UNIVERSIDADE FEDERAL DOS VALES DO JEQUITINHONHA E MUCURI
DEPARTAMENTO DE COMPUTAÇÃO
CURSO DE GRADUAÇÃO EM SISTEMAS DE INFORMAÇÃO

COMPORTAMENTO DO MODELO CASSIOPEIA COM O USO DE SUMÁRIOS HUMANOS NOS IDIOMAS PORTUGUÊS E INGLÊS

Jésyka Milleny Azevedo Gonçalves

Diamantina - MG
2013

UNIVERSIDADE FEDERAL DOS VALES DO JEQUITINHONHA E MUCURI
DEPARTAMENTO DE COMPUTAÇÃO
CURSO DE GRADUAÇÃO EM SISTEMAS DE INFORMAÇÃO

COMPORTAMENTO DO MODELO CASSIOPEIA COM O USO DE SUMÁRIOS HUMANOS NOS IDIOMAS PORTUGUÊS E INGLÊS

Autor:
Jésyka Milleny Azevedo Gonçalves

Orientador:
Marcus Vinícius Carvalho Guelpeli

Trabalho de Conclusão de Curso apresentada ao Curso de Sistemas de Informação da Universidade Federal dos Vales do Jequitinhonha e Mucuri – UFVJM, como parte dos requisitos exigidos para a obtenção do título de Bacharel em Sistemas de Informação.

Diamantina - MG
2013

COMPORTAMENTO DO MODELO CASSIOPEIA COM O USO DE SUMÁRIOS HUMANOS NOS IDIOMAS PORTUGUÊS E INGLÊS

Jésyka Milleny Azevedo Gonçalves

Trabalho de Conclusão de Curso apresentada ao Curso de Sistemas de Informação da Universidade Federal dos Vales do Jequitinhonha e Mucuri – UFVJM, como parte dos requisitos exigidos para a obtenção do título de Bacharel em Sistemas de Informação.

APROVADO em:_____/_____/_____

_____
Profª. Caroline Queiroz Santos – UFVJM

_____
Profª. Geruza de Fatima Tomé – UFVJM

_____
Profº. Marcus Vinícius Carvalho Guelpeli- UFVJM

Dedico este trabalho às pessoas mais importantes da minha vida: meus pais, Kincas e Lu, aos meus irmãos, Hugo e Bruna e a meu noivo Wesley, que confiaram no meu potencial para esta conquista. Saibam que não conquistaria nada se não estivessem ao meu lado. Amo vocês!

Jesyka Milleny Azevedo Gonçalves

AGRADECIMENTOS

Como já dizia Anitelli: "*Sonho parece verdade quando a gente esquece de acordar*".

Hoje, vivo uma realidade que parece um sonho, mas foi preciso muito esforço, determinação, paciência, perseverança, ousadia e maleabilidade para chegar até aqui, e nada disso eu conseguiria sozinha. Minha eterna gratidão a todos aqueles que colaboraram para que este sonho pudesse ser concretizado.

Grata a Deus pelo dom da vida, pelo seu amor infinito, sem Ele nada sou. Agradeço aos meus pais, Kincas e Lu, meus maiores exemplos. Obrigada por cada incentivo e orientação, pela amizade, pela garra, pelas orações em meu favor, pela preocupação para que estivesse sempre andando pelo caminho correto. Vocês são incríveis!!! Falta-me a melhor forma de agradecê-los.

Aos meus irmãos, Hugo e Bruna, por todo amor e carinho. A função de irmã mais velha, que protege e ampara, não me cai bem frente à admiração que tenho por vocês. À Lidiane, minha cunhada linda! Te agradeço tanto pela amizade e confiança presente entre nós…Vocês fazem os meus dias mais alegres!!!

Ao meu noivo,Wesley, e sua família, por todo amor, carinho, paciência e compreensão a mim dedicados. Você é pra mim, Lelyn, um exemplo de pessoa a ser seguida! Agradeço a Deus todos os dias por ter me presenteado ao colocá-lo em minha vida. Esta caminhada não seria a mesma sem você.

Ao meu professor e orientador, Marcus Guelpeli que, com muita paciência e atenção, dedicou do seu valioso tempo para me orientar em cada passo deste trabalho. Aprendi tanta coisa com você neste tão pouco tempo que trabalhamos juntos! Estou certa de que a minha formação acadêmica e pessoal não estaria completa sem os seus diálogos e conselhos… os levarei para sempre comigo.Você é admirável!

À minha professora Caroline Queiroz pela contribuição na minha vida acadêmica e por tanta influência na minha vida profissional. Depois que nos conhecemos, várias portas se abriram, e a maioria delas, foi por você ter confiado em mim. Muito Obrigada!

À todos os meus professores de graduação, pelos ensinamentos passados. À minha família, aos meus colegas de classe e aos meus amigos que me apoiaram e me fizeram saber que, antes de tudo, a vida sem vocês não teria graça alguma.

Obrigada a todos que, mesmo não estando citados aqui, tanto contribuíram para a conclusão desta etapa que tanto almejei.

RESUMO

Este estudo propõe a modificação na etapa de pré-processamento do modelo Cassiopeia, onde originalmente foi usado a sumarização automática. Neste trabalho foi utilizado sumários humanos que foram comparados com sumários automáticos. Foi verificado qual proporciona, ao Cassiopeia, um resultado para melhor agrupar os textos. A análise é realizada com os domínios , médico e jurídico, nos idiomas português e inglês. O desempenho será medido em termos das métricas internas e externas e comprovado através de testes estatísticos. Ao final tem-se os resultados da eficiência ou não do uso de sumários humanos no agrupamento do modelo Cassiopeia.

Palavras-Chave: Sumarização, Agrupamento, Clusterização, Cassiopeia, Métricas Internas, Métricas Externas, Testes Estatísticos.

ABSTRACT

This study proposes the modification in the pre-processing model Cassiopeia, which was originally used in automatic summarization. In our study, the use of human summaries, comparing and checking which provides, in Cassiopeia, a result with the best grouping. The analysis is performed with the same areas of the original model, medical and legal, in Portuguese and English. The performance will be measured in terms of internal and external metrics and proven by statistical tests. At the end of the results is the efficiency of the use or non-human in cluster summaries model Cassiopeia.

Keywords: Summarization, Grouping, Clustering, Cassiopeia, Metric Internal, External Metrics, Statistical Tests.

LISTA DE ILUSTRAÇÕES

Figura 1: Diferentes tipos de agrupamentos……19

Figura 2: Tipos de agrupamentos quanto a sua posição……..19

Figura 3: Curva de Zipf……25

Figura 4: Curva de Zipf com os Cortes de Luhn……..26

Figura 5: Modelo Cassiopeia……31

Figura 6: Seleção de Atributos no Cassiopeia……..34

Figura 7: Dendograma do Método Hierárquico Aglomerativo……36

Figura 8: Agrupamento pelo Método *"Cliques*".....37

Figura 9: Adaptação ao Modelo Cassiopeia……37

Figura 10: Regra de três simples para obtenção do percentual de compressão……40

Figura 11: Modelo de Sumarização e Agrupamento gerados para o domínio Médico no idioma Portugues……..41

Figura 12: Modelo de Sumarização e Agrupamento gerados para o domínio Médico no idioma Portugues……..42

Figura 13: Modelo de Sumarização e Agrupamento gerados para o domínio Médico no idioma Portugues……..43

Figura 14: Resultados obtidos pelo modelo Cassiopeia usando a medida externa F-Measure no idioma Português do domínio Jurídico……44

Figura 15: Resultados obtidos pelo modelo Cassiopeia usando a medida interna Coeficiente Silhouette no idioma Português do domínio Jurídico .…….……..…44

Figura 16: Resultados obtidos pelo modelo Cassiopéia usando a medida externa F-Measure no idioma Português do domínio Médico.……..…..45

Figura 17: Resultados obtidos pelo modelo Cassiopeia usando a medida interna Coeficiente Silhouette no idioma Português do domínio Médico……...…….……45

Figura 18: Resultados obtidos pelo modelo Cassiopeia usando a medida externa F-Measure no idioma Inglês do domínio Médico…...……46

Figura 19: Resultados obtidos pelo modelo Cassiopeia usando a medida interna Coeficiente Silhouette no idioma Inglês do domínio Médico. ……..…….46

LISTA DE TABELAS

Tabela 1. Teste Estatístico dos resultados usando a medida interna Coeficiente Silhouette no domínio Jurídico no idioma Portugues……48

Tabela 2. Teste Estatístico dos resultados usando a medida extena F-Measure no domínio Jurídico no idioma Portugues……...……48

Tabela 3. Teste Estatístico dos resultados usando a medida interna Coeficiente Silhouette no domínio Médico no idioma Portugues……..…49

Tabela 4. Teste Estatístico dos resultados usando a medida extena F-Measure no domínio Médico no idioma Portugues……...…….49

Tabela 5. Teste Estatístico dos resultados usando a medida interna Coeficiente Silhouette no domínio Médico no idioma Inglês……50

Tabela 6. Teste Estatístico dos resultados usando a medida extena F-Measure no domínio Médico no idioma Inglês……..50

LISTA DE FIGURAS

Figura 1: Resultados obtidos pelo modelo Cassiopeia usando a medida externa F-Measure no idioma Português do domínio Jurídico…….44

Figura 2: Resultados obtidos pelo modelo Cassiopeia usando a medida interna Coeficiente Silhouette no idioma Português do domínio Jurídico .…….……..…..….44

Figura: Resultados obtidos pelo modelo Cassiopéia usando a medida externa F-Measure no idioma Português do domínio Médico.……..…..45

Figura 4: Resultados obtidos pelo modelo Cassiopeia usando a medida interna Coeficiente Silhouette no idioma Português do domínio Médico……...……45

Figura 5: Resultados obtidos pelo modelo Cassiopeia usando a medida externa F-Measure no idioma Inglês do domínio Médico…...…….46

Figura 6: Resultados obtidos pelo modelo Cassiopeia usando a medida interna Coeficiente Silhouette no idioma Inglês do domínio Médico. ……..…….46

LISTA DE ABREVIATURAS E SIGLAS

SA	**S***umário* **A***utomático*
SH	**S***umário* **H***umano*
RI	**R***ecuperação de* **I***nformação em Texto*
MT	**M***ineração de* **T***exto*

SUMÁRIO

CAPÍTULO 1 - INTRODUÇÃO 12

1.1 Motivação ..... 14~~15~~

1.2 Problema ..... 14~~15~~

1.3 Hipótese 15

1.4 Contribuição 15

1.5 Metodologia da Pesquisa 15

1.6 Estrutura da Proposta 16

CAPÍTULO 2 – FUNDAMENTAÇÃO TEÓRICA 17

2.1 Agrupamento 17

2.2 Métricas para análise de agrupamentos textuais 20

2.2.1 Métricas Externas ..... 20~~21~~

2.2.2 Métricas Internas 22

2.3 Problema da Alta Dimensionalidade 23

2.3.1 Lei de Zipf 24

2.3.2 Corte de Luhn 25

2.4 Sumarização 26

2.4.1 Sumarização automática 27

2.4.2 Sumárização Humana 28

2.5 Testes Estatísticos ..... 28~~29~~

CAPÍTULO 3 – MODELOS 30

3.1 Funcionamento do Modelo Cassiopeia 30

3.1.1 Etapa de Pré-Processamento 31

3.1.2 Etapa de Processamento 31

3.1.3 Etapa de Pós-Processamento 35

3.2 Modelo Proposto 36

CAPÍTULO 4 – METODOLOGIA 37

4.1 Corpus 37

4.2 Sumarizadores 37

4.2.1 Copernic Summarizer 37

4.2.2 Intellexer Summarizer Pro 38

4.2.3 Supor 2 38
4.3 Cálculo da Compressão 38
4.4 Modelo de Sumarização e Agrupamento 40
CAPITULO 5 - RESULTADOS 42
5.1 Experimentos 42
5.1.1 Domínio Jurídico 42
5.1.2 Domínio Médico 44
5.2 Comprovação da Hipótese 46
5.3 Análise dos Testes Estatísticos 47
5.3.1 Teste estatístico para o domínio Jurídico 47
5.3.2 Teste estatístico para o domínio Médico 48
CAPITULO 6 – CONCLUSÕES 50
6.1 Contribuições 50
6.2 Limitações 51
6.3 Trabalhos Futuros 51
REFERÊNCIAS BIBLIOGRÁFICAS 52
APÊNDICES 57
APÊNDICE A 58
APÊNDICE B 67
APÊNDICE C 74

CAPÍTULO 1 - INTRODUÇÃO

Atualmente são muitas as informações textuais disponibilizadas na internet. Com o avanço das tecnologias e da alta velocidade de propagação de dados, tem-se valorizado muito a área de mineração de textos – MT, pois a dificuldade de processamento e assimilação de todas estas informações exige o aperfeiçoamento de técnicas de agrupamento de dados para fins de refinamento de pesquisas com intuito de trazer ao usuário resultados mais concisos e precisos, pois sabe-se que a capacidade humana de leitura e registro é limitada.

Ramos e Brascher (2009) afirmam que o crescimento acelerado da internet e a amplitude com que a informação é gerada e compartilhada pelos usuários possibilita o surgimento de uma nova dinâmica de reaproveitamento e produção de novos conhecimentos, pois se torna impossível a assimilação de todas as informações disponíveis. Sendo assim, selecionar as informações que melhor correspondem aos interesses do público facilita seu processamento e sua recuperação.

Chen (2001) afirma que 80% do conteúdo da internet esta em formato textual. Lyman (2000) afirma, em um levantamento realizado em 2000, que o repositório de conteúdo na internet duplicaria anualmente, e nessa época, o autor tinha estimado que no ano de 2000 estaria em torno de dois bilhões o número de páginas disponíveis. Porém Smyth et al. (2004) revelaram a existência de 10 bilhões de documentos no ano de 2004. De acordo com Gantz e Reinsel (2010), para o final de 2007 a estimativa foi de 487 exabytes de informação digital. De acordo com o relatório de Bohn et al. (2010), de 2007 a 2012 o crescimento foi de 5 vezes o total, chegando a 1.2 zettabytes. Verificou-se também que 90% das informações armazenadas por uma empresa eram também de dados não estruturados, ou seja, em formato textual (KUECHLER, 2007).

Levy (2005) acredita que o problema de se lidar com muita informação é que se perde um tempo que poderia ser bem melhor empregado pensando, refletindo ou raciocinando. A superação dos desafios de como obter conhecimento a partir desse excesso de informações pode significar vantagem competitiva para as instituições e a para as pesquisas.

Ao analisar estes e outros estudos, Guelpeli (2012) propôs um novo modelo para a sumarização e o agrupamento de textos, denominado Cassiopeia, contribuindo para área de Recuperação da Informação – RI, uma subárea da MT. Segundo Rezende *et al*. (2011), uma

coleção de textos pode ter milhares de termos que, em parte, são redundantes e pouco informativos para RI. Isso ocasiona uma solução computacional pouco eficiente, no momento de recuperação dos textos o que, segundo Rezende *et al*. (2011), torna o processo lento e com pouca qualidade.

Howland e Park (2007) afirmam que para atenuar a alta dimensionalidade e os dados esparsos é necessário reduzir o número de palavras, sendo crucial o tratamento dos dados no pré-processamento da RI. Uma solução adotada para atenuar esta questão é a eliminação das stopwords na etapa de pré-processamento. Esse procedimento é encontrado na maioria dos trabalhos, dentre os quais destacam-se (JONES e WILLET, 1997), (WIVES, 2004), (ARANHA, 2007), (NOGUEIRA, 2009), (OLIVEIRA, 2009) e (REZENDE *et al*., 2011). Segundo Aranha (2007), as *stopwords* são palavras que aparecem em todos os tipos de textos e não são capazes de colaborar para a recuperação de textos relativos a um assunto específico.

Wives (2004) afirma que mesmo após a retirada as *stopwords* ainda encontra-se textos com um número muito grande de palavras, sendo necessária então uma nova fase denominada seleção de atributos. Segundo Nogueira (2009), esta seleção é um fator decisivo para a boa qualidade do melhor desempenho para a RI.

Após a diminuição destes atributos, assim como outros autores, Feldman e Sanger (2006) propõem uma etapa de organização por meio da técnica de agrupamento de textos. Esta faz com que os textos do mesmo grupo tenham alta similaridade e sejam dissimilares aos documentos de outros grupos. De acordo com Loh (2001), Wives (2004) e Lopes (2011), esta técnica não deve sofrer intervenção humana, justificando-se o fato pelo grande volume de informações a ser analisado. Ao utilizar-se a estrutura hierárquica obtem-se uma organização *topdown*, ou seja, cada grupo superior apresenta textos com assuntos mais genéricos e os subgrupos, temas mais específicos (GUELPELI, 2012).

O trabalho de Guelpeli (2012) apresenta então o modelo Cassiopeia como um agrupador de textos hierárquico, que utiliza a sumarização de textos em seu pré-processamento para possibilitar um desempenho relativamente maior que outros modelos já descritos, visando melhorar a precisão na recuperação dos documentos, a coesão e acoplamento dos grupos de documentos formados, gerar agrupamentos a partir de domínios distintos, ou seja, ser independente do idioma. Nestes foram realizados testes, dentre eles com o uso de textos nos idiomas português e inglês, nos domínios jornalístico, médico e jurídico,

textos com ou sem *stopwords*, e com a utilização de vários sumarizadores automáticos diferentes para a comprovação da sua hipótese.

Ao estudar o trabalho de Guelpeli (2012), observa-se que não foi feito avaliações utilizando um sumário humano – SH.

A sumarização textual, como tarefa de produção de textos, possui uma restrição fundamental: a de transmitir a mensagem essencial, de forma concisa. A sumarização humana não se difere deste padrão, podendo ser definida como a tarefa de redução do tamanho de um texto-fonte, com a preservação íntegrade seu conteúdo mais relevante. Contudo a sumarização humana, apesar de ser a ideal, é cara, demorada, não produtiva e suscetível a erros e inconsistências devido ao julgamento humano, a avaliação de sistemas de sumarização é um tema que vem sendo muito investigado.

Foi desenvolvido ainda um trabalho por Delgado *et al*., (2012), que usou apenas o domínio jornalístico no idioma português e inglês. Surge então a questão: Com a utilização de SH em seu pré-processamento, existiria um ganho considerável em relação aos SA nos agrupamentos gerados pelo modelo Cassiopeia nos domínios jurídico e médico e nos idiomas português e inglês? Com base nesta indagação, este trabalho realizou testes com a utilização de SH e SA, a fim de obter resultados dos quais possam ser analisados e comparados, para que possa-se garantir a utilização do melhor sumário para o modelo Cassiopeia.

1.1 Motivação

Ao se estudar o modelo proposto por Guelpeli (2012) e ampliado por Delgado *et al.*, (2012) não encontraram avaliações do modelo Cassiopeia ao utilizar um sumário humano em seu pré-processamento nos domínios jurídico e médico e nos idiomas português e inglês. Acredita-se então, que possa ser realizada uma análise de seu comportamento mediante a uma nova abordagem de sua primeira etapa.

1.2 Problema

Não se sabe ao certo se os agrupamentos realizados pelo modelo Cassiopeia, proposto por Guelpeli (2012), são ou não mais eficientes quando são utilizados sumários humanos em

sua etapa de pré-processamento nos domínios jurídico e médico e nos idiomas português e inglês ao invés de sumários automáticos como proposta inicial pelo autor.

1.3 Hipótese

Os sumários humanos não são mais eficientes que os sumários automáticos para a geração de agrupamentos pelo modelo Cassiopeia.

1.4 Contribuição

A contribuição deste trabalho será para a área de Recuperação da Informação, pois pode-se garantir a utilização do melhor sumário, seja automático ou humano, para obter um melhor desempenho dos agrupamentos de textos gerados pelo modelo Cassiopeia.

1.5 Metodologia da Pesquisa

A metodologia adotada para realização desta pesquisa compreende: leitura bibliográfica, métodos quantitativos com testes de hipótese em bases públicas em inglês e português, análise do desempenho do agrupamento através das métricas externas e internas, visando dar suporte a toda análise realizada no trabalho, baseada na área de mineração de texto, dentro da subárea de recuperação de informação em texto, com foco em agrupamento de texto. Serão consultados (GUELPELI, 2012), (WIVES, 2004), (LOPES, 2004), (MARIA et al., 2008), (RIBEIRO, 2009) e (HOURDAKIS et al, 2010).

Também será apresentada uma adaptação do modelo Cassiopeia (GUELPELI, 2012), com uma alteração na sua etapa de pré-processamento. Esta etapa será detalhada no Capítulo 3. A metodológica será tratada no Capítulo 4. Serão aplicadas as métricas externas, com medidas de Recall, Precision e F-Measure, e métricas internas, com medidas de Coesão, Acoplamento e Coeficiente Silhouette, as quais serão descritas no Capítulo 5. Também serão apresentados testes estatísticos para o auxílio da comprovação da hipótese.

1.6 Estrutura da Proposta

Capítulo 2 – Fundamentação Teórica

O capítulo 2 apresenta o estudo da arte do Agrupamento de Textos e os principais conceitos dessa área, como Sumarização, Lei de Zipf e Corte de Luhn. Os quais são base de estudo para tal trabalho.

Capítulo 3 - Modelo

O Capítulo 3 apresenta o modelo Cassiopeia proposto por Guelpeli (2012), assim como sua funcionalidade em suas etapas de pré-processamento, processamento e pós-processamento. També é apresentado neste a adaptação ao modelo, em sua etapa de pré-processamento, com o uso da sumarização humana.

Capítulo 4 - Metodologia

O Capítulo 4 apresenta a metodologia adotada neste trabalho. Serão descritos os métodos de elaboração dos testes, o corpus utilizado nos experimentos, os sumarizadores utilizados, bem como os critérios para sua escolha, o método para calcular a compressão de cada texto com relação ao seu respectivo texto fonte e ainda o modelo de sumarização e agrupamento proposto neste trabalho.

Capitulo 5 - Resultados

O Capítulo 5 apresentará os resultados obitidos no experimento, será realizada a comprovação da hipótese, assim como a aplicação dos testes estatísticos e a exibição dos resultados obtidos.

Capítulo 6 - Conclusões

No Capítulo 6 serão discutidas as considerações alcançadas com os resultados obtidos, as limitações, contribuições e os trabalhos futuros.

CAPÍTULO 2 – FUNDAMENTAÇÃO TEÓRICA

Neste capítulo serão abordados os principais conceitos para fundamentação deste trabalho. Inicialmente, um conceito mais amplo de agrupadores e seus métodos seguidos de uma abordagem mais contextualizada, o de agrupadores de textos e suas propriedades. Este apresenta ainda o conceito de sumarização, sumarização automática e seus elementose tipos de acordo com a literatura. Serão apresentadas as abordagens mais frequentes para o problema da alta dimensionalidade, a Curva de Zipf e os Cortes de Luhn, utilizadas no modelo Cassiopéia. Por fim, serão tratados os conceitos e fundamentações dos Testes Estatísticos, utilizados neste trabalho para validação da hitpótese.

2.1 Agrupamento

Abordada por Wives (2004) como aglomeração, clusterização ou simplesmente agrupamento, este processo significa colocar 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 outro qualquer do mesmo grupo (BERKHIN, 2002).

Segundo Wives (2004) os algoritmos de agrupamento têm as seguintes propriedades: agregação de pontos no espaço (densidade), grau de dispersão dos pontos presentes no agrupamento, o (variância) raio ou diâmetro (dimensão), disposição dos pontos no espaço (forma) e isolamento dos agrupamentos no espaço (separação). Delgado *et al.,* (2012) define

a) **Densidade** – define o agrupamento como sendo uma densa agregação de pontos no espaço quando comparado a outras áreas que possuam poucos pontos ou nenhum. Pode ser compreendida como a quantidade de pontos do agrupamento;

b) **Variância** – o grau de dispersão dos pontos presentes no agrupamento, em relação ao seu centro ou em relação uns com os outros (ou seja, a proximidade entre pontos);

c) **Dimensão** – só é útil quando o agrupamento possui forma arredondada (hiper-esfera) e indica seu ‘raio’ ou diâmetro. Em Agrupamentos que possuem outras

formas, é mais útil o conceito de *conectividade*, que mede a distância entre outros pontos de um agrupamento;

d) **Forma** – Arranjo, organização ou disposição dos pontos no espaço, Normalmente os agrupamentos possuem a forma circular, elipsóide ou alongada;

e) **Separação** – É o grau de sobreposição ou isolamento dos agrupamentos no espaço. Os agrupamentos podem possuir grandes espaços vazios entre eles, ou podem estar tão próximos que seus limites são difíceis de definir (estando, portanto, sobrepostos);

A partir dessas propriedades surgem diferentes tipos de agrupamentos de acordo com sua forma, podendo ser hiperesféricos (Figura 1-a), alongados (Figura 1-b), curvilíneos (Figura 1-c) ou possuir estruturas mais diferenciadas (Figura 1-d) (ALDENDERFER e BLASHFIELD, 1984), (FASULO, 1999), (BERKHIN, 2002) e (WIVES, 2004).

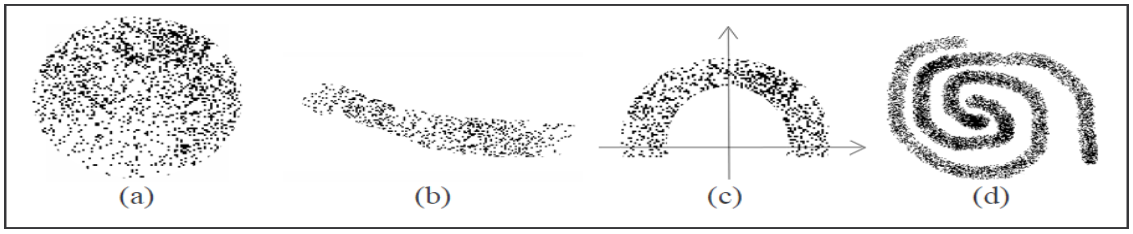

Figura 1: Diferentes tipos de agrupamentos. (Wives, 2004)

Segundo Guelpeli (2012) os métodos de agrupamento são ainda classificados em duas principais classes de acordo com o tipo de partição feita nos objetos: Métodos Hierárquicos (Figura 2-a) e Métodos Não-Hierárquicos (Figura 2-b). Guelpeli (2012) adota, por ser mais usual na literatura, a taxonomia proposta por Aldenderfer e Blashfield (1984).

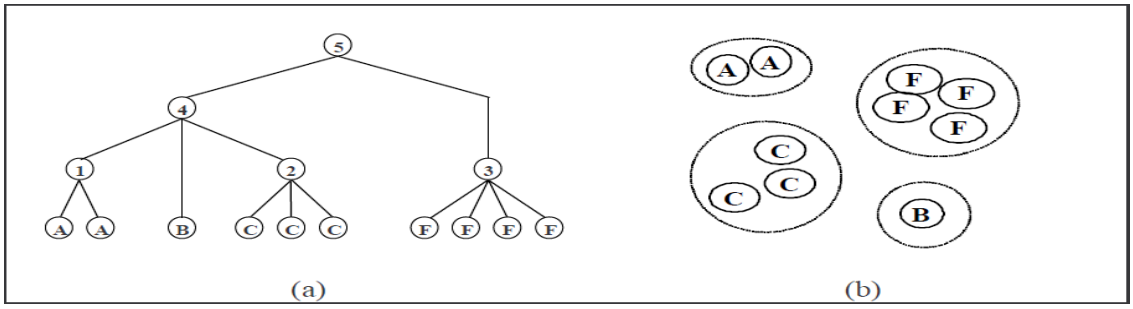

Figura 2: Tipos de agrupamentos quanto à sua partição. (Wives, 2004)

Segundo Aldenderfer e Blashfield (1984) os agrupamentos têm a seguinte taxonomia, quanto à sua configuração: hierárquicos aglomerativos, hierárquicos divisivos, de particionamento iterativo, de busca em profundidade, fator-analítico, de amontoamento e grafoteoréticos. Quando esses métodos são aplicados a um conjunto de dados, geram resultados diferentes (ALDENDERFER e BLASHFIELD, 1984), (KARYPIS, 1999), (EVERITT, 2001), (BERKHIN, 2002), (WIVES, 2004), (SILVA et al., (2005), (RIBEIRO, 2009), (METZ E MONARD, 2009) e (LOPES, 2011).

O método hierárquico aglomerativo é o mais popular e trabalha juntando os objetos em agrupamentos cada vez maiores, incluindo não só elementos, mas os próprios agrupamentos já identificados (WIVES, 2004). Por esta característica e pelo custo computacional, esse método foi o escolhido por Guelpeli (2012) para fundamentação de sua tese. Este método é melhor abordado na Seção 3.1.2 deste trabalho.

Segundo Wives (2004), no método hierárquico divisivo, todos os objetos são organizados em um único grupo que vai sendo dividido em grupos menores, até que cada objeto esteja em um agrupamento separado. Esse método não é muito usado devido ao seu custo computacional e, segundo Kaufman e Rousseeuw (1990), sua complexidade cresce exponencialmente em relação ao tamanho do conjunto de dados.

No método de particionamento iterativo acontece o particionamento do conjunto de dados cujos agrupamentos fazem iterações com os conjuntos. Já para o método de busca em profundidade, de acordo com Wives (2004), a busca é feita por regiões de alta densidade de pontos no espaço, densidade esta que se dá por identificação de zonas de baixa densidade, separadas umas das outras.

No método de fator analítico, os agrupamentos são organizados através da análise de fatores extraídos de uma matriz de similaridades que contém os graus de similaridade entre todos os elementos de um conjunto de dados (WIVES, 1999). No método de amontoamento, os agrupamentos criados vêm sobrepostos, permitindo que os objetos sejam colocados em mais de um, simultaneamente.

O método grafoteorético, segundo Wives (2004), baseia-se em teoremas e axiomas da teoria dos grafos. Tem capacidade mais dedutiva, com fundamentação teórica maior que a dos anteriores. Os outros métodos são, ainda, segundo Wives (2004), mais heurísticos.

2.2 Métricas para análise de agrupamentos textuais

A avaliação de agrupamentos poder ser distribuída, segundo Halkidi *et al*. (2001), em três grandes categorias de métricas: externas ou supervisionadas; internas ou não supervisionadas e relativas.

Segundo Guelpeli (2012), a métrica relativa tem como objetivo encontrar o melhor conjunto de grupos que um algoritmo de agrupamento pode definir, a partir de certas suposições e parâmetros. A avaliação de um agrupamento é realizada por comparações entre esse agrupamento, gerados pelo mesmo algoritmo, mas com diferentes parâmetros de entrada. Como a métrica tem a função de avaliar e comparar os agrupamentos gerados pelo próprio algoritmo e não focar na comparação entre o método proposto e outros na literatura, esta não foi a avaliação escolhida por Guelpeli (2012) para analisar seu trabalho. Sendo assim, pode-se dizer que as métricas mais adequadas são as internas e as externas.

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 Tan et al. (2006), 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 Tan et al. (2006) e Aranganayagil e Thangavel (2007), para este fim, são *Coesão, Acoplamento e Coeficiente de Silhouette.*

Com o objetivo de mesurar os resultados dos experimentos utilizados para este trabalho, foram escolhidas as métricas externas e internas, sendo definidas as seguintes medidas:

2.2.1 Métricas Externas

Recall(R): **Equação 1:**

$$R = \frac{n(A)}{n(A \cup D)} \qquad (1)$$

O *Recall* mede a proporção de objetos corretamente alocados a um agrupamento, em relação total de objetos da classe associada a este agrupamento (RIJSBERGEN, 1979) e (MANNING *et al*., 2008).

Onde $n(A)$ é o número de elementos do subconjunto A de acertos e $n(D)$ é o número de elementos do subconjunto D de falsos negativos[1] e $n(A \cup D)$ é o número total de elementos da classe correspondente (Guelpeli, 2012).

Precision(P): **Equação 2:**

$$P = \frac{n(A)}{n(A \cup B)} \tag{2}$$

A *Precision* mede a proporção de objetos corretamente alocados a um agrupamento, em relação ao total de objetos deste agrupamento (RIJSBERGEN, 1979) e (MANNING *et al*., 2008).

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 $n(A \cup B)$ é o número total de elementos do grupo. (Guelpeli, 2012).

F-Measure(F): **Equação 3:**

$$F = 2 * \frac{Precision(P) * Recall(R)}{Precision(P) + Recall(R)} \tag{3}$$

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 um (RIJSBERGEN, 1979) e (MANNING et al., 2008).

[1] Falsos negativos são elementos que deveriam ter sido alocados a um grupo e que foram alocados a outros.

Cada uma das medidas descritas é calculada para cada um dos grupos obtidos, fornecendo assim a qualidade de cada grupo. A medida de avaliação, para todo o agrupamento, é obtida através do cálculo da média entre cada uma das medidas de todos os grupos. (Guelpeli, 2012).

2.2.2 Métricas Internas

Coesão(C): **Equação 4**:

$$C = \frac{\sum_{i>j} Sim(P_i, P_j)}{\frac{n(n-1)}{2}} \quad (4)$$

A *Coesão* mede a similaridade entre os elementos do mesmo agrupamento. Quanto maior a similaridade entre eles, maior a coesão deste agrupamento (KUNZ e BLACK, 1995).

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*. (Guelpeli, 2012).

Acoplamento (A): **Equação 5**:

$$A = \frac{\sum_{i>j} Sim(C_i, C_j)}{\frac{n_a(n_a-1)}{2}} \quad (5)$$

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).

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 Silhouette(S): **Equação 6:**

$$S = \frac{b(i)-a(i)}{\max(a(i),b(i))} \quad (6)$$

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 de outro grupo. Assim, essa medida combina as medidas de coesão e acoplamento (ARANGANAYAGIL E THANGAVEL, 2007) e (ZOUBI E RAWI, 2008).

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).

O *Coeficiente Silhouette* de um grupo é a média aritmética dos coeficientes calculados para cada elemento pertencente ao grupo, sendo apresentado na Equação 7 a seguir, onde o valor de *S* situa-se na faixa de 0 a 1. (Guelpeli, 2012).

$$\bar{s} = \frac{1}{N}\sum_{i=1}^{N} S \quad (7)$$

2.3 Problema da Alta Dimensionalidade

O problema da alta dimensionalidade introduzido na literatura por Richard Bellman (BELLMAN, 1961) refere-se ao problema causado pelo aumento exponencial no volume, decorrente da adição de dimensões extras a um espaço matemático, ou seja, dimensões de uma região do espaço em celulas regulares que crescem exponencialmente com a dimensão do espaço (GUELPELI, 2012).

Beyer et al. (1999) afirmam que o uso de um elevado número de atributos gera a alta dimensionalidade e que para manter a capacidade de discriminação do atributo é necessário manter baixa a dimensionalidade dos dados. Na RI, segundo Guelpeli (2012) este problema pode ser descrito na forma de atributos de um *corpus*, ou seja, a relação entre o número de documentos da coleção, a quantidade de palavras distintas que aparece no total da coleção, e a que aparece em cada documento.

Nogueira (2009) cita algumas técnicas de redução da alta dimensionalidade e dos dados esparsos, porém a técnica mais usual da literatura, de acordo com Quoniam (2001),

Cummins e O'Riordan (2005) e Nogueira (2009) é o corte de Luhn (LUHN, 1958), que se baseia na Lei de Zipf, conhecida como Princípio do Menor Esforço.

2.3.1 Lei de Zipf

A Lei de Zipf é conhecida como Princípio do Menor Esforço. A Curva de Zipf mostrada na Figura 3 é uma distribuição estatística e especifica utilizada em agrupamento, a qual se encontra em raros fenômenos estocásticos. Um deles é a distribuição da frequência da ocorrência de palavras em um texto, em que nas ordenadas f, se tem um valor dessa frequência, e nas abcissas r, o valor da posição de ordenação relativa dessa palavra, em termos da sua frequência em relação ao das outras palavras do texto. Para a curva de Zipf de uma dada amostra específica, tem se $f.r = k$, em que k será uma constante específica para essa amostra. Quanto mais próximo do eixo Y mais frequente serão as palavras enquanto que, quanto mais próximo do eixo X, menos frequente serão as palavras, chegando até a frequência de uma ocorrência em todo o texto.

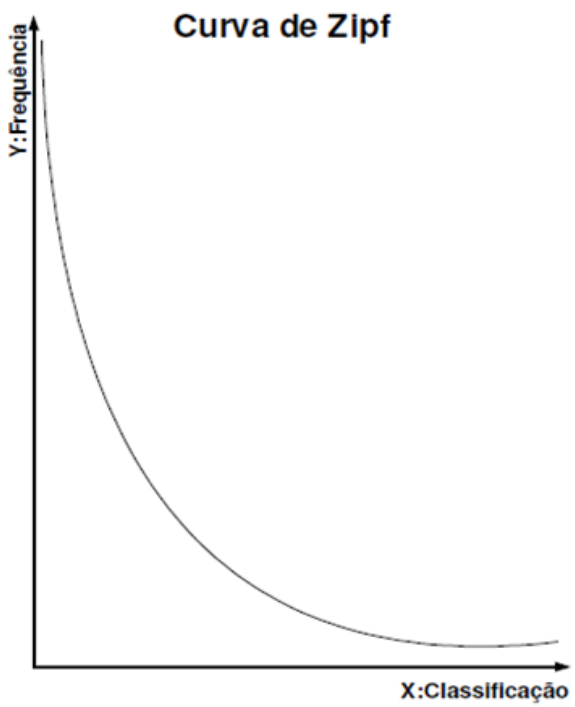

Figura 3: A Curva de Zipf.

2.3.2 Corte de Luhn

Luhn (1958) propôs uma técnica para encontrar termos relevantes, assumindo que os mais significativos para discriminação de um conteúdo de um documento estão em um pico imaginário entre dois cortes como mostra a Figura 4.

Foi então proposto o primeiro corte na Curva de Zipf, que tem como finalidade retirar as *stopwords*. As *stopwords* são palavras com mais frequências, que para Pardo (2002) não trazem muita informatividade para o texto, são palavras como pronomes, interjeições e artigos.

Após este primeiro corte a quantidade de palavras é relativamente menor, porém se encontra outro grande problema, as palavras com menos frequência, que são as palavras específicas, encontradas apenas uma única vez nos documentos, as quais fazem com que em uma representação matricial, contribuam para um grande número de dados esparsos, também conhecidas como *ruídos*. Com isso foi proposto o segundo corte na Curva de Zipf, eliminando estas palavras que apareciam em apenas uma vez em todo o texto.

Segundo Guelpeli (2012) com o primeiro e o segundo corte, surge então o pico imaginário, um processo heurístico e fonte de estudos de pesquisas atuais (Guelpeli ,2012). Para Quoniam (2001), a Curva de *Zipf*, com o corte de *Luhn* (Figura 4), possuí três áreas distintas. Na área I, encontram-se as informações com maior frequência; na área II, as informações interessantes e, na área III os ruídos.

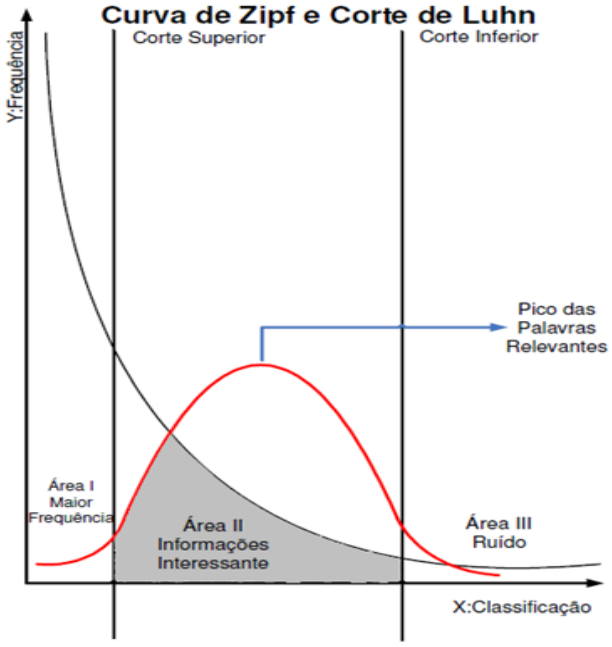

Figura 4: Curva de Zipf com os cortes de Luhn. (Guelpeli, 2012)

2.4 Sumarização

A sumarizar significa reduzir o tamanho de um determinado texto, ou seja, resumir. Segundo Pardo (2007) sumários são textos reduzidos que transmitem as ideias principais e mais relevantes de um texto original, de forma clara e objetiva e sem perda de Por definição, sumários são textos reduzidos, que transmitem as ideias principais e mais sem perda da informatividade.

Com o avanço das tecnologias surge o aumento do volume de informações disponíveis nos veículos de comunicação e em consequência a incapacidade dos leitores de absorver a totalidade de conteúdos dos textos originais (GUELPELI, 2012). Logo, o sumário é um texto resumido criado com a finalidade de passar em poucas linhas a ideia do principal do autor.

Hutchins (1987) classifica sumários científicos em três tipos: indicativos, informativos e sumários de crítica (*evaluative*). Para ele, sumários indicativos contêm apenas os tópicos essenciais de um texto, não necessariamente contendo detalhes de resultados, argumentos e conclusões. Sumários informativos, por sua vez, são considerados substitutos do texto, devendo conter todos os seus aspectos principais. Assim, se o texto-fonte correspondente for organizado em função de, por exemplo, dados, métodos, hipóteses e conclusões, um sumário informativo deveria conter as informações principais de cada um desses tópicos. Sumários de crítica assumem a função de avaliadores que, por exemplo, apresentam uma análise comparativa do conteúdo do texto-fonte com o contexto de outros trabalhos relacionados a esse conteúdo, na área específica em foco. Segundo Hutchins, seria mais simples produzir automaticamente sumários indicativos, devido à complexidade de se modelar adequadamente a sumarização humana para os demais tipos de sumários.

Existem dois tipos de denominação para os sumários, segundo Mani (2001), *extract* (extrato) ou *abstract* (resumo). Um extrato é um texto onde as sentenças utilizadas são copiadas e organizadas, de acordo com o texto fonte. O resumo é composto por sentenças reescritas ou rearranjadas, não se limitando a somente à cópia das sentenças do texto fonte.

A sumarização é parte integrante do modelo Cassiopeia, pois como melhor detalhado na seção 3.1, em sua etapa de pré-processamento são inseridos textos sumarizados para geração dos clusteres, sendo esta a técnica adotada por Guelpeli (2012) para redução da dimensionalidade. Este autor utiliza os sumários automáticos (Seção 2.4.1) em seu modelo,

mas a partir deste trabalho surge a necessidade de entendimento de sumários humanos (Seção 2.4.2) descritos a seguir.

2.4.1 Sumarização automática

A sumarização automática vem sendo explorada desde a década de 50, quando começaram a surgir os primeiros métodos para a produção de extratos, sendo o método das palavras-chave (Luhn, 1958) o mais significativo então. Entretanto, como métodos "cegos", fazendo uso de técnicas superficiais, os resultados apresentavam inúmeros problemas de coesão e coerência (Hutchins, 1987), razão pela qual a área ficou praticamente estagnada nas décadas seguintes, voltando a ser objeto de interesse com o advento da Internet e, portanto, com o aumento considerável de documentos disponíveis on-line e com a necessidade de se "digerir" informações em larga escala (isto é, em grande quantidade e no menor tempo possível).

Jones (1993) atribui a ausência de progresso significativo na área até meados da década de 90 à dificuldade de se modelar adequadamente o processo, resultando na impossibilidade de se obter sumários automáticos de qualidade. Contudo, ela observa que, para aplicações restritas, é possível explorar critérios estruturais ou composicionais em certa profundidade, baseada nos aspectos lingüísticos do texto fonte.

A a abordagem profunda toma do processamento humano da linguagem (e, portanto, também da psicologia cognitiva) a base interpretativa para a compreensão do texto-fonte e posterior extração/produção do(s) sumário(s) correspondente(s). A base fundamental dessa metodologia é que um bom sistema automático de Processamento de Linguagens Naturais - PNL deve fazer uso intensivo de regras gramaticais e de habilidades de inferência lógica, além de manipular grandes bases de conhecimento de mundo, além do próprio conhecimento lingüístico (Sampson, 1987).

Segundo essa abordagem, a sumarização automática é baseada em teorias lingüísticas formais. Entendido dessa forma, o sistema computacional deveria simular a inteligência humana, para obter um processamento eficiente da língua. Porém, a grande maioria de pesquisadores entende que a tarefa de simular a inteligência humana para um domínio aberto, no PLN, ainda está fora do alcance.

Baseados em tal argumento, alguns pesquisadores utilizam as técnicas estatísticas como métodos superficiais, que seriam de mais simples aplicação, não necessitando de algoritmos complexos, podendo ser aplicados para quaisquer conjuntos de entrada, sejam estruturas lingüísticas bem formadas ou não. Por outro lado, tais métodos, ditos “cegos”, não levam em consideração todo o conhecimento potencial da língua, tampouco os critérios de compreensão ou critérios lingüísticos necessários a essa tarefa (KLAVANS E RESNICK, 1996),

2.4.2 Sumárização Humana

Segundo Hutchins (1987) o ponto central da sumarização é reconhecer, em um texto, o que é relevante e o que pode ser descartado, para compor um sumário. Esse é também um dos pontos mais problemáticos e sujeitos a controvérsias. A importância de uma sentença ou trecho de um texto pode depender de vários fatores: dos objetivos do autor do sumário, dos objetivos ou interesse de seus possíveis leitores e da importância relativa (e subjetiva) que o próprio autor (ou leitor) atribui às informações textuais.

Estes fatores inferem diretamente na produção do sumário humano, pois o que pode ser relevante para determinada pessoal pode não ser para outra. O que realmente aproxima-se do sumário ideal é a produção deste ser realizada pelo próprio autor do texto-fonte, pois assim a subjetividade e o raciocínio referente ao assunto principal a ser abordado já está completamente definido.

2.5 Testes Estatísticos

Os testes estatísticos têm por objetivo comparar condições experimentais, podendo auxiliar e fornecer respaudo científico àquelas que tenham validade e aceitabilidade no meio científico (GUELPELI, 2012).

Segundo Callegari e Jacques (2007) os testes estatísticos podem ser divididos em paramétricos e não-paramétricos. Nos testes paramétricos, os valores da variável estudada devem ter distribuição normal ou aproximação normal. Já os não paramétricos, também chamados de distribuição livre, não têm exigências quanto ao conhecimento da distribuição da variável na população.

Guelpeli (2012) utiliza os testes estatísticos de ANOVA, de Friedman, e o coeficiente de concordância Kendall, pois estes foram considerados os mais adequados para verificar se existe diferença significativa na distribuição em todas as amostras analisadas em seus experimentos.

Existem vários softwares estatísticos tais como: *Statistica*, *Statgraphics*, *SPSS*, *Minitab*, *SAS*, *SPHINX*, *WINKS*, entre outros. No entanto são softwares geralmente de custo elevado e envolvem um aprendizado específico de usabilidade.

Neste trabalho foi utilizado, para realizar os testes estatísticos dos experimentos e comprovação da hipótese, o software *StatPlus*® (http://www.analystsoft.com/en/products/statplus/) uma versão *Trial*, este software foi escolhidos porque contém os testes estatísticos, ANOVA de Friedman e o coeficiente de concordância de Kendall, também adotados no trabalho de Guelpeli (2012).

CAPÍTULO 3 – MODELOS

Será apresentado neste capítulo o modelo Cassiopeia, proposto por Guelpeli (2012) e a alteração em seu pré-processamento proposta neste trabalho. Para melhor compreensão, serão descritas cada uma de suas três etapas (Pré-processamento, Processamento e Pós-processamento), o método hierárquico aglomerativo, o algoritmo Cliques e o cálculo utilizado para obtenção do percentual de compressão dos sumários utilizados.

3.1 Funcionamento do Modelo Cassiopeia

O modelo Cassiopeia, proposto na Figura 5 foi idealizado para ser um agrupador de texto hierárquico, com um novo método para definição do corte na curva de Zipf. Seu funcionamento abrange três macroetapas (pré-processamento, processamento e pós-processamento) descritas a seguir para um melhor entendimento das funcionalidades deste modelo.

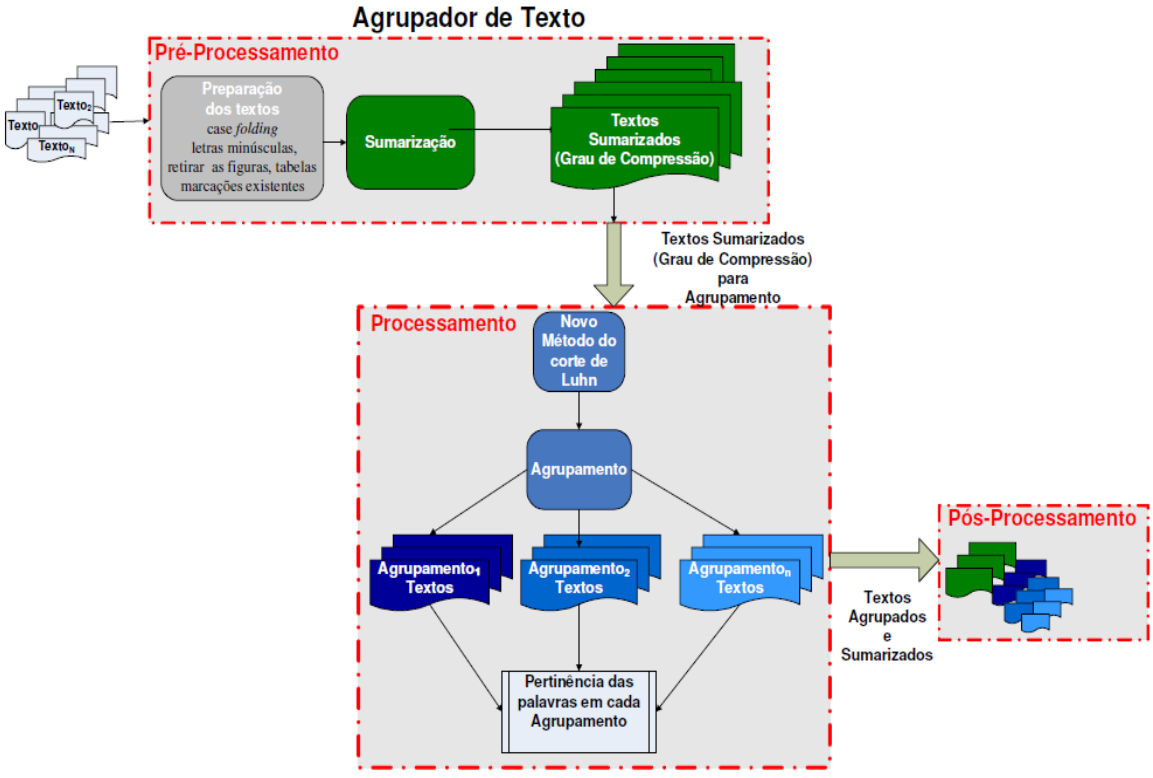

Figura 5: Modelo Cassiopeia.

3.1.1 Etapa de Pré-Processamento

Todo o processo se inicia com a entrada de textos, referenciados neste trabalho como textos-fonte. Estes são tratados para o processo computacional na etapa de pré-processamento, utilizando-se a técnica *case folding* (WITTEN *et al*., 1994) que coloca todas as letras minúsculas, descarta todas as ilustrações, tabelas e marcações, entre outros cuidados implicando um formato compatível para serem processados.

Para viabilizar ainda mais o processamento, nesta etapa, é utilizado o processo de sumarização com a finalidade de diminuir o número de palavras, obtendo-se a ideia principal do texto-fonte através da criação de um resumo com as palavras mais significativas. De acordo com Guelpeli (2012) a sumarização possibilita o uso de um espaço amostral que consegue atenua a questão da alta dimensionalidade e dos dados esparsos, além de viabilizar a permanência das *stopwords*, dando ao Cassiopeia a independência do idioma.

A redução de palavras não relevantes é considerável, justamente, pela definição do grau de compressão do sumário, ou seja, o percentual de sentenças a serem extraídas do texto-fonte, sendo este definido pelo usuário dependendo do nível de conhecimento que tenha sobre o assunto e/ou seu grau de interesse sobre este.

Esta é a etapa que mais consome tempo de toda mineração de textos segundo Goldschmidt e Passos (2005). A preocupação, não apenas da limpeza dos textos e da preparação para o processo computacional, mas na redução do número de palavras visando manter as sentenças com maior grau de informatividade proporciona um ganho qualitativo e quantitativo para o processo de agrupamento.

3.1.2 Etapa de Processamento

Quando terminada a etapa de pré-processamento começa-se a etapa de processamento que utiliza o processo de agrupamento de textos hierárquicos e um algoritmo para agrupar os textos com similaridade.

O agrupamento hierárquico é usado 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 (RIZZI et al., 2000) e (LOH, 2001). Como os

agrupamentos, não são previamente definidos, segundo Alsumait e Domeniconi (2007), o processo não ser interativo.

A utilização da sumarização de textos na etapa anterior permitiu a Guelpeli (2012) propor uma nova abordagem para o método de Corte de Luhn, onde se insere um corte médio na distribuição da frequência das palavras (Figura 6). Para viabilizar tal procedimento, foram utilizados centroides para representar o espaço amostral e para organização dos textos nos agrupamentos. Já para garantir a similaridade dos agrupamentos, utilizou-se o método hierárquico aglomerativo e o algoritmo *Cliques,* ambos descritos na seção 3.1.3.

Identificação e Seleção dos Atributos

A identificação das palavras no documento, pelo Cassiopeia, é realizada utilizando sua frequência relativa. Segundo Delgado e Dias (2012) quanto mais um termo aparecer em um documento, mais importante é, para aquele documento. A frequencia relativa é calculada de acordo com a Equação 8, fórmula que normaliza o resultado da frequência absoluta das palavras, implicando que todos os documentos sejam representados por vetores de mesmo tamanho.

$$F_r X = \frac{F_{abs} X}{N} \tag{8}$$

Onde, de acordo com Guelpeli (2012), F_rX é igual à frequência relativa de X, $F_{abs}X$ 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.

Após adquirir como base o peso das palavras é calculada a média sobre todas as palavras no documento. O modelo usa um tamanho máximo de 50 posições para os vetores de palavras, realizando um corte que representa a frequencia média das palavras obtidas com os cálculos para organização dos vetores de palavras (Figura 6). De acordo com Wives (2004) esta truncagem de 50 posições é o suficiente para a representação de um vetor com "boas características". O modelo Cassiopeia divide este vetor de 50 palavras com 25 posições à direita e 25 posição à esquerda da frequência média calculada para realização da ordenação do vetor.

Com este processo os agrupamentos são organizados na forma *top-down*, ou seja, hierárquica. O seu reagrupamento ocorre até o momento em que os centroides de cada

agrupamento estejam estáveks, ou seja, não sofram mais alterações com a inclusão de novos textos.

1 - Algoritmo do método Cassiopeia:

1. Estabelecer frequência média do conjunto P de palavras do documento baseado na Curva de Zipf.

2. $f(P,N) = \frac{\sum_{n=1}^{N} F_r P_n}{N}$ *Escolher as 25 palavras à esquerda da média e as 25 palavras à direita da média.*

Após a aplicação do algoritmo Cassiopeia surge a necessidade da garantia da similaridade entre os centroides e entre os textos agrupados, logo é aplicado o método hierárquico aglomerativo (Figura 7) e o algoritmo Cliques, descritos a seguir.

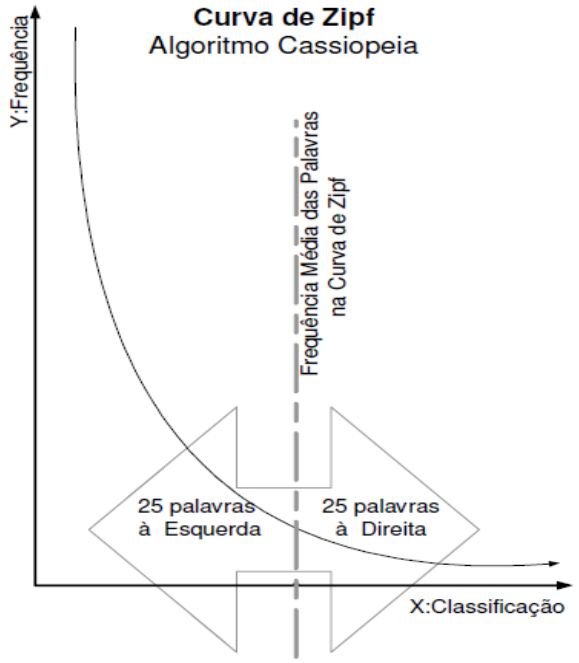

Figura 6: Seleção de atributos no modelo Cassiopeia

Método Hierárquico Aglomerativo e o Algoritmo Cliques

No método hierárquico aglomerativo os agrupamentos são recursivamente gerados considerando alguma medida de similaridade. Sendo assim os agrupamentos são em número reduzido, com baixo grau de similaridade no início, mas com o decorrer do processo, estes

vão aumentando, assim como o grau de similaridade entre os documentos de cada agrupamento (SILVA *et al*., 2005).

> ***Algoritmo 2- Algoritmo Aglomerativo:***
>
> *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*

O agrupamento hierárquico tem por objetivo gerar os agrupamentos, conforme apresentado na Figura 7, onde, a cada iteração um novo agrupamento será gerado (DELGADO e DIAS, 2012).

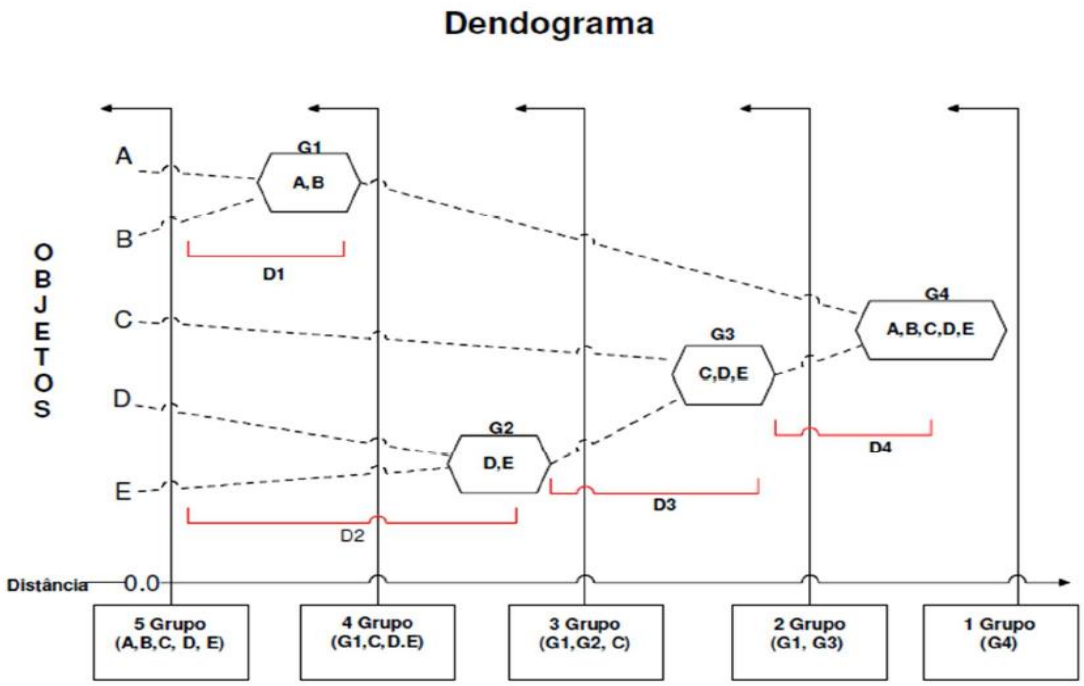

Figura 7: Dendograma do método hierárquico aglomerativo

O Algoritmo *Cliques*, segundo Wives (1999), exige que todos os objetos do *agrupamento* sejam similares entre si. A Figura 8 apresenta o resultado de um agrupamento utilizando este algoritmo da forma correta e da incorreta. Este produz os melhores resultados, contudo é o mais demorado já que todos os objetos são comparados uns com os outros (DELGADO e DIAS, 2012).

Algoritmo 3: 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.*

Segundo Korfhage (1997) o algoritmo Cliques, dentre vários outros, é o que oferece um melhor resultado, pois gera grupos de elementos muito coesos (se o grau de similaridade mínimo for elevado). Isso porque para cada elemento adicionado ao grupo, verifica-se o grau de similaridade com todos os outros elementos do mesmo grupo. Caso este elemento não tenha grau de similaridade maior do que o mínimo, com um dos objetos ele não é adicionado ao grupo. Porém, devido a esta verificação, este algoritmo torna-se o mais lento.

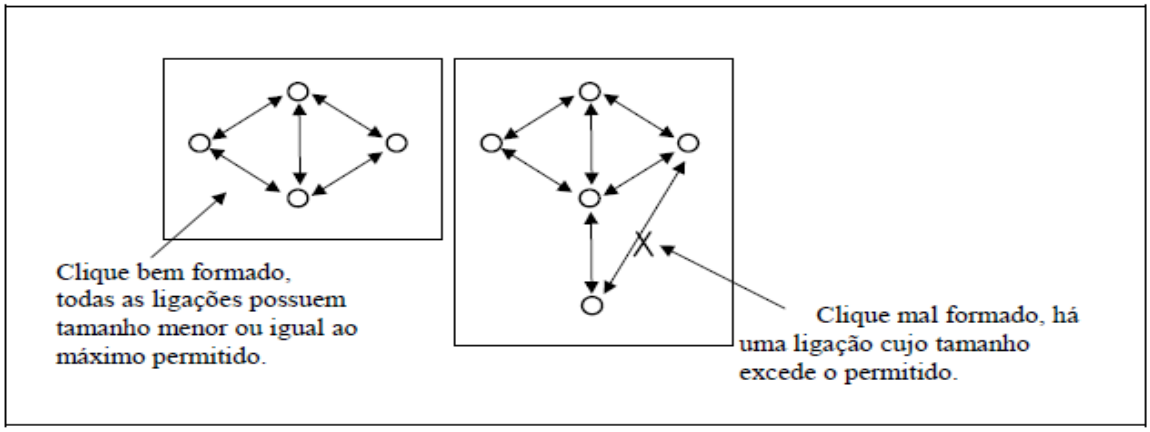

Figura 8: Agrupamento pelo método "cliques".

3.1.3 Etapa de Pós-Processamento

Terminada a etapa de processamento, cada um dos agrupamentos ou subagrupamentos terá, por similaridade, um conjunto de textos-fonte com os sumários correspondentes, com alto grau de informatividade e com suas idéias principais, característica esta da sumarização. A organização dos textos na estrutura hierárquica obtida no pós-processamento é importante para a área da recuperação da informação (RI), pois a estrutura gerada possibilita maior grau de informatividade nos textos agrupados, para atenuar a sobrecarga de informação. A hierarquia dos textos obtidos no pós-processamento possibilita uma generalização, e/ou especificação dos textos agrupados por similaridade, contribuindo, assim, para um ganho

expressivo, pois todos os textos ficam resumidos com altograu de informatividade, devido à sumarização (GUELPELI, 2012).

3.2 Modelo Proposto

O presente trabalho propõe a utilização também de sumários humanos na etapa de pré-processamento do modelo Cassiopéia, como ilustrado na Figura 9~~X~~ abaixo. Esta modificação é sugerida para a possibilidade de verificação e comparação para garantia da utilização do melhor sumário, seja ele humano ou automático.

