**B**ulas de **M**edicamentos
C_1	***C**orpus* **1** – “Como este medicamento funciona”
C_2	***C**orpus* **2** – “O que devo saber antes de usar”
C_3	***C**orpus* **3** – “C_1+C_2”
DM	***D**ata **M**ining*
KDD	***K**nowledge **D**iscovery in **D**atabases*
KDT	***K**nowledge **D**iscovery in **T**exts*
SGBD	**S**istema de **G**erenciamento de **B**anco de **D**ados
TM	***T**ext **M**ining*

SUMÁRIO

1. INTRODUÇÃO

Com o advento das novas Tecnologias de Informação, o grande volume de informação não estruturada vem crescendo exponencialmente nos dias atuais, pois a Internet é um repositório de informação e consiste na principal fonte onde as pessoas pesquisam e fornecem diversos tipos de assuntos e interesses. Esse crescimento está acontecendo de forma desordenada e grande parte dessas informações está no formato textual. Essa estrutura quase anárquica trouxe consigo um grande problema de organização (de informações), que surge mediante dificuldade do ser humano em armazenar grandes quantidades de informações.

Segundo Silva (2012), parte considerável dessa informação encontra-se na forma de textos nos mais diversos formatos. Desde a década de noventa, estudos já apontavam que 80% da informação encontrava-se na forma textual. A cada ano são produzidos aproximadamente 968 mil livros, 80 mil revistas, 40 mil periódicos, bilhões de documentos. Redes sociais, sites, e blogs, dependendo do foco de análise, devem ser considerados como importantes fontes de informação textual devido principalmente a sua dinamicidade.

Visando recuperar informações relevantes e solucionar o problema do crescimento desordenado de dados, as técnicas de *Data Mining* (DM) surgiram para utilização em dados estruturados, e as técnicas para TM estão sendo constantemente desenvolvidas e aprimoradas para tratar dados não estruturados, representando uma área interessante e pouco explorada nas últimas décadas (CASTRO, SIMÕES, MATTOS, 2009). Dados estruturados possuem relações, atributos e esquemas, sendo esses dados previamente preparados para uma melhor recuperação das informações. Já os dados não estruturados, não possuem uma estrutura definida, são normalmente caracterizados por documentos, textos, imagens e vídeos. A grande maioria dos dados atuais da Web e das empresas segue esse formato.

O TM tem como principal finalidade, extrair padrões em dados não estruturados, desta forma ela executa vários processos em várias etapas que transforma ou organiza uma quantidade de documentos em uma estrutura

sistematizada, normalmente com algum critério de similaridade. Isso possibilita, posteriormente, a sua utilização de forma eficiente e inteligente. Portando, o TM surgiu do DM e consiste em um conjunto de processos e técnicas que visam descobrir nos textos conhecimento relevante, visto que há muito potencial implícito nesse tipo de informação não estruturada.

Ao lidar com muita informação ocorre perda de tempo que poderia ser mais bem empregado pensando, refletindo ou raciocinando (LOH, 2001). Dentre essas informações textuais não estruturadas, estão documentos como relatórios, atas de reuniões, históricos pessoais, e-mails, planilhas, documentações e artigos. Entretanto, até pouco tempo atrás, a organização dessas informações textuais não eram tidas como importantes, até que estudos provaram o contrário, mostrando que estas informações não estruturadas, a partir de uma organização eficaz, poderiam ser utilizadas de forma inteligente em vários campos, possibilitando até mesmo, tirar alguma vantagem competitiva ou dar suporte a tomada de decisões.

As bases textuais que foram objeto de estudo neste trabalho, são as BM. Foi observada também com o advento da Internet, há grande quantidade dessas encontradas em diversos sites específicos da área de saúde. As BM oferecem grande número de informações e são bastante importantes tanto para pacientes, quanto para profissionais de saúde. Sendo assim, técnicas de TM podem auxiliar a organização dessas bulas em agrupamentos com algum critério de similaridade, para uma possível utilização dos processos de Descoberta de Conhecimento em Textos (*Knowledge Discovery in Texts* - KDT).

Uma *corpora*[1] de 707 BM foram selecionadas pelo site Bulário.net (http://bulario.net/alfa/), pré-processadas, processadas e analisadas através do pós-processamento. Existe no Bulário.net duas categorias, uma de bulas destinadas aos pacientes e outra de bulas de profissionais da saúde. Dentro dessas duas classes estão dispostas as bulas. Dentro das bulas existem muitas informações do tipo: “O que devo saber antes de usar”, “Como este medicamento funciona”, “Indicações”, “Contra Indicações” e “Como usar este medicamento”. Neste trabalho, a classe escolhida corresponde às bulas destinadas aos pacientes. Cada *corpus*[2] dentro

[1] Conjunto de informações sobre um determinado assunto, o qual servirá como base para algum estudo.
[2] Coletânea ou conjunto de documentos sobre determinado tema.

dessa categoria é composto de um tipo, que foi separado em arquivos txt. Para este trabalho foi escolhido o tipo “Como este medicamento funciona” que será denominado por questões metodológicas de C_1 e “O que devo saber antes de usar” denominado de C_2. Para uma análise mais profunda dos resultados, os *corpus* C_1 e C_2 foram processados juntos, criando um novo *corpus* denominado de C_3.

O processamento é feito por TM através do Modelo Cassiopeia (Um Modelo de Agrupamento de Textos Baseado em Sumarização).

O Modelo Cassiopeia, utiliza um algoritmo de TM, que tem como principal finalidade gerar agrupamentos, ou seja, *clusters* (grupos) de documentos textuais que apresentam algum tipo similaridade.

O presente trabalho tem como objetivo analisar os agrupamentos textuais gerados no processamento do Modelo Cassiopeia, através das métricas internas ou não supervisionadas.

1.1. MOTIVAÇÃO

No exame da literatura levantada, há diversos trabalhos correlatos na área de TM com utilização de técnicas de Agrupamentos. Todos esses trabalhos têm como principal finalidade a descoberta de conhecimento útil através de grandes bases de dados textuais existentes na Internet e em outras fontes. Eles apresentaram resultados interessantes e eficazes fazendo com que a área de TM tornasse um objeto de estudo bastante interessante e não mais desconhecida.

Foi observada também com o advento da Internet, a grande quantidade de BM encontradas em diversos sites específicos da área de saúde e que apresentam informações interessantes para a utilização de TM. Não se encontrou, na bibliografia pesquisada, referência ao uso de BM em processos de TM.

Portanto, a motivação deste trabalho foi avaliar o Modelo Cassiopeia sobre as BM, para que os resultados obtidos possam ser utilizados em trabalhos futuros, por profissionais da área da saúde, através de técnicas de KDT.

1.2. PROBLEMA

Será que o Modelo Cassiopeia comporta-se adequadamente usando as métricas internas (coesão, acoplamento e coeficiente de silhouette) para análise dos *clusters* no domínio farmacêutico usando BM?

1.3. HIPÓTESE

O Modelo Cassipoeia apresentará resultados satisfatórios das métricas internas (coesão, acoplamento e coeficiente silhouette) sobre os *clusters* no domínio farmacêutico usando BM.

1.4. CONTRIBUIÇÃO

Com o uso do Modelo Cassiopeia, pode-se ter como contribuição para futuros trabalhos a partir dos agrupamentos gerados, descoberta de novos conhecimentos implícitos em BM.

1.5. METODOLOGIA DE PESQUISA

A metodologia de pesquisa adotada para a realização deste trabalho compreende primeiramente em leituras bibliográficas, visando o desenvolvimento dos assuntos foco do trabalho, sendo esses, TM e Agrupamento Textual.

Será apresentado o Modelo Cassiopeia, constituído de três etapas: pré-processamento, processamento e pós-processamento. Essas etapas serão detalhadas no Capítulo 2.

Para o processamento no Modelo Cassiopeia, serão utilizados *corpus* no domínio farmacêutico com BM. As etapas de seleção da *corpora*, estatísticas da *corpora*, pré-processamento, processamento e pós-processamento sobre essas bulas, serão apresentadas detalhadamente no Capítulo 3.

Para análise dos resultados do processamento, serão utilizadas as métricas internas para análise de agrupamento textual, Coesão, Acoplamento e Coeficiente de Silhouette. Os resultados e discussões serão apresentados nos Capítulos 4 e 5, respectivamente.

1.6. ESTRUTURA PROPOSTA

Capítulo 2: Fundamentação Teórica

Neste capítulo serão descritos os principais conceitos que fundamentam este trabalho. Serão apresentados conceitos sobre Descoberta de Conhecimento, Descoberta de Conhecimento em Bases de Dados (*Knowledge Discovery in Databases* - KDD), KDT, agrupamento, agrupamento textual, Modelo Cassiopeia e por fim as medidas de avaliação dos agrupamentos textuais.

Capítulo 3: Metodologia

Neste capítulo, serão apresentados como a *corpora* de BM foi selecionada, estatísticas da *corpora,* pré-processamento, processamento e pós-processamento das BM a partir das etapas do Modelo Cassiopeia.

Capítulo 4: Resultados

O capítulo mostrará o relatório dos resultados obtidos no experimento através de gráficos.

Capítulo 5: Discussão dos Resultados

O capítulo discutirá a análise crítica dos resultados obtidos no experimento.

Capítulo 6: Conclusões

Neste capítulo serão discutidas as limitações, as contribuições e os trabalhos futuros.

2. FUNDAMENTAÇÃO TEÓRICA

Neste capítulo serão descritos os principais conceitos que fundamentam este trabalho. Inicialmente será apresentado o surgimento da técnica de Descoberta de Conhecimento. A partir da introdução da história da Descoberta de Conhecimento, será apresentada a evolução de KDD para KDT. Em agrupamento, será descrito o seu conceito, o que é um agrupamento de informações textuais e suas fases. Uma visão geral do Modelo Cassiopeia, pois este que contém o algoritmo que possibilita o agrupamento de textos e será usado na metodologia como ferramenta para obter os agrupamentos. Medidas para avaliação dos agrupamentos textuais, dando uma visão geral nas métricas externas, composta das medidas *Recall*, *Precision* e *F-Measure* e aprofundando nas métricas internas, composta das medidas Coesão, Acoplamento e Coeficiente Silhouette, pois esta será usada para as análises dos resultados.

2.1. DESCOBERTA DE CONHECIMENTO

A descoberta de conhecimento não é uma área nova da computação. Ela surgiu na Inteligência Artificial (*Artificial Intelligence* - AI), na década de 90 e preocupava-se não só em descobrir conhecimento, mas sim, também, em descobrir formas de aquisição e armazenamento deste conhecimento (WIVES e LOH, 2006).

O aumento da quantidade de informação gerada está em evidencia uma vez que diversos estudiosos analisam este fenômeno e meios de extrair conhecimento de toda essa informação. Segundo Silva (2012, p.21),

> "A velocidade e a amplitude com que o conhecimento gerado passou a ser compartilhado provocaram o surgimento de uma dinâmica de reaproveitamento e produção de novos conhecimentos, bem como o aparecimento de novas necessidades de tratar a informação."

Com o tempo e o advento dos Sistemas de Gerenciamento de Bancos de Dados (SGBDs), os pesquisadores da área de Sistemas de Informação (mais

especificamente, de Bancos de Dados) passaram a pesquisar novas aplicações para as informações armazenadas nestes bancos de dados. Ou seja, pensavam que, além das informações tradicionais armazenadas nestes sistemas, poderiam descobrir informações implícitas (que não estavam disponíveis de forma clara) que também pudessem ser úteis para as empresas (WIVES e LOH, 2006).

Com o advento das novas tecnologias, o surgimento e popularização da Internet, novos tipos de informações foram surgindo com o uso de emails, bate-papo, sites, blogs, e essas informações encontradas na Internet são conhecidas como informações não estruturadas, diferente das informações encontradas em SGBDs que são estruturadas,

Logo, novas técnicas surgiram para dar suporte à descoberta de conhecimento a estes dados não estruturados, os textos. Isso ocasionou o surgimento de um novo ramo de pesquisas na área de descoberta de conhecimento: o ramo KDT que estuda técnicas específicas para esse tipo de informação.

Com isso, basicamente, existem duas grandes abordagens utilizadas na área de Descoberta de Conhecimento, a Descoberta de Conhecimento em dados Estruturados e a Descoberta de Conhecimento em Dados não Estruturados como mostra a Figura 1.

Figura 1: Tipos de Descoberta de Conhecimento.

2.2. DESCOBERTA DE CONHECIMENTO EM BASES TEXTUAIS

O termo “descoberta de conhecimento em textos” foi utilizado pela primeira vez por Feldman e Dagan (1995), para designar o processo de encontrar algo interessante em coleções de textos (artigos, histórias de revistas e jornais, mensagens de e-mail, páginas Web.). Pode-se então definir KDT como sendo o processo de extrair padrões ou conhecimento, interessantes e não triviais, a partir de documentos textuais (LOH, 2001).

A partir das técnicas do KDD iniciou-se o desenvolvimento de métodos de KDT.

Dados estruturados possuem relações, atributos e esquemas, sendo esses dados previamente preparados para uma melhor recuperação das informações. Já os dados não estruturados, não possuem uma estrutura definida, são normalmente caracterizadas por documentos, textos, imagens e vídeos. A grande maioria dos dados atuais da Web e empresas segue esse formato.

Então, pode-se considerar o processo de KDT similar ao KDD, porém o KDT trabalha com um *corpus* em linguagem natural, buscando padrões e tendências, classificando e comparando documentos (SILVA e ROVER, 2011).

Análise de dados armazenados 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 característica da não estruturação. Logo, são necessárias técnicas e ferramentas específicas para tratamento deste tipo de dados.

O KDT faz uso de técnicas e ferramentas inteligentes que auxiliam o processo de análise em bases de dados textuais de grande volume, visando obter conhecimento útil, beneficiando tanto o usuário que utiliza documentos da Web, quanto de qualquer domínio que faça uso intensivo de bases textuais não estruturadas (CASTRO, SIMÕES, MATTOS, 2009).

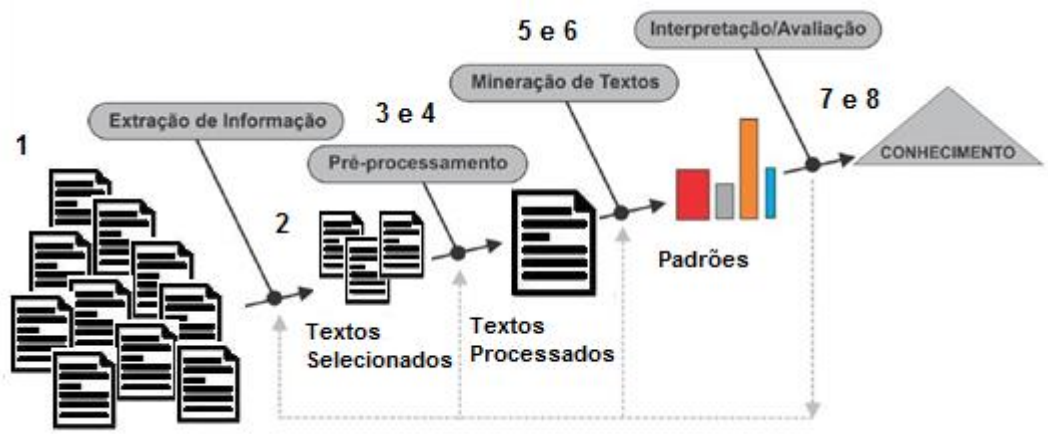

Figura 2: Uma visão geral do processo KDT.

Segundo Cardoso, et al (2012), Muitos métodos e ferramentas que fazem alguma espécie de KDT, e que podem ser enquadrados na área, realizam alguns procedimentos que são similares na área de KDD. Fazendo-se uma análise técnica dos artigos e teses correlacionados com KDT consegue-se identificar que há uma metodologia para o processo KDT. As etapas que compõem essa metodologia são representadas na Figura 2 e descritas a seguir. Apesar de ser uma metodologia, nem todos os modelos de descoberta de conhecimento seguem essas etapas. Em alguns casos, determinadas fases são simplesmente suprimidas. Em outros, fases adicionais são utilizadas. Porém as etapas mais importantes para o processo, segundo (CARDOSO et al, 2012, p.2) são:

1. "Definição de objetivos: compreensão do domínio, identificação do que deve ou o que não deve ser descoberto;
2. Seleção dos textos (etapa de textos selecionados): a identificação de quais informações, dentre as bases textuais existentes, devem ser efetivamente consideradas. A utilização de muitas informações pode influenciar negativamente ao resultado de uma descoberta, além de tornar mais demorado;
3. Limpeza dos textos (etapa de pré-processamento): tem o intuito de remover ruídos e preparar os textos. Esses métodos vão desde a limpeza de caracteres indesejados, passando pela correção da ortográfica e morfológica e indo até a análise semântica e normalização do vocabulário;
4. Redução ou projeção dos textos (etapa de pré-processamento, escolha das características relevantes para a análise): nem todas as características (palavras ou trechos) são importantes. Dependendo do objetivo da análise, partes de um documento ou conjunto de documentos podem ser importantes do que outras. Somente estas devem ser analisadas para que o resultado não seja inútil e para que o processamento seja mais eficiente (rápido);
5. Escolha da técnica, método ou tarefa de TM: existem vários métodos de descoberta, cada um deles capaz de descobrir algo diferente.
6. TM: a aplicação dos métodos escolhidos;

7. Interpretação dos resultados;
8. Consolidação do conhecimento descoberto (documentos ou gráficos).”

Portanto, os benefícios da TM pode se estender a qualquer domínio que utilize textos, sendo que suas principais contribuições estão relacionadas à busca de informações específicas em documentos, a análise qualitativa e quantitativa de grandes volumes de textos e a melhor compreensão do conteúdo disponível em documentos textuais.

Ao utilizar os recursos de TM, um usuário não solicita exatamente uma busca, mas sim uma análise de um documento. Entretanto, este não recupera o conhecimento em si. É importante que o resultado da consulta seja analisado e contextualizado para posterior descoberta de conhecimento (MIRANDA, 2009).

2.3. DESCOBERTA DE CONHECIMENTO POR AGRUPAMENTO TEXTUAL

A técnica de agrupamento procura separar automaticamente elementos em classes que serão identificadas durante o processo (não há classes pré-definidas).

O agrupamento auxilia o processo de descoberta de conhecimento, facilitando a identificação de padrões (características comuns dos elementos) nas classes. Isso porque a divisão em classes facilita a compreensão humana das observações e o desenvolvimento subsequente de teorias científicas.

O agrupamento também pode ser utilizado para estruturar e sintetizar o conhecimento, quando este é incompleto, quando há muitos atributos a serem considerados ou para extrair categorias dos textos.

As interpretações das classes identificadas devem ser feitas pelo usuário, os quais necessitam um pouco de conhecimento do domínio. Entretanto, diferentes usuários são levados às mesmas interpretações. Estas interpretações avaliam as associações entre termos (revelando o contexto semântico quase em linguagem natural), a estrutura das coleções por análise da terminologia e a informação qualitativa sobre diferenças e similaridades entre componentes ou classes.

2.4. AGRUPAMENTO

Agrupamento é o nome dado para o grupo de técnicas computacionais cujo propósito consiste em separar objetos em grupos, baseando-se nas características que estes objetos possuem. A ideia básica consiste em colocar em um mesmo grupo, objetos que sejam similares de acordo com algum critério pré-determinado.

A análise de agrupamento é uma ferramenta útil para a análise de dados em muitas situações. Esta técnica pode ser usada para reduzir a dimensão de um conjunto de dados, reduzindo uma ampla gama de objetos à informação do centro do seu conjunto (LINDEN, 2009).

A técnica de agrupamento é utilizada quando não se conhece previamente as categorias nas quais os objetos podem ser classificados. Quando estas categorias são conhecidas, utiliza-se classificação.

As técnicas de agrupamento de dados normalmente são dependentes de domínio, ou seja, dificilmente se encontrará uma técnica genérica, aplicável de maneira satisfatória a todos os tipos de dados e em todos os contextos. Por esse motivo, existem diversas técnicas e estas são de grande utilidade em diversas áreas (SANTOS, 2009).

O problema de agrupamento é de interesse em qualquer área em que se deseje agrupar dados, sejam estes relativos às compras efetuadas em um supermercado, às especificações físicas e químicas de petróleos, aos sintomas de doenças, às características de seres vivos, às funcionalidades de genes, aos documentos existentes na Web, à composição de solos, aos aspectos da personalidade de indivíduos, às transações bancárias realizadas por clientes de um banco, entre outros.

2.5. AGRUPAMENTO DE INFORMAÇÕES TEXTUAIS

Segundo Ramos e Brascher (2009), o agrupamento de documentos constitui o grande diferencial da técnica de TM, visto que identifica associações entre

documentos aparentemente sem nenhuma relação. Ou seja, são apresentadas possibilidades de extração de conhecimentos totalmente novos (Figura 3).

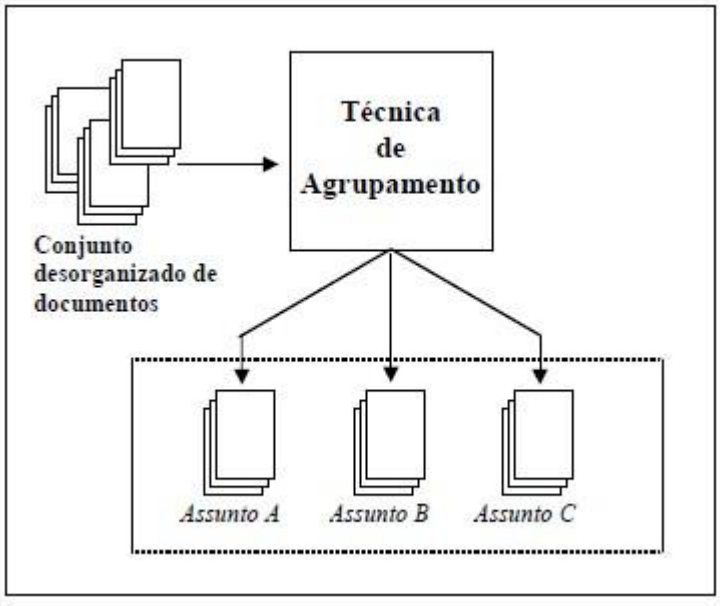

Figura 3: Objetivo do agrupamento de informações textuais.

O agrupamento permite que novas classes sejam descobertas, já que consegue agrupar documentos mesmo que estes não pertençam a assuntos conhecidos. Isso porque não há necessidade de conhecimento prévio sobre os assuntos (ou os possíveis assuntos) dos documentos. Os assuntos ou as classes dos documentos são descobertos após o agrupamento, durante o processo de análise dos grupos obtidos (WIVES, 2012).

Por esse motivo, diz-se que o processo de agrupamento é utilizado com base em dados não rotulados (*unlabeled data*), ou seja, não conhecidos e sem modelos estatísticos que os descrevam. Assim, o processo de agrupamento consegue agrupar uma coleção de padrões desconhecidos (não classificados) em *clusters* que possuem algum significado para o usuário.

Esse tipo de análise é dito não supervisionado, pois, ao contrário da classificação propriamente dita, não há como comparar os resultados com modelos conhecidos para saber se o processamento está sendo feito de forma adequada ou não (WIVES, 2012).

Como o objetivo do agrupamento é organizar os documentos em *clusters* de objetos similares, ele está baseado na identificação da similaridade entre os documentos. Uma vez identificada a similaridade, eles são atribuídos a um *cluster* de documentos que possuem alguma relação de similaridade. Por consequência, documentos pertencentes a um mesmo *cluster* tendem a ser mais similares entre si do que em relação a outros documentos pertencentes a outros *clusters.*

2.5.1. Fases do processo de agrupamento de informações textuais

O processo de agrupamento de textos, conforme ilustra a Figura 4, é dividido em três fases: pré-processamento, processamento e pós-processamento.

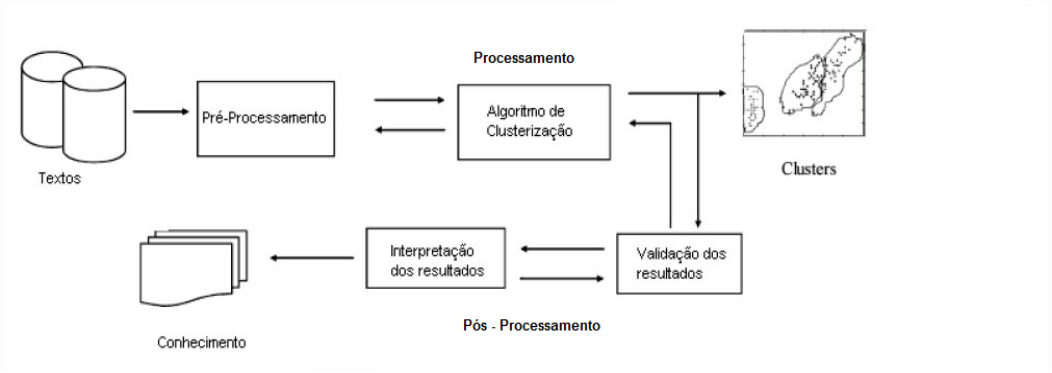

Figura 4: Etapas do processo de agrupamento textual.

Considerando que a etapa de coleta de documentos tenha sido cumprida e, portanto, os documentos estejam disponíveis, é necessário realizar o pré-processamento desses documentos. O pré-processamento é a atividade que mais necessita de tempo e dedicação, pois determina a boa qualidade dos resultados dos *clusters*. O principal objetivo da etapa de pré-processamento de textos é estruturar os documentos para serem submetidos a algum algoritmo de agrupamento textual. De modo geral, a etapa de pré-processamento tem por finalidade melhorar a qualidade dos documentos já disponíveis e organizá-los.

A etapa do processamento, nada mais é que o uso do algoritmo de agrupamento escolhido para determinado contexto em si. Através do processamento, os *clusters* são criados.

A última etapa, pós- processamento há a validação e a interpretação dos resultados através dos *clusters* gerados. O conhecimento extraído na fase do processamento pode gerar uma grande quantidade de padrões.

2.6. MODELO CASSIOPEIA

O Modelo Cassiopeia é um agrupador de textos criados por Guelpeli (2012), será usado neste trabalho e fará parte da metodologia apresentada no Capítulo 3.

O Modelo Cassiopeia como boa parte dos agrupadores está dividido em três macro etapas que são: pré-processamento, processamento e pós-processamento, como ilustrado na Figura 5.

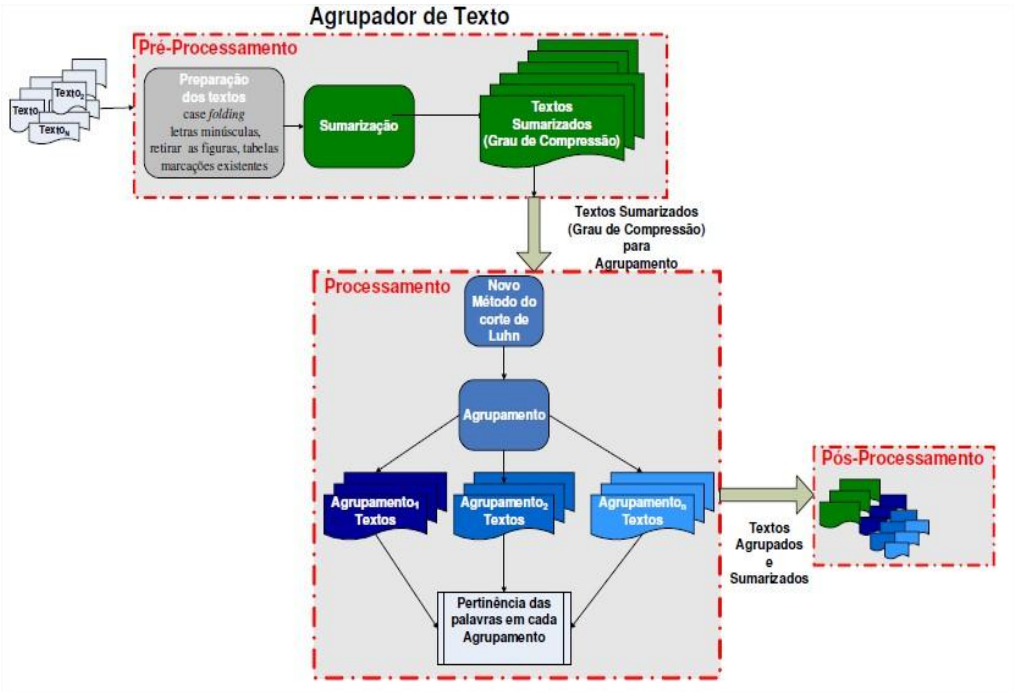

Figura 5: Modelo Cassiopeia.

As três macro etapas do Modelo Cassiopeia começam com a entrada dos textos pré-processados. Utiliza-se a técnica *case folding* para o pré-processamento,

que corresponde aos descartes de figuras, tabelas e marcações existentes, para os textos ficarem de acordo com o processo computacional.

Após o pré-processamento dos textos, há a etapa do processamento que utiliza agrupamento de textos hierárquicos e um algoritmo para unir os textos com similaridade. À medida que novos textos são agrupados ocorre o reagrupamento, podendo surgir agrupamentos, subagrupamentos ou até mesmo a fusão destes (GUELPELI, 2012).

Segundo Guelpeli (2012), terminada a etapa de processamento, começa a de pós-processamento, na qual cada um dos agrupamentos ou subagrupamentos 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.

2.6.1. Pré-processamento

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).

No trabalho de Guelpeli, 2012, é usado o processo de sumarização, cuja finalidade é diminuir o número de palavras, viabilizando o processamento. Com o processo de sumarização, obtém-se a parte mais importante, ou seja, a ideia principal do texto-fonte, através da criação de um resumo com as palavras mais significativas.

2.6.2. Processamento

Após o pré-processamento dos textos, há a etapa do processamento onde acontece o uso do algoritmo de agrupamento.

Segundo Guelpeli (2012), 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.

2.6.2.1. Identificação dos atributos

Quanto mais um termo aparecer em um documento, mais importante é, para aquele documento. O Modelo Cassiopeia utiliza frequência relativa, que identifica as características de uma palavra dentro de um documento, definindo a importância deste termo, de acordo com a frequência com que é encontrado. A frequência relativa é calculada por meio da **Equação 1:**

$$F_{\mathrm{r}}X = \frac{F_{\mathrm{abs}}X}{\mathrm{N}} \qquad (1)$$

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).

2.6.2.2. 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 6 e Figura 7). 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 método do corte proposto no Modelo Cassiopeia, ou seja, a seleção dos atributos, segundo (GUELPELI, 2012, p.51):

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 descrescente 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 frequencia 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.”

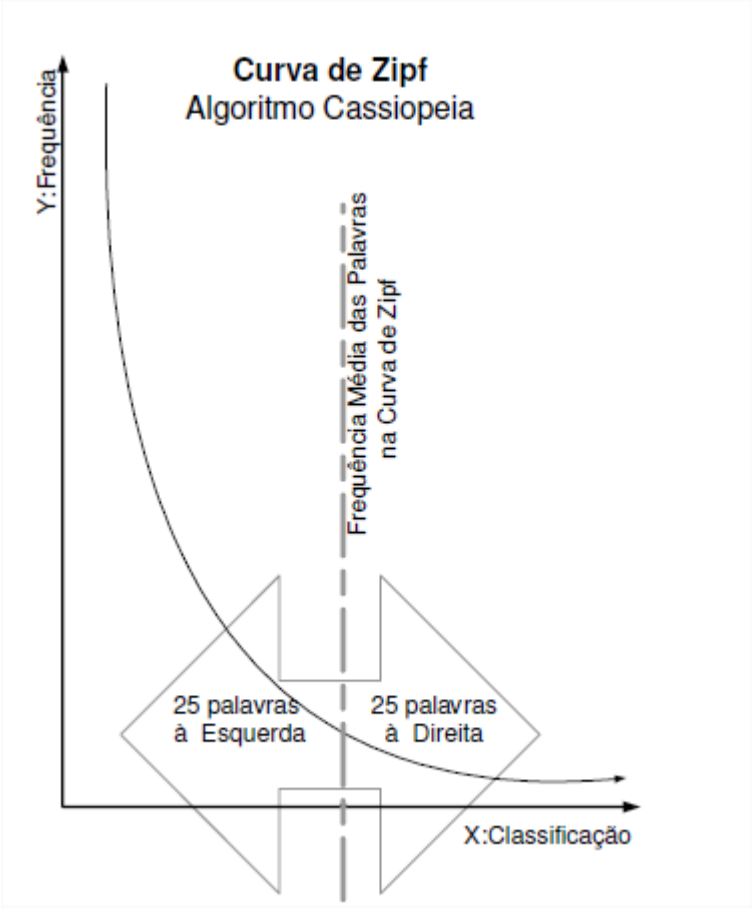

Figura 6: Seleção dos Atributos no Modelo Cassiopeia.

1 - Algoritmo do Método Cassiopeia:

1. *Estabelecer frequência média do conjunto P de palavras do documento.*
   Equação 2:

$$f(P,N) = \frac{\sum_{n=1}^{N} F_r P_n}{N} \qquad (2)$$

Onde: N é igual ao número total de palavras no documento; FrPn é igual à frequência relativa de Pn; onde P é o conjunto de palavras no documento e Pn refere-se a quantidade de vezes que uma palavra aparece no documento e f(P,N) é a frequência média das palavras na distribuição (GUELPELI, 2012).

2. *Escolher as 25 palavras à esquerda da média e as 25 palavras à direita da média.*

Figura 7: Algoritmo do Método Cassiopeia.

Portanto, sabendo-se o vetor de palavras que mais se aproximam da frequência média nos documentos, os agrupamentos são formados a partir da similaridade entre essas palavras e com isso, centroides para cada agrupamento são criados, cujas palavras são de alta relevância.

2.6.2.3. Uso do método hierárquico aglomerativo e do algoritmo *Cliques*

O Modelo Cassiopeia, utiliza o método hierárquico para organizar seus textos em agrupamentos que são particionados, sucessivamente, produzindo uma representação hierárquica. Este tipo 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. A principal vantagem e a característica determinante para escolha do método a ser usado no Modelo Cassiopeia é a facilidade de lidar com qualquer medida de similaridade utilizada, ou seja, o algoritmo *Cliques* e a sua consequente aplicabilidade a qualquer tipo de atributo (GUELPELI, 2012).

No método hierárquico aglomerativo (Figura 8), descrito na Figura 9, os agrupamentos são recursivamente criados, considerando alguma medida de similaridade. Sendo assim, no início, os agrupamentos são em número reduzido, com baixo grau de similaridade, mas com o decorrer do processo, eles vão aumentando e tornando-se dissimilares, com alto grau de similaridade entre os documentos de cada agrupamento (GUELPELI, 2012).

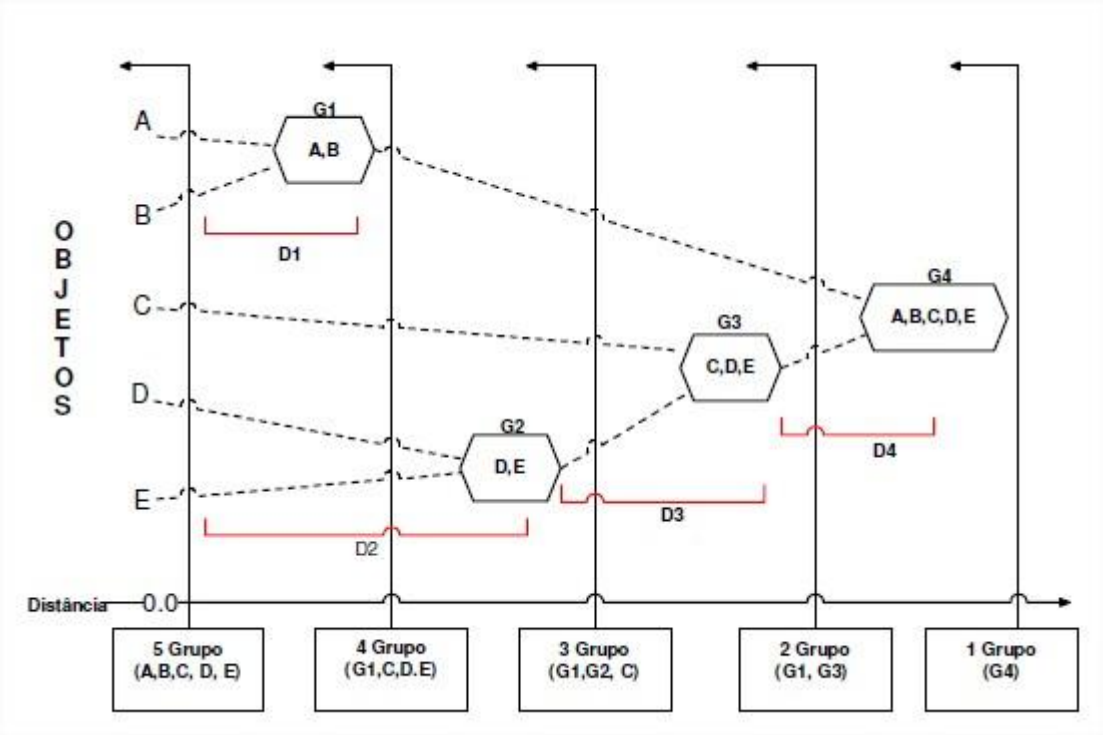

Figura 8: Dendograma do Método Hierarquico Aglomerativo.

2 – Algoritmo Aglomerativo:

1. *Procure pelo par de clusters com 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. *Voltar ao passo 1 até que reste apenas um cluster*

Figura 9: Algoritmo Aglomerativo.

No Modelo Cassiopeia, o algoritmo utilizado para agrupamento textual de modo hierárquico, é o algoritmo *Cliques.*

Devido à sua capacidade de construir agrupamentos mais coesos, o algoritmo *Cliques* (Figura 10) é o mais adequado e usado no agrupamento de texto. Os textos só são adicionados a um agrupamento, caso seu grau de similaridade seja maior do que o limiar definido para todos os textos já presentes nesse agrupamento. A Figura 11 descreve os passos do algoritmo *Cliques* (GUELPELI, 2012).

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, pode surgir agrupamentos, subagrupamentos ou até mesmo a fusão destes.

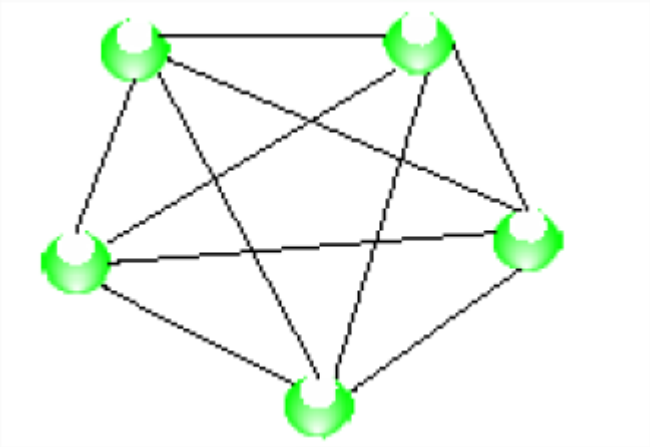

Figura 10: Grafo do Algoritmo *Cliques*.

2 - *Algoritmo Cliques:*

1. *Seleciona 1º 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 agrupado.*
4. *Voltar ao passo 2, enquanto houver objetos.*
5. *Para os elementos não alocados, repetir o passo 1.*

Figura 11: Algoritmo *Cliques*.

2.6.3. Pós-processamento

Nesta etapa, o modelo terá como saída *links* para os textos fontes agrupados por similaridade e também os textos sumarizados.

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, justificando assim a criação dessa estrutura.

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, a estrutura possibilitará ter o texto-fonte e o seu sumário correlato, ou seja, com alto grau de informatividade.

2.7. MÉTRICAS PARA ANÁLISE DE AGRUPAMENTO TEXTUAL

Segundo Guelpeli (2012) a avaliação de agrupamentos pode ser distribuída em duas grandes categorias de métricas: externas ou supervisionadas e internas ou não supervisionadas.

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 humano. Para esse tipo na opinião de Guelpeli (2012), são usadas medidas como: *Precisão*, *Recall* e como medida harmônica destas duas, o *F-Measure*.

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

2.7.1. Métricas internas

Coesão (*C*): **Equação 3**:

$$\frac{\sum_{i>j} Sim(P_i, P_j)}{\frac{n(n-1)}{2}} \qquad (3)$$

A *Coesão* mede a similaridade entre os elementos do mesmo agrupamento. Quanto maior a similaridade entre eles, maior a coesão deste agrupamento (GUELPELI, 2012).

Onde *Sim* (*Pi, Pj*) é o calculo 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* (GUELPELI, 2012).

Acoplamento (*A*): **Equação 4:**

$$\frac{\sum_{i>j} Sim(C_i, C_j)}{\frac{n_a(n_a - 1)}{2}} \qquad (4)$$

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 (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* (GUELPELI, 2012).

Coeficiente de Silhouette (*S*): **Equação 5:**

$$S = \frac{b(i) - a(i)}{max(a(i), b(i))} \qquad (5)$$

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* (GUELPELI, 2012).

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*) (GUELPELI, 2012).

3. METODOLOGIA

Neste capítulo serão apresentadas as partes que compõem a metodologia usada para o desenvolvimento do trabalho. Com isso, para inicio, foram feitas pesquisas referentes ao problema abordado com bibliografia atualizada, de forma a se familiarizar com o problema e com as terminologias adotadas. Será mostrado como as BM foram selecionadas para dar início à técnica de agrupamento textual. Estatísticas da *corpora* foram criadas para facilitar posteriormente a análise dos resultados no pós processamento. Por fim, será descrito como as bulas serão pré-processadas para ficarem de acordo ao processo computacional do Modelo Cassiopeia, processadas e pós-processadas.

3.1. SELEÇÃO DO *CORPUS*

Corpus é um conjunto de dados textuais coletados criteriosamente para ser objeto de pesquisa. Podem ser documentos textuais de domínio diversos. *Corpora* nada mais é do que o plural, o conjunto desse *corpus*. Neste trabalho a *corpora* criada é referente ao domínio farmacêutico, especificamente bulas de medicamento. A *corpora* possui 707 bulas coletadas pelo site Bulário.net (http://bulario.net/alfa/).

Existem muitos tipos de site que são específicos para obter BM, como por exemplo, o site Tua Saúde (http://www.tuasaude.com/), Wikibula (http://www.wikibula.com.br/), MedicinaNet (http://www.medicinanet.com.br/), Bulas.med.br (http://www.bulas.med.br/), ANVISA (http://portal.anvisa.gov.br/), Bulário.net, , onde pode-se encontrar bulas com diferentes especificações. Foi escolhido o Bulário,net, pois contêm bulas mais fáceis de coletar e de serem pré processadas, essas bulas são retiradas do Bulário Eletrônico da ANVISA, pois apresenta uma base de dados mais completa, contendo mais de 3000 dos principais laboratórios farmacêuticos.

Apesar do site da ANVISA possuir bulas mais completas e uma maior quantidade, o Bulário.net foi usado devido a dificuldade enfrentada de pré processar e coletar todas as bulas do site da ANVISA.

Existe no Bulário.net duas categorias, uma de bulas destinadas aos pacientes e outra de bulas de profissionais da saúde conforme mostra a Figura 12.

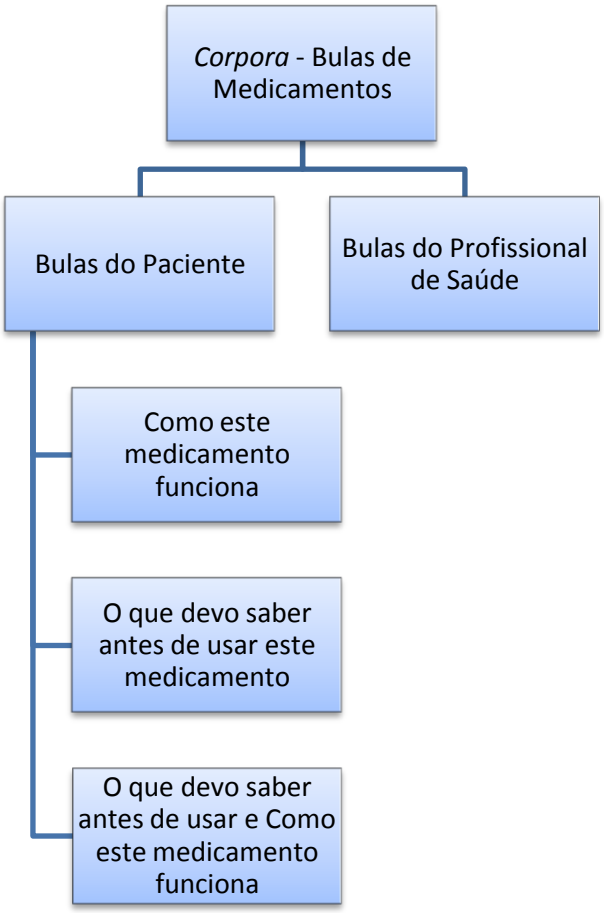

Figura 12: Esquema da *corpora* usado na metodologia.

Dentro dessas duas classes estão dispostas as bulas. Dentro das bulas existem muitas informações do tipo: “O que devo saber antes de usar”, “Como este medicamento funciona”, “Indicações”, “Contra Indicações”, “Como usar este medicamento”, entre outras.

Neste trabalho, a classe escolhida corresponde às bulas destinadas aos pacientes. Cada *corpus* dentro dessa categoria é composto de um tipo, que foi separado em arquivos txt. Para este trabalho foi escolhido o tipo “Como este medicamento funciona” (Figura 13) que será denominado por questões metodológicas de C_1 e “O que devo saber antes de usar” (Figura 14) denominado de C_2. Esta escolha deve-se a importância que esses dois tipos têm.

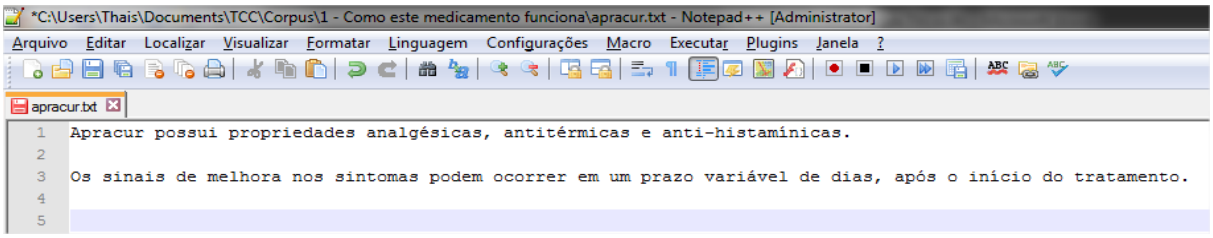

Figura 13: “Como este medicamento funciona” da bula do medicamento Apracur.

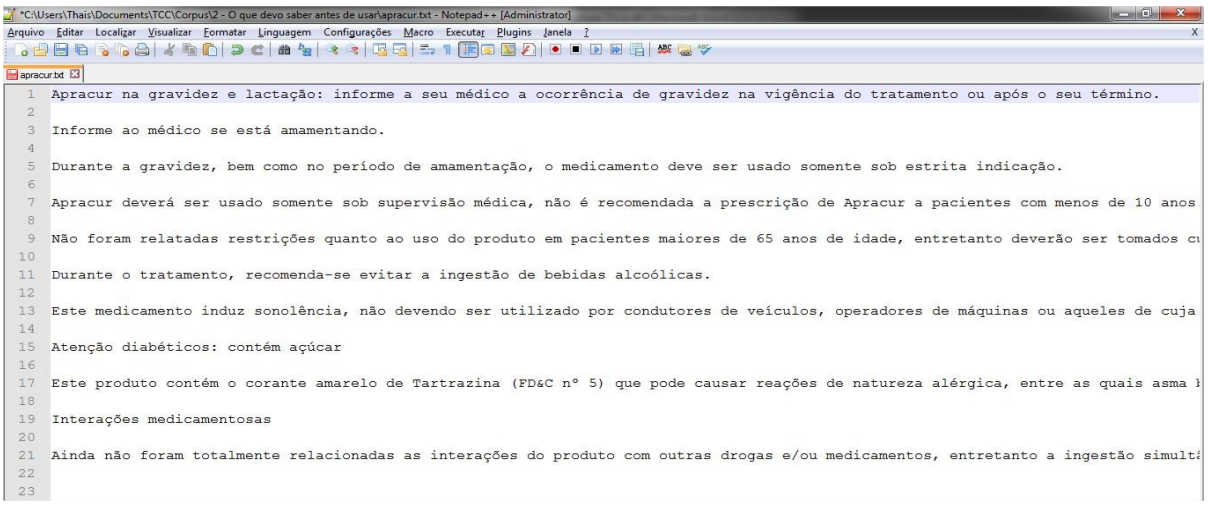

Figura 14: "O que devo saber antes de usar" da bula do medicamento Apracur.

Por questões de estratégia de conhecimento foi ainda criado o *corpus* que é a junção de C_1 e C_2 denominado aqui de C_3 . Esta junção é necessária para analisar a quantidade de palavras nos agrupamentos e bem como a qualidade dessas palavras.

3.2. ESTATÍSTICAS DA *CORPORA*

As estatísticas da c*orpora,* apresentadas nas Tabelas 1, 2 e 3, foram calculados com a utilização do software *Get FineCount* 2.6., cuja última versão é de 30 de setembro de 2013.

A Tabela 1 apresenta a estatística do *corpus* C_1. As linhas têm valores mínimos, máximos, totais e médias relacionadas a cada coluna. Um item que faz parte da coluna, muito importante para este trabalho, por exemplo, é o número de

palavras máximo e mínimo. Este *corpus* tem um texto com 5 palavras, no mínimo, e outro com 892, no máximo. No total dos 707 textos, são 53.034 palavras, com média de 75,1 por texto.

Tabela 1: Estatística dos 707 textos-fonte no domínio farmacêutico, compostos pelo *corpus* C_1.

Itens	Nº de Palavras	Nº de Palavras + Numeral	Nº (%)	Caracteres	Caracteres + Espaços	Sentenças
Mínimos	5	5	0%	36	40	1
Máximos	892	927	21,43%	5247	6236	43
Totais	53034	55252	4,01%	305041	363179	3153
Médias	75,1	78,15	-	431,46	513,7	4,46

A Tabela 2 apresenta a estatística do *corpus* C_2. As linhas têm valores mínimos, máximos, totais e médias relacionadas a cada coluna. Um item que faz parte da coluna, muito importante para este trabalho, por exemplo, é o número de palavras máximo e mínimo. Este *corpus* tem um texto com 12 palavras, no mínimo, e outro com 2.241, no máximo. No total dos 707 textos, são 297.512 palavras, com média de 420,8 por texto.

Tabela 2: Estatística dos 707 textos-fonte no domínio farmacêutico, compostos pelo *corpus* C_2.

Itens	Nº de Palavras	Nº de Palavras + Numeral	Nº (%)	Caracteres	Caracteres + Espaços	Sentenças
Mínimos	12	13	0%	54	72	1
Máximos	2241	2248	14,93%	12690	15197	118
Totais	297512	299421	0,64%	1766471	2097227	17103
Médias	420,8	423,5	-	2498,5	2966,38	24,2

A Tabela 3 apresenta a estatística do *corpus* C_3. As linhas têm valores mínimos, máximos, totais e médias relacionadas a cada coluna. Um item que faz

parte da coluna, muito importante para este trabalho, por exemplo, é o número de palavras máximo e mínimo. Este *corpus* tem um texto com 5 palavras, no mínimo, e outro com 2.241, no máximo. No total dos 1.414 textos, são 350.546 palavras, com média de 248 por texto.

Tabela 3: Estatística dos 1414 textos-fonte no domínio farmacêutico, compostos pelo *corpus* $C_{3.}$

Itens	Nº de Palavras	Nº de Palavras + Numeral	Nº (%)	Caracteres	Caracteres + Espaços	Sentenças
Mínimos	5	5	0%	36	40	1
Máximos	2241	2248	21,43%	12690	15197	118
Totais	350546	354673	1,16%	2071512	2460406	20256
Médias	248	250,8	-	1465	1740	14,3

3.3. USO DO MODELO CASSIOPEIA

Nesta seção, será mostrada como as etapas do Modelo Cassiopeia, pré-processamento, processamento e pós-processamento, foram utilizadas na metodologia do trabalho.

3.3.1. Pré-processamento

Segundo Guelpeli (2012), preparar os textos para o processo computacional é uma atividade difícil e trabalhosa. É a atividade mais importante em todo o processo. Com isso, a fase de pré-processamento foi a mais demorada.

Nesta etapa, os *corpus* passaram por um processo de “limpeza de dados”. Foram retiradas algumas figuras e tabelas presentes nas BM que não forneceriam qualidade para o processamento no Modelo Cassiopeia.

A técnica sumarização, também presente no pré- processamento do Modelo

Cassiopeia, não foi utilizada no trabalho, pois os *corpus* são pequenos não havendo necessidade do uso da técnica.

3.3.2. Processamento

Terminado a fase de pré-processamento inicia-se o processamento com o Modelo Cassiopeia (GUELPELI, 2012). Para realizar o processamento, foram escolhidas as métricas internas para análise de agrupamento textual. Cada um dos *corpus* C_1, C_2 e C_3, foram processados ao longo de 50 passos pelo Modelo Cassiopeia.

3.3.3. Pós-processamento

No pós-processamento, as bulas vão estar agrupadas em *clusters*. Começa então a fase final da pesquisa que é análise e avaliação dos resultados gerados pelo Modelo Cassiopeia.

Foram gerados gráficos contendo os valores em arquivos .csv utilizados no Excel no processamento do Modelo Cassiopeia. Com esses valores, foram gerados as médias acumuladas para cada um dos *corpus*. As métricas Acoplamento, Coesão e Coeficiente de Silhouette, mensuraram a qualidade dos agrupamentos gerados.

A análise e comparação dos gráficos serão vistas com detalhe no Capítulo 4.

4. RESULTADOS

Neste capítulo, serão apresentados os resultados referentes às métricas internas ou não supervisionadas (Coesão, Acoplamento e Coeficiente Silhouette*)*, dos *corpus* de C_1, C_2 e C_3.

Apenas os resultados das métricas internas serão vistas, pois, no presente trabalho não há um supervisor para analisar os resultados das métricas externas.

4.1. MÉTRICA INTERNA: COESÃO, ACOPLAMENTO E COEFICIENTE DE SILHOUETTE

As Figuras 15, 16, 17, 18, 19 e 20, apresentam valores do eixo x correspondentes ao número de passos, ou número de vezes que cada *corpus* foi processado no algoritmo do Modelo Cassiopeia. Os eixos y apresentam os valores acumulados das métricas internas

Os resultados da Figura 15 demonstram que os valores da média acumulada de acoplamento do *corpus* $C_{1:}$ “Como este medicamento funciona”, apresentam valores maiores do que no $C_{2:}$ “O que devo saber antes de usar”. O valor de C_1 está entre 0,05 e 0,06, enquanto que C_2 entre 0,03 e 0,04. Os valores de acoplamento consistem na similaridade entre os *clusters* gerados, então, quanto menor for esse valor de similaridade e quanto mais próximo de 0 (zero), melhor é o resultado.

Tabela 4: Resultados obtidos pelo Modelo Cassiopeia, usando Média Acumulada do Acoplamento para C_1 e C_2.

Média Acumulada do Acoplamento		
Cálculos	Como este medicamento funciona	O que devo saber antes de usar
Variância	0.0000196327	0.0000186122
Média	0,0526	0,0376

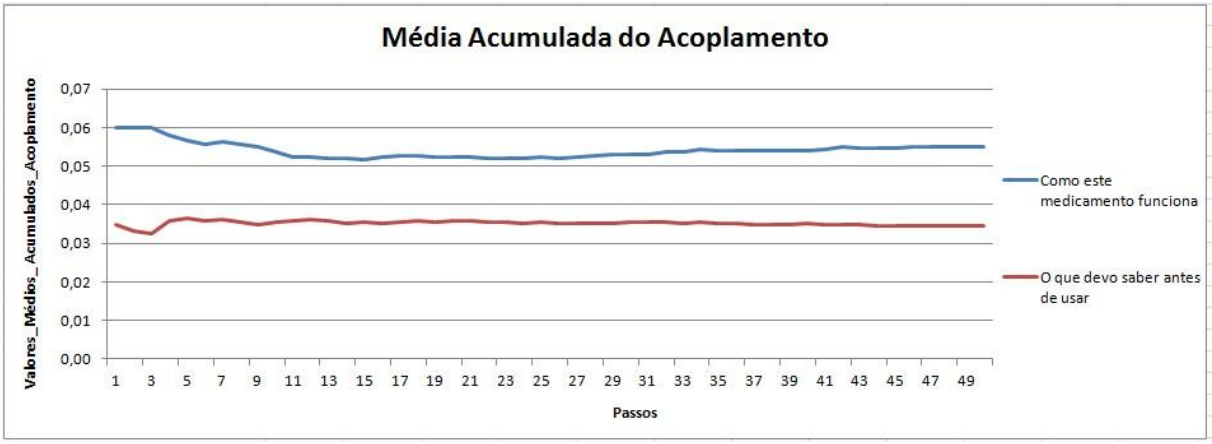

Figura 15: Resultados obtidos pelo Modelo Cassiopeia, usando Média Acumulada do Acoplamento para C_1 e C_2.

Os resultados da Figura 16 demonstram que os valores da média acumulada da coesão do *corpus* C_1, apresentam valores menores do que o *corpus* C_2. O valor C_1 está entre 0,35 e 0,36, enquanto que no *corpus* C_2 esta entre 0,42 e 0,43. Os valores de coesão consistem na similaridade entre documentos dentro de um mesmo *cluster*, então, quanto maior for esse valor de similaridade e quanto mais distante do 0 (zero), melhor é o resultado. Em comparação com os resultados anteriores de média acumulada do acoplamento, pode-se perceber que quanto menor for a similaridade entre os *clusters* gerados, maior é a similaridade entre os documentos dentro de um *cluster*. Ou seja, o *corpus* C_2, apresentou os melhores resultados de acoplamento e coesão.

Tabela 5: Resultados obtidos pelo Modelo Cassiopeia, usando Média Acumulada da Coesão para C_1 e C_2.

Média Acumulada da Coesão		
Cálculos	Como este medicamento funciona	O que devo saber antes de usar
Variância	0,000002	0.0000075102
Média	0,3502	0,4292

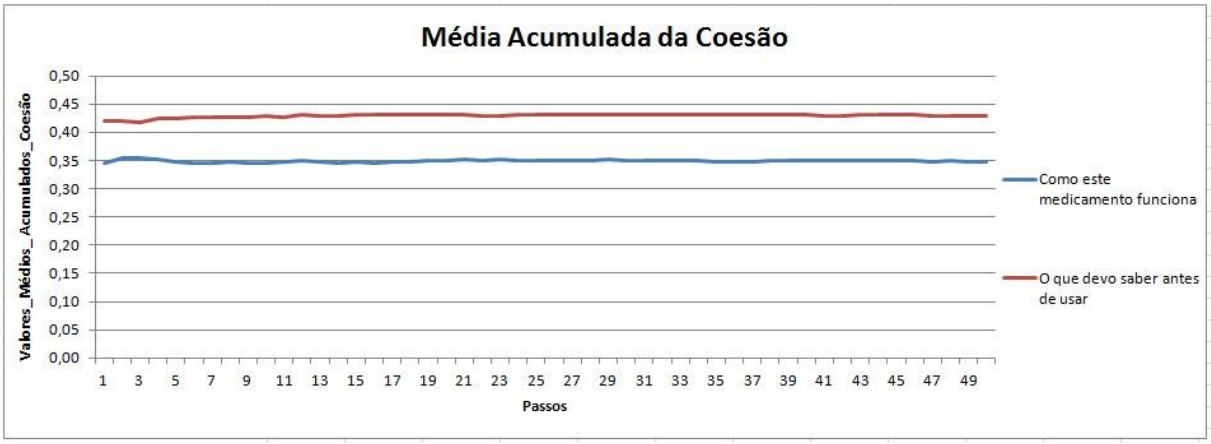

Figura 16: Resultados obtidos pelo Modelo Cassiopeia, usando Média Acumulada da Coesão para C_1 e C_2.

Os resultados da Figura 17 demonstram que os valores da média acumulada do coeficiente de silhouette do *corpus* C_2, apresentam valores maiores do que no C_1. O valor da média acumulada de C_2 é de 0,98, enquanto de C_1 está entre 0,95 e 0,96. De acordo com a fundamentação teórica descrita no Capítulo 2, os melhores valores do coeficiente de silhouette, são os que estão mais próximos de 1 (um). Coeficiente de Silhouette consiste na média harmônica entre o acoplamento e a coesão, ou seja, de quanto um objeto é similar aos demais membros do seu grupo, e de quanto este mesmo objeto é distante dos de outro grupo. O melhor resultado está no *corpus* C_2 relacionado com seus baixos valores no acoplamento e altos valores na coesão.

Tabela 6: Resultados obtidos pelo Modelo Cassiopeia, usando Média Acumulada do Coeficiente de Silhouette para C_1 e C_2.

Média Acumulada do Coeficiente de Silhouette		
Cálculos	Como este medicamento funciona	O que devo saber antes de usar
Variância	0.000002	0
Média	0,9502	0,98

Figura 17: Resultados obtidos pelo Modelo Cassiopeia, usando Média Acumulada do Coeficiente de Silhouette para C_1 e C_2.

Para uma análise mais profunda dos resultados, os *corpus* C_1 e C_2, foram processados juntos, criando um novo *corpus* denominado de C_3.

Os resultados da Figura 18 demonstram que com a combinação dos *corpus*, a média acumulada do acoplamento apresentou uma pequena melhora em comparação aos números de C_2, que apresentou os melhores resultados nas análises anteriores.

Tabela 7: Resultados obtidos pelo Modelo Cassiopeia usando Média Acumulada do Acoplamento para C_3.

Média Acumulada do Acoplamento	
Cálculos	O que devo saber antes de usar e Como este medicamento funciona
Variância	0.0000039184
Média	0.0204

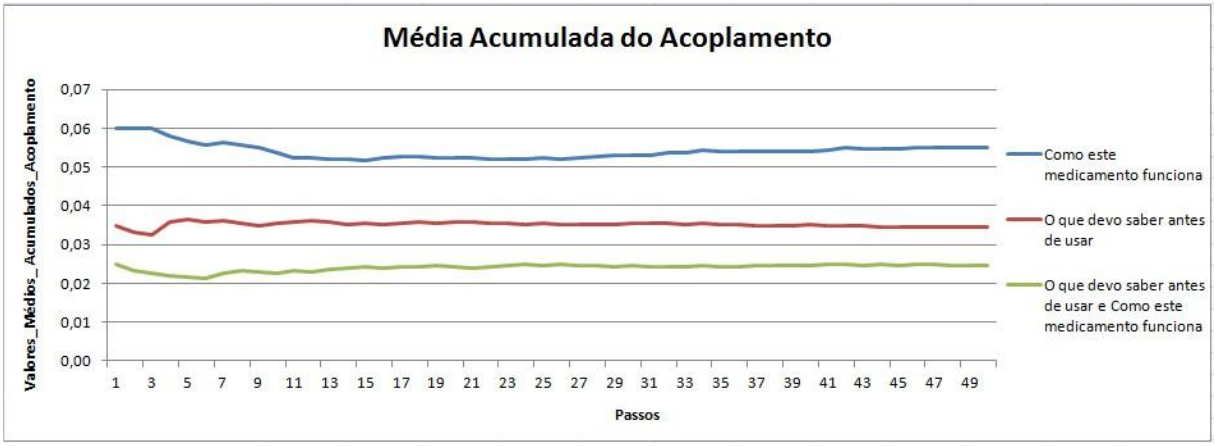

Figura 18: Resultados obtidos pelo Modelo Cassiopeia usando Média Acumulada do Acoplamento para C_3.

Os resultados da Figura 19 demonstram que mesmo com a combinação dos *corpus*, a média acumulada da coesão não apresentou uma melhora em comparação aos números de C_2, apesar dos valores ficarem bem próximos

Tabela 8: Resultados obtidos pelo Modelo Cassiopeia usando Média Acumulada da Coesão C_3.

Média Acumulada da Coesão	
Cálculos	O que devo saber antes de usar e Como este medicamento funciona
Variância	0.0000292245
Média	0.4056

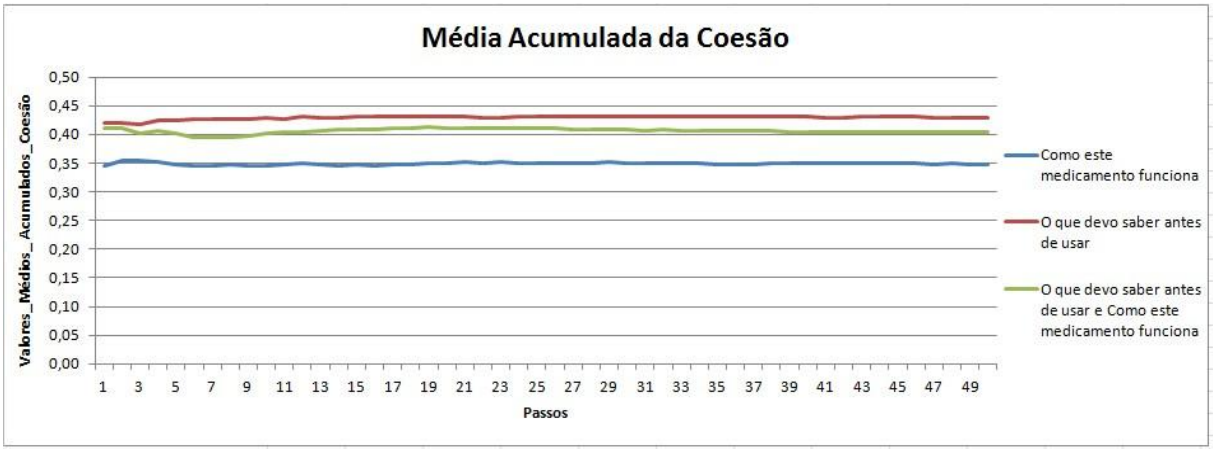

Figura 19: Resultados obtidos pelo Modelo Cassiopeia usando Média Acumulada da Coesão para C_3.

Os resultados da Figura 20 demonstram também que mesmo com a combinação dos *corpus*, a média acumulada do coeficiente de silhouette não apresentou uma melhora em comparação aos números de C_2, os valores foram praticamente os mesmos.

Tabela 9: Resultados obtidos pelo Modelo Cassiopeia usando Média Acumulada do Coeficiente de Silhouette para C_3.

Média Acumulada do Coeficiente de Silhouette	
Cálculos	O que devo saber antes de usar e Como este medicamento funciona
Variância	0
Média	0.98

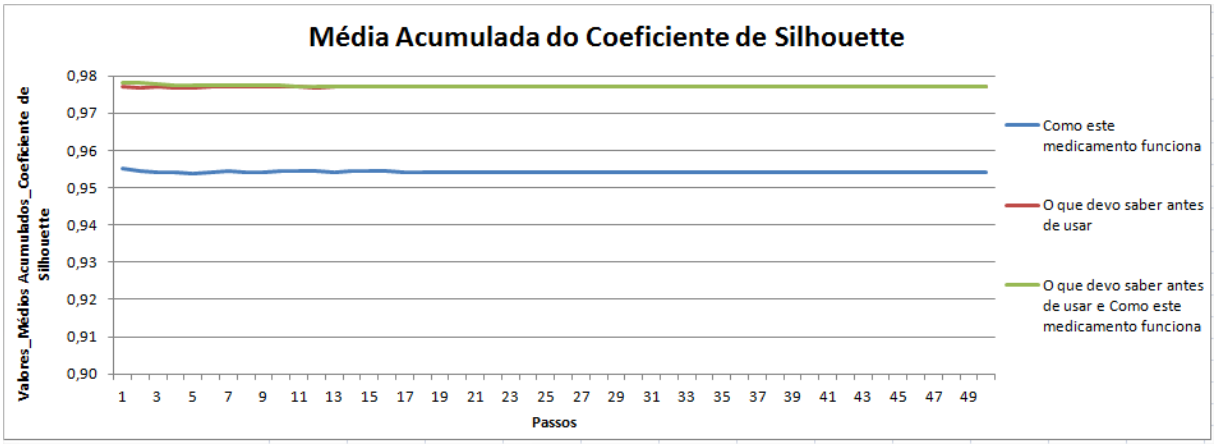

Figura 20: Resultados obtidos pelo Modelo Cassiopeia usando Média Acumulada do Coeficiente de Silhouette para C_3.

5. DISCUSSÃO DOS RESULTADOS

Neste capítulo, algumas discussões serão descritas a partir de análises dos resultados obtidos no Capítulo 4.

Para discussão dos resultados, foram utilizadas as Tabelas 1, 2 e 3 da Seção 3.2.

Nas estatísticas do C_3, o número de palavras apresentou um total de 350.546 que corresponde à soma das palavras de C_1 + C_2. De acordo com as Figuras 18, 19 e 20, do Capítulo 4, seus valores de acoplamento não melhoraram significativamente, os de coesão apresentaram valores um pouco menores e em coeficiente de silhouette os valores foram praticamente iguais em relação ao C_2. A diferença de palavras entre o C_3 com o C_2, é igual ao número de palavras no C_1, 53.034.

A partir dessa análise, pode-se perceber que a diferença de palavras entre o C_2 e o C_1, apresentam números maiores do que a diferença entre o C_3 e o *corpus* com os melhores resultados, C_2. Com isso, para os valores superarem aos melhores de acoplamento, coesão e coeficiente de silhouette significativamente, não basta aumentar o número de palavras em um processamento como foi feito em C_3, as palavras precisam ter qualidade. No presente trabalho, as palavras dentro das BM acrescentadas à um *corpus* precisam conter algum tipo de similaridade com as bulas que já estão presentes.

Como o C_1 contém informações de como um determinado medicamento funciona e o C_2 informações sobre o que o paciente deve saber antes de usar, a junção em C_3 não foi o suficiente para melhorar os resultados em comparação ao C_2, que contém números menores de palavras, porém com conteúdos similares.

Serão demonstradas na Figura 21 e na Figura 22 as partes “Como este medicamento funciona” e “O que devo saber antes de usar”, da bula do medicamento Neosaldina.

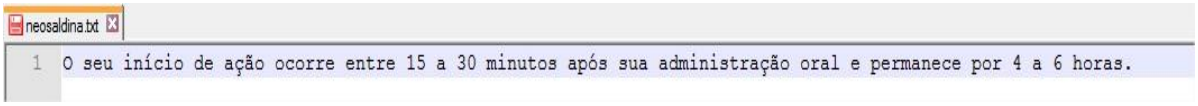

Figura 21: “Como este medicamento funciona” da bula Neosaldina.

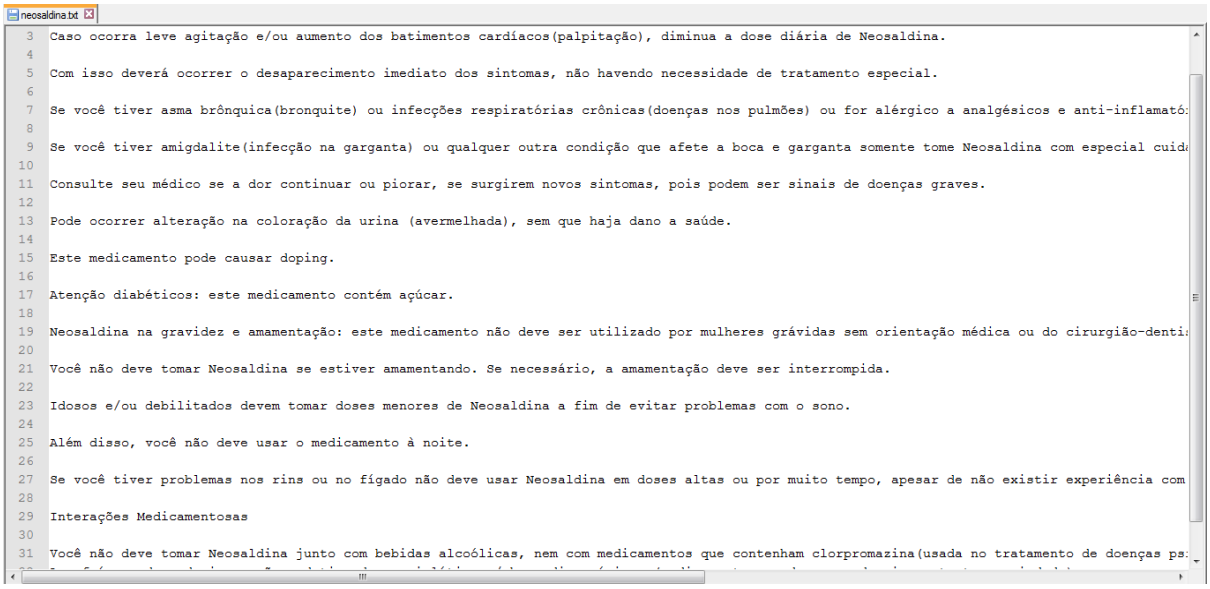

Figura 22: “O que devo saber antes de usar” da bula Neosaldina.

Através dessas figuras, o objetivo foi mostrar a diferença entre os números de palavras e as informações contidas nas duas partes das bulas de medicamento e reforçar o porquê do processamento de C_3 não ter superado os resultados de C_2.

Com isso, pode-se assimilar os bons resultados do processamento no Modelo Cassiopeia, não só à quantidade de palavras, mas também à qualidade e similaridade das informações nas BM.

6. CONCLUSÃO

Em um contexto no qual grande parte das informações contidas na Internet e nas organizações está em formato textual, faz-se necessário o desenvolvimento de técnicas computacionais para a organização destas bases textuais e a exploração do conhecimento nelas contido. Para tal fim, tarefas eficazes e eficientes de organização do conhecimento textual podem ser aplicadas.

No presente trabalho, a técnica computacional de organização de bases textuais utilizada foi a do algoritmo de Agrupamento Textual do Modelo Cassiopeia, sobre BM. O Modelo Cassiopeia, utiliza um algoritmo de TM que tem como principal finalidade gerar agrupamentos, ou seja, *clusters* (grupos) de documentos textuais que apresentam algum tipo similaridade.

No decorrer do desenvolvimento deste trabalho, foi descrita uma metodologia onde o principal objetivo foi analisar os agrupamentos textuais gerados no processamento do Modelo Cassiopeia, através das métricas internas ou não supervisionadas.

A partir das análises dos valores das métricas internas (acoplamento, coesão e coeficiente de silhouette), o Modelo Cassiopeia apresentou valores pertinentes quanto ao número e qualidade de palavras dos *clusters*. No C_1, onde as BM continham quantidades menores de palavras (53.034), seus valores das métricas internas foram visivelmente piores do que do C_2, que continha 297.512. Já no C_3 com 350.546 palavras (C_1+C_2), seus resultados não apresentaram melhora em comparação ao C_2, pois apesar de conter um número maior de palavras, as informações de C_1 acrescentadas com as de C_2 que poderiam melhorar ainda mais os resultados, não continham qualidade quanto à similaridade de informações.

Com isso, pode-se concluir que o Modelo Cassiopeia atendeu aos propósitos de pré-processamento, processamento e pós processamento, fornecendo valores de acoplamento, coesão e coeficiente de silhouette satisfatórios em relação à quantidade de palavras e a qualidade dessas dentro de um *cluster.*

A área de pesquisa envolvendo técnica de TM é vasta e esta em constante

evolução. A riqueza dos textos e a complexidade dos problemas relacionados à linguagem e à dimensão dos dados são desafios sempre presentes quando se trata de descoberta de conhecimento em textos. No entanto, a área tende a continuar com seu crescimento rápido devido à enorme quantidade de documentos publicados diariamente na Internet, e pela necessidade de transformar as informações contidas nestes documentos em conhecimento útil e inovador.

6.1. CONTRIBUIÇÃO

Com o uso do Modelo Cassiopeia, tem-se como contribuição para futuros trabalhos a partir dos agrupamentos gerados, descoberta de novos conhecimentos implícitos em BM.

6.2. DIFICULDADES E LIMITAÇÕES

No início, as pesquisas referentes à fundamentação teórica do trabalho, eram voltadas para os processos KDT. Porém, houve dificuldade de obter uma co-orientação de um profissional na área da saúde, para acompanhar e analisar os resultados no pós-processamento. Com isso, a fundamentação teórica voltou-se para a Descoberta de Conhecimento por Agrupamento Textual, ou seja, possibilitar a descoberta de conhecimento através dos agrupamentos criados e analisados por métricas internas.

Outra parte do trabalho que começou a ser desenvolvida e após limitações de prazo dificultou a continuação, foram as etapas de Seleção do *Corpus* e Pré Processamento. No início, foram coletadas 2000 BM para pacientes e mais 2000 para profissionais da saúde pelo site da Anvisa. Essas bulas se encontravam no formato PDF e para que elas pudessem ser processadas, foi preciso transformá-las em formato .txt. Com isso, uma grande parte do tempo de desenvolvimento, foi consumida. Logo após houve problemas em processar essas bulas no formato .txt, consumindo mais tempo para encontrar os erros. Foi decidido então, começar o

processo de Seleção do *Corpus* e Pré Processamento do início e de uma forma diferente para que não houvesse mais perda de tempo.

No trabalho foram usados uma *corpora* de apenas 707 BM para pacientes, formando-se três *corpus*. Também, apenas as métricas internas foram usadas, por não se conseguir um especialista na área farmacêutica para acompanhar o trabalho. Acredita-se que estes sejam fatores limitantes, considerando que as BM poderiam ser em número maior, com maior diversificação de *corpus* e considerando também que os valores poderiam ser melhores analisados utilizando mais métricas.

6.3. TRABALHOS FUTUROS

Com a dificuldade e limitação de se conseguir um especialista e com os agrupamentos já gerados no processamento do Modelo Cassiopeia, surgiu a possibilidade de trabalhos futuros, utilizando KDT por um profissional da saúde interessado em encontrar informações novas e implícitas nas BM dentro de cada *cluster*. Com isso, as métricas externas também seriam utilizadas.

Outra possibilidade de trabalhos futuros é a utilização de uma *corpora* maior, com mais BM coletadas no site Anvisa e com mais c*orpus*. No Capítulo 5, a quantidade e qualidade de palavras precisam existir para que os valores de acoplamento, coesão e coeficiente de silhouette, fossem significativos. Com isso, a combinação de mais *corpus* das Bulas do Paciente, Bulas do Profissional de Saúde e até a combinação das duas classes juntas, poderiam ser feitas para que os resultados melhorassem. Por exemplo, combinando *corpus* com informações de “Indicação” e “Contra Indicação”, os valores apresentariam uma melhoria em relação a apenas um desses *corpus* processado sozinho, pois a combinação de mais palavras e com conteúdo similar, ajudariam a melhorar os valores.

A Figura 23, é uma sugestão de esquema da *corpora* para trabalhos futuros.

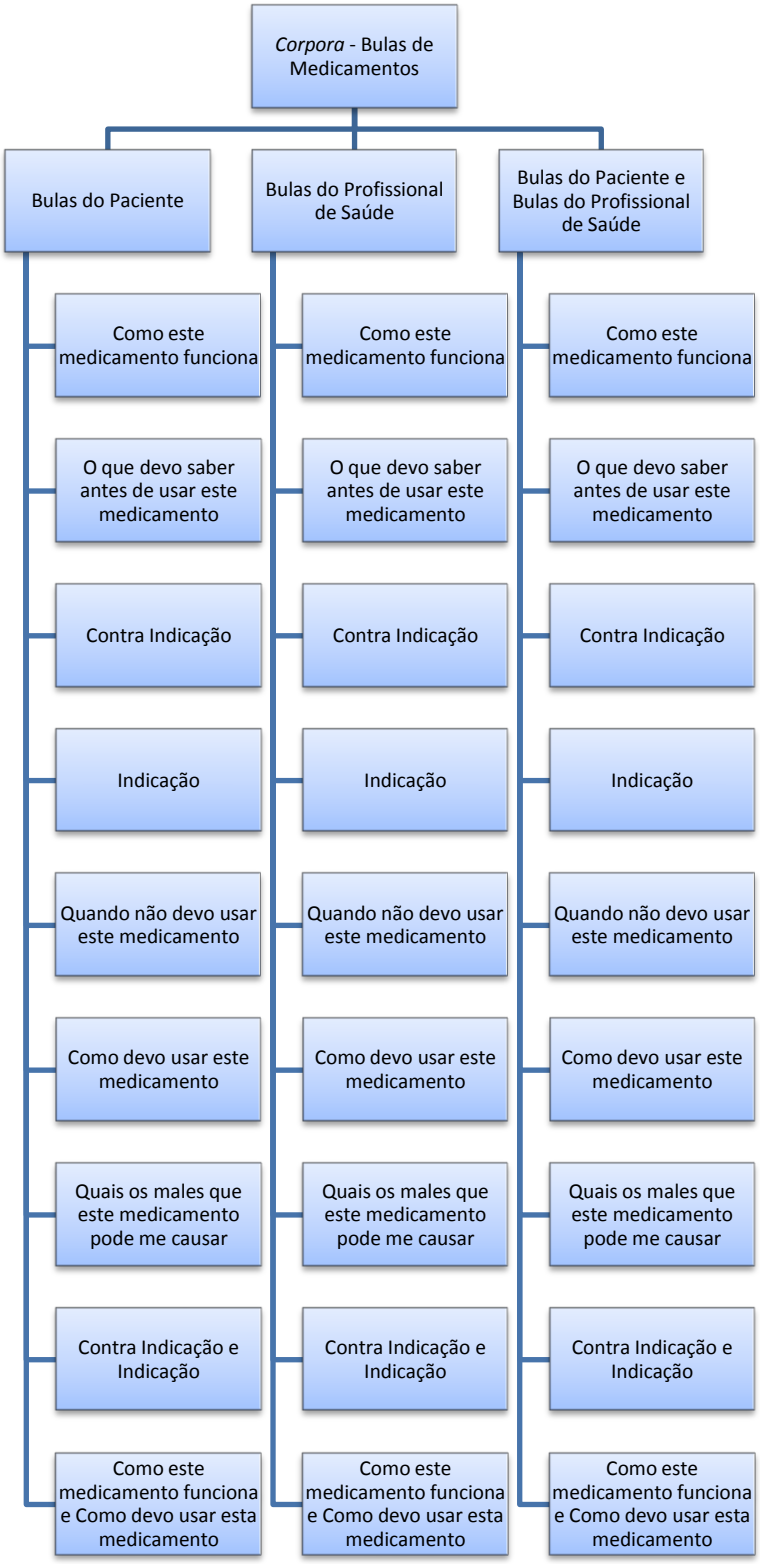

Figura 23: Esquema da *corpora* para trabalhos futuros.

REFERÊNCIAS

1. CARDOSO, Diogo *et al.* **Descoberta de Conhecimento em Texto – Análise Semântica.** Universidade Federal de Santa Catarina, Departamento de Informática e Estatística, Florianópolis, 2012.

2. CASTRO, Flaviana S.; SIMÕES, Priscyla W. T. A.; MATTOS, Merisandra C. **Mineração de Textos na Saúde por meio da Utilização da Ferramenta Eurekha**. Revista de Iniciação Científica (Criciúma). v. 7, p. 1, 2009.

3. FELDMAN, R. & DAGAN, I. (1998). **Mining Text using keyword distributions".** Journal of Intelligent Information Systems, v.10, n.3, pp. 281-300.

4. GUELPELI, Marcus 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.

5. LINDEN, R. **Técnicas de agrupamento**. Revista de Sistemas de Informação da FSMA, 2009. Disponível em: <http://www.fsma.edu.br/si/edicao4/FSMA_SI_2009_2_Tutorial.pdf>. Acesso em: 6 jan. 2014.

6. LOH, Stanley. **Abordagem Baseada em Conceitos para Descoberta de Conhecimento em Textos.** 2001. 195 f. Tese (Doutorado) - Curso de Programa de Pós Graduação em Computação, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2001. Cap. 8.

7. MIRANDA, Aline R. O. **DESCOBERTA DE CONHECIMENTO EM TEXTO APLICADA A UM SISTEMA DE ATENDIMENTO AOS USUÁRIOS DE UM PLANO DE ASSISTÊNCIA À SAÚDE.** 2009. 93 f. Dissertação (Mestrado) -

Curso de Engenharia Civil, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2009. Cap. 5.

8. RAMOS, Hélia S. C.; BRASCHER, Marisa. **Aplicação da descoberta de conhecimento em textos para apoio à construção de indicadores informétricos para a área de C&T**. Ciência da Informação, Brasília, v.38, n.2, p.56-68, maio/ago. 2009.

9. SANTOS, Daniela S. **Bee clustering : um algoritmo para agrupamento de dados inspirado em inteligência de enxames.** 2009. 89 f. Dissertação (Mestrado) - Curso de Ciência da Computação, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2009. Cap. 7.

10. SILVA, Edna R. G.; ROVER, Aires J. **O Processo de descoberta do conhecimento como suporte à análise criminal: minerando dados da Segurança Pública de Santa Catarina**. In: International Conference on Information Systems and Technology Management, 2011, São Paulo. Anais da International Conference on Information Systems and Technology Management. São Paulo: FEA, 2011. v. 8.

11. SILVA, Thales N. **Uma Arquitetura para Descoberta de Conhecimento a partir de Bases Textuais.** 2012. 78 f. TCC (Graduação) - Curso de Tecnologias da Informação e Comunicação, Universidade Federal de Santa Catarina, Araranguá, 2012. Cap. 6.

12. WIVES, Leandro K. **Utilizando conceitos como descritores de textos para o processo de identificação de conglomerados (clustering) de documentos.** 2004. 136 f. Tese (Doutorado) - Curso de Ciência da Computação, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2012. Cap. 6.

13. WIVES, Leandro K.; LOH, Stanley. **Tecnologias de descoberta de conhecimento em informações textuais (ênfase em agrupamento de**

informações). In: Oficia de Inteligência Artificial (OIA). Pelotas, RS. EDUCAT, 2006. p.28-48.

APÊNDICE A

O Apêndice A mostra os outros gráficos gerados para análises no pós-processamento. Gráficos dos valores acoplamento, coesão e coeficiente de silhouette, dos *corpus* C_1, C_2 e C_3.

VALORES DE ACOPLAMENTO, COESÃO E COEFICIENTE DE SILHOUETTE DO *CORPUS* C_1.

A Figura 24 mostra o gráfico dos resultados de acoplamento de C_1, obtidas nos agrupamentos de textos ao longo das 50 interações no Modelo Cassiopeia.

Figura 24: Valores de acoplamento do C_1.

A Figura 25 mostra o gráfico dos resultados de coesão de C_1, obtidas nos agrupamentos de textos ao longo das 50 interações no Modelo Cassiopeia.

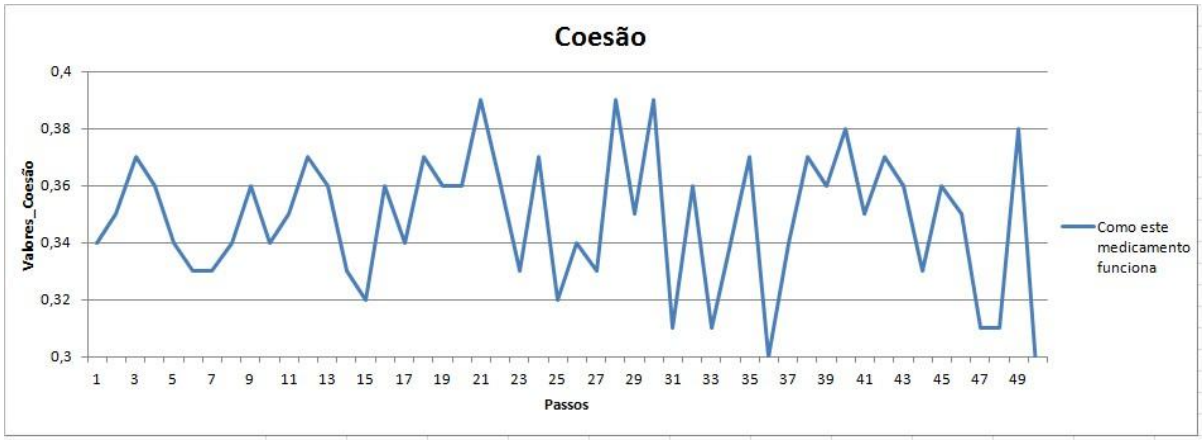

Figura 25: Valores de coesão do C_1.

A Figura 26 mostra o gráfico dos resultados de coeficiente de silhouette de C_1, obtidas nos agrupamentos de textos ao longo das 50 interações no Modelo Cassiopeia.

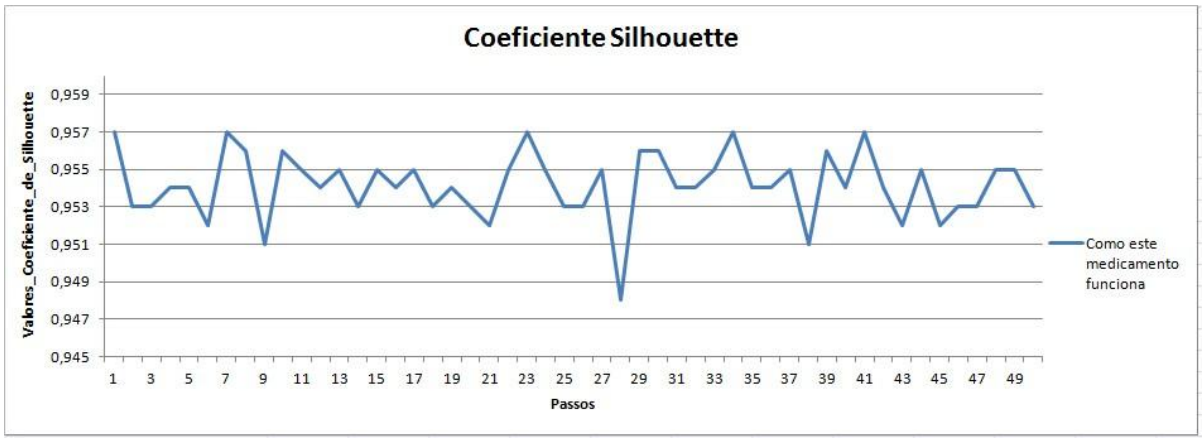

Figura 26: Valores de coeficiente de silhouette do C_1.

A Figura 27 mostra o gráfico dos resultados de acoplamento, coesão e coeficiente de silhouette de C_1, obtidas nos agrupamentos de textos ao longo das 50 interações no Modelo Cassiopeia.

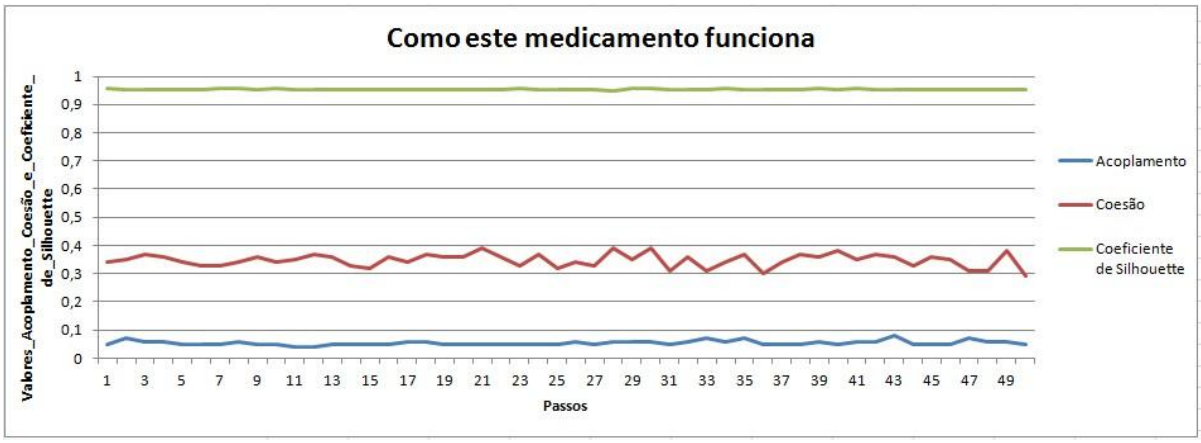

Figura 27: Valores de acoplamento, coesão e coeficiente de silhouette do C_1.

VALORES DE ACOPLAMENTO, COESÃO E COEFICIENTE DE SILHOUETTE DO *CORPUS* C_2.

A Figura 28 mostra o gráfico dos resultados de acoplamento de C_2, obtidas nos agrupamentos de textos ao longo das 50 interações no Modelo Cassiopeia.

Figura 28: Valores de acoplamento do C_2.

A Figura 29 mostra o gráfico dos resultados de coesão de C_2, obtidas nos agrupamentos de textos ao longo das 50 interações no Modelo Cassiopeia.

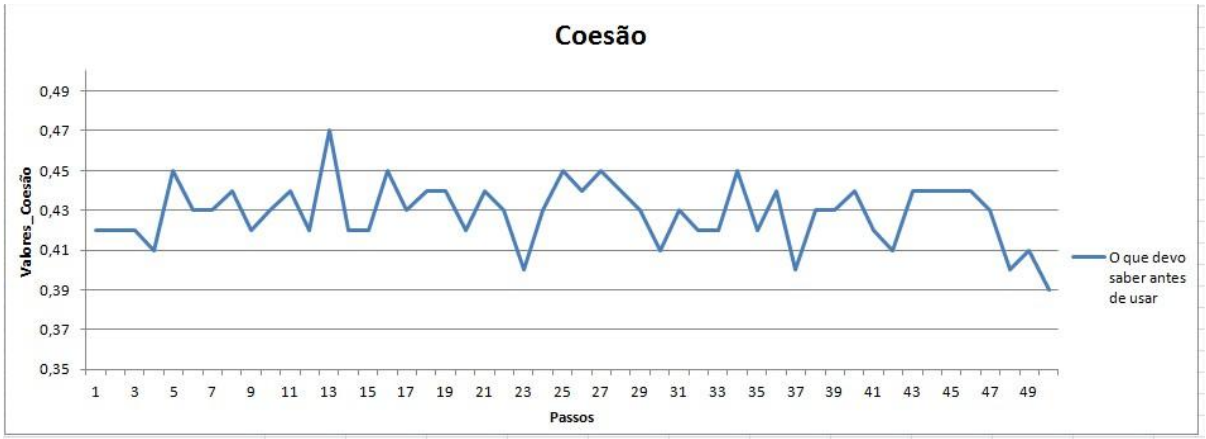

Figura 29: Valores de coesão do C_2.

A Figura 30 mostra o gráfico dos resultados de coeficiente de silhouette de C_2, obtidas nos agrupamentos de textos ao longo das 50 interações no Modelo Cassiopeia.

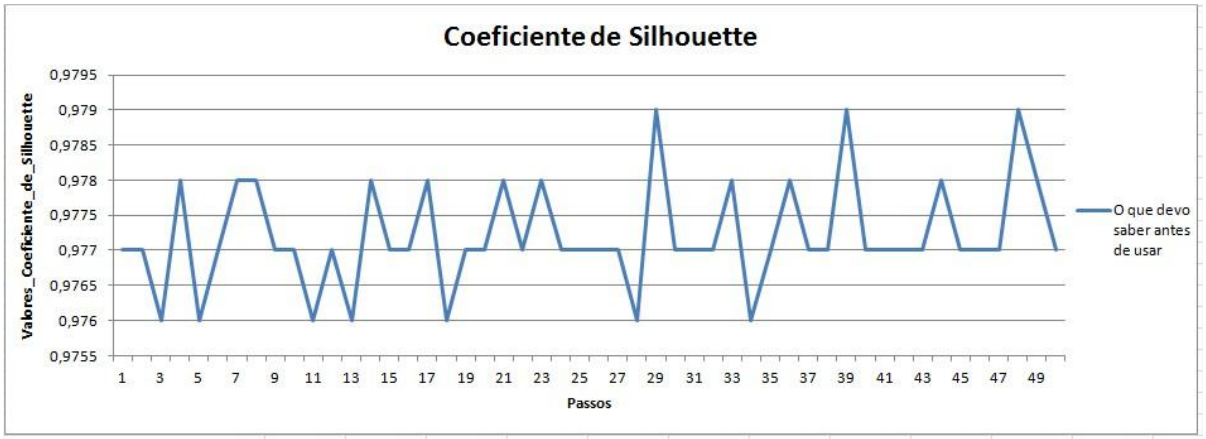

Figura 30: Valores de coeficiente de silhouette do C_2.

A Figura 31 mostra o gráfico dos resultados de acoplamento, coesão e coeficiente de silhouette de C_2, obtidas nos agrupamentos de textos ao longo das 50 interações no Modelo Cassiopeia.

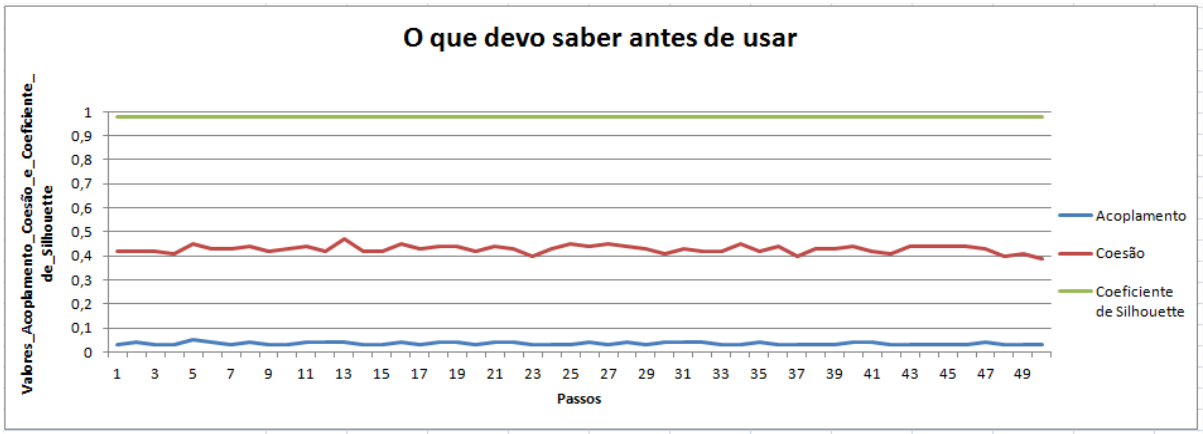

Figura 31: Valores de acoplamento, coesão e coeficiente de silhouette do C_2.

VALORES DE ACOPLAMENTO, COESÃO E COEFICIENTE DE SILHOUETTE DO *CORPUS* C_3.

A Figura 32 mostra o gráfico dos resultados de acoplamento de C_3, obtidas nos agrupamentos de textos ao longo das 50 interações no Modelo Cassiopeia.

Figura 32: Valores de acoplamento do C_3.

A Figura 33 mostra o gráfico dos resultados de coesão de C_3, obtidas nos agrupamentos de textos ao longo das 50 interações no Modelo Cassiopeia.

Figura 33: Valores de coesão do C_3.

A Figura 34 mostra o gráfico dos resultados de coeficiente de silhouette de C_3, obtidas nos agrupamentos de textos ao longo das 50 interações no Modelo Cassiopeia.

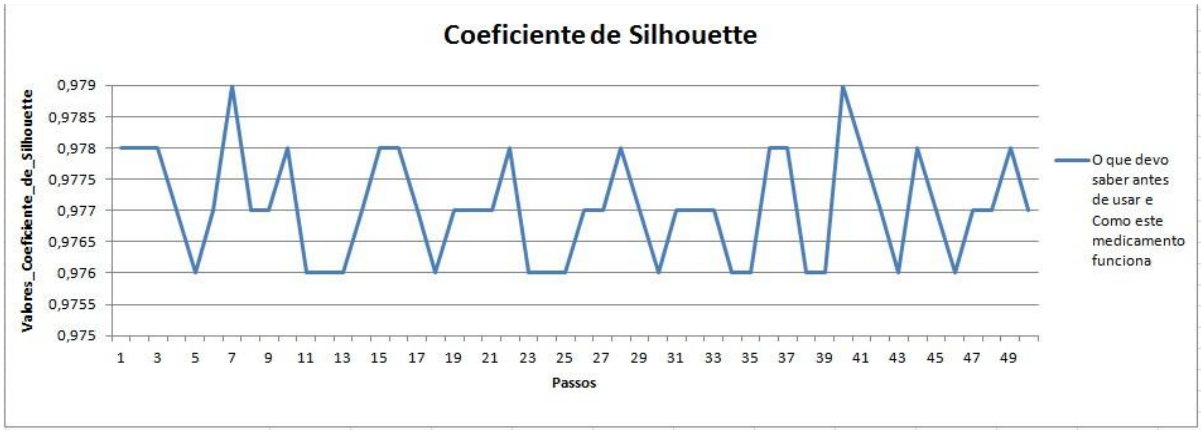

Figura 34: Valores de coeficiente de silhouette do C_3.

A Figura 35 mostra o gráfico dos resultados de acoplamento, coesão e coeficiente de silhouette de C_3, obtidas nos agrupamentos de textos ao longo das 50 interações no Modelo Cassiopeia.

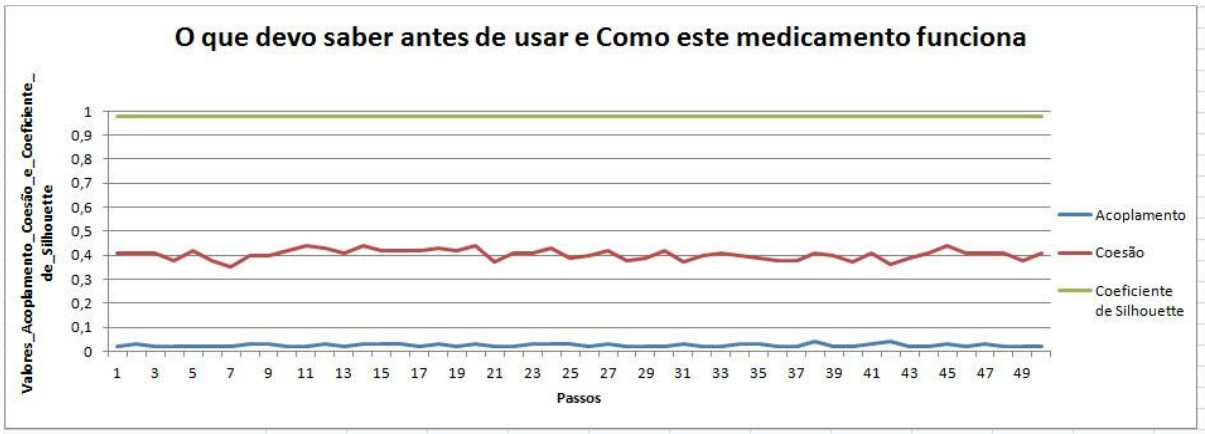

Figura 35: Valores de acoplamento, coesão e coeficiente de silhouette do C_3.

VALORES DE ACOPLAMENTO, COESÃO E COEFICIENTE DE SILHOUETTE DOS *CORPUS* C_1, C_2 e C_3.

A Figura 36 mostra o gráfico dos resultados de acoplamento de C_1, C_2 e C_3, obtidas nos agrupamentos de textos ao longo das 50 interações no Modelo Cassiopeia.

Figura 36: Valores de acoplamento do C_1, C_2 e C_3.

A Figura 37 mostra o gráfico dos resultados de coesão de C_1, C_2 e C_3, obtidas nos agrupamentos de textos ao longo das 50 interações no Modelo Cassiopeia.

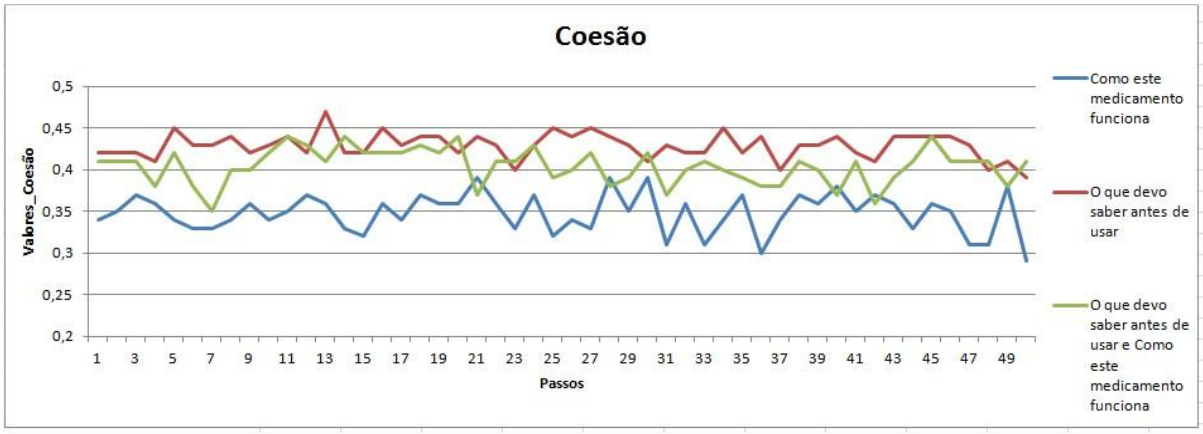

Figura 37: Valores de coesão do C_1, C_2 e C_3.

A Figura 38 mostra o gráfico dos resultados de coeficiente de silhouette de C_1, C_2 e C_3, obtidas nos agrupamentos de textos ao longo das 50 interações no Modelo Cassiopeia.

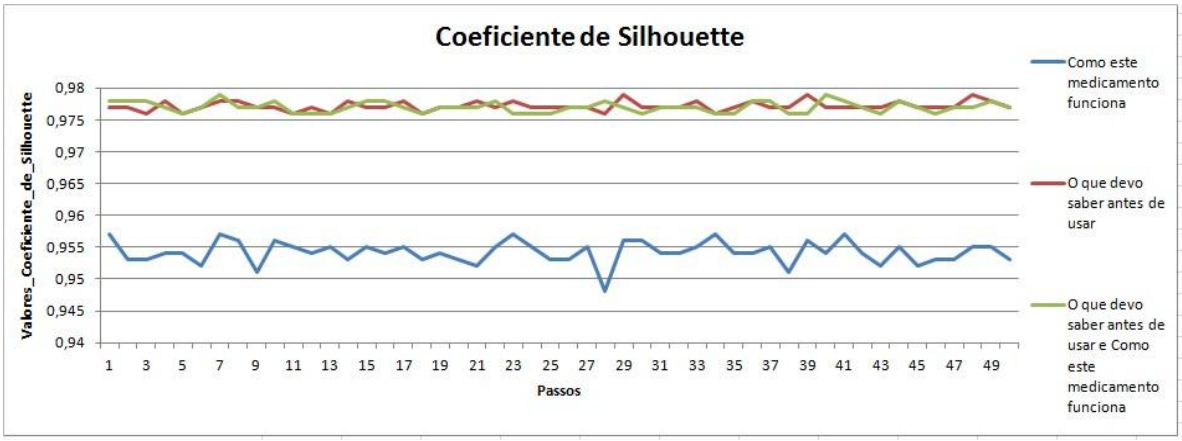

Figura 38: Valores de coeficiente de silhouette do C_1, C_2 e C_3.