MARCUS VINICIUS CARVALHO GUELPELI

CASSIOPEIA: UM MODELO DE AGRUPAMENTO DE TEXTOS BASEADO EM SUMARIZAÇÃO.

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense, como requisito parcial para obtenção do Grau de Doutor. Área de Concentração: Inteligência Artificial.

Orientadora: Ph.D. Ana Cristina Bicharra Garcia
Coorientador: Ph.D. António Horta Branco

Niterói
2012

Ficha Catalográfica elaborada pela Biblioteca da Escola de Engenharia e Instituto de Computação da UFF

G925 Guelpeli, Marcus Vinicius Carvalho
Cassiopeia : um modelo de agrupamento de textos baseado em sumarização / Marcus Carvalho Guelpeli. – Niterói, RJ : [s.n.], 2012.
220 f.

Tese (Doutorado em Computação) - Universidade Federal Fluminense, 2012.
Orientadores: Ana Cristina Bicharra Garcia, António Horta Branco.

1. Recuperação da informação. 2. Sumarização de texto. 3. Cassiopeia. I. Título.

CDD 005.74

MARCUS VINICIUS CARVALHO GUELPELI.

CASSIOPEIA: UM MODELO DE AGRUPAMENTO DE TEXTOS BASEADO EM SUMARIZAÇÃO.

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense, como requisito parcial para obtenção do Grau de Doutor. Área de Concentração: Inteligência Artificial.

MARCUS VINICIUS CARVALHO GUELPELI.

CASSIOPEIA: UM MODELO DE AGRUPAMENTO HIERÁRQUICO DE TEXTOS BASEADO EM SUMARIZAÇÃO.

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense, como requisito parcial para obtenção do Grau de Doutor. Área de Concentração: Inteligência Artificial.

Aprovada em Março de 2012.

BANCA EXAMINADORA:

Ana Cristina Bicharra Garcia, Ph.D – Orientadora
UFF – Universidade Federal Fluminense

Antônio Horta Branco, Ph.D – Coorientador
Universidade de Lisboa

Bianca Zadrozny, Ph.D.
UFF – Universidade Federal Fluminense

Aura Conci, Ph.D
UFF – Universidade Federal Fluminense

Kate Cerqueira Revoredo, DSc
UNIRIO- Universidade Federal do Estado do Rio de Janeiro

Cristiano Maciel, DSc.
UFMT – Universidade Federal de Mato Grosso

Adriana Santarosa Vivacqua, DSc.
UFRJ - Universidade Federal Rio de Janeiro

Ao meu filho João Miguel Pinto Guelpeli por ser fonte de inspiração na minha vida e com sua presença iluminada me dar força nos momentos de fragilidade.

A minha adorada esposa, Alison Cristine Pinto Guelpeli pelo apoio, compreensão, carinho, dedicação, incentivo e por sua grande coragem em deixar sua vida profissional e junto comigo e com nosso filho partir para uma grande aventura, fora do nosso país.

AGRADECIMENTOS

É impossível desenvolver uma tese sozinho. Este trabalho é uma junção de forças positivas e negativas que acabaram por formar uma grande e forte corrente.

De alguma forma todos que estão aqui lembrados fazem parte dos elos dessa grande corrente, pois colaboraram para que pudesse estar concluída a minha tese.

Gostaria inicialmente de agradecer as duas pessoas mais importantes da minha vida. Meu filho João Miguel e minha esposa Alison Cristine, pelo enorme apoio que sempre tive de vocês, sem os quais não teria tido força suficiente para chegar ao final dessa árdua missão.

Ao meu filho do coração Didi (Edgleisson Cunha) por tudo que você é para nossa família, pelo apoio de sempre e pelo enorme desafio de ter ficado no Brasil olhando tomando conta de nossa casa e cuidado dos meus outros filhos enquanto estivemos em Portugal.

Quantas vezes nas noites solitárias no escritório, trabalhando na tese, tive os seus olhares e sua grande companhia. Os meus filhos caninos que tanto amo Laisa (*in memória*), Aika (*in memória*), Uclan (*in memória*), Thor e mais recente o Bob.

Aos meus orientadores, a professora Ana Cristina e ao professor António Branco, pelo enorme tempo de dedicação a minha orientação.

Aos professores Luís Correia e Flávia Bernardini que estimo muito, pois contribuíram para o enriquecimento e evolução do meu trabalho.

Amigos de doutorado, Cristiano Maciel e Alessandro Copetti, pela acolhida, carinho e a explícita demonstração de companheirismo ao longo dessa jornada.

Em Portugal aos amigos do NLX, Sara Silveira, João Silva, Catarina Carvalheiro, Clara Pinto, Silvia Pereira, Sérgio Castro, Ruben Reis, Patrícia Nunes, Rosa del Gaudio e Francisco Costa, onde deixei um forte laço de amizade e fui muito bem acolhido, aprendi muito e tive uma convivência importante para minha vida pessoal.

A minha coordenadora e amiga de NEAD-UBM, professora Estela Oliveira que em todo o momento do meu doutorado compreendeu as minhas urgências e saídas do meu posto de trabalho para resolver problemas no doutorado.

Um grande parceiro de código e troca de ideias do inicio até o final, o amigo Rafael Santiago. No início do meu doutorado até o momento que foi para seu doutorado em Portugal, o irmão Ricardo Maia. E mais recente, o amigo Marcelo Arantes.

Aos experientes amigos e professores Ladário da Silva e Denner Martins, por acreditarem na minha capacidade de fazer um doutoramento, pelo incentivo e onde por muitas vezes fui ouvir seus sábios conselhos.

Nas diversas vezes em que tive que fazer simulações, sempre contei com apoio e consentimento do coordenador do NTI Alexsander em usar os laboratórios do UBM e nesses laboratórios contei sempre com a ajuda dos amigos desse setor Tarcisio Filho e Clério.

Aos meus colegas do UBM, Eduardo Arbex, Anderson Simeão, Cida Coelho e Ana Lúcia Marquez que de alguma forma me apoiaram nessa longa jornada com suas palavras e ações amigas.

A todos os colegas do AddLabs aqui representados por Fernando Pinto, Cida e as secretárias Tânia Lattanzi e Adriana.

As secretárias da Pós da UFF, representadas pela Ângela Dias, Teresa Cancela e Maria Freitas.

Aos professores da UFF aos quais tive a oportunidade de apreender com suas aulas.

A todos os colegas do IC da UFF representados pelo amigo Mário Mestria.

Algumas instituições foram fundamentais para realização do meu doutorado. Primeiramente o Centro Universitário de Barra Mansa - UBM que sempre me apoiou em todos os aspectos do início ao fim do meu doutoramento.

A CAPES pelo apoio e financiamento do doutorado sanduíche em Portugal.

A UFF e a Universidade de Lisboa onde pude me aperfeiçoar.

Por fim, esta bebida amiga, o café que manteve acordado e as madrugadas que vi passar diante dos meus olhos que tanto me ajudaram com seu silêncio.

A todos vocês meu muito obrigado!

Objetivo Alcançado!

“Continue com fome. Continue bobo. E eu sempre desejei isso para mim mesmo. E agora, quando vocês se formam e começam de novo, eu desejo isso para vocês. Continuem com fome. Continuem bobos.”

(Final do discurso proferido para turma de formando da Universidade de Stanford, Steve Jobs, 2005).

RESUMO

Esta tese contribui para a área de recuperação de informação (*information retrieval*) que utiliza a técnica de agrupamento de documentos (*document clustering*) para apoiar a busca e recuperação de textos em grandes bases textuais. A maioria dos agrupadores textuais são desenvolvidos para domínios específicos. Quando usados em domínios diferentes, apresentam uma quebra do seu desempenho, dificultando a recuperação de textos. Essa dependência do domínio está intimamente ligada à escolha dos atributos que são usados para fazer o agrupamento e à definição do conjunto de palavras que devem ser retiradas (*stopwords*). Os atributos e as *stopwords* constituem a base para a definição do corte de Luhn. Esta tese explora a hipótese de que haverá melhoria do desempenho dos agrupadores de documentos através da inclusão de sumarização dos documentos, na fase de pré-processamento, e do uso do processo de agrupamento hierárquico dos documentos, na fase de processamento. Como contribuição fundamental, foi ainda definido um novo método para o corte de Luhn, em vista de se melhorar o desempenho do agrupamento de textos, com o ganho adicional de passar-se a ter independência tanto do domínio como do idioma. Para a avaliação desta hipótese foi desenvolvido o modelo Cassiopeia, integrando estas novidades metodológicas. O modelo foi testado com corpora de bases públicas, oriundos dos domínios jornalístico, jurídico e médico, e para os idiomas português e inglês. Os resultados obtidos mostram um grande avanço no desempenho do agrupamento de documentos usando o Cassiopeia. Esse avanço foi medido em termos das usuais métricas de precisão, recuperação de informação, coesão e acoplamento. Em consequência deste avanço, obtém-se a atenuação da sobrecarga de informação no momento da recuperação de textos.

Palavras-chave: recuperação de informação, agrupamento de texto, sumarização de texto, corte de Luhn.

ABSTRACT

This thesis contributor to the field of information retrieval, using document clustering techniques to support text searches and retrievals within larger text databases. Most text clusters are developed for specific domains. When used in different domains, they may suffer a loss in performance, which renders text retrieval difficult. This domain dependency is intimately linked to the choice of attributes used to create the clustering and the definition of the word set that should be removed (stopwords). Attributes and stopwords compose the foundation for defining Luhn's cut-off. This paper explores the hypothesis that the performance of document clustering will improve by including document summarization during the pre-processing stage, and by including the use of hierarchical document clustering during the processing stage. A major contribution, furthermore, was the definition of a new method for Luhn's cut-off, with the intention of improving text clustering performance and with the additional advantage of independence of both domain and language. In order to evaluate this hypothesis, the Cassiopeia model was developed by integrating these methodological innovations. This model was tested with corpora of public databases, originating from the journalistic, legal and medical domains, and for both Portuguese and English languages. The results reveal great progress in the performance of document clustering using the Cassiopeia, measured with respect to the usual precision measurements, information retrieval, cohesion, and coupling. Consequently, information overload is mitigated during text retrieval.