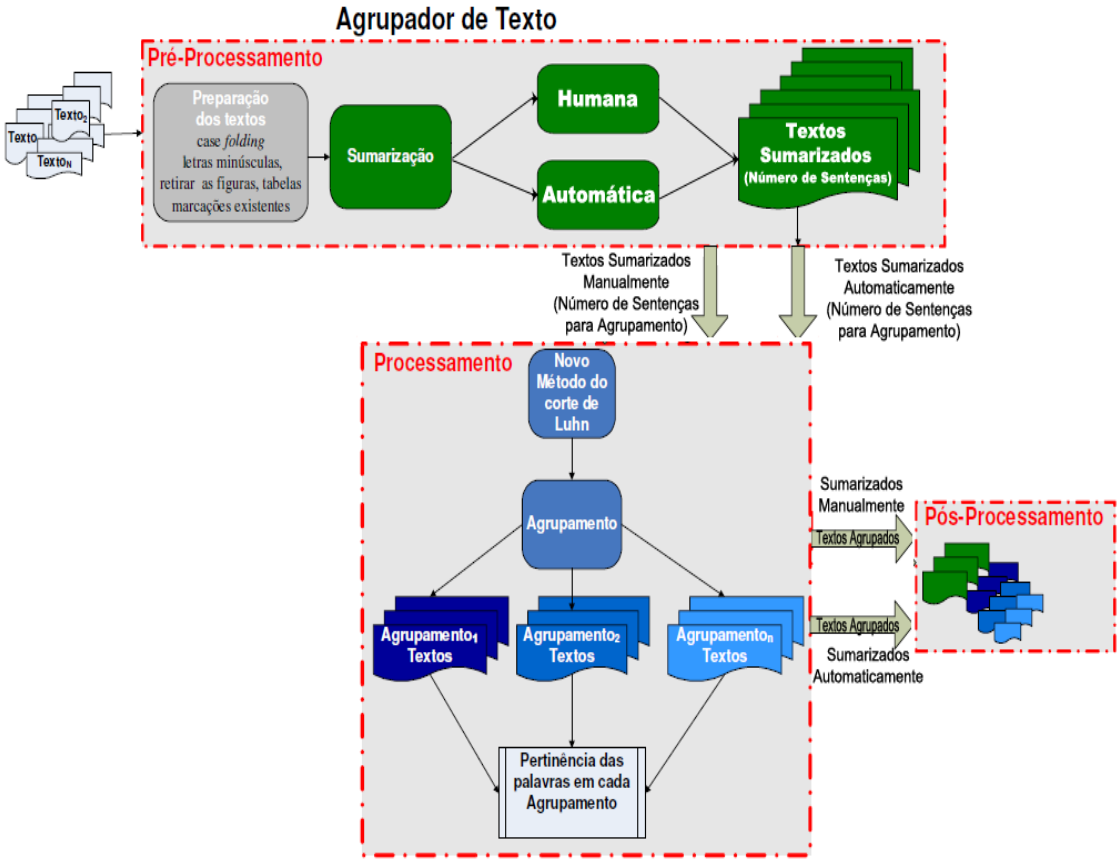

Figura 9: Adaptação ao modelo Cassiopeia

CAPÍTULO 4 – METODOLOGIA

A metodologia deste trabalho usa a corpora, descritos na seção 4.1. Foram escolhidos os sumarizadores citados na seção 4.2. Para o cálculo do percentual compressão de cada sumário, foi utilizada a aplicação da regra de três a qual será explicada na seção 4.3. O modelo de sumarização e agrupamento desenvolvido será descrito na seção 4.4. Para a análise dos resultados foram escolhidas as métricas externas e internas que são utilizadas na mensuração do processo de agrupamento e comentadas detalhadamente na seção 2.2.

4.1 Corpus

A corpora utilizada é o original no estudo de Guelpeli (2012). Serão utilizados cem amostras de textos fontes no domínio médico no idioma português, cem amostras de textos fontes no domínio médico no idioma inglês, cem amostras de textos fontes no domínio jurídico no idioma português, cem amostras de sumários humanos no domínio médico no idioma português, cem amostras de sumários humanos no domínio médico no idioma inglês, cem amostras de sumários humanos no domínio jurídico no idioma português. Será realizado também a sumarização dos textos fontes, de ambos os domínios, para gerar cem extratos para cada sumarizador utilizado.

4.2 Sumarizadores

A sumarização será feita para os textos no idioma inglês com o uso de dois sumarizadores profissionais da língua inglesa, o *Copernic Summarizer* e o *Intellexer Summarizer Pro*. Nos textos no idioma português foi utilizado um sumarizador da língua portuguesa, o *SuPor 2* e o sumarizador profissional da língua inglesa *Intellexer Summarizer Pro,* que também tem a possibilidade de sumarizar no idioma português.

4.2.1 Copernic Summarizer

O sumarizador profissional Copernic Summarizer foi desenvolvido pela Copernic Inc. e está disponível em: *http://www.copernic.com/en/products/ summarizer/*. Neste trabalho foi utilizado sua versão *Trial.*

4.2.2 Intellexer Summarizer Pro

O sumarizador profissional Intellexer Summarizer Pro foi desenvolvido pela EffectiveSoft e pode ser encontrado no link: *http://summarizer.intellexer.com/*. Neste trabalho foi utilizado sua versão *Trial.*

4.2.3 Supor 2

O sumarizador acadêmico da língua portuguesa SuPor2 foi desenvolvido por Daniel Saraiva Leite, na Universidade de São Carlos, sob o apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico e do Programa Institucional de Bolsas de Iniciação Científica.

Segundo Leite e Rino (2006), o SuPor-2 é uma versão modificada do SuPor (Ambiente para Sumarização Automática de Textos em Português), um sumarizador extrativo que depende de informações fornecidas pelo usuário e de seu conhecimento para treino e combinação de diversos métodos de extração de informações relevantes. Foram mantidos do SuPor tanto o método de classificação das informações, quanto os métodos de SA extrativa, a saber: Tamanho da Sentença, Posição das Sentenças, Freqüência das Palavras, Nomes Próprios, Cadeias Lexicais, Importância dos Tópicos e Mapa de Relacionamentos.

A modificação introduzida visou facilitar a escolha de características relevantes para o treinamento e, assim, tornar o sistema mais independente do conhecimento do usuário, pois não era possível descobrir sistematicamente qual a melhor combinação de métodos para a SA de determinado texto. Tampouco o SuPor permitia recuperar informações que levassem à análise das situações em que seu desempenho flutuava devido a escolhas indevidas de características durante seu treinamento. Em outras palavras, o SuPor 2 resulta da modificação do módulo específico de aprendizado do SuPor.

4.3 Cálculo da Compressão

A sumarização será realizada utilizando a compressão por percentual de palavras. Esta será calculada com base na comparação entre número de palavras do texto fonte e seu sumário humano, para os idiomas Português e Inglês nos domínios Médico e Jurídico.

Após isto, é realizada aplicação a Equação 9, que nos permitirá obter o percentual de compressão individual para cada sumário humano.

Equação 9:

$$C = \frac{P_{sh} \times 100}{P_{tf}} \quad (9)$$

Onde:

C é o percentual de compressão;

P_{sh} é o número de palavras do sumário humano;

P_{tf} é o número de palavras do texto fonte.

Será realizada então a sumarização automática de cada texto fonte com o mesmo percentual de compressão do sumário humano, garantindo assim a integridade dos testes. Guelpeli (2012) aplica ao modelo Cassiopeia sumários com compressões distintas para geração de seus agrupamentos, logo neste trabalho é possivel analisar o desempenho do Cassiopeia em outra metodologia.

A Equação 9 foi gerada com uma aplicação matemática da regra de três simples, onde se relacionam quatro valores divididos entre dois pares de mesma grandeza e unidade (Figura 9). As grandezas aplicadas são o número de palavras e o percentual de compressão, as quais são diretamente proporcionais.

Número de Palavras	Percentual de Compressão
P_{tf}	100
P_{sh}	C

Figura 10: Regra de três simples para obtenção do percentual de compressão.

Para extrair o número de palavras de cada texto fonte e de cada sumário humano, para realização dos cálculos da Equação 9, é utilizado o software FineCount 2.6 o qual realiza a contagem de palavras dentre outras funções.

Com a aplicação da Equação 9, é obtido o percentual de compressão individual de cada sumário humano (Apêndice A), podendo assim efetuar uma sumarização automática de cada texto fonte de forma que seu sumário automático e o sumário humano possuam exatamente o mesmo percentual de compressão, garantindo a integridade dos testes.

4.4 Modelo de Sumarização e Agrupamento

A sumarização automática foi realizada no corpus em português e inglês de todos os textos fontes com os sumarizadores apresentados na seção 4.2, obtendo assim, sumários automáticos com equivalência no número de palavras em relação aos sumários humanos. Foi feito então os agrupamentos no Cassiopeia, onde são utilizadas as métricas de coesão, acoplamento, coeficiente silhouette e precision, recall, f-measure, descritas na seção 2.2 e com seus resultados apresentados na seção 5.1. Com estes valores será possível saber se existe um ganho com o uso de sumários humanos no modelo Cassiopeia. As Figuras 11, 12 e 13, a seguir, esboçam os Modelos de Sumarização e Agrupamento gerados para cada domínio em seu idioma específico.

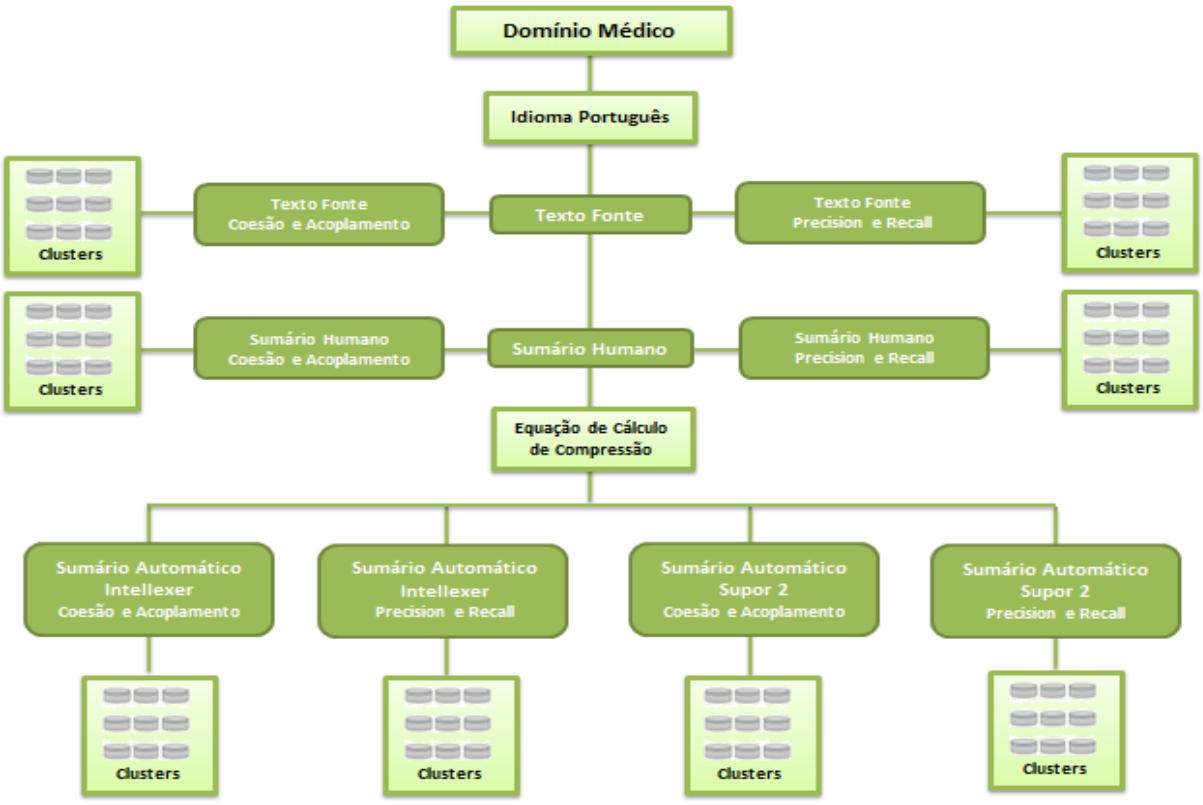

Figura 11. Modelos de Sumarização e Agrupamento gerados para o Domínio Médico no idioma Português.

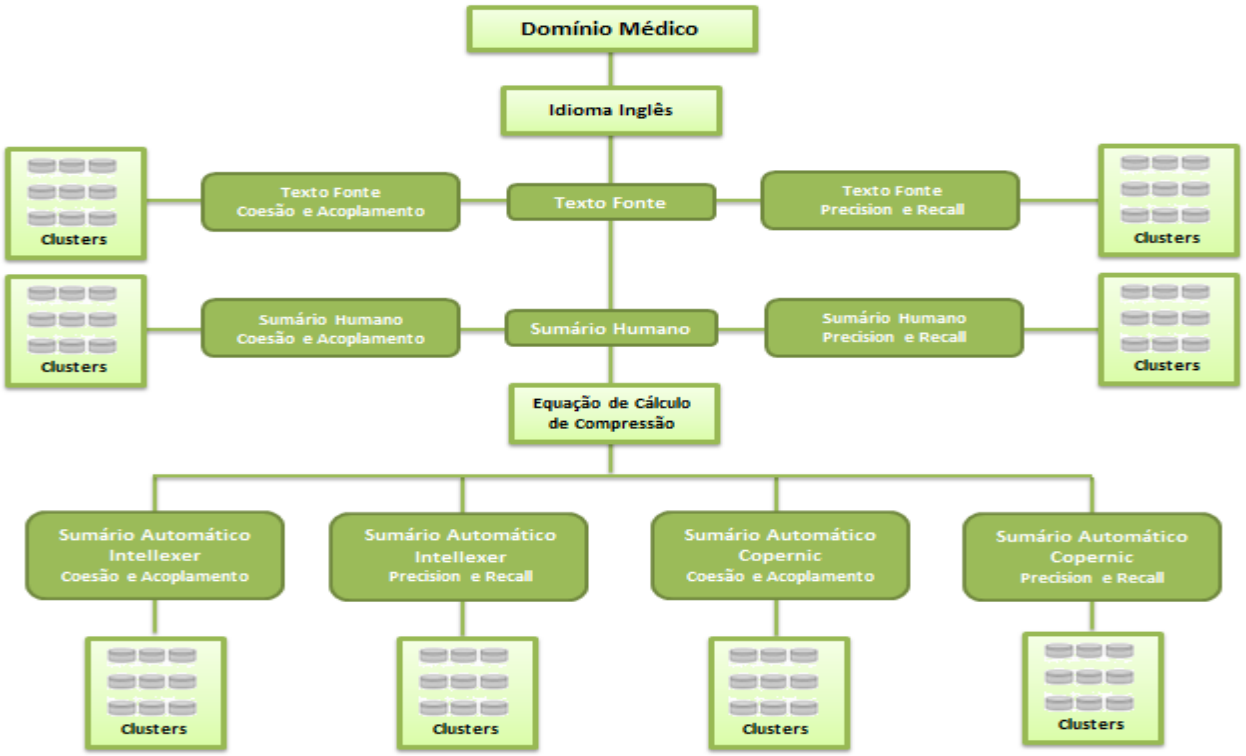

Figura 12. Modelos de Sumarização e Agrupamento gerados para o Domínio Médico no idioma Inglês.

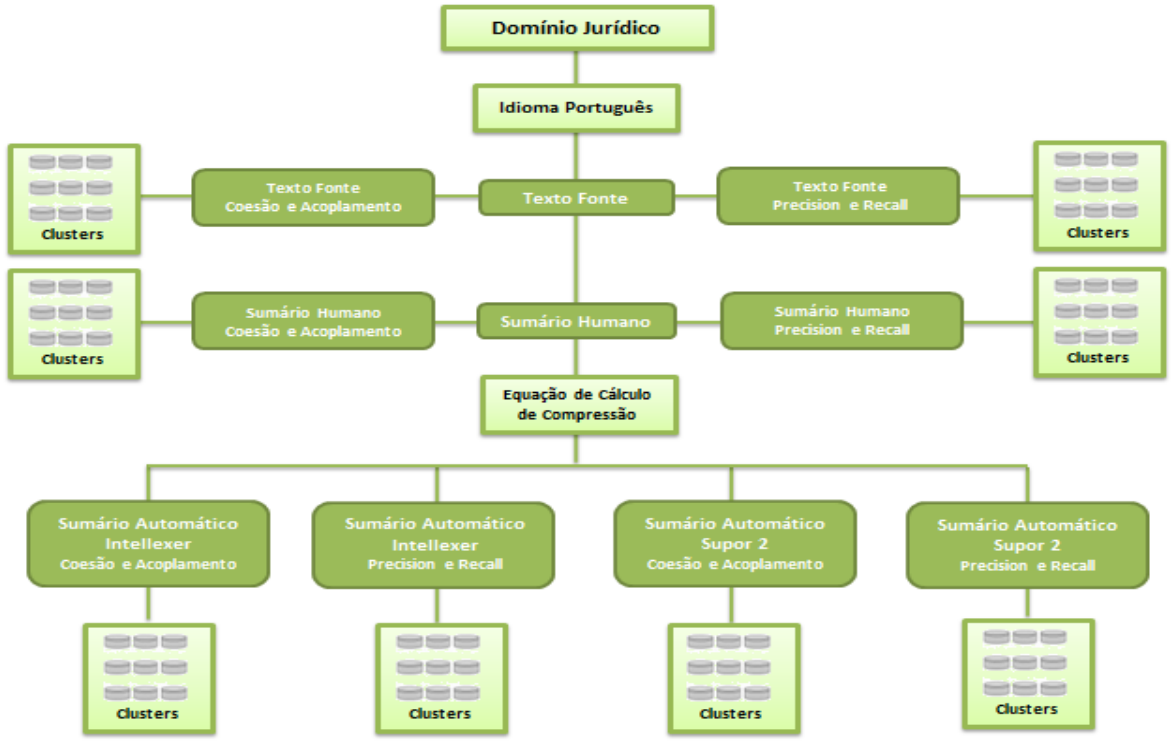

Figura 13. Modelos de Sumarização e Agrupamento gerados para o Domínio Jurídico no idioma Português

CAPITULO 5 - RESULTADOS

Neste capítulo serão apresentados os resultados das métricas obtidos pelo agrupamento utlizando o modelo Cassiopeia, assim como os testes estatísticos e a comprovação da hipótese deste trabalho.

5.1 Experimentos

Nos testes foram realizados o agrupamento do texto-fonte (sem sumarização) e dos textos sumarizados que foram gerados pelos sumarizadores descritos na seção 4.2. Foram utilizados, para os testes, cem textos para os textos fontes em cada domínio, cem sumários humanos de cada texto fonte e cem textos de cada sumarizador. Os mesmos foram submetidos duas vezes ao Cassiopeia, uma para geração das métricas Externas (Recall, Precision e F-Measure) e outra para métricas Internas (Coesão, Acoplamento e Coeficiente Silhouette), permitindo assim, a mensuração dos agrupamentos gerados.

Com estes resultados foram gerados gráficos com as médias acumuladas de cada uma das métricas para cada domínio e idioma específico, sendo apresentados nas seções 5.1.1 e 5.1.2 o domínio jurídico e médico, respectivamente, as médias ponderadas de F-Measure e de Coeficiente Silhouette. As demais métricas são apresentadas no Apêndice B. Nestes o texto-fonte será identificado através de uma linha preta, o sumário humano como linha verde e as demais linhas representam os sumários automáticos gerados pelos sumarizadores escolhidos anteriormente.

5.1.1 Domínio Jurídico

O Figura 14 mostra os resultados dos testes no domínio jurídico no idioma português, com medida externa F-Measure. Observa-se que o sumário automático Intellexer obteve valores de F-Measure superiores ao sumário humano.

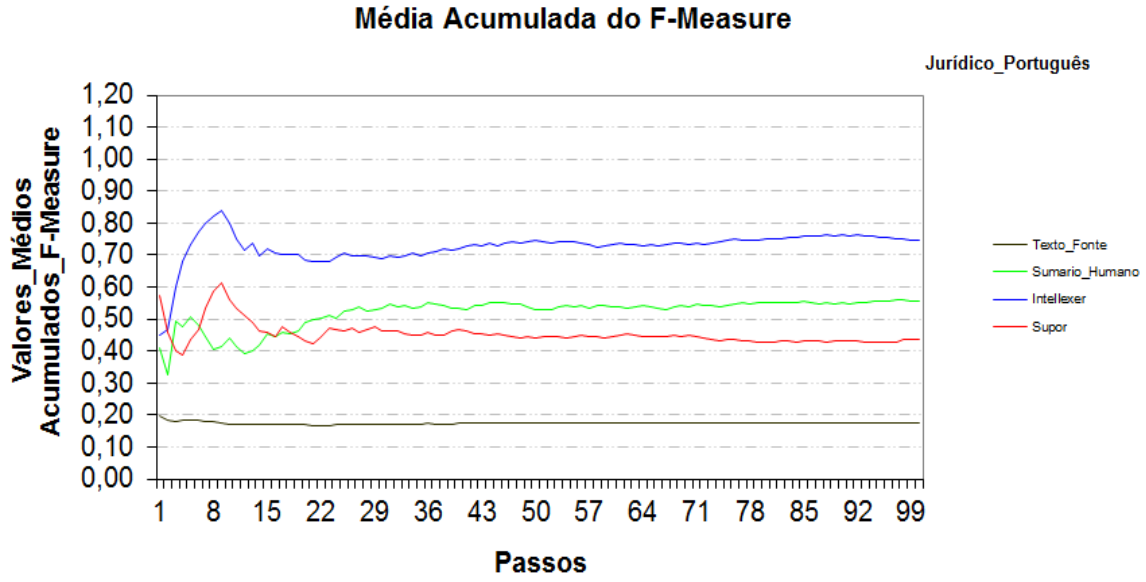

Figura 14. Resultados obtidos pelo modelo Cassiopeia usando a medida externa F-Measure no idioma Português do domínio Jurídico.

O Figura 15 mostra os resultados dos testes no domínio jurídico no idioma português, com medida interna Coeficiente Silhouette. Observa-se que os sumários automáticos Intellexer e Supor obtiveram valores de Coeficiente Silhouette superiores ao sumário humano.

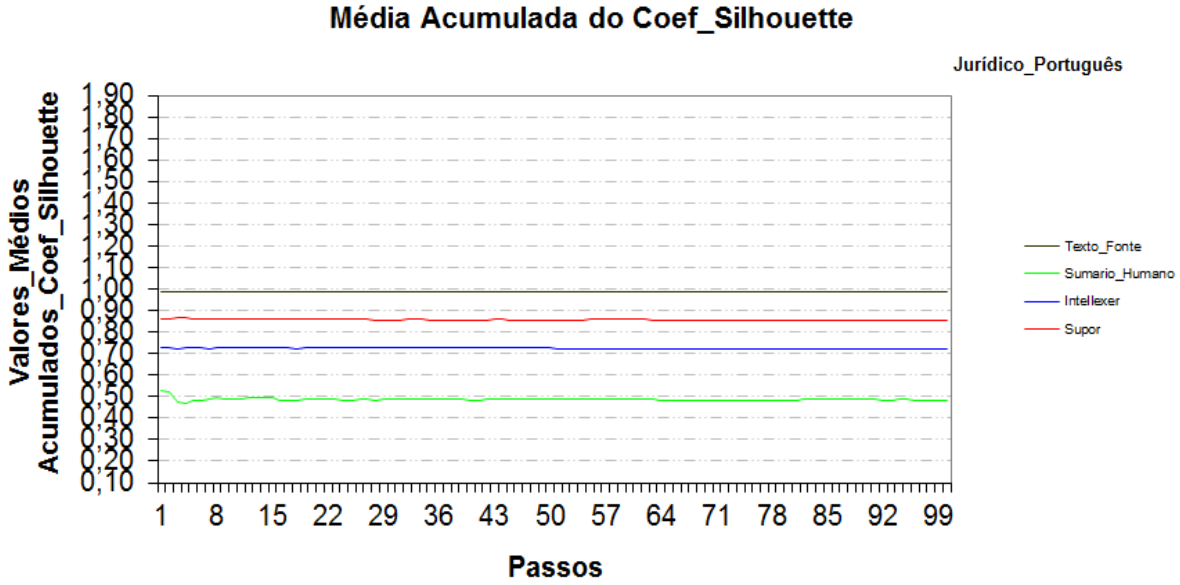

Figura 15. Resultados obtidos pelo modelo Cassiopeia usando a medida interna Coeficiente Silhouette no idioma Português do domínio Jurídico.

5.1.2 Domínio Médico

O Figura 16 mostra os resultados dos testes no domínio médico no idioma Português, com medida externa F-Measure. Observa-se que os sumários automáticos Intellexer e Supor obtiveram valores de F-Measure superiores ao sumário humano.

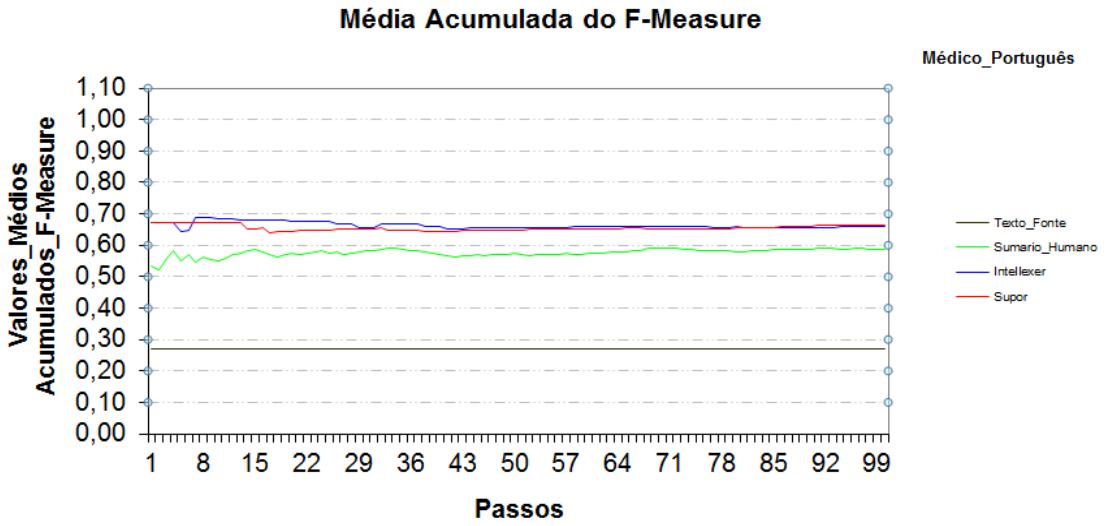

Figura 16. Resultados obtidos pelo modelo Cassiopéia usando a medida externa F-Measure no idioma Português do domínio Médico.

O Figura 17 mostra os resultados dos testes no domínio médico no idioma português, com medida interna Coeficiente Silhouette. Observa-se que os sumários automáticos Intellexer e Supor obtiveram valores de Coeficiente Silhouette superiores ao sumário humano.

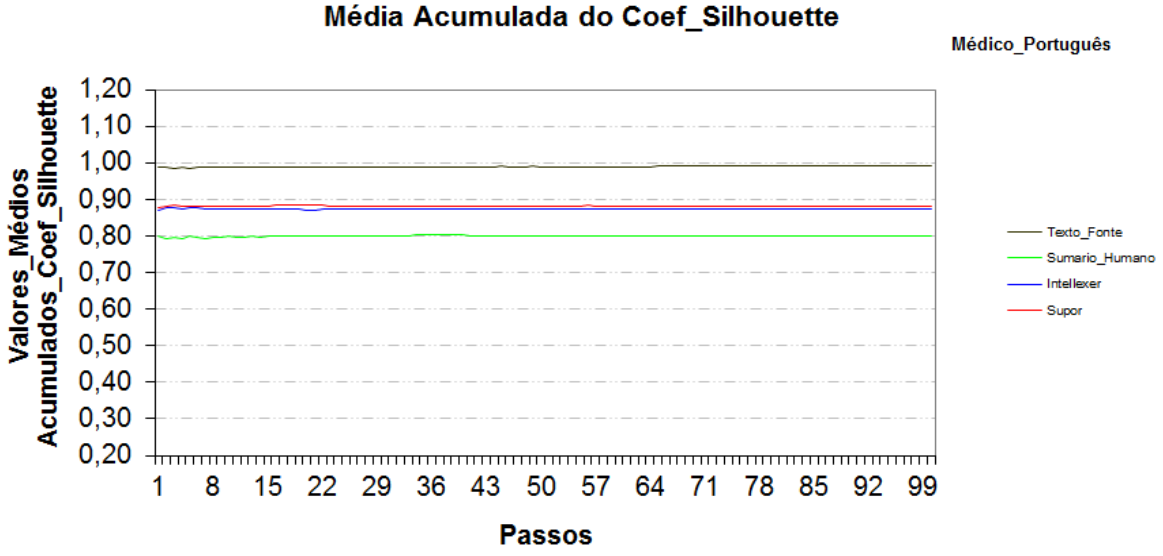

Figura 17 Resultados obtidos pelo modelo Cassiopeia usando a medida interna Coeficiente Silhouette no idioma Português do domínio Médico.

O Figura 18 mostra os resultados dos testes no domínio médico no idioma Inglês, com medida externa F-Measure. Observa-se que o sumário automático Copernic obteve valores de F-Measure superiores ao sumário humano.

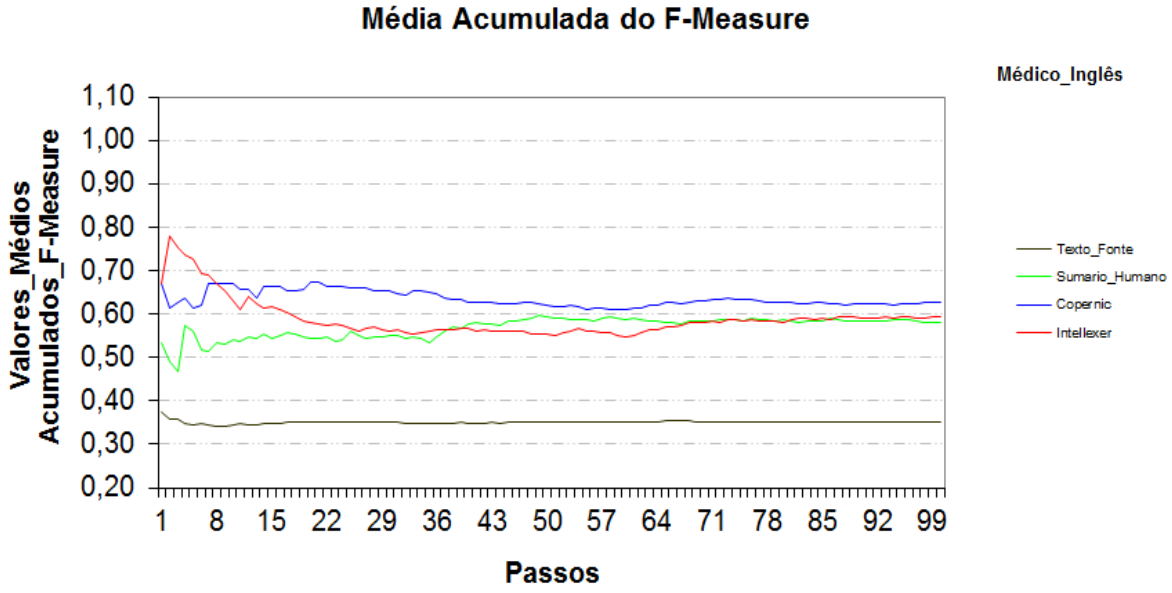

Figura 18. Resultados obtidos pelo modelo Cassiopeia usando a medida externa F-Measure no idioma Inglês do domínio Médico.

O Figura 19 mostra os resultados dos testes no domínio médico no idioma Inglês, com medida interna Coeficiente Silhouette. Observa-se que o sumário automático Copernic obteve valores de Coeficiente Silhouette superiores ao sumário humano.

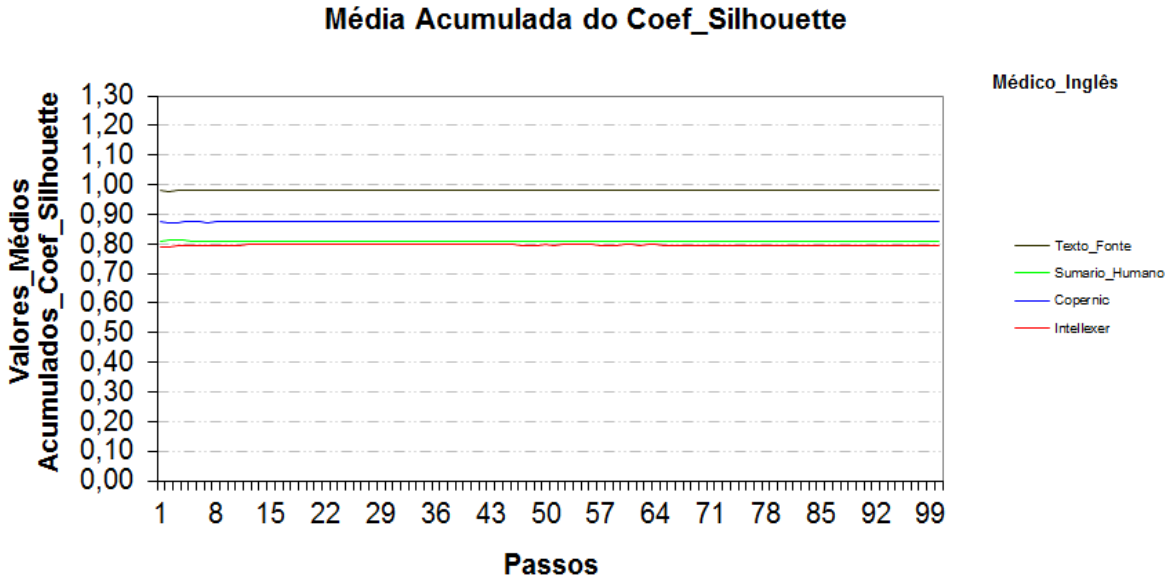

Figura 19. Resultados obtidos pelo modelo Cassiopeia usando a medida interna Coeficiente Silhouette no idioma Inglês do domínio Médico.

5.2 Comprovação da Hipótese

Para melhor entendimento pode-se afirmar que a hipótese nula consiste que os sumários humanos conseguem obter um melhor desempenho que os sumários automáticos no agrupamento no modelo Cassiopeia.

Representação da hipótese nula através da **Equação 10:**

$$H_0: k_{sumarios\ humanos} = k_{sumarios\ automaticos} \tag{10}$$

Onde:

H_0 = hipótese nula;

$k_{sumarios\ humanos}$ = *k* amostras de sumários humanos;

$k_{sumarios\ automaticos}$ = *k* amostras de sumários automáticos;

Quando a hipótese nula, H_0, for rejeitada, outra hipótese, a alternativa H_1 deve ser aceita, onde o agrupamento dos sumários automáticos garante um melhor desempenho no modelo Cassiopeia em relação ao sumário humanos, independentemente de seu idioma.

Representação da hipótese alternativa através da **Equação 11:**

$$H_0: k_{sumarios\ humanos} < k_{sumarios\ automaticos} \tag{11}$$

A hipótese alternativa foi baseada nas amostras obtidas com os testes de agrupamento utilizando o modelo Cassiopeia nos texto-fontes, sumários automáticos e sumário humano, usando as medidas de *Recall, Precision, F-Measure, Coesão, Acoplamento e Coeficiente Silhouette.*

Para comprovação foi utilizado o teste ANOVA de Friedman, que considera que as diversas amostras são, estatisticamente, idênticas, na sua distribuição (hipótese de nulidade, ou de H_0). A hipótese alternativa (H_1) aponta como elas são significativamente diferentes, na sua distribuição e o teste de concordância de Kendall normaliza o teste estatístico de Friedman, com a finalidade de gerar uma avaliação de concordância, ou não, com Ranques estabelecidos. (Guelpeli, 2012).

5.3 Análise dos Testes Estatísticos

Por uma questão de organização os testes estatísticos com as métricas Coesão, Acoplamento, Recall e Precision serão colocados no Apêndice C. As tabelas com as métricas de Coeficiente Silhouette e F-Measure são apresentadas abaixo.Nelas estão esboçados os valores obtidos para o teste ANOVA de Friedman e para o de Concordância Kendall, obtidos com o software StatPlus 2009.

5.3.1 Teste estatístico para o domínio Jurídico

A Tabela 01 comprova que os sumários automáticos obtiveram valores de Coeficiente Silhouette superior ao sumário humano no idioma português.

Tabela 1. Teste Estatístico dos resultados usando a medida interna Coeficiente Silhouette no domínio Jurídico no idioma Português

Comparando amostras múltiplas relacionadas			
N	100	*Graus de liberdade*	3
qui-quadrado	300	*p-nível*	0
Coef. de concordância de Kendall	1	*Ordem médio*	1
	Ordem médio	**Soma de ordens**	**Média**
Texto-Fonte	4	400	0,99
Sumario_Humano	1	100	0,4873
Intellexer	2	200	0,7247
SuPor2	3	300	0,8586

A Tabela 02 comprova que o sumarizador automático Intellexer obteve valores de F-Measure superior ao sumário humano no idioma português.

Tabela 2. Teste Estatístico dos resultados usando a medida extena F-Measure no domínio Jurídico no idioma Português

Comparando amostras múltiplas relacionadas			
N	100	*Graus de liberdade*	3
qui-quadrado	284,045	*p-nível*	0
Coef. de concordância de Kendall	0,9468	*Ordem médio*	0,9463
	Ordem médio	**Soma de ordens**	**Média**
Texto_Fonte	1	100	0,1743
Sumario_Humano	2,865	286,5	0,5208
Intellexer	3,99	399	0,727
Supor	2,145	214,5	0,4531

5.3.2 Teste estatístico para o domínio Médico

A Tabela 3 comprova que os sumários automáticos obtiveram valores de Coeficiente Silhouette superior ao sumário humano no idioma português.

Tabela 3. Teste Estatístico dos resultados usando a medida interna Coeficiente Silhouette no domínio Médico no idioma Portugues

Comparando amostras múltiplas relacionadas			
N	100	*Graus de liberdade*	3
qui-quadrado	298,5678	*p-nível*	0
Coef. de concordância de Kendall	0,9952	*Ordem médio*	0,9952
	Ordem médio	**Soma de ordens**	**Média**
Texto_Fonte	4	400	0,9898
Sumario_Humano	1	100	0,7997
Intellexer	2,025	202,5	0,8705
Supor	2,975	297,5	0,88

A Tabela 04 comprova que os sumários automáticos obtiveram valores de F-Measure superior ao sumário humano no idioma português.

Tabela 4. Teste Estatístico dos resultados usando a medida extena F-Measure no domínio Médico no idioma Portugues

Comparando amostras múltiplas relacionadas			
N	100	*Graus de liberdade*	3
qui-quadrado	292,5738	*p-nível*	0
Coef. de concordância de Kendall	0,9752	*Ordem médio*	0,975
	Ordem médio	**Soma de ordens**	**Média**
Texto_Fonte	1	100	0,27
Sumario_Humano	2	200	0,5766
Intellexer	3,86	386	0,6644
Supor	3,14	314	0,6532

A Tabela 05 comprova que o sumarizador automático Copernic obteve valores de Coeficiente Silhouette superior ao sumário humano no idioma português.

Tabela 5. Teste Estatístico dos resultados usando a medida interna Coeficiente Silhouette no domínio Médico no idioma Inglês

Comparando amostras múltiplas relacionadas			
N	100	*Graus de liberdade*	3
qui-quadrado	300	*p-nível*	0
Coef. de concordância de Kendall	1	*Ordem médio*	1
	Ordem médio	**Soma de ordens**	**Média**
Texto-Fonte	4	400	0,98
Sumario_Humano	2	200	0,81
Copernic	3	300	0,8755
Intellexer	1	100	0,7985

A Tabela 06 comprova que os sumários automáticos obtiveram valores de F-Measure superior ao sumário humano no idioma português.

Tabela 6. Teste Estatístico dos resultados usando a medida extena F-Measure no domínio Médico no idioma Inglês

Comparando amostras múltiplas relacionadas			
N	100	*Graus de liberdade*	3
qui-quadrado	264,1833	*p-nível*	0
Coef. de concordância de Kendall	0,8806	*Ordem médio*	0,8794
	Ordem médio	**Soma de ordens**	**Média**
Texto_Fonte	1	100	0,3501
Sumario_Humano	2,42	242	0,5687
Copernic	3,93	393	0,6344
Intellexer	2,65	265	0,5884

CAPITULO 6 – CONCLUSÕES

A internet se consolida a cada dia como o principal veículo de distribuição e armazenamento de informações, por este motivo o estudo de informações textuais tem demonstrado ser um importante foco de pesquisa contribuindo muito com a área de Mineração de Textos e incentivando, cada vez mais, a aprimoração de técnicas que possibilitem uma melhor manipulação e recuperação destas informações, sendo este um importante diferencial para as pessoas e para as empresas.

O modelo Cassiopeia consiste no uso da sumarização no pré-processamento para redução da quantidade de atributos, mantendo-se assim uma lista de *stopwords* que o torna independente do idioma e ainda apresentando um nova proposta para o corte de Luhn que o torna independente também do domínio dos textos inseridos. Este modelo apresenta um grande avanço em termos de precisão, recuperação da informação, coesão e acoplamento.

A abordagem deste trabalho, ao comparar a utilização dos SA e dos SH no pré-processamento, permitiu comprovar ainda mais estas afirmações especificadas por Guelpeli (2012) sobre seu modelo, pois pode-se observar nos resultados que as métricas de F-Measure e Coeficiente Silhouette representam um desempenho 100% de eficácia quando se é utilizado os sumários automáticos, ou seja, os agrupamentos gerados, quando utilizados os SA, são de melhor qualidade, mais rápidos e menos custosos. Esta afirmação ainda foi comprovada no Capítulo 5, onde são apresentados os testes estatísticos que confirmam ainda mais a hipótese, assim como os resultados obtidos nos experimentos.

A conclusão deste trabalho demonstra que os agrupamentos gerados pelo modelo Cassiopeia, quando se é utilizada a técnica de sumarização automática são promissores, para os agrupamentos do modelo Cassiopeia, em comparação aos gerados com a utilização de sumários humanos dentro do domínio médico e jurídico, nos idiomas português e inglês.

6.1 Contribuições

Este trabalho traz como contribuição a utilização dos sumários de melhor desempenho, independente dos idiomas dos textos. Afirma-se que no modelo Cassiopeia a utilização dos sumários automáticos são superiores aos sumários gerados por humanos, trazendo assim, como benefício, a garantia da melhor utilização do modelo e a redução do tempo utilizado para geração de sumários.

6.2 Limitações

As limitações encontradas para realização deste trabalho estão relacionadas com o encontro de sumarizadores automáticos de qualidade e de custo viável para os idiomas português e inglês.

No idioma inglês existem muitos softwares encontrados que poderiam ser utilizados se não implicassem em um alto custo financeiro. Já para o idioma português não existe grande diversidade de sumarizadores o que traz uma grande dificuldade na criação destes sumários automáticos para este idioma.

6.3 Trabalhos Futuros

Sugere-se como trabalhos complementares a criação de uma corpora com domínios diferentes dos utilizados por Guelpeli (2012) e a comparação dos resultados obtidos com estes em diferentes idiomas.

REFERÊNCIAS BIBLIOGRÁFICAS

ALDENDERFER, M. S.; BLASHFIELD, R. K. **Cluster Analysis**. Beverly Hills, CA: Sage, 1984.

ALSUMAIT, L. and DOMENICONI, C. **Text Clustering with Local Semantic Kernels**. Book survey of text mining: clustering, classification, and retrieval Second . Editors BERRY, M. E CASTELLANO, M. Editin, Springer, Part I Clustering, pp 87- 108, 2007.

ARANGANAYAGIL, S. and THANGAVEL, K. **Clustering Categorial Data Using Silhouette Coefficient as a Relocating Measure.** In International conference on computational Intelligence and multimedia Applications, ICCIMA, 2007, Sivakasi, India. Proceedings. Los Alamitos: IEEE 2007. p13-17.

ARANHA, C. N. **Uma Abordagem de Pré-Processamento Automático para Mineração de Textos em Português: Sobe o Enfoque da Inteligência Computacional**. Tese de Doutorado. PUC-Rio de Janeiro, Brasil, 2007.

BERKHIN, P. **Survey of Clustering Data Mining Techniques.** Accrue Software, San Jose, CA, 2002.

BEYER K, GODSTEIN J, RAMAKRISHNAN R, SHAFT U. **When is "Nearest Neighbor" Meaningful?** In: Beeri C, Buneman P, editors. International Conference on Database Theory (ICDT); 1999 January 10-12, Jerusalem, Israel: Springer Verlag;. p. 217-235, 1999.

BOHN, R, E., BARU, C., SHORT, J. E. **How Much Information? 2010 Report on**

Enterprise Server Information. Global Information Industry Center UC San Diego 950 Gilman Drive, Mail Code 0519. La Jolla, CA 92093-0519. http://hmi.ucsd.edu/howmuchinfo.php. Acesso em: 23 JUL 2013.

CALLEGARI-JACQUES, S. M. **Bioestatística: Princípios e Aplicações**. Porto Alegre: Artmed, p,264, 2007.

CHEN, H. **Knowledge management system: a text mining perspective**. Artificial intelligence Lab, Department of MIS, University of Arizona, Knowledge computing Corporatin, Tucson, Arizona, 2001.

CUMMINS, R., O'RIORDAN, C. E**volving General Term-Weighting Schemes for Information Retrieval: Tests on Larger Collections.** Journal Artificial Intelligence Review http://dl.acm.org/citation.cfm?id=1107370archive Volume 24 Issue 3-4, November 2005 Kluwer Academic Publishers Norwell, MA, USA, 2005.

DELGADO, C.C.N.; e DIAS, H. D.; **Utilização de Sumários Humanos no Modelo Cassiopéia** – Trabalho de Conclusão de Curso – Centro Universitário de Barra Mansa. Graduação em Engenharia da Computação, Barra Mansa, Brasil, 2012.

EVERITT, B.S. and DUNN, G. **Applied multivariate analysis.** Book 2nd. ed. London: Arnold, 2001.

FASULO, D. **An Analysis of Recent Work on Clustering Algorithms.** Technical Report, Dept. of Computer Science and Engineering, Univ. of Washington, 1999.

FELDMAN, R. e SANGER J. **The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data.** The book Cambridge University Press, 2006.

GANTZ J. F., REINSEL, D. **The digital universe decade - are you ready?** External Publication of IDC (Analyse the Future) Information and Data, pp. 1–16, 2010.

GOLDSCHMIDT, R., PASSOS, E. **Data Mining: Um Guia Prático**. Livro Editora Campus - Rio de Janeiro: Elsevier, 2005.

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.

HOURDAKIS, N.; ARGYRIOU, M.; EURIPIDES G. M.; e PETRAKIS, E.E.M. **Hierarchical Clustering in Medical Document Collections: the BIC-Means Method** Journal of Digital Information Management,Volume 8,Issue 2, April, 2010, Pages71-77,2010.

HOWLAND, P. e PARK, H. C**luster-Preserving Dimension Reduction Methods for Document Classification.** Book survey of text mining: clustering, classification, and retrieval Second . Editors BERRY, M. E CASTELLANO, M. Edition, Springer, Part I Clustering, pp 3- 24, 2007.

JONES, K. S. e WILLET, P. **Readings in Information Retrieval**. Book Edit Morgan Kaufmann, An Imprint of Elsevier Science, 1997. http://books.google.com.br/books?hl=ptBR&lr=&id=TRc2tBJrsd0C&oi=fnd&pg=PR11&dq=Readings+in+Information+Retrieval&ots=dg7ubtpJkh&sig=cz3HhQes5ryqS6-IqT-rasxzdU#v=onepage&q&f=false. Acesso em: 23 AGO. 2013.

HALKIDI, M. BATISTAKIS, Y., VARZIRGIANNIS, M. **On clustering validation techniques**. Journal of Intelligent Information Systems, 17(2-3):107-145, 2001.

HUTCHINS, J. **Summarization: Some problems and Methods.** In: Jones. Meaning: The frontier of informatics. Cambridge. London, pp. 151-173, 1987.

KARYPIS, G., HAN, E.H. S.; and KUMAR, V. **CHAMELEON: A Hierarchical clustering algorithm using dynamic modeling .** To Appear in the IEEE Computer, 32(8):68–75, 1999.

KAUFMAN, L. and ROUSSEEUW, P. **Finding Groups in Data: An Introduction to Cluster Analysis**. New York: Wiley Interscience, 1990.

KLAVANS ,J., RESNIK, P. **The Balancing Act: Combining Symbolic and Statistical Approaches to Language**. (Selected Papers of the 32nd Meeting of the ACL). Las Cruces, New Mexico.1996

KORFHAGE, Robert R. **Information retrieval and storage**. [S.l.]: John Wiley & Sons, 1997.

KUECHLER, W. L. **Business applications of unstructured.** Magazine Communications of ACM New York, NY, USA, vol. 50, no. 10, pp. 86–93, 2007.

KUNZ, T., BLACK, J.P.: **Using Automatic Process Clustering for Design Recovery and Distributed Debugging.** IEEE Trans. Software Eng.515 527,1995.

LEVY, D.M. **To grow in wisdom: vannevar bush, information overload, and the life of leisure.** In JCDL(2005) p.281-286, 2005.

LYMAN, P. e VARIAN, H. **How much information**, URL: <http://www2.sims.berkeley.edu/research/projects/how-much-info/>USA: University of California, 2000. Acesso em: 23 AGO. 2013.

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

LOPES, M. C. S. **Mineração de dados textuais utilizando técnicas de clustering para o idioma português** Tese de Doutorado COPPE/UFRJ -Rio de Janeiro, Brasil, 2004.

LOPES, G. A. W. **Um Modelo de Rede Complexa para Análise de Informações Textuais.** Dissertação apresentada ao Curso de Mestrado em Inteligência Artificial Aplicada à Automação Industrial do Centro Universitário da FEI, São Paulo, 2011.

LUHN, H. P. **The automatic creation of literature abstracts.** IBM Journal of Research and Development, 2, pp. 159-165,1958.

MANNING, C. D., RAGHAVAN, P., SCHUTZE, H. **Introduction to Information Retrieval**, Cambridge University Press. 2008.

MARIA, S.; MORAES, W. e LIMA, V. L. s. **Abordagem não supervisionada para Extração de Conceitos a partir de Textos** In: Tecnologia da Informação e da Linguagem Humana- TIL, 2008, Vila Velha. Todas as Palavras da Sentença como Métrica param um Sumarizador Automático. Vila Velha: WebMedia, 2008. p. 359-363.

METZ, J e MONARD, M. C. **Clustering hierárquico: uma metodologia para auxiliar na interpretacão dos clusters** XXV Congresso da Sociedade Brasileira de Computação-ENIA 2009 p. 1170-1173, 2009.

NOGUEIRA, B. M. **Seleção não-supervisionada de atributos para Mineração de Textos**. 2009. Dissertação (Mestrado em Ciências de Computação e Matemática

Computacional) - Instituto de Ciências Matemáticas e de Computação,São Paulo, Brasil, 2009.

OLIVEIRA, I. M. **Estudo de uma metodologia de mineração de textos científicos em Língua portuguesa.** Tese de Mestrado apresentada ao Programa de Pós Graduação em Engenharia Civil, COPPE da Universidade Federal do Rio de Janeiro, Brasil,2009.

PARDO, T.A.S. **GistSumm: Um Sumarizador Automático Baseado na Idéia Principal de Textos.** Série de Relatórios do NILC. NILC-TR-02-13, 2002.

PARDO, T.A.S. **Sumarização Automática de Textos Científicos: Estudo de Caso com o Sistema GistSumm**. Série de Relatórios do NILC. NILC-TR-07-11, 2007.

QUONIUM , L., TARAPANOFF, K., ARAUJO, R. H. J. e ALVARES, L. **Inteligência obtida pela aplicação de *Data Mining* em bases de tese francesa sobre o Brasil.** Ci. Inf., Brasília, v. 30, n. 2, p. 20-28, maio/ago. 2001

RAMOS, H. S. C., BRASCHER, M. **Aplicação da descoberta de conhecimento em textos para apoio à construção de indicadores infométricos para a área de C&T.** Ciência da Informação, V. 38, n. 2, p. 56-68, 2009.

REZENDE, S. O.; MARCACINI, R. M.; MOURA, M. F., **"O uso da Mineração de Textos para Extração e Organização não Supervisionada de Conhecimento",** Revista de Sistemas de Informação da FSMA n 7 (2011).

RIBEIRO, M. N. **Seleção Local de Características em Agrupamento Hierárquico de Documentos** Dissertação de Mestrado Curso de Mestradoem Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco, Fevereiro, 2009.

LEITE, D. S. E RINO, L. H. M. **SuPor: extensões e acoplamento a um ambiente para mineração de dados.** Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional NILC - ICMC-USP, NILC-TR-06-07-Agosto, 2006 Universidade Federal de São Carlos, SP, Brasil.

RIJSBERGEN, C. J. **Information Retrieval.** Book London: Butterworths, 1979

RIZZI, C.; WIVES, L. K.; ENGEL, P. M.; OLIVEIRA, J. P. M. **Fazendo Uso da Categorização de Textos em Atividades Empresariais**. In: International Symposium on Knowledge Management/Document Management (ISKM/DM 2000), III, Nov, 2000.

SAMPSON, G., **Probabilistic models of analysis. In G. Leech, R. Garsude and G. Sampson (eds.)**, *The computational analysis of English*, pp. 16-29. Longman. Harrow. 1987.

SPARCK J, K. **What might be in a summary? In: Knorz, Krause and Womser-Hacker (eds.).** *Information Retrieval 93: Von der Modellierung zur Anwendung*, Universitatsverlag Konstanz, pp. 9-26.1993

SILVA, C. M., VIDIGAL, M. C.; VIDIGAL, P. S.; SCAPIM, C. A.; DAROS, E., SILVÉRIO, L. **Genetic diversity among sugarcane clones (Saccharum spp.)**. Acta Scientiarum. Agronomy, v.27, p.315-319, 2005.

SMYTH, B., BALFE, E., FREYNE, J., BRIGGS, P., COYLE, M., BOYDELL O. **Exploiting Query Repetition and Regularity in an Adaptive Community-Based Web Search Engine.** User Modeling and User-Adapted Interaction, Springer ISSN 0924-1868- DOI 10.1007/s11257-004-5270-4.v. 14, n. 5, p. 383-423, 2004.

TAN, P. N.; STEINBACH, M.; and KUMAR, V.**Introduction to Data Mining**.Addison-Wesley, 2006.

WITTEN, I.H., MOFFAT, A.; BELL, T.C. (1994). **Managing Gigabytes**. Van Nostrand Reinhold. New York.

WIVES, L.K. **Um estudo sobre Agrupamento de Documentos Textuais em Processamento de Informações não Estruturadas Usando Técnicas de Clustering**. Porto Alegre: UFRGS, 1999. Dissertação (Mestrado em Ciência da Computação), Instituto de Informática, Universidade Federal do Rio Grande do Sul, 1999.

WIVES, L. K. **Utilizando conceitos como descritores de textos para o processo de identificação de conglomerados (clustering) de documentos** – Tese (doutorado) – Universidade Federal do Rio Grande do Sul.Programa de Pós-graduação em Computação, Porto Alegre, BR – RS, Brasil, 2004.

ZOUBI, M. B. and RAWI, M. An Efficient Approach for Computing Silhouette Coefficients. **Journal of Computer Science** Volume 4 Page No.: 252 – 255, 2008.

APÊNDICES

APÊNDICE A

Lista de percentual de compressão segundo a Equação 9 para o domínio Jurídico no idioma Portugues

		Texto Original	***Texto Sumarizado***	
Nº	**Nome do Arquivo**	**Número de Palavras**	**Número de Palavras**	**%**
1	Ambiental01.txt	1665	12	1
2	Ambiental02.txt	805	35	4
3	Ambiental03.txt	2179	44	2
4	Ambiental04.txt	685	40	6
5	Ambiental05.txt	579	41	7
6	Ambiental06.txt	4169	41	1
7	Ambiental07.txt	2129	20	1
8	Ambiental08.txt	838	37	4
9	Ambiental09.txt	1628	37	2
10	Ambiental10.txt	2150	18	1
11	Civil01.txt	3002	65	2
12	Civil02.txt	10647	61	1
13	Civil03.txt	1162	170	15
14	Civil04.txt	4183	179	4
15	Civil05.txt	16196	283	2
16	Civil06.txt	2591	31	1
17	Civil07.txt	2531	27	1
18	Civil08.txt	2885	43	1
19	Civil09.txt	3475	257	7
20	Civil10.txt	1404	42	3
21	Constitucional01.txt	4537	76	2
22	Constitucional02.txt	1464	350	24
23	Constitucional03.txt	2365	178	8
24	Constitucional04.txt	8272	140	2
25	Constitucional05.txt	1169	120	10
26	Constitucional06.txt	1810	41	2
27	Constitucional07.txt	2179	46	2
28	Constitucional08.txt	1361	29	2
29	Constitucional09.txt	857	26	3
30	Constitucional10.txt	1646	29	2
31	Consumidor01.txt	757	23	3
32	Consumidor02.txt	2445	40	2
33	Consumidor03.txt	1365	22	2
34	Consumidor04.txt	2454	31	1
35	Consumidor05.txt	772	20	3
36	Consumidor06.txt	1956	31	2
37	Consumidor07.txt	1286	32	2
38	Consumidor08.txt	814	42	5
39	Consumidor09.txt	2406	19	1

40	Consumidor10.txt	3149	44	1
41	Consumidor11.txt	1012	41	4
42	Consumidor12.txt	724	44	6
43	Consumidor13.txt	1055	34	3
44	Consumidor14.txt	2086	50	2
45	Consumidor15.txt	550	39	7
46	Familia01.txt	2640	36	1
47	Familia02.txt	1724	46	3
48	Familia03.txt	1530	19	1
49	Familia04.txt	2478	35	1
50	Familia05.txt	2730	18	1
51	Familia06.txt	3739	43	1
52	Familia07.txt	1376	19	1
53	Familia08.txt	2970	32	1
54	Familia09.txt	3241	9	0
55	Familia10.txt	2600	51	2
56	Familia11.txt	3511	41	1
57	Familia12.txt	1777	23	1
58	Familia13.txt	3107	25	1
59	Familia14.txt	968	28	3
60	Familia15.txt	1592	28	2
61	Penal01.txt	2200	81	4
62	Penal02.txt	5648	195	3
63	Penal03.txt	3712	333	9
64	Penal04.txt	4529	204	5
65	Penal05.txt	13099	343	3
66	Penal06.txt	3487	28	1
67	Penal07.txt	1972	59	3
68	Penal08.txt	1678	39	2
69	Penal09.txt	2220	54	2
70	Penal10.txt	3265	37	1
71	Previdenciario01.txt	1598	17	1
72	Previdenciario02.txt	1087	30	3
73	Previdenciario03.txt	3787	162	4
74	Previdenciario04.txt	3818	20	1
75	Previdenciario05.txt	2169	44	2
76	Previdenciario06.txt	2447	14	1
77	Previdenciario07.txt	3656	33	1
78	Previdenciario08.txt	2697	30	1
79	Previdenciario09.txt	2498	50	2
80	Previdenciario10.txt	2494	34	1
81	Processual01.txt	7278	311	4
82	Processual02.txt	6102	88	1
83	Processual03.txt	4831	162	3

84	Processual04.txt	2823	216	8
85	Processual05.txt	3690	145	4
86	Trabalhista01.txt	652	17	3
87	Trabalhista02.txt	2013	37	2
88	Trabalhista03.txt	2467	25	1
89	Trabalhista04.txt	1982	32	2
90	Trabalhista05.txt	2012	35	2
91	Trabalhista06.txt	2591	31	1
92	Trabalhista07.txt	2631	21	1
93	Trabalhista08.txt	3245	40	1
94	Trabalhista09.txt	3279	35	1
95	Trabalhista10.txt	1863	32	2
96	Trabalhista11.txt	1860	44	2
97	Trabalhista12.txt	2639	11	0
98	Trabalhista13.txt	2983	28	1
99	Trabalhista14.txt	830	25	3
100	Trabalhista15.txt	848	31	4

Lista de percentual de compressão segundo a Equação 9 para o domínio Médico no idioma Portugues

		Texto Original	***Texto Sumarizado***	
Nº	**Nome do Arquivo**	**Número de Palavras**	**Número de Palavras**	**%**
1	Cardiologia01.txt	4389	169	4
2	Cardiologia02.txt	3309	340	10
3	Cardiologia03.txt	7181	164	2
4	Cardiologia04.txt	1553	264	17
5	Cardiologia05.txt	4063	370	9
6	Cardiologia06.txt	2658	247	9
7	Cardiologia07.txt	1915	321	17
8	Cardiologia08.txt	3029	139	5
9	Cardiologia09.txt	3168	176	6
10	Cardiologia10.txt	4389	169	4
11	Dermatologia01.txt	4956	188	4
12	Dermatologia02.txt	2150	255	12
13	Dermatologia03.txt	672	71	11
14	Dermatologia04.txt	1920	98	5
15	Dermatologia05.txt	2363	122	5
16	Dermatologia06.txt	4899	124	3
17	Dermatologia07.txt	5710	219	4
18	Dermatologia08.txt	2518	194	8
19	Dermatologia09.txt	1909	229	12

20	Dermatologia10.txt	3192	214	7
21	Epidemiologia01.txt	2258	123	5
22	Epidemiologia02.txt	3078	486	16
23	Epidemiologia03.txt	3052	287	9
24	Epidemiologia04.txt	1809	151	8
25	Epidemiologia05.txt	4668	148	3
26	Epidemiologia06.txt	3698	65	2
27	Epidemiologia07.txt	864	48	6
28	Epidemiologia08.txt	2568	250	10
29	Epidemiologia09.txt	2286	60	3
30	Epidemiologia10.txt	5514	157	3
31	Geriatria01.txt	4565	282	6
32	Geriatria02.txt	2766	88	3
33	Geriatria03.txt	3352	290	9
34	Geriatria04.txt	1214	54	4
35	Geriatria05.txt	2150	218	10
36	Geriatria06.txt	8581	183	2
37	Geriatria07.txt	2792	181	6
38	Geriatria08.txt	3846	229	6
39	Geriatria09.txt	7203	221	3
40	Geriatria10.txt	2447	387	16
41	Ginecologia01.txt	2725	255	9
42	Ginecologia02.txt	4217	300	7
43	Ginecologia03.txt	2715	212	8
44	Ginecologia04.txt	2381	231	10
45	Ginecologia05.txt	2804	241	9
46	Ginecologia06.txt	3490	308	9
47	Ginecologia07.txt	1752	199	11
48	Ginecologia08.txt	7951	124	2
49	Ginecologia09.txt	6127	147	2
50	Ginecologia10.txt	6376	126	2
51	Hematologia01.txt	1769	205	12
52	Hematologia02.txt	1836	219	12
53	Hematologia03.txt	3354	304	9
54	Hematologia04.txt	3222	238	7
55	Hematologia05.txt	1337	240	18
56	Hematologia06.txt	2985	134	4
57	Hematologia07.txt	5353	260	5
58	Hematologia08.txt	3325	345	10
59	Hematologia09.txt	1936	416	21
60	Hematologia10.txt	3531	199	6
61	Neurologia01.txt	6545	124	2
62	Neurologia02.txt	4819	133	3
63	Neurologia03.txt	3258	182	6

64	Neurologia04.txt	1753	154	9
65	Neurologia05.txt	3787	182	5
66	Neurologia06.txt	2461	233	9
67	Neurologia07.txt	1466	172	12
68	Neurologia08.txt	3059	240	8
69	Neurologia09.txt	2816	32	1
70	Neurologia10.txt	2242	80	4
71	Oncologia01.txt	3926	284	7
72	Oncologia02.txt	3015	101	3
73	Oncologia03.txt	5782	169	3
74	Oncologia04.txt	1891	349	18
75	Oncologia05.txt	2328	231	10
76	Oncologia06.txt	1888	176	9
77	Oncologia07.txt	3883	279	7
78	Oncologia08.txt	2627	180	7
79	Oncologia09.txt	2091	106	5
80	Oncologia10.txt	6085	164	3
81	Ortopedia01.txt	2010	139	7
82	Ortopedia02.txt	1561	166	11
83	Ortopedia03.txt	3263	267	8
84	Ortopedia04.txt	2328	231	10
85	Ortopedia05.txt	1826	264	14
86	Ortopedia06.txt	3060	92	3
87	Ortopedia07.txt	2295	259	11
88	Ortopedia08.txt	4674	307	7
89	Ortopedia09.txt	4029	297	7
90	Ortopedia10.txt	3147	306	10
91	Pediatria01.txt	4924	166	3
92	Pediatria02.txt	2415	136	6
93	Pediatria03.txt	4442	128	3
94	Pediatria04.txt	5260	167	3
95	Pediatria05.txt	6072	279	5
96	Pediatria06.txt	3456	288	8
97	Pediatria07.txt	4373	266	6
98	Pediatria08.txt	3574	240	7
99	Pediatria09.txt	4232	190	4
100	Pediatria10.txt	2707	343	13

Lista de percentual de compressão segundo a Equação 9 para o domínio Médico no idioma Inglês

		Texto Original	***Texto Sumarizado***	
Nº	**Nome**	**Número de Palavras**	**Número de Palavras**	**%**
1	Cardiologia01.txt	2274	233	10
2	Cardiologia02.txt	2040	242	12
3	Cardiologia03.txt	3111	247	8
4	Cardiologia04.txt	2156	230	11
5	Cardiologia05.txt	3735	286	8
6	Cardiologia06.txt	3376	271	8
7	Cardiologia07.txt	2610	214	8
8	Cardiologia08.txt	1919	114	6
9	Cardiologia09.txt	927	102	11
10	Cardiologia10.txt	1129	37	3
11	Dermatologia01.txt	1304	72	6
12	Dermatologia02.txt	472	71	15
13	Dermatologia03.txt	781	75	10
14	Dermatologia04.txt	1467	176	12
15	Dermatologia05.txt	2473	238	10
16	Dermatologia06.txt	1810	165	9
17	Dermatologia07.txt	3737	264	7
18	Dermatologia08.txt	5672	36	1
19	Dermatologia09.txt	2702	151	6
20	Dermatologia10.txt	3131	206	7
21	Epidemiologia01.txt	4974	159	3
22	Epidemiologia02.txt	8905	166	2
23	Epidemiologia03.txt	3279	78	2
24	Epidemiologia04.txt	1171	169	14
25	Epidemiologia05.txt	2869	178	6
26	Epidemiologia06.txt	3998	140	4
27	Epidemiologia07.txt	8657	153	2
28	Epidemiologia08.txt	2768	182	7
29	Epidemiologia09.txt	5080	152	3
30	Epidemiologia10.txt	5032	173	3
31	Geriatria01.txt	2976	149	5
32	Geriatria02.txt	3810	142	4
33	Geriatria03.txt	1536	239	16
34	Geriatria04.txt	1967	215	11
35	Geriatria05.txt	3474	249	7
36	Geriatria06.txt	3065	140	5
37	Geriatria07.txt	2205	258	12

38	Geriatria08.txt	2393	270	11
39	Geriatria09.txt	2104	150	7
40	Geriatria10.txt	3560	253	7
41	Ginecologia01.txt	1886	236	13
42	Ginecologia02.txt	2316	153	7
43	Ginecologia03.txt	2441	160	7
44	Ginecologia04.txt	2861	288	10
45	Ginecologia05.txt	1607	123	8
46	Ginecologia06.txt	4306	174	4
47	Ginecologia07.txt	659	107	16
48	Ginecologia08.txt	1076	153	14
49	Ginecologia09.txt	2196	240	11
50	Ginecologia10.txt	2565	280	11
51	Hematologia01.txt	4161	247	6
52	Hematologia02.txt	1410	182	13
53	Hematologia03.txt	3716	233	6
54	Hematologia04.txt	2590	446	17
55	Hematologia05.txt	1430	156	11
56	Hematologia06.txt	885	183	21
57	Hematologia07.txt	930	141	15
58	Hematologia08.txt	2974	148	5
59	Hematologia09.txt	3806	254	7
60	Hematologia10.txt	1861	73	4
61	Neurologia01.txt	867	391	45
62	Neurologia02.txt	1784	245	14
63	Neurologia03.txt	1555	193	12
64	Neurologia04.txt	1855	236	13
65	Neurologia05.txt	1510	202	13
66	Neurologia06.txt	1792	155	9
67	Neurologia07.txt	757	128	17
68	Neurologia08.txt	3162	90	3
69	Neurologia09.txt	1653	140	8
70	Neurologia10.txt	1218	115	9
71	Oncologia01.txt	1815	128	7
72	Oncologia02.txt	2829	128	5
73	Oncologia03.txt	4149	145	3
74	Oncologia04.txt	1815	349	19
75	Oncologia05.txt	4162	151	4
76	Oncologia06.txt	3450	148	4
77	Oncologia07.txt	2974	148	5
78	Oncologia08.txt	2616	144	6
79	Oncologia09.txt	2800	131	5
80	Oncologia10.txt	4272	160	4
81	Ortopedia01.txt	3077	232	8

82	Ortopedia02.txt	3472	243	7
83	Ortopedia03.txt	1482	292	20
84	Ortopedia04.txt	1377	156	11
85	Ortopedia05.txt	1584	227	14
86	Ortopedia06.txt	4063	236	6
87	Ortopedia07.txt	1542	99	6
88	Ortopedia08.txt	1878	328	17
89	Ortopedia09.txt	3140	204	6
90	Ortopedia10.txt	2552	259	10
91	Pediatria01.txt	2754	232	8
92	Pediatria02.txt	2831	171	6
93	Pediatria03.txt	1792	292	16
94	Pediatria04.txt	3463	268	8
95	Pediatria05.txt	3113	232	7
96	Pediatria06.txt	2677	263	10
97	Pediatria07.txt	4373	149	3
98	Pediatria08.txt	2451	285	12
99	Pediatria09.txt	4415	172	4
100	Pediatria10.txt	2773	252	9

APÊNDICE B

Domínio Jurídico

Métricas Externas para o Idioma Português

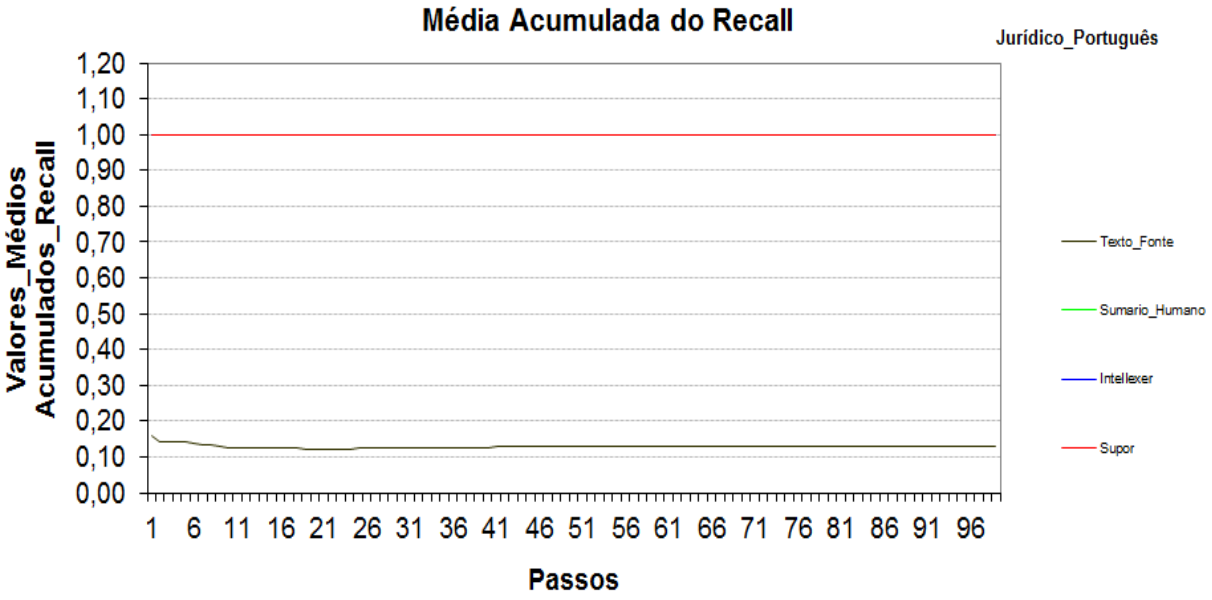

Gráfico 1: resultados dos testes no domínio Jurídico no idioma português, com medida externa Recall.

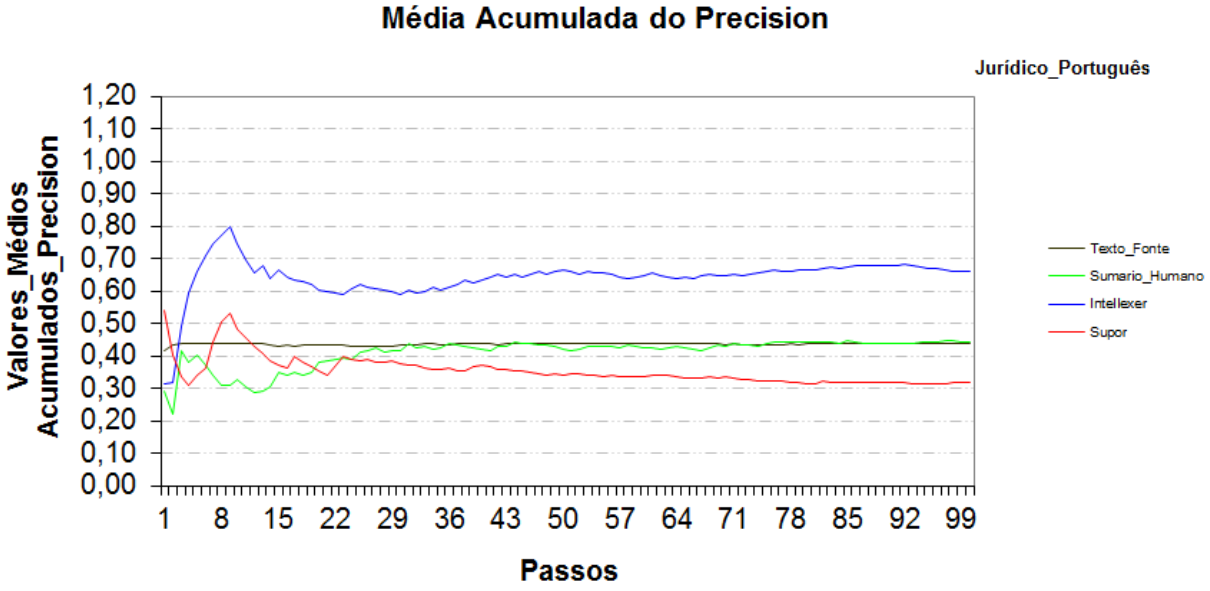

Gráfico 2: resultados dos testes no domínio Jurídico no idioma português, com medida externa Precision.

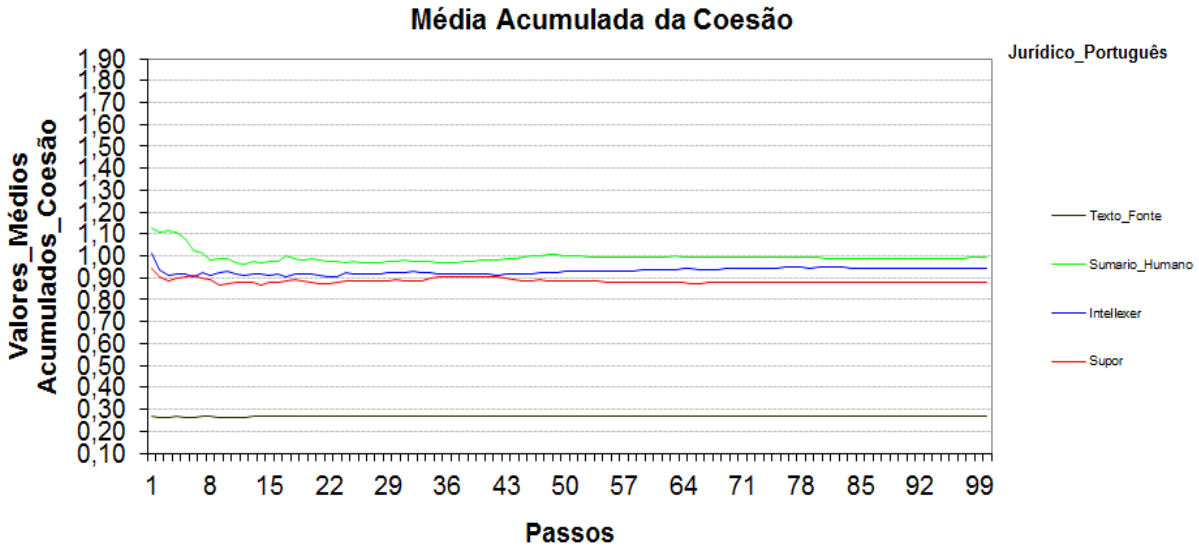

Gráfico 03: resultados dos testes no domínio Jurídico no idioma português, com medida interna Coesão.

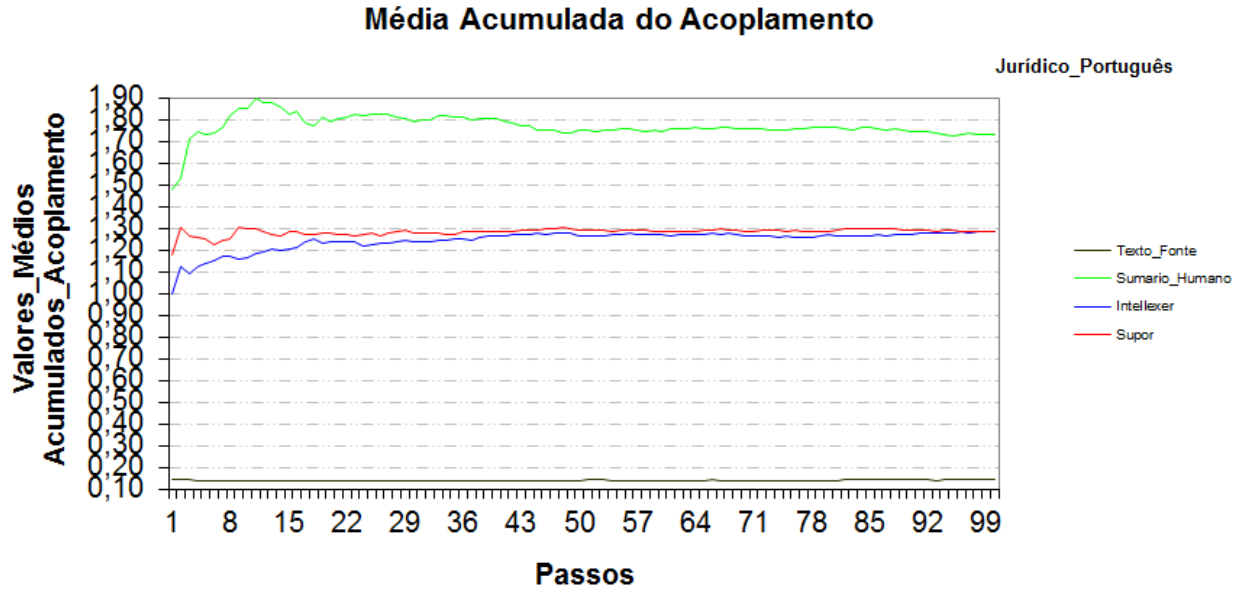

Gráfico 4: resultados dos testes no domínio Jurídico no idioma português, com medida interna Acoplamento.

Domínio Médico

Métricas Externas para o Idioma Português

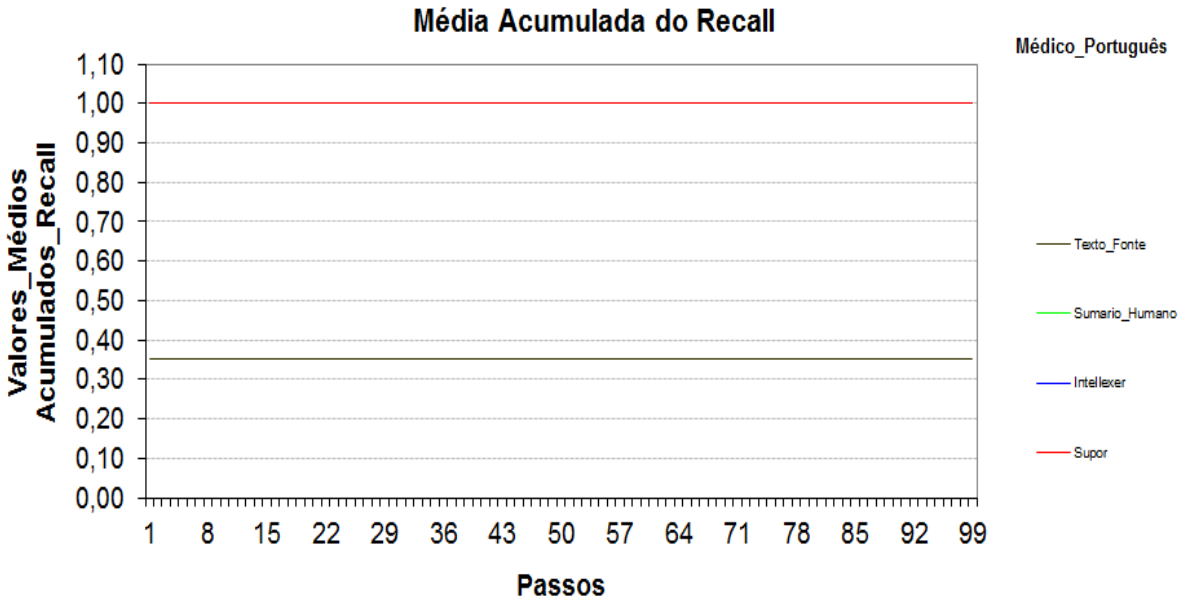

Gráfico 5: resultados dos testes no domínio Médico no idioma Português, com medida externa Recall.

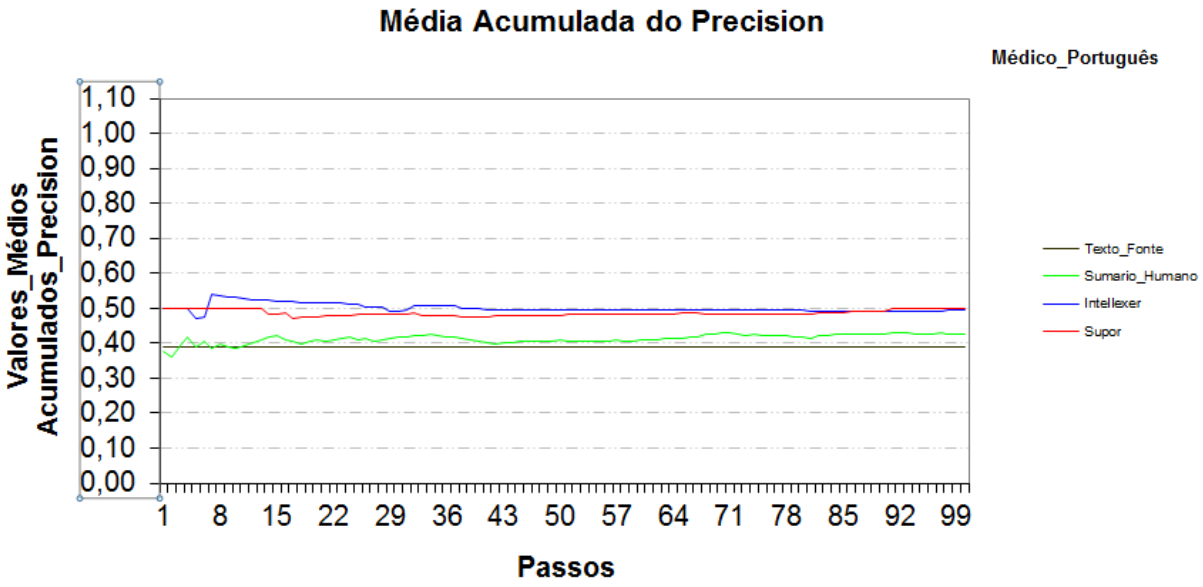

Gráfico 6: resultados dos testes no domínio Médico no idioma Português, com medida externa Precision.

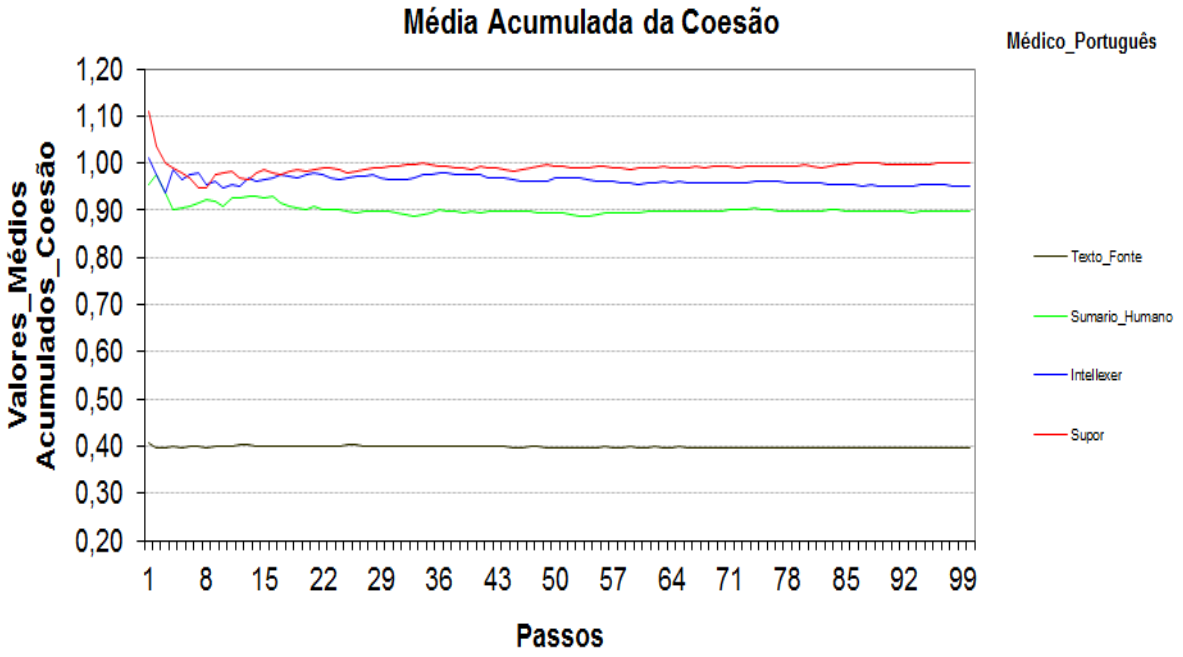

Gráfico 7: resultados dos testes no domínio Médico no idioma português, com medida interna Coesão.

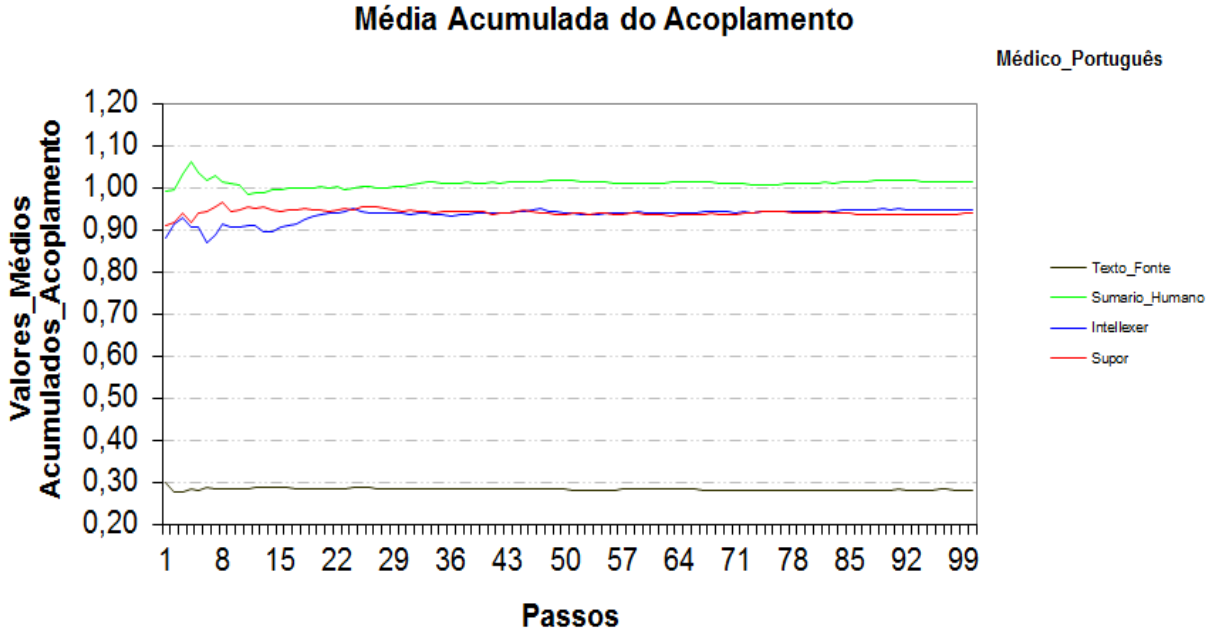

Gráfico 8: resultados dos testes no domínio Médico no idioma português, com medida interna Acoplamento.

Métricas Externas para o Idioma Inglês

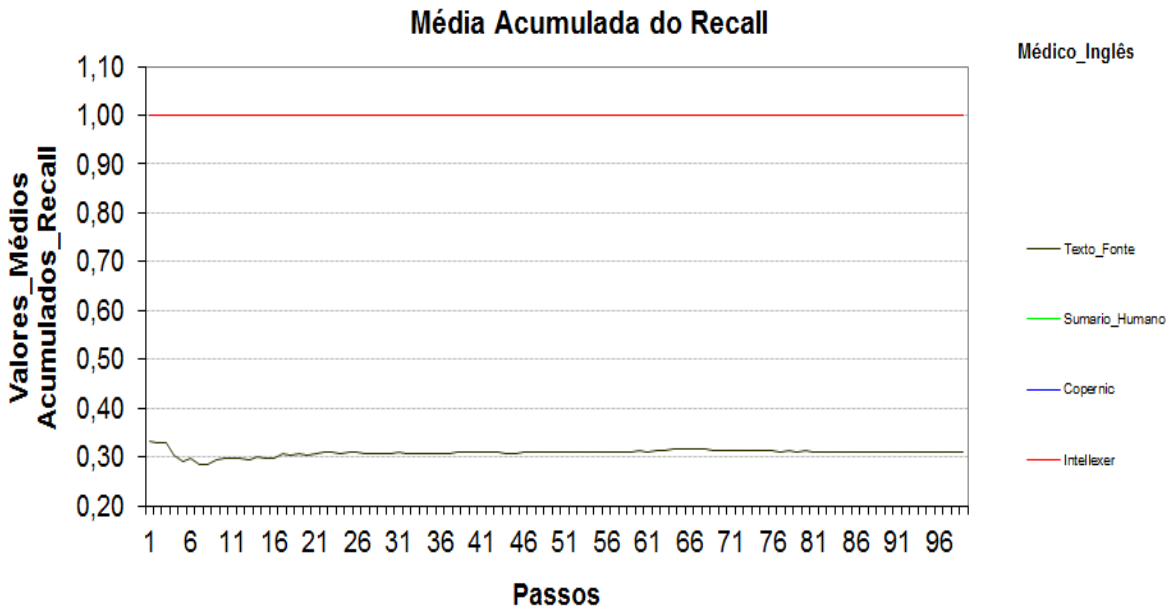

Gráfico 10: resultados dos testes no domínio Médico no idioma Inglês, com medida externa Recall.

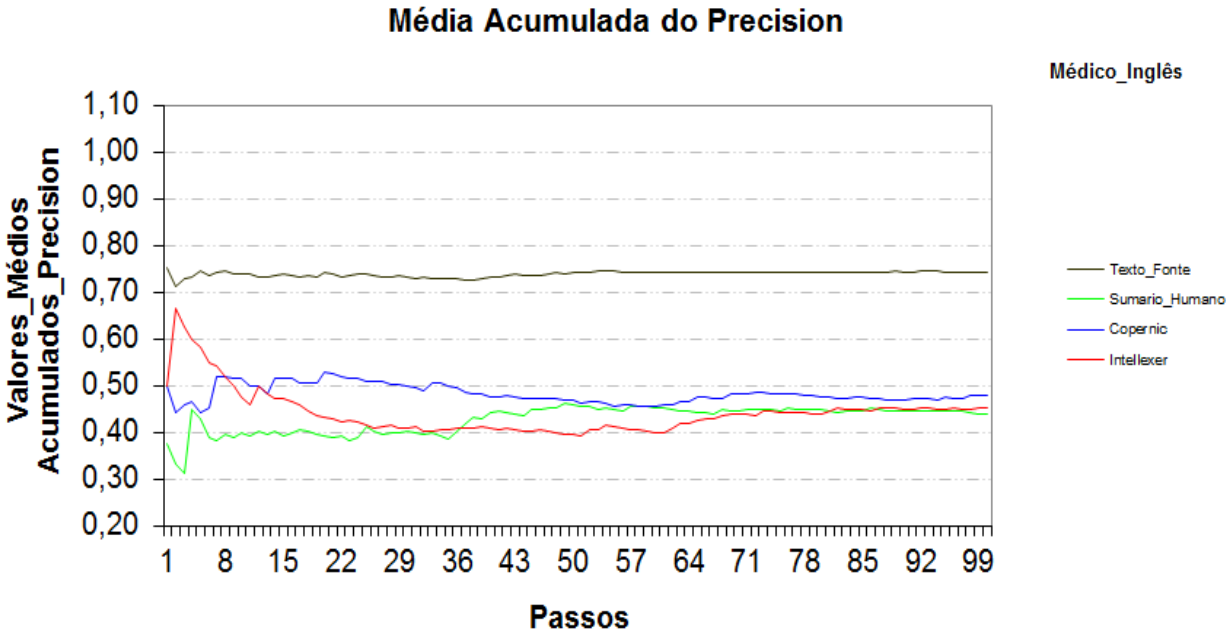

Gráfico 11: resultados dos testes no domínio Médico no idioma Inglês, com medida externa Precision.

Métricas Internas para o Idioma Inglês

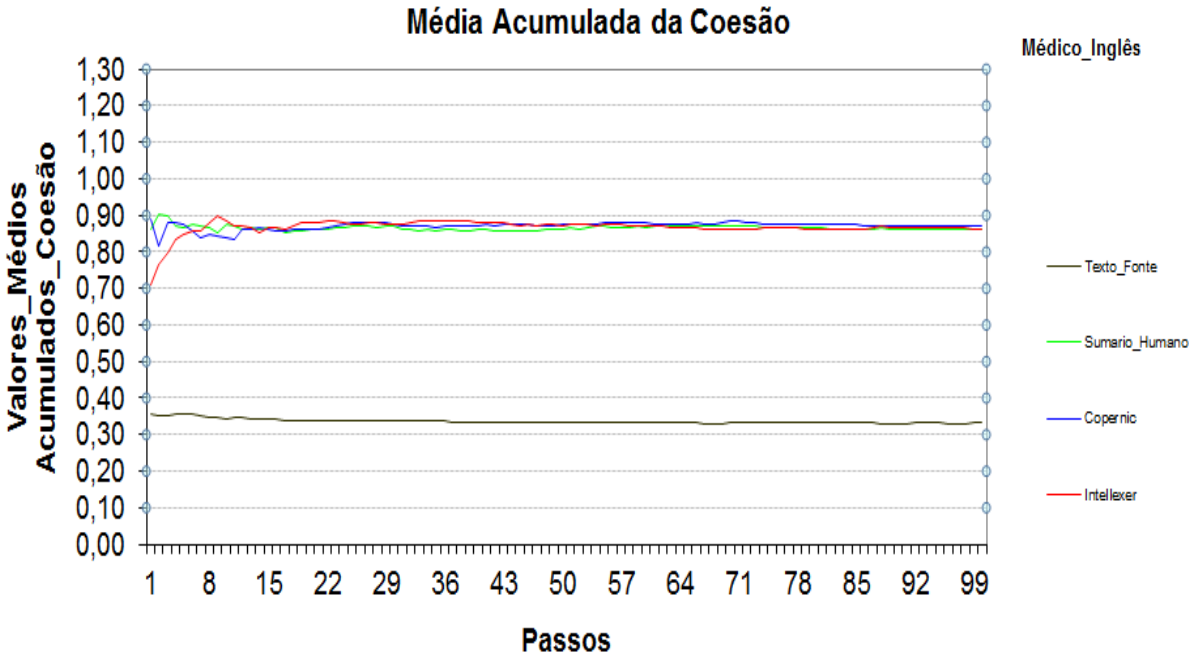

Gráfico 12: resultados dos testes no domínio Médico no idioma Inglês, com medida interna Coesão.

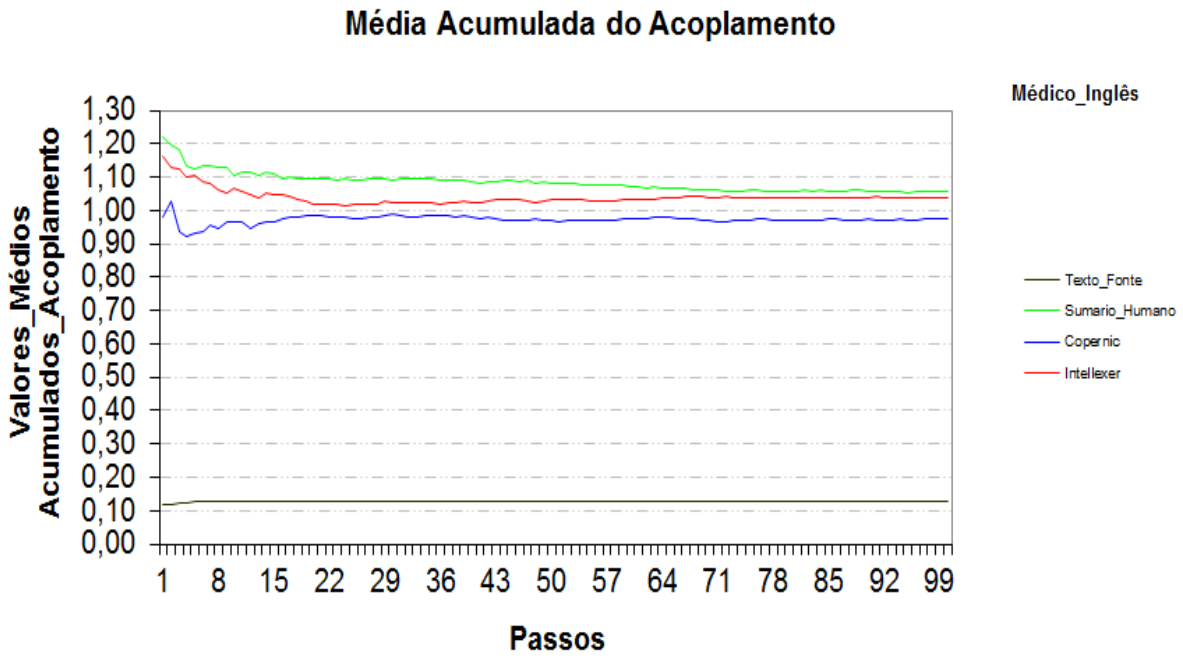

Gráfico 13: resultados dos testes no domínio Médico no idioma Inglês, com medida interna Acoplamento.

APÊNDICE C

SOFTWARES COM OS TESTES ESTATÍSTICOS

Existem vários softwares estatísticos tais como: *Statistica*, *Statgraphics*, *SPSS*, *Minitab*, *SAS*, *SPHINX*, *WINKS*, entre outros. No entanto são softwares geralmente de custo elevado e envolvem um aprendizado específico de usabilidade.

Neste trabalho foi utilizado para realizar os testes estatísticos dos experimentos e comprovação da hipótese o seguinte software *StatPlus*® (http://www.analystsoft.com/en/products/statplus/) uma versão *Trial*, este software foi escolhidos porque contém os testes estatísticos ANOVA de Friedman e o coeficiente de concordância de Kendall adotado neste trabalho.

Tabela 1: Teste Estatístico dos resultados usando a medida Coesão no idioma português para o domínio Jurídico.

Comparando amostras múltiplas relacionadas			
N	100	*Graus de liberdade*	3
qui-quadrado	299,7027	*p-nível*	0
Coef. de concordância de Kendall	0,999	*Ordem médio*	0,999
	Ordem médio	**Soma de ordens**	**Média**
Texto-Fonte	1	100	0,2695
Sumario_Humano	4	400	0,9932
Intellexer	2,995	299,5	0,9306
SuPor2	2,005	200,5	0,8852

Tabela 2: Teste Estatístico dos resultados usando a medida Acoplamento no idioma português para o domínio Jurídico.

Comparando amostras múltiplas relacionadas			
N	100	*Graus de liberdade*	3
qui-quadrado	298,8434	*p-nível*	0
Coef. de concordância de Kendall	0,9961	*Ordem médio*	0,9961
	Ordem médio	**Soma de ordens**	**Média**
Texto-Fonte	1	100	0,1402
Sumario_Humano	4	400	1,7679
Intellexer	2,02	202	1,2443
SuPor2	2,98	298	1,2818

Tabela 3: Teste Estatístico dos resultados usando a medida Recall no idioma português para o domínio Jurídico.

Comparando amostras múltiplas relacionadas			
N	100	*Graus de liberdade*	3
qui-quadrado	300	*p-nível*	0
Coef. de concordância de Kendall	1	*Ordem médio*	1
	Ordem médio	**Soma de ordens**	**Média**
Texto_Fonte	1	100	0,1291
Sumario_Humano	3	300	1
Intellexer	3	300	1
Supor	3	300	1

Tabela 4: Teste Estatístico dos resultados usando a medida Precision no idioma português para o domínio Jurídico.

Comparando amostras múltiplas relacionadas			
N	100	*Graus de liberdade*	3
qui-quadrado	249,9276	*p-nível*	0
Coef. de concordância de Kendall	0,8331	*Ordem médio*	0,8314
	Ordem médio	**Soma de ordens**	**Média**
Texto_Fonte	2,78	278	0,4379
Sumario_Humano	2,02	202	0,4102
Intellexer	3,96	396	0,6422
Supor	1,24	124	0,354

Tabela 5: Teste Estatístico dos resultados usando a medida Coesão no idioma português para o domínio Médico.

Comparando amostras múltiplas relacionadas			
N	100	*Graus de liberdade*	3
qui-quadrado	296,198	*p-nível*	0
Coef. de concordância de Kendall	0,9873	*Ordem médio*	0,9872
	Ordem médio	**Soma de ordens**	**Média**
Texto_Fonte	1	100	0,4001
Sumario_Humano	2,01	201	0,903
Intellexer	3,025	302,5	0,9633
Supor	3,965	396,5	0,9914

Tabela 6: Teste Estatístico dos resultados usando a medida Acoplamento no idioma português para o domínio Médico.

Comparando amostras múltiplas relacionadas			
N	100	*Graus de liberdade*	3
qui-quadrado	285,0443	*p-nível*	0
Coef. de concordância de Kendall	0,9501	*Ordem médio*	0,9496
	Ordem médio	**Soma de ordens**	**Média**
Texto_Fonte	1	100	0,2812
Sumario_Humano	4	400	1,0094
Intellexer	2,435	243,5	0,935
Supor	2,565	256,5	0,9417

Tabela 7: Teste Estatístico dos resultados usando a medida Recall no idioma português para o domínio Médico.

Comparando amostras múltiplas relacionadas			
N	100	*Graus de liberdade*	3
qui-quadrado	300	*p-nível*	0
Coef. de concordância de Kendall	1	*Ordem médio*	1
	Ordem médio	**Soma de ordens**	**Média**
Texto_Fonte	1	100	0,35
Sumario_Humano	3	300	1
Intellexer	3	300	1
Supor	3	300	1

Tabela 8: Teste Estatístico dos resultados usando a medida Precision no idioma português para o domínio Médico.

Comparando amostras múltiplas relacionadas			
N	100	*Graus de liberdade*	3
qui-quadrado	282,1951	*p-nível*	0
Coef. de concordância de Kendall	0,9407	*Ordem médio*	0,9401
	Ordem médio	**Soma de ordens**	**Média**
Texto_Fonte	1,04	104	0,39
Sumario_Humano	1,96	196	0,4138
Intellexer	3,82	382	0,5004
Supor	3,18	318	0,4854

Tabela 9: Teste Estatístico dos resultados usando a medida Coesão no idioma inglês para o domínio Médico.

Comparando amostras múltiplas relacionadas			
N	100	*Graus de liberdade*	3
qui-quadrado	218,7428	*p-nível*	0
Coef. de concordância de Kendall	0,7291	*Ordem médio*	0,7264
	Ordem médio	**Soma de ordens**	**Média**
Texto-Fonte	1	100	0,3351
Sumario_Humano	2,525	252,5	0,865
Copernic	3,415	341,5	0,871
Intellexer	3,06	306	0,8666

Tabela 10: Teste Estatístico dos resultados usando a medida Acoplamento no idioma inglês para o domínio Médico.

Comparando amostras múltiplas relacionadas			
N	100	*Graus de liberdade*	3
qui-quadrado	300	*p-nível*	0
Coef. de concordância de Kendall	1	*Ordem médio*	1
	Ordem médio	**Soma de ordens**	**Média**
Texto-Fonte	1	100	0,1296
Sumario_Humano	4	400	1,0844
Copernic	2	200	0,9723
Intellexer	3	300	1,0401

Tabela 11: Teste Estatístico dos resultados usando a medida Recall no idioma inglês para o domínio Médico.

Comparando amostras múltiplas relacionadas			
N	100	*Graus de liberdade*	3
qui-quadrado	300	*p-nível*	0
Coef. de concordância de Kendall	1	*Ordem médio*	1
	Ordem médio	**Soma de ordens**	**Média**
Texto_Fonte	1	100	0,309
Sumario_Humano	3	300	1
Copernic	3	300	1
Intellexer	3	300	1

Tabela 12: Teste Estatístico dos resultados usando a medida Precision no idioma inglês para o domínio Médico.

Comparando amostras múltiplas relacionadas			
N	100	*Graus de liberdade*	3
qui-quadrado	256,4772	*p-nível*	0
Coef. de concordância de Kendall	0,8549	*Ordem médio*	0,8535
	Ordem médio	**Soma de ordens**	**Média**
Texto_Fonte	4	400	0,7374
Sumario_Humano	1,505	150,5	0,4272
Copernic	2,9	290	0,4815
Intellexer	1,595	159,5	0,4415