Keywords: information retrieval, text clustering, text summarization, Luhn's cut-off.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

C_StpWrd	**C***om* **St***o***pw***o***rd**
DF	**D***ocument* **F***requency*
EI	***E****xtração de* ***I****nformação em Texto*
KDT	***K****nowledge* ***D****iscovery in****T****exts*
LC	***L****inguística de* ***C****orpus*
MT	***M****ineração* de ***T****exto*
NILC	***N****úcleo* ***I****nterinstitucional de* ***L****ingüística* ***C****omputacional*
PC	***P****roblemas de* ***C****lusterização*
PCA	***P****roblema de* ***C****lusterização* ***A****utomática*
RI	***R****ecuperação de* **I***nformação em texto*
SA	***S****umarização* ***A****utomática*
S_StpWrd	**S***em* **St***o***pw***o***rd**
TAC	***T****ext Analysis* ***C****onference*
TF	**T***erm* **F***requency*
TF-IDF	**T***erm* **F***requency-* **I***nverse* **D***ocument* **F***requency*

SUMÁRIO

CAPÍTULO 1 – INTRODUÇÃO

Atualmente, são muitas as informações textuais disponibilizadas na internet, chegando de várias maneiras e com diferentes finalidades, o que torna impossível assimilar todas. Selecionar as que melhor correspondem aos interesses do público facilita o processamento e a recuperação dessas informações.

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. Sendo assim, faz-se necessário o tratamento dessas informações, já que a capacidade humana de leitura e registro é limitada.

O estudo de Chen (2001) mostrou que 80% do conteúdo da internet estava em formato textual. Esses números foram motivados pelos avanços nas tecnologias da informação. As informações a cada momento são mais acessíveis. Um levantamento, realizado no ano de 2000, por Lyman (2000), indicou 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 páginas disponíveis. Smyth *et al.* (2004) revelaram a existência de 10 bilhões de documentos no ano de 2004. Shaw (2005) estimou em, aproximadamente, treze bilhões o número de páginas em 2005. De acordo com Gantz e Reinsel (2010), para o final de 2007 a estimativa foi de 487 exabytes de informação digital. De acordo com relatório de Bohn *et al.* (2010), de 2007 a 2010, o crescimento foi 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).

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

Analisando o volume de informação, este trabalho contribui para a área de Recuperação de Informação, uma subárea da Mineração de Texto (*Text Mining*) – MT. Na literatura, observam-se três grandes áreas da MT, segundo Loh (2001), Wives (2004) e Hotho *et al.* (2005): Extração de Informação em Texto – EI; Recuperação de Informação em Texto – RI e Descoberta de Conhecimento em Texto (*Knowledge Discovery from Texts*) – KDT.

Segundo Rezende *et al.* (2011), uma coleção de textos pode ter milhares de termos[1] 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.

Com a finalidade de representar o espaço de busca na RI, existe, na literatura, uma taxonomia de modelos, usados na representação do espaço amostral, incluindo o booleano (WARTIK, 1992), o espaço-vetorial (SALTON *et al.* 1997), o probabilístico (VAN RIJSBERGEN, 1992), o difuso (SUBASIC e HUETTNER, 2001), o da busca direta (BAEZA e RIBEIRO, 1999) e os lógicos indutivos (HU e ATWELL, 2003).

A escolha do modelo para representação do espaço amostral influencia a questão da alta dimensionalidade e os dados esparsos em RI. Isso ocorre porque toda palavra, no documento, pode ser considerada como um atributo. Na consideração de um espaço-vetorial, no qual cada palavra representa uma dimensão, têm-se tantas dimensões quantas palavras diferentes. Assim, a escolha do espaço amostral tem uma influência na *performance* da RI. A maioria das representações do espaço amostral em RI é feita através da matriz documento-termo, o que leva a uma solução com alta dimensionalidade e dados esparsos, que serão abordados no capítulo 2, seção 2.4.

Para atenuar a alta dimensionalidade e os dados esparsos, segundo Howland e Park (2007), o tratamento dos dados no pré-processamento da RI é crucial, buscando-se reduzir o número de palavras, no intuito de minimizar a alta dimensionalidade e os dados esparsos. Esses estudos são encontrados na literatura, como em (CRESTANI e RIJSBERGEN, 1997), (JONES e WILLET, 1997), (HOWLAND e PARK, 2007), (NOGUEIRA, 2009), (OLIVEIRA, 2009), (TRAINA e SILVA, 2010), (REZENDE *et al.*, 2011), (BOTELHO, 2011), (LOPES, 2011), entre outros, que não serão abordados nesta tese.

Uma das soluções adotadas para atenuar a questão da alta dimensionalidade e dos dados esparsos bastante comum, mas não a única na literatura, vem a ser a proposta de 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 *stopword*s são palavras que aparecem em todos os tipos de textos e não são capazes de

[1] Neste trabalho, palavra, termo ou atributo serão usados como sinônimos. Atributo, para computação, é uma propriedade característica, informação ou parâmetro que os objetos podem ter e compartilhar. Em dados estruturados, é forma de dimensionar um objeto. Em texto, a palavra é a única forma dimensionável, assim, para dados não estruturados (formato textual), a palavra é um atributo.

colaborar para a recuperação de textos relativos a um assunto específico. As *stopwords* serão discutidas na seção 2.4 e na subseção 3.1.1.

Mesmo após a retirada das *stopwords*, ainda existe um número muito grande de atributos (WIVES, 2004). O que ocorre, na maioria dos sistemas, é a execução de uma fase denominada seleção dos atributos. Segundo Nogueira (2009), a seleção de atributos tem um fator decisivo para a boa qualidade do melhor desempenho da RI. Há variações de métodos e formas de definir a seleção dos atributos, ou seja, através de algum tipo de heurística.

Após a diminuição desses atributos, uma forma usual, segundo Feldman e Sanger (2006), vem a ser a etapa de organização, por meio da técnica de agrupamento de textos, das informações textuais em RI.

De acordo com Loh (2001), Wives (2004) e Lopes (2011), a técnica de agrupamento não deve sofrer intervenção humana, justificando-se o fato pelo grande volume de informações a ser analisado. A técnica de agrupamento faz com que os textos de um mesmo grupo tenham alta similaridade, mas sejam dissimilares em relação aos documentos de outros grupos. Para organizar esses documentos, usa-se uma estrutura hierárquica que, no final do processamento, vai comportar os textos distribuídos em grupos e subgrupos e cada texto deve estar relacionado com o mesmo tema e/ou assunto, ou seja, devem ter similaridade. A estrutura hierárquica deve possibilitar uma organização *topdown*. Cada grupo superior apresenta textos com assuntos mais genéricos e os subgrupos, temas mais específicos. Para criar os agrupamentos, na estrutura hierárquica, Loh (2001) e Wives (2004) citam em seus trabalhos o algoritmo a ser usado na criação da estrutura e o que garante a similaridade dos textos. Esses algoritmos serão apresentados no capítulo 3.

Nos agrupadores textuais usados em RI, foi observado um problema estudado neste trabalho. Eles não são bem avaliados fora do domínio específico para o qual foram projetados, ou seja, são dependentes do domínio, e seu desempenho, fora deste domínio, compromete a estrutura hierárquica gerada e, consequentemente, as avaliações dos agrupamentos são ruins, ocasionando sobrecarga de informação para a RI. A fundamentação teórica sobre este tema será apresentada no capítulo 2, e no capítulo 3, serão formalizados os algoritmos usados na criação da estrutura hierárquica e na garantia da similaridade entre os documentos textuais.

Este trabalho apresenta o modelo Cassiopeia para agrupamento de documentos, visando a melhorar a precisão, na recuperação de documentos, a coesão e o acoplamento dos grupos de documentos formados.

O modelo Cassiopeia foi criado para possibilitar o agrupamento dos textos com maior qualidade nas avaliações em domínios distintos e/ou antagônicos, independentes do idioma, avaliados pelas métricas externas e internas. Nos experimentos serão usados *copora,* em domínios

e idiomas diferentes. Este modelo se apresenta como uma proposta poliestruturada, que agrega a abordagem de sumarização e agrupamento de texto. Atenua a alta dimensionalidade e dados esparsos, definindo um novo método do corte de Luhn, e dessa forma melhora os agrupamentos, viabiliza a representação do espaço amostral, usando a solução de vetores denominados centroides, e minimiza o problema da sobrecarga de informação em RI.

1.1 MOTIVAÇÃO

No exame da literatura levantada, dentro da área de RI e na subárea de agrupamento de texto, uma nova questão de pesquisa não foi previamente considerada por outros pesquisadores. Por outro lado, como foi visto anteriormente, os agrupadores textuais, quando usados em domínios para os quais não foram projetados apresentam maus resultados, além de forte dependência dos idiomas para os quais foram desenvolvidos.

Grande parte desses agrupadores usam, como solução para representação do seu espaço amostral, uma matriz de documento-termo, gerando assim uma alta complexidade de armazenamento, e ocasionando um espaço de alta dimensionalidade e com dados esparsos.

A seleção dos atributos é um outro ponto de motivação para desenvolvimento desse trabalho, e está relacionada à qualidade das palavras escolhidas, e é fundamental para todo o processo. A escolha dos atributos a serem usados é fator determinante e decisivo para a boa avaliação dos agrupadores de textos e, consequentemente, um estudo bastante significativo. A dependência da intervenção humana em partes do processo, para tentar definir os melhores atributos, cria uma subjetividade nessa escolha.

Não se encontrou, na bibliografia pesquisada, referência ao uso, em conjunto, das abordagens de sumarização, na etapa de pré-processamento, e agrupamento, na etapa de processamento, ou seja, uma proposta poliestruturada, como solução dos problemas observados e citados.

Acredita-se que existe um grande potencial de pesquisa no desenvolvimento de agrupadores de texto que não tenham essas restrições, por isso, é um desafio trazer soluções para esta lacuna, contribuindo para a área de RI, com ganhos para os agrupadores de texto.

1.2 PROBLEMA

Os agrupadores de textos não apresentam boas avaliações em *corpus* com domínios diferentes daqueles para o qual foram especificamente desenvolvidos, ou seja, agrupamentos com baixa coesão e alto acoplamento. Há, ainda, a complexidade do espaço na representação matricial, causando alta dimensionalidade e dados esparsos, a dependência do idioma e a interferência

humana, que traz subjetividade para o processo e impacto na precisão da recuperação de informação em RI.

1.3 HIPÓTESE

A melhoria do desempenho de agrupadores em bases textuais incluem, na etapa de pré-processamento, a sumarização de textos, e na etapa de processamento, usam o processo de agrupamento hierárquico, com um novo método para definição do corte de Luhn, sem restrição de um domínio, com independência do idioma e uma solução vetorial para o espaço amostral.

1.4 CONTRIBUIÇÕES

Com o uso do modelo Cassiopeia, que será apresentado neste trabalho, serão citadas algumas contribuições que têm como finalidade:

- criar um modelo que melhore o desempenho da coesão e o acoplamento dos agrupadores textuais;
- criar um modelo que possibilite independência do idioma;
- criar um modelo que melhore o desempenho da precisão na recuperação de informação;
- criar um modelo que possibilite agrupar textos com qualidade, em bases textuais, em domínios distintos e/ou antagônicos;
- criar um novo método para seleção de atributos;
- criar um modelo poliestruturado (técnica de sumarização e agrupamento);
- diminuir, com o uso da sumarização, a quantidade de atributos e
- melhorar, com o uso da sumarização, a qualidade da seleção dos atributos.

1.5 METODOLOGIA DE PESQUISA

A metodologia adotada para realização desta pesquisa compreende: leitura bibliográfica, métodos quantitativos com testes de hipótese em bases públicas em inglês e português, melhoria do desempenho das métricas externas e internas, visando a dar suporte a toda análise realizada na tese, baseada na área de MT, dentro da subárea de RI, com foco em agrupamento de texto. Serão consultados (WIVES 2004), (LOPES, 2004), (MARIA *et al.*, 2008), (RIBEIRO, 2009) e (HOURDAKIS *et al.*, 2010).

Também será apresentado o modelo Cassiopeia, que foi implementado, constituído de três etapas: pré-processamento, processamento e pós-processamento. Essas etapas serão detalhadas no Capítulo 3. Para mensuração do modelo, a implementação do teste de viabilidade, realizado em *corpora*, constituído de um *corpus* jornalístico, um jurídico e um médico, em português, e um *corpus* jornalístico e um médico em inglês. Todas as explicações e referências serão tratadas no capítulo 4. Serão aplicadas as métricas externas, com medidas como *Recall*, *Precision* e *F-Measure* e internas, como Coesão, Acoplamento e Coeficiente de Silhouette.

Os resultados parciais de publicações da pesquisa, já realizadas em datas anteriores, apresentam-se descritas, em detalhes, no Apêndice F.

1.6 ESTRUTURA DA PROPOSTA

Capítulo 2 – Fundamentação Teórica

Neste capítulo, serão apresentados conceitos sobre agrupadores, métodos de agrupamento, agrupadores de texto e seus algoritmos, métricas para análise da *performance* dos agrupadores de texto, formas de identificação de atributos, problema da alta dimensionalidade, sumarização automática de texto e formalização dos testes estatísticos usados no trabalho.

Capítulo 3- Modelo Cassiopeia

Será apresentada a arquitetura do modelo *Cassiopeia,* e suas etapas de pré-processamento, com uso de sumarização de texto, processamento com a seleção de atributos, usando o método Cassiopeia, e com uso do processo de agrupamento hierárquico de texto e pós-processamento, nos quais serão apresentados os agrupamentos por similaridade e os textos sumarizados.

Capítulo 4: Método para a Elaboração dos Testes

Neste capítulo, será apresentada a metodologia. Serão analisados os *corpora* utilizados nos experimentos, os sumarizadores da literatura e os profissionais, bem como os critérios para sua escolha e seus algoritmos, e a justificativa dos dados coletados.

Capítulo 5: Resultados

O capítulo discutirá a análise crítica dos resultados obtidos no experimento, a comprovação da hipótese, a aplicação e os resultados obtidos nos testes estatísticos e trabalhos correlatos.

Capítulo 6: Conclusões

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

CAPÍTULO 2 – FUNDAMENTAÇÃO TEÓRICA

Neste capítulo serão apontados os principais conceitos que fundamentam este trabalho. Inicialmente, um conceito mais amplo de agrupadores e os seus métodos, depois um mais contextualizado, o de agrupadores de texto, que é o foco da tese. Em agrupadores de texto serão mostradas as medidas de similaridade mais adotadas na literatura e representadas as técnicas de limiar de similaridade e truncagem, ambas criadas com a finalidade de restringir o número de atributos. As fases que compõem o processo de agrupamento de texto e suas funcionalidades, assim como os algoritmos de agrupamentos textuais mais adotados na literatura vão ser descritos e explicados. As medidas para avaliação dos agrupamentos textuais aqui empregadas, como a métrica externa, composta das medidas *Recall*, *Precision* e *F-Measure* e a métrica interna, composta das medidas Coesão, Acoplamento e Coeficiente Silhouette, serão mostradas e formalizadas. Serão apresentados, fundamentados e descritos: a seleção de atributos e os principais métodos encontrados na literatura, o problema da alta dimensionalidade, com as abordagens mais frequentes, como a Curva de Zipf e os cortes de Luhn, e os problemas gerados com esses métodos. O conceito de sumarização, sumarização automática, a ferramenta utilizada para avaliação de sumários automáticos também serão objeto de estudo, além dos testes estatísticos adotados para validar a hipótese, que serão fundamentados e discutidos neste capítulo.

2.1 AGRUPADORES

A palavra agrupamento é comumente usada na literatura como *clusterização,* tradução do termo *clustering,* e os grupos formados por esse processo são conhecidos como *clusters*. Na literatura, a denominação do processo é ampla: aglomeração, *clusterização* ou, simplesmente, agrupamento (WIVES, 2004).

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

Os algoritmos de agrupamento, segundo Wives (2004), têm as seguintes propriedades: agregação de pontos no espaço (densidade), grau de dispersão dos pontos presentes no agrupamento (variância), raio ou diâmetro — só presente em agrupamentos com forma arredondada — (dimensão), disposição dos pontos no espaço (forma), e isolamento dos agrupamentos no espaço (separação). A partir dessas propriedades surgem diferentes tipos de agrupamentos, que podem ser hiperesféricos, alongados, curvilíneos ou possuir estruturas mais

diferenciadas (ALDENDERFER e BLASHFIELD, 1984), (FASULO, 1999), (BERKHIN, 2002) e (WIVES, 2004).

2.1.1 MÉTODOS DE AGRUPAMENTO

É bastante comum, na literatura, classificar os métodos de agrupamento em duas principais classes: a dos métodos hierárquicos (ou de partição hierárquica) e a dos não-hierárquicos. Essas duas classes têm como referência o tipo de partição feita nos objetos, mas há outros fatores que fazem com que os métodos sejam classificados de maneira mais refinada e específica. Neste trabalho será adotada a taxonomia de Aldenderfer e Blashfield (1984), por ser uma classificação mais usual dentro da literatura.

Segundo estes autores, os agrupamentos têm a seguinte taxonomia, quanto à sua configuração: hierárquicos aglomerativos (*hierarchical aglomerative*), hierárquicos divisivos (*hierarchical divisive*), de particionamento iterativo (*iterative partitioning*), de busca em profundidade (*density search*), fator-analítico (*factor analytic*), de amontoamento (*clumping*) e grafoteoréticos (*graph-theoretic*). 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, e será detalhado no capítulo 3. Já no método hierárquico divisivo, segundo Wives (2004), 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 (WIVES, 2004). A complexidade do hierárquico divisivo, segundo Kaufman e Rousseeuw (1990), cresce, exponencialmente, em relação ao tamanho do conjunto de dados, sendo proibitiva sua aplicação em conjuntos de dados grandes.

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 AGRUPADORES DE TEXTOS

De acordo com Fan *et al.* (2006), o agrupamento de texto é um processo totalmente automático que reparte uma coleção em grupos de textos de conteúdos similares, cujo objetivo é ter maior conhecimento sobre esses textos e suas relações. Assim, este processo consegue reunir uma coleção de padrões desconhecidos (não classificados) em agrupamentos que possuam algum significado.

Formalmente, o problema de agrupamento de texto pode ser definido como: dada uma base de texto T, devem-se agrupar os elementos de T de maneira que os textos mais similares sejam colocados no mesmo grupo, e os menos similares, em grupos distintos. Sendo assim, dado um conjunto com *n* elementos T= $\{T_1, T_2,..., T_n\}$, obtém-se um conjunto de *k* agrupamentos G= $\{G_1, G_2,..., G_k\}$, cujos elementos de um determinado agrupamento G_i são similares entre si, mas não são similares aos elementos contidos em um conjunto G_j qualquer, onde i≠j. Dessa forma, pode-se definir:

$$\bigcup_{i=1}^{k} G_i = T \;, \quad G_i \neq \varnothing, para 1 \leq i \leq k \quad G_i \cap G_j = \varnothing, para 1 \leq i, j \leq k, i \neq j$$

2.2.1 MEDIDA DE SIMILARIDADE EM AGRUPAMENTO DE TEXTO

Para definir a similaridade entre os textos, utiliza-se uma "medida" de similaridade que, definida, vai mensurar os valores dos atributos de cada texto, ou seja, quanto menor a distância entre tais valores, mais similares são esses textos.

Os tipos de medida de similaridade utilizados na literatura são as de distância, os coeficientes de correlação, os de associação, e as medidas probabilísticas de similaridade (ALDENDERFER e BLASHFIELD, 1984).

As medidas de distância são: a Euclidiana, a mais usual (ALDENDERFER e BLASHFIELD, 1984), função cosine (SALTON e MACGILL, 1983) e a distância Manhattan, ou métrica *city block* (KAUFMAN e ROUSSEEUW, 1990).

Nos coeficientes de correlação, os mais importantes são: o de Pearson, o de Jaccard, o de associação simples (*simple matching coefficient*) e o de Gower.

Para medida probabilística, não existe um cálculo, mas um aferimento direto, no dado bruto. Já para a medida fuzzy, Wives (2004) cita o seu trabalho, no qual usa as funções, inclusão simples (*set theoretic inclusion*) de Cross (1994), e a média por operadores difusos, de Oliveira (1996).

No processo de recuperação de informação, os documentos podem ser descritos por um conjunto de termos representativos do corpus em questão (vocabulário do domínio). O grau de similaridade entre dois documentos é calculado em função da distância entre o conjunto de termos de cada um.

2.2.2 AS TÉCNICAS DE LIMIAR DE SIMILARIDADE E TRUNCAGEM NO AGRUPAMENTO DE TEXTO

Mesmo com a utilização dessas medidas de similaridade, conforme Wives (2004), ainda é necessário o uso de um limiar (*threshold*) e de uma técnica de truncamento. A técnica de limiar estabelece um critério de corte, um limiar, ou *threshold*. O corte pode ser feito a partir da importância da palavra no texto, estipulando-se um valor mínimo ou máximo. A técnica de truncamento é estabelecida pelo limite máximo de palavras representativas do texto. De acordo com Wives (1999), na maioria dos casos, 50 palavras são suficientes. Então, pelo método da truncagem, esse número pré-estabelecido é utilizado para formar o índice de cada texto, desconsiderando-se as demais palavras. Para as duas técnicas, são necessários o cálculo do peso das palavras e o seu ordenamento decrescente.

2.2.3 FASES DO AGRUPAMENTO DE TEXTO

O agrupamento de texto têm três fases: pré-processamento, processamento e pós-processamento.

Para Goldschmidt e Passos (2005), a etapa de pré-processamento consome 60% de todo o processo, e é uma etapa vital, tanto para a economia de tempo como para o bom funcionamento das seguintes. Preparar os textos para o processo computacional é uma atividade difícil e trabalhosa, que não é nova. No trabalho de Pyle (1999), é apresentada uma lista de alguns desafios.

Nogueira (2009) considera a fase de pré-processamento como a parte mais crítica, pois determina a boa qualidade dos agrupadores textuais. Existem, nessa fase, técnicas usadas para reduzir os atributos, como por exemplo, a retirada das *stopword*, a radicalização de termos com

stemming, disponível nas línguas portuguesa e inglesa; outras técnicas encontradas na literatura podem ser vistas em detalhes, em (WIVES, 2004), (GOLDSCHMIDT e PASSOS, 2005), (HOWLAND e PARK, 2007), (ARANHA, 2007) e (NOGUEIRA, 2009).

Nogueira (2009) diz que, após a etapa de pré-processamento, os agrupadores de texto precisam de uma estruturação de documentos, para torná-los processáveis pelos algoritmos de agrupamento. Segundo Feldman e Sanger (2006), o modelo mais usual para representação de dados textuais é o modelo espaço-vetorial, no qual cada documento é um vetor em um espaço multidimensional, e cada dimensão é um termo da coleção, obtendo-se, assim, uma matriz de documento-termo (Tabela 1). Nessa representação, para Forman (2009), existem muitos termos que ultrapassam o número de textos em mais de uma ordem de magnitude, gerando o problema em MT, da alta dimesionalidade e os dados esparsos, consequentemente, interferindo em todas as subáreas que usam esta representação matricial, dentre elas, a já citada RI. Assim, é imprescindível, nas etapas de pré-processamento e de processamento, que haja redução de dimensionalidade. O problema da alta dimensionalidade será tratado com mais detalhes, na seção 2.4 deste capítulo.

Na etapa de processamento, a redução de dimensionalidade será viabilizada, conforme Wives (2004), com técnicas de seleção de atributos que identificam os pesos das palavras. Uma das maneiras usuais de selecionar atributos, segundo a proposta de Wives(2004), vem a ser a medida da importância de cada palavra[2], identificando seu "peso" ou "força" de representatividade na coleção de textos. Há, na literatura, diferentes formas de calcular o peso das palavras, e as mais relevantes serão discutidas adiante, na seção 2.3. No processamento, depois da seleção dos atributos, ocorre a formação da matriz documento-termo, com os cálculos de distância citados anteriormente, quando a representação adotada for a do espaço vetorial, o que acontece na maioria dos trabalhos de agrupamento de textos (NOGUEIRA, 2009).

Na fase de pós-processamento, são usadas as medidas de validação de agrupamento, discutidas e apresentadas na subseção 2.2.5. Pode-se também utilizar um módulo para exploração visual de hierarquias de tópicos, no qual cada agrupamento textual representa um tema, uma categoria ou mesmo um tópico, que pode ser explorado através das generalizações ou especificações (LOPES, 2004).

[2] Por palavra entende-se qualquer sequência ou cadeia de caracteres (*string*) delimitada por espaços ou pontos: "...unidade constituída por grafemas, delimitada por espaços em branco e/ou sinais de pontuação" (FERREIRA, 1999).

2.2.4 ALGORITMOS DE AGRUPAMENTOS DE TEXTO

Na visão algorítmica, os textos são analisados em agrupamentos, e é preciso que grupos constituídos por eles tenham certa coesão entre si, sendo esse o grande dificultador para os algoritmos, por ser bastante complexa a análise textual, que envolve uma série de questões, de âmbito linguístico, cultural, social, situacional, político, como coerência e coesão dos textos, e/ou por estarem diretamente relacionadas com o autor e o momento em que o texto foi escrito (FÁVERO, 2000). Dessa forma, obter agrupamentos com textos muito diferentes não seria admissível, pela falta de coesão, significado ou sentido entre eles. Os algoritmos de agrupamento estão, intimamente realcionados aos métodos de organização escolhidos e descritos na seção 2.1.1.

Para o método hierárquico aglomerativo Kowalski (1997), Wives (2004) e Larose (2004) citam quatro algoritmos: ligação simples (*single linkage*), ligação completa (*complete linkage*), ligação mediana ou valor médio (*average linkage*), e *Ward*.

O algoritmo de ligação simples utiliza-se do critério de vizinho mais próximo, no qual a distância entre dois grupos é determinada pela distância do par de documentos mais próximos, cada um pertencente a um desses grupos. Esse algoritmo, cuja característica é a união de grupos, apresenta um problema conhecido como "efeito da corrente", quando ocorre a união indevida de grupos, influenciada pela presença de ruídos na base de dados (LAROSE, 2004).

No algoritmo ligação completa, ao contrário do ligação simples, o critério utilizado é o de vizinho mais distante. A distância entre dois grupos é a maior entre um par de documentos, e cada documento pertence a um grupo distinto. Esse método dificulta a formação do "efeito da corrente", como ocorre no grupo de ligação simples, e tende a formar grupos mais compactos e em formatos esféricos (LAROSE, 2004).

O algoritmo de valor médio apresenta a definição da distância entre dois grupos. Mostra como é a média das distâncias entre todos os pares de documentos em cada grupo, e cada par é composto por um documento de cada grupo. Esse método elimina muitos problemas relacionados à dependência do tamanho dos grupos, mantendo próxima a variabilidade interna entre eles (LAROSE, 2004).

No *Ward*, há uma variação dos anteriores, buscando menor variância entre os agrupamentos, juntando os elementos cuja soma dos quadrados ou o erro desta soma seja mínimo (ALDENDERFER e BLASHFIELD, 1984). Segundo Wives (2004), este algoritmo gera grupos hiperesféricos de tamanhos semelhantes.

O método hierárquico divisivo tem dois tipos de categoria de algoritmos, os monotéticos e os politéticos. Os monotéticos apresentam apenas uma variável, testada a cada divisão do

agrupamento, que ocorre na presença ou ausência de cada uma das variáveis, que são binárias, registrando, assim, a existência ou não de um objeto. O resultado apresenta um agrupamento cujos elementos têm os mesmos valores para uma variável. Essa abordagem divisiva monotética também é conhecida como análise de associações (WIVES 2004). Já os algoritmos da categoria politéticos, para cada repetição do método divisivo, todas as variáveis entram no processo de cálculo de distância (WIVES, 2004).

Para o método particionado iterativo, o algoritmo mais utilizado e conhecido é o k-médias (*k-means*), que, faz uma comparação entre o valor de cada linha, por meio da distância, para gerar os agrupamentos, utilizando, geralmente, a distância euclidiana, para calcular quão "longe" uma ocorrência está da outra. A maneira de calcular essa distância vai depender da quantidade de atributos. Após o cálculo das distâncias, o algoritmo calcula centroides, para cada um dos agrupamentos, e conforme vai iterando, o valor de cada centroide é refinado pela média dos valores de cada atributo, de cada ocorrência que pertence a esse centroide. Com isso, o algoritmo gera *k* centroides e coloca as ocorrências em uma matriz, de acordo com a distância dos centroides. Wives (2004) explica que o grande problema vem da necessidade de o usuário ter que definir o número de agrupamentos. Ressalte-se que Fan *et al.* (2006) e Alsumait e Domeniconi (2007), nos agrupamentos textuais, consideram os processos como não interativos, ou seja, mostram que não existe a possibilidade de se prever o número de agrupamentos, como o caso do algoritmo *k*-médias, que necessita definir, previamente, um número de agrupamentos.

No método de busca em profundidade, os algoritmos podem ser divididos em duas categorias, os de ligação simples, discutidos anteriormente e os de distribuições multivariadas de probabilidade. Esses últimos são baseados em um modelo estatístico que assume serem membros de grupos diferentes, portadores de distribuições de probabilidades diferentes, para cada variável.

No método de fator analítico, deve ser calculado, *a priori*, através de algum coeficiente de similaridade. Os objetos são então alocados em agrupamentos, de acordo com a sua carga em cada fator (*factor loading*[3]). Já com o método de amontoamento, os algoritmos permitem sobreposição, ou seja, podem associar os objetos a um ou mais grupos. Esse amontoamento mostra que um objeto está em um ou mais grupos ou, ainda, que pertence a todos os grupos com um grau de pertinência/probabilidade (WITTEN e FRANK, 2005).

Dois são os principais algoritmos do método grafoteorético: o *cliques* (Figura 6), descrito formalmente no algoritmo 3, e o estrela (*star*), com suas variações que, segundo Wives (2004), é o melhor estrela (*best star*), e estrelas completas (*full stars*). Neste último, os agrupamentos têm forma similar à de uma estrela, ou seja, um elemento central, que possui relação com todos os

[3] *Factor loading* são os coeficientes de correlação entre as variáveis e fatores.

elementos da estrela, e diversos outros, ligados a ele, formando as pontas. Os elementos nas extremidades não têm relação com os outros, sendo esse fato um dos problemas desse algoritmo, pois eles podem não ser similares, mas existem variações para atenuar esse problema. O algoritmo melhor estrela aloca um elemento à estrela mais similar, pois não ignora os elementos já adicionados a outras estrelas, conseguindo realocá-los. No estrelas completas, os elementos são alocados em mais de um agrupamento.

2.2.5 MÉTRICAS PARA ANÁLISE DE AGRUPAMENTO TEXTUAL

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

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, não será usada neste trabalho, e seria mais adequada depois de comprovada a eficiência do algoritmo aqui testado nas avaliações de seus próprios agrupamentos, com parâmetros diferentes. Este trabalho está focado na comparação do método aqui proposto com outros da literatura. 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 seguites medidas:

2.2.5.1 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 ao 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[4] e $n(A \cup D)$ é o número total de elementos da classe correspondente.

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.

F-Measure(*F*): **Equação 3:**

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

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).

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 cáculo da média entre cada uma das medidas de todos os grupos.

2.2.5.2 MÉTRICAS INTERNAS

Coesão(*C*): **Equação 4:**

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

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

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* (P_i, P_j) é o calculo da similaridade entre os textos *i* e *j* pertencentes ao agrupamento *P, n* é o número de textos no agrupamento *P*, e P_i e P_j são membros do agrupamento *P*.

Acoplamento (*A*): **Equação 5:**

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

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* (C_i, C_j) é o cálculo da similaridade do texto *i* pertencente ao agrupamento *P* e o texto *j* não pertence a *P*, C_i centroide do agrupamento *P* e C_j é centroide do agrupamento P_i e n_a é o número de agrupamentos presentes em *P*.

Coeficiente Silhouette(S): **Equação 6**:

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

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 dos de um 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*).

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

$$\overline{s} = \frac{1}{\mathrm{N}} \sum_{i=1}^{N} S \tag{7}$$

2.3 SELEÇÃO DE ATRIBUTOS

A seleção dos atributos está relacionada diretamente ao problema da alta dimensionalidade, discutido na seção 2.4. Na literatura, há grandes divergências na questão de nomenclatura. Neste trabalho, será denominada seleção de atributos, e foi assim descrita nos trabalhos de Loh (2001), Wives (2004), Ventura (2008) e Nogueira(2009). A seleção de atributos está em evidência em muitos trabalhos, dentre os quais, WIVES (2004), HOTHO *et al.*, (2005), ARANHA (2007), MOSTAFA *et al.,* (2007), VENTURA (2008), NOGUEIRA (2009) e NOGUEIRA (2010), e muitos que não foram mencionados neste trabalho e estão relacionados à área de seleção automática de atributos, em documentos textuais.

Conhecida também como abordagens estatísticas, na literatura, segundo Ventura (2008), a seleção de atributos utiliza, unicamente, métodos estatísticos. Essas abordagens têm como principal vantagem a rapidez, na implementação, em comparação com métodos não estatísticos, por não dependerem de informação simbólico-morfossintática específica (VENTURA, 2008). Esse fato influencia, por exemplo, os algoritmos de agrupamento que são extremamente sensíveis à alta dimensionalidade e a dados esparsos. A alta dimensionalidade exige um tempo maior de processamento, e muitas vezes, torna a solução inviável (NOGUEIRA, 2009) e (REZENDE, 2007). Esse assunto será discutido mais amplamente na seção 2.4., daí a importância da seleção de atributos para este trabalho, já que tem como principal objetivo reduzir a dimensionalidade dos atributos, mantendo-os com maior capacidade de representar a coleção de documentos.

Para a compreensão dos métodos de seleção de atributos é importante entender o significado de frequência de termo (*term frequency*) – TF e a frequência de documento (*document frequency*) - DF. O TF contabiliza a frequência absoluta de um determinado termo, ao longo da coleção de documentos, e o DF contabiliza o número de documentos em que um determinado termo aparece (NOGUEIRA, 2009). Para efeito de notação, considera-se que o índice *i* é utilizado para o i-ésimo documento e o índice *j* para o j-ésimo atributo.

Nas próximas seções serão definidos os métodos de seleção de atributos utilizados na literatura.

2.3.1 RANKING PELA FREQUÊNCIA DE TERMO (RTF)

O RTF utiliza TF como medida de melhor desempenho para um determinado atributo, dando maior valor àquele que apresenta a maior frequência ao longo da coleção. Assim, o *ranking* pela frequência de termo é definido como na **Equação 8**:

$$TF_j = \sum_{i=1}^{N} f_{ij} \quad (8)$$

Onde TF_j é a *TF* para j-ésimo atributo e o f_{ij} é a frequência do j-ésimo atributo no i-ésimo documento.

2.3.2 RANKING PELA FREQUÊNCIA DE DOCUMENTOS (RDF)

O RDF é baseado na frequência de documentos, e calcula o número de documentos em que os termos aparecem. No RDF considera-se a possibilidade de certos termos que aparecem em poucos documentos não serem relevantes para a coleção. Nesse caso, poderão ser desprezados. Formalmente, o cálculo é obtido através da **Equação 9:**

$$DF_j = \sum_{i=1}^{N} (1 | f_{ij} \neq 0) \quad (9)$$

Onde f_{ij} é a frequência do j-ésimo atributo no i-ésimo documento.

2.3.3 FREQUÊNCIA INVERSA DE DOCUMENTOS (TFIDF)

A ideia é que os termos que mais ocorrem no documento são mais relevantes que os menos frequentes. No entanto, um termo muito frequente, também pode ocorrer em quase todo o conjunto de documentos. Quando isso acontece, esses termos não são úteis para uma boa discriminação, e assim é introduzido o valor do inverso da frequência de documentos (*inverse document frequency*) - IDF para um atributo, como explicado na **Equação 10:**

$$IDF_j = \log(\frac{N}{DF_j}) \quad (10)$$

Onde DF_j é a frequência de documento do j-ésimo atributo.

O TFIDF é a multiplicação do TF pela IDF que pode ser obtido pela **Equação 11**:

$$TFIDF_j = \sum_{i=1}^{N} f_{ij} * IDF_j \quad (11)$$

Onde f_{ij} é frequência do j-ésimo atributo no i-ésimo documento e o IDF_j é o inverso da frequência de documentos do j-ésimo atributo.

2.4 PROBLEMA DA ALTA DIMENSIONALIDADE

O problema da alta dimensionalidade e dados esparsos não é novo. Conhecido na literatura como *Curse of dimensionality*, isto é, maldição da dimensionalidade, foi introduzido por Richard Bellman (BELLMAN, 1961), em seu livro *Adaptive Control Processes: A Guided Tour*. Refere-se ao problema causado pelo aumento exponencial no volume, decorrente da adição de dimensões extras a um espaço matemático, ou seja, divisões de uma região do espaço em células regulares, que crescem exponencialmente com a dimensão do espaço.

No trabalho de Beyer *et al.* (1999), constata-se que o uso de um elevado número de atributos gera a alta dimensionalidade, e os autores afirmam que, para manter a capacidade de discriminação do atributo, é necessário manter baixa a dimensionalidade dos dados. Para Aggarwal (2001), o uso de um grande número de atributos leva à maldição da alta dimensionalidade.

Na RI, o problema da maldição da alta dimensionalidade 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.

Como já foi descrito no capítulo 1, a representação do espaço de busca na RI segue uma taxonomia de modelos que são representações do espaço amostral; o espaço-vetorial (SALTON *et al.* 1997) é uma formalização comum na literatura para área de RI. Com denominação de matriz, esse espaço amostral é uma representação vetorial bidimencional que contém os graus de similaridade entre todos os elementos de um conjunto de dados, e deve ser calculada a *priori*, através de algum coeficiente de similaridade

Para Manning *et al*.(2008) a matriz é um documento-termo $C = D \times t$. Mostrada na Tabela 1, é uma representação da coleção de documentos de texto no espaço amostral. Formalmente, cada uma de suas linhas representa um documento (D), e cada uma das colunas, um termo(t) na coleção. Onde D é $\overrightarrow{D_i} = (a_{i1}, a_{i2}, \ldots, a_{iM})$, em que d_i corresponde ao i-*ésimo* documento, t_j representa o j-*ésimo* termo e a_{ij} é um valor que relaciona o i-*ésimo* documento com o j-*ésimo* termo, e pode ser calculado usando um determinado termo que está presente ou não, em um dado documento, ou mesmo em um valor que indica a importância ou distribuição do termo ao longo da coleção de documentos.

Afirmam Manning *et al*.(2008) que, mesmo para uma coleção de documentos de tamanho modesto, a matriz C tem várias dezenas de milhares de linhas e colunas, o que caracteriza a alta dimensionalidade e os dados esparsos.

	t_1	t_2	…	t_M
d_1	a_{11}	a_{12}	…	a_{1M}
d_2	a_{21}	a_{22}	…	a_{2M}
⋮	⋮	⋮	⋱	⋮
d_N	a_{N1}	a_{N2}	…	a_{NM}

Tabela 1: Matriz de documento-termo (MANNING *et al.*, 2008).

Carvalho *et al.* (2007), em seu trabalho, afirmam que o problema gerado pela altas dimensionalidades dificulta muito a RI. Como afirma Nogueira(2009), há necessidade do uso de métodos de redução de palavras, a fim de condensar a informação pertinente para a etapa de processamento da RI.

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.[5] A Curva de Zipf (Figura 1) é uma distribuição estatística específica, que se encontra em raros fenômenos estocásticos. Acontece que um deles é a distribuição da frequência da ocorrência de palavras num texto, em que nas ordenadas *f,* se tem o valor dessa frequência, e nas abscissas *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 Zipf de uma dada amostra específica, tem-se *f. r* = *k*, em que *k* será uma constante específica para essa amostra.

Luhn propôs que na Figura 1, se pode definir um limite superior e um inferior, de corte, denominado limiares de corte de Luhn, (Figura 2). Com isso, Luhn propôs uma técnica para encontrar termos relevantes, assumindo que os mais significativos para discriminar o conteúdo do documento estão em um pico imaginário, posicionado no meio dos dois pontos de corte, de acordo com a Figura 2. Porém, segundo Sayão (2007), certa arbitrariedade está envolvida na determinação dos pontos de corte, bem como na Curva imaginária, os quais são estabelecidos por tentativa e erro (VAN RIJSBERGEN, 1979).

[5] A frequência de ocorrência de alguns eventos está relacionada à função de ordenação. Para o uso textual, ao somar a frequência de palavras(f) e ordenar (r) de forma decrescente, surge a Curva de Zipf (vide figura 1).

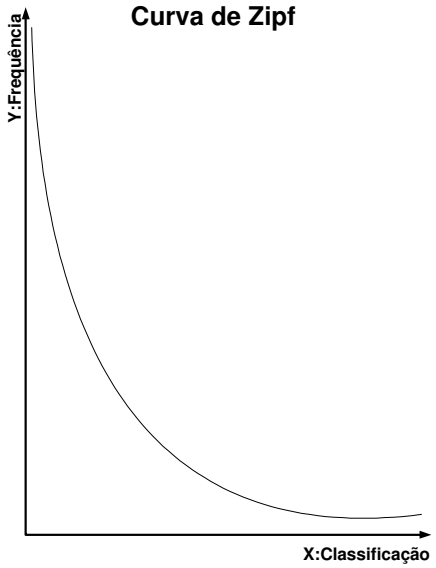


Figura 1: Curva de Zipf.

O primeiro corte de Luhn, conhecido como corte superior, tem por finalidade retirar as *stopwords*, ou seja, as palavras mais frequentes de um texto. Segundo Vianna (2004), a não utilização desse corte causaria um problema de alta dimensionalidade e dos dados esparsos que crescem, exponencialmente, de acordo com sua base de texto, gerando um problema crucial na área de MT e, consequentemente, na de RI. Aranha (2007) considera que as *stopwords* aparecem em todos os documentos, ou na maioria deles, por isso não são capazes de colaborar na seleção de documentos relativos a um assunto específico. Para Pardo (2002), as *stopwords* são classes fechadas de palavras que não carregam significados, tais como artigos, pronomes, interjeições e preposições. A retirada das *stopwords*, segundo Aranha (2007), é um processo manual. O projetista do sistema avalia quais palavras devem ou não estar contidas na lista de *stopwords* (o que varia de idioma para idioma, ou até mesmo entre sistemas). O grande problema do fornecimento desta lista de *stopword*, é que o sistema fica dependente do idioma, ou seja, de uma intervenção humana.

O segundo corte serve para diminuir o número de palavras muito específicas, encontradas apenas uma única vez nos documentos, e fazem com que na representação matricial, contribuam para um número grande de dados esparsos.

Com o primeiro e o segundo corte, surge o pico imaginário, um processo heurístico e fonte de estudos e pesquisas atuais. Para Quoniam (2001), a Curva de *Zipf* , com o corte *Luhn* (Figura 2), tem três áreas distintas. Na área I, encontram-se as informações triviais ou básicas, com maior frequência; na área II, as informações interessantes e, na área III, os ruídos.

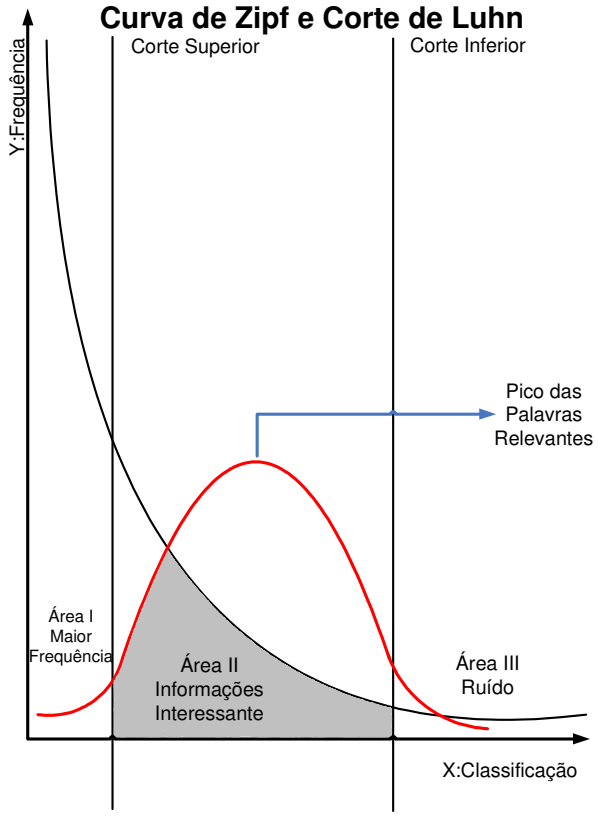


Figura 2: Curva de Zipf com os cortes de Lunh.

2.5 SUMARIZAÇÃO AUTOMÁTICA DE TEXTO

A sumarização, cujo uso no modelo Cassiopeia será explicado na subseção 3.1.1, é a técnica usada neste trabalho para redução da dimensionalidade.

Por definição, sumários são textos reduzidos, que transmitem as ideias principais e mais relevantes de um texto original, de forma clara e objetiva, sem perda da informatividade[6] (PARDO, 2007). Essa necessidade de simplificar e resumir acontece, devido ao aumento do volume de informações disponíveis nos meios de comunicação (principalmente a Internet) e ao pouco tempo para leitura de textos de diversas naturezas. Em consequência desse processo, ocorre a incapacidade dos leitores de absorver a totalidade de conteúdo dos textos originais. Dessa forma, o sumário é um resumo, cujo objetivo é captar a ideia principal do autor e passá-la em poucas linhas para o leitor.

A sumarização vem sendo formalizada desde 1950, e o marco inicial dessas pesquisas foi o método de Luhn, o da palavra-chave (LUHN, 1958). A partir daí, surgiram outros trabalhos na

[6] Informatividade de um texto é medida de acordo com o conhecimento de mundo das pessoas a quem ele se destina. Ou seja, um texto tem um alto grau de informatividade, quando a sua compreensão mais ampla depender do repertório cultural do leitor. Um texto é mais informativo, quanto menor for sua previsibilidade, e vice-versa. Para que haja sucesso na interação verbal, é preciso que a informatividade seja adequada ao interlocutor.

área, como: a importância da primeira e da última frase do texto original (BAXENDALE, 1958) e (ARMAN e AKBARZADEH, 2006), escolha computacional das frases com maior potencial de transmitir significância do texto original (EDMUNDSON, 1969), relevância da restrição do domínio (POLLOCK e ZAMORA, 1975), classificação dos sumários em indicativos, informativos e de críticas (HUTCHINS, 1987), hierarquização dos papéis semânticos em cada frase (PAICE e JONES, 1993), uso da abordagem híbrida (MAYBURY, 1993), relações retóricas de cada frase do texto (MARCU, 1997), uso do conhecimento simbólico e de técnicas estatísticas para sumarização (HOVY e LIN, 1997), e por fim, a taxonomia criada por Hutchins (1987), fator determinante para estabelecer sua aplicabilidade, e uma avaliação consistente do processo (SPARCK, 1999).

Como constatado nos trabalhos que delimitam a sumarização, verifica-se uma taxonomia metodológica, cujo processo é denominado superficial; também é descrita como empírica ou estatística, e uma outra abordagem, conhecida como profunda ou fundamental (HEARST, 1993), (HEARST, 1997), (MITAL *et al.*, 1999), (LARROCA *et al.*, 2000), (PARDO, 2001) e (FANEGO, 2008).

Segundo Pardo *et al.* (2001), os trabalhos têm adotado metodologias híbridas, ou seja, o uso da abordagem superficial e profunda, variando os métodos de cada uma. Em relatórios mais atuais, liberados pelo *Text Analysis Conference* – TAC nos anos de 2009 e 2010, há uma outra tendência na evolução dos sumarizadores para utilização de aprendizado, tanto na abordagem profunda, como na superficial.

A sumarização automática – SA, segundo Fanego (2008) e Leite (2010), é extrativa, seguindo a abordagem empírica, conhecida também como abordagem superficial. Essa técnica usa os métodos estatísticos ou superficiais que identificam os segmentos mais relevantes do texto fonte, produzindo os sumários através da justaposição das sentenças extraídas, sem qualquer modificação em relação à ordem do texto original.

Outro fator importante a ser salientado, nessa área, é a tendência dos sumarizadores mais atuais de não dependerem do fornecimento do percentual de compressão. Os melhores, apresentados no maior evento da área, o TAC, foram os que utilizaram algum processo de aprendizado e não necessitaram do fornecimento do grau de compressão pelo usuário, adaptando-se às necessidades de informação dos usuários. Em relatórios de 2009 de 2010, divulgados pelo TAC, constatou-se que os vencedores da conferência apresentam essa abordagem.

Como o processo de sumarização faz parte integrante do modelo Cassiopeia, o entendimento dessa área foi crucial para o desenvolvimento deste trabalho, o que justifica a seção 2.5, incluindo também a explicação do processo de avaliação dos sumários. Para entender melhor

esse processo, foram de muita valia o estudo e a compreensão do pacote de avaliação de sumários, *Recall-Oriented Understudy for Gisting Evaluation* – ROUGE,[7] de Lin e Hovy (2003), adotado em conferências internacionais dedicadas ao tema, como a TAC, realizadas anualmente nos Estados Unidos da América, patrocinadas pelo sistema de defesa americano. Para usar o ROUGE, na criação dos *corpora,* foi necessário obter sumários manuais (resumos fornecidos por autores e sumários automáticos). O ROUGE foi utilizado porque faz a comparação entre os sumários manuais e os automáticos, e segundo Pardo e Rino (2006), é uma medida que verifica a cobertura de sumários automáticos em relação aos manuais, comparando a coocorrência de n-gramas. Para os *corpora* de textos jornalísticos, foram escolhidos os *leads*[8] e os textos científicos, para o *corpus* dos domínios jurídico e médico, os resumos de trabalhos científicos, para os textos em português, e os *abstracts* para os textos em inglês. Para os sumários automáticos foram usados os sumarizadores, apresentados na seção 4.2. A formação dos *corpora* desta tese será discutida em detalhes no capítulo 4, seção 4.1.

2.6 TESTES ESTATÍSTICOS

Os testes estatísticos, fundamentalmente utilizados em pesquisas, que têm como objetivo comparar condições experimentais, podem auxiliar e fornecer respaldo científico àquelas que tenham validade e aceitabilidade no meio científico. Podem ser divididos em paramétricos e não paramétricos.

Conforme Callegari e Jacques (2007), 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. A estatística não paramétrica representa um conjunto de ferramentas de uso mais apropriado em pesquisas. Neste tipo, não se conhece bem a distribuição da população e seus parâmetros, o que reforça o estudo e a importância da análise de pesquisas através desses testes.

A escolha dos testes estatísticos usados nos experimentos deste trabalho basearam-se nas amostras obtidas nas simulações feitas com o modelo Cassiopeia, (Cap. 3), usando os *corpora* e sumarizadores automáticos (Cap. 4, seções 4.1 e 4.2). Tal decisão seguiu o diagrama do anexo A,

[7] ROUGE utiliza sumários de referência humanos para comparação com os automáticos, o que permite também a fácil reprodução da avaliação, o baixo custo de executá-la, comparado com uma avaliação manual, além de evitar os erros humanos, geralmente cometidos.

[8] *Lead* vem à frente de uma notícia e dá sua direção primeira. Ele deve fornecer ao leitor as informações básicas sobre o assunto que será tratado na matéria. Seguindo a lógica da "pirâmide invertida" (segundo a qual, as informações nas notícias devem aparecer por ordem de importância), o *lead* informa, já no primeiro parágrafo, quais são os fatos pontuais relatados por aquela notícia. Para isso, o *lead* deve responder a seis perguntas: "O quê?" "Quem?" "Quando?" "Onde?" "Como?" e "Por quê?".
.

proposto por Callegari e Jacques (2007), e os relatórios da conferência TAC, uma conferência internacional de avaliação de SA, a mais relevante dentro da área. O diagrama do anexo A e os relatórios (TAC, 2005 e 2006) indicam que os testes estatísticos de ANOVA, de Friedman, e o coeficiente de concordância de Kendall são os mais adequados para verificar se existe diferença significativa na distribuição em todas as amostras analisadas nos experimentos.

Analisando o diagrama do anexo A e as amostras obtidas na simulação do modelo Cassiopeia, verificou-se que as amostras não seguiram uma distribuição normal. Assim, os testes usados foram os não paramétricos, com *K* amostras independentes, com *k* variáveis e com dados ordinais. Dessa forma, o diagrama do anexo A indica o teste ANOVA de Friedman e o coeficiente de concordância de Kendall, cujas populações foram assumidas como as mesmas, e para isso verificou-se se houve diferença significativa entre as amostras.

Esses dois testes possuem pequenas diferenças na natureza, no entanto, exigem entradas similares. ANOVA de Friedman é uma alternativa não paramétrica para *one-way*, análise de variância de medidas repetidas. A estatística de concordância Kendall é semelhante à de Spearman R (não paramétrico de correlação entre duas variáveis), exceto que ANOVA, de Friedman, expressa a relação entre múltiplos casos.

O ANOVA de Friedman busca comparar os resultados de três ou mais amostras relacionadas, numa distribuição bivariada. Esse teste ordena os resultados para cada um dos casos e depois calcula a média das ordens, para cada amostra (CALLEGARI e JACQUES, 2007).

O teste de coeficiente de concordância de Kendall tem a finalidade de normalizar o teste estatístico ANOVA de Friedman. É um teste não paramétrico, que gera uma avaliação de concordância ou não, com *ranks* estabelecidos nos experimentos, e assim, mede a diferença entre a probabilidade de as classificações estarem na mesma ordem e a de estarem em ordens diferentes. Quanto mais próximo de zero, menor é a concordância, e quanto mais próximo de um, maior é a concordância (CALLEGARI e JACQUES, 2007).

2.6.1 TESTE ESTATÍSTICO ANOVA DE FRIEDMAN

Para calcular ordens para cada amostra, a estatística de teste ANOVA de Friedman ordena as *k* observações da menor para a maior, de forma separada, em cada um dos *b* blocos, e atribui os *ranks* {*1, 2,..., k*} para cada bloco da tabela de observações (CALLEGARI e JACQUES, 2007). Assim, a posição esperada de qualquer observação sob H_0 é $(k + 1) /2$. Sendo $r(Xij)$ o *rank* da observação, *Xij* define a soma de todos os *ranks* da coluna j (ou seja, de cada tratamento) mostrado na **Equação 12:**

$$R_j = \sum_{i=1}^{b} r(X_{ij}),1 \le j \le k \tag{12}$$

Se H_0 é verdadeira, o valor esperado de *Rj* é $E(Rj) = b(k+1)/2$. Dessa forma, a estatística é mostrada na **Equação 13:**

$$\sum_{j=1}^{b} (R_j - \frac{b(k+1)}{2})^2 \tag{13}$$

A estatística do teste de Friedman será dada pela **Equação 14:**

$$SFr = \frac{12b}{k(k+1)} \sum_{i=1}^{k} (\frac{R_j}{b} - \frac{k+1}{2})^2 = \left[\frac{12}{bk(k+1)} \sum_{j=1}^{k} R_j^2 \right] - 3b(k+1) \tag{14}$$

Se $Fj(t) = F(t+\tau j)$ é a função de distribuição do tratamento *j*, com *j* = 1, 2,..., *k*, no teste de Friedman estar-se-á interessado em testar a hipótese H_0: $\tau_1 = \tau_2 = ... = \tau_k$ contra a hipótese alternativa de que $\tau_1, \tau_2, ..., \tau_k$ não são todas iguais. Neste caso, em nível de significância α, rejeita-se a hipótese H_0 se $SFr \ge SFr(\alpha)$, caso contrário, não se rejeita a hipótese nula, em que a constante $SFr(\alpha)$ é escolhida de modo que a probabilidade de erro do tipo I [9] seja igual a α.

A aproximação para amostras grandes, em que H_0, a estatística *SFr* tem, quando *n* tende ao infinito, uma distribuição qui-quadrado χ^2 com *k-1* graus de liberdade. Nesse caso, utilizando a aproximação qui-quadrado, rejeita-se H_0 se $SFr \ge \chi^2_{k-1,\alpha}$, caso contrário não se rejeita H_0, onde $\chi^2_{k-1,\alpha}$ é tal que $P - valor = P[\chi^2_{k-1} \ge \chi^2_{k-1,\alpha}] = \alpha$.

No caso de observações repetidas entre as *k* observações de um mesmo bloco, uma modificação para a estatística *SFr* é necessária. Nesse caso, substitui-se *SFr* pela **Equação 15:**

$$SFr' = \frac{12\sum_{j=1}^{k} R_j^2 - 3b^2k(k+1)^2}{bk(k+1) - \left[\frac{1}{k-1}\right] \sum_{i=1}^{n} \left\{ (\sum_{j=1}^{g_1} t_{i,j}^3) - k \right\}} \tag{15}$$

Onde g_i denota o número de grupos de observações repetidas no i-*ésimo* bloco e $t_{i,j}$ é o tamanho do j-ésimo grupo de observações repetidas no *i-ésimo* bloco. Em particular, se não há observações repetidas entre as observações no *i-ésimo* bloco, então $g_i = k$ e $t_{i,j} = 1$ para cada $j = 1,..., k$. Se em todos os blocos não existem observações repetidas, então *SFr´* se reduz a SFr. Sendo assim, o p-valor é calculado desta forma $P[\chi^2_{k-1} \ge EFr | H_0]$.

[9] Erro tipo I consiste em rejeitar H_0 quando a mesma é verdadeira.

2.6.2 TESTE ESTATÍSTICO COEFICIENTE DE CONCORDÂNCIA KENDALL

O coeficiente de concordância de Kendall, segundo Callegari e Jacques (2007), é uma medida da relação entre vários conjuntos de postos de *N* objetos ou indivíduos.

Quando se tem *k* conjuntos de postos, pode-se determinar a associação entre eles utilizando-se o coeficiente de concordância de Kendall.

O coeficiente de concordância de Kendall expressa a associação simultânea (relacionamento) entre as séries *r* de *rankings* (por exemplo, os casos de amostras correlacionadas). Basicamente, esse coeficiente mede a diferença entre a probabilidade de as classificações estarem na mesma ordem e a de estarem em ordens diferentes (CALLEGARI E JACQUES, 2007).

Assim, o cálculo de $M_{Kendall}$ é mais simples e $M_{Kendall}$ tem uma relação linear com o valor médio de *r,* relativo a todos os grupos, denotando $M_{rKendall,}$ o valor médio dos coeficientes de correlação por postos de Kendall, mostrado na **Equação 16:**

$$M_{rKendall} = \left(\frac{k * M_{Kendall} - 1}{k - 1} \right) \qquad (16)$$

Outro processo consiste em imaginar como se apresentariam dados se não houvesse concordância alguma entre os conjuntos de postos e, em seguida, como se apresentariam se houvesse concordância perfeita.

O coeficiente de concordância seria, então, um índice de divergência entre a concordância efetiva, acusada pelos dados, e a concordância máxima possível (perfeita). De modo $M_{Kendall}$ é um coeficiente desta natureza.

Verificou-se que o grau de concordância entre os *K* julgamentos é refletido pelo grau de variância entre as *N* somas de postos $M_{Kendall,}$ coeficiente de concordância, que é uma função desse grau de variância. Mostrada na **Equação 17:**

$$M_{Kendall} = \left(\frac{S}{\frac{1}{12} k^2 (N^3 - N)} \right) \qquad (17)$$

Onde

S= soma dos quadrados dos desvios observados a contar da média dos $R_{j,}$ isto é, $S = \sum (R_j - \sum \frac{R_j}{N})^2$;

k=número de conjuntos de postos;

N=número de entidades (objetos ou indivíduos a que se atribuem postos);

$\frac{1}{12}k^2(N^3 - N)$=valor máximo possível da soma dos quadrados dos desvios, isto é, *S* que ocorreria o caso de concordância perfeita entre os *k* conjuntos de postos.

O efeito dos empates foi reduzir o valor de $M_{Kendall}$. Se a proporção de empates é pequena, então o efeito é desprezível.

Se a proporção de empates é grande, pode introduzir-se uma correção, que aumentará ligeiramente o valor de $M_{Kendall}$ em relação ao que se apresenta sem correção. O elemento corretivo é mostrado na **Equação 18:**

$$E_c = \frac{\sum(c^3 - c)}{12} \tag{18}$$

Onde

c= número de observações num grupo empatadas em relação a um dado posto.

A **Equação 19** ficaria introduzindo o elemento E_c:

$$M_{Kendall} = \left(\frac{S}{\frac{1}{12}k^2(N^3 - N) - \sum E_c}\right) \tag{19}$$

CAPÍTULO 3 – MODELO CASSIOPEIA

O modelo Cassiopeia, ilustrado na Figura 3, foi proposto para ser um agrupador de texto hierárquico, com novo método para definição do corte de Luhn.

Inicialmente, para melhor entendimento, uma visão geral de seu funcionamento, e a seguir, um detalhamento de cada uma das três macroetapas (pré-processamento, processamento e pós-processamento) que compõem este modelo. Para cada uma dessas etapas serão descritas as funcionalidades do modelo.

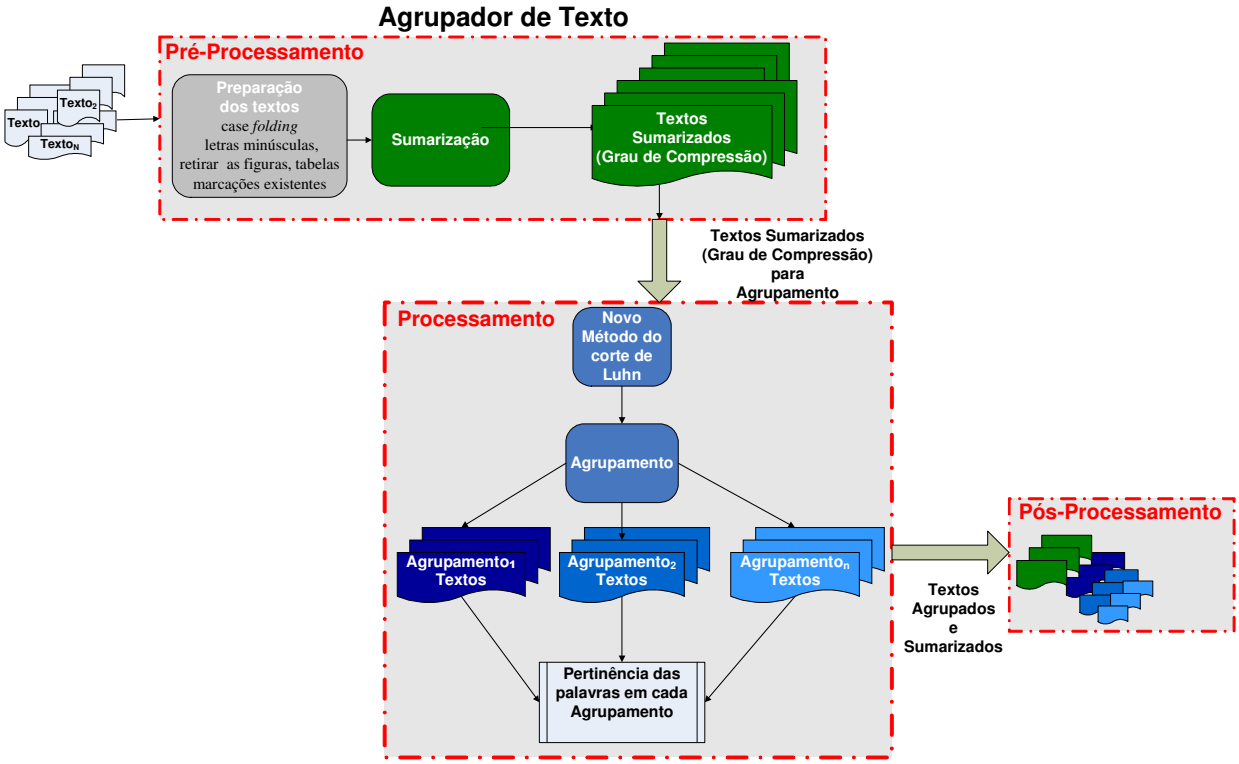


Figura 3: Modelo Cassiopeia.

O processo começa com a entrada de textos, que passam pela etapa de pré-processamento, na qual são preparados para o processo computacional, utilizando-se a técnica *case folding* (WITTEN *et al.*, 1994), que coloca todas as letras em minúsculas, além de outros cuidados, como descarte de todas as figuras, tabelas e marcações existentes. Os textos ficam em formato compatível para serem processados. Nesta etapa, é usado o processo de sumarização, cuja finalidade é diminuir o número de palavras, viabilizando o processamento. Com o processo de sumarização, obtém-se a parte mais importante, ou seja, a ideia principal do texto-fonte, através da criação de um resumo com as palavras mais significativas. O sumário, além de ser mais

conciso do que o texto-fonte, tem um número muito menor de atributos. Essa redução possibilita o uso de um espaço amostral que consegue atenuar a questão da alta dimensionalidade e dos dados esparsos. A sumarização também consegue viabilizar a permanência das *stopwords*, o que possibilita que o modelo Cassiopeia seja independente do idioma.

Terminada a etapa de pré-processamento, começa a de processamento, que usa o processo de agrupamento de textos hierárquicos e um algoritmo para juntar os textos com similaridade. Os agrupamentos criados nesta etapa têm um vetor de palavras que representam os centroides destes, cujas palavras são de alta relevância para cada agrupamento, e pertinentes em relação à frequência média das palavras nos textos agrupados. À medida que novos textos são agrupados ocorre o reagrupamento, podendo surgir agrupamentos, subagrupamentos ou até mesmo a fusão destes (LOH, 2001). As palavras colocadas nos centroides são calculadas pela média da sua frequência no texto e selecionadas, conforme a Figura 4.

Os vetores de palavras ou centroides dos agrupamentos, por questão de dimensionalidade, adotam uma truncagem que, segundo Wives (2004), é de 50 posições, não sendo necessário um valor maior. Essas palavras devem estar ordenadas da maior para a menor, com base na frequência média de cada uma. Esses agrupamentos estão organizados de uma forma *top-down*, ou seja, hierárquica, o que será explicado mais adiante. O seu reagrupamento ocorre até o momento em que os centroides de cada agrupamento estejam estáveis, ou seja, não sofram mais alterações, com a inclusão de novos textos. Para determinar a similaridade desses textos nos agrupamentos é utilizado o algoritmo *Cliques* (Figura 6) descrito pelo algoritmo 3.

Terminada a etapa de processamento, começa a de pós-processamento, na qual cada um dos agrupamentos ou subagrupamentos terá, por similaridade, um conjunto de textos-fonte com os sumários correspondentes, que têm alto grau de informatividade e contêm as ideias principais dos textos-fonte, característica da sumarização. A organização dos textos na estrutura hierárquica, obtida no pós-processamento, será importante para a área de RI, pois com os resultados alcançados, apresentados no Capítulo 5, mostra que seria um ganho usar os agrupamentos de texto do modelo Cassiopeia.

3.1 MODELO CASSIOPEIA - PRÉ-PROCESSAMENTO

A etapa de pré-processamento é a que consome mais tempo de toda a mineração de texto, segundo Goldschmidt e Passos (2005). É essencial, tanto para a economia de tempo como para o bom funcionamento das etapas seguintes da RI, principalmente a de processamento, que depende, fundamentalmente, da quantidade e da qualidade das palavras mantidas depois desta etapa

(ARANHA, 2007). Preparar os textos para o processo computacional é uma atividade difícil e trabalhosa.

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

3.1.1 TRATAMENTO DA ALTA DIMENSIONALIDADE NO MODELO CASSIOPEIA

A sumarização é a proposta deste modelo para diminuição do volume de palavras no pré-processamento. A redução de palavras com a sumarização, gera uma ganho qualitativo, quantitativo, e viabiliza o uso das *stopwords*, tornando o modelo independente do idioma.

Nessa técnica, surge uma contribuição do modelo para reduzir a dimensionalidade e os dados esparsos. A redução de palavras não relevantes é considerável, justamente pela definição do grau de compressão do sumário (percentuais de sentenças a serem extraídas do texto original). Esse grau de compressão nos sumarizadores é, normalmente, definido pelo usuário, dependendo do nível de conhecimento que tenha sobre o assunto abordado e/ou seu grau de interesse no texto.

A necessidade de fornecer o grau de compressão é um desabonador, no uso da sumarização. Na etapa de pré-processamento, é necessária a intervenção humana, para dar o parâmetro a ser utilizado por cada um dos usuários, na redução do texto. Mas como foi relatado, na seção 2.5, a tendência dos sumarizadores mais atuais é usar o processo de aprendizado, que no modelo Cassiopeia, será discutido na seção 6.3.

A taxa de compressão causa um impacto nos resultados do agrupamento do modelo referido. Por isso, a consequência do aumento do grau de compressão, dentro do agrupamento de textos é fundamental para análise da qualidade dos agrupamentos gerados. Dessa forma, os experimentos servirão para mostrar qual é a taxa de compressão que melhora os agrupamentos de texto.

Outra vantagem de se utilizar a sumarização é o nível de informatividade no modelo Cassiopeia. Enquanto a lista de *stopwords* apenas visa à retirada dessa classe de palavras, com intuito de diminuir o número de palavras, unicamente para viabilizar a etapa de processamento, a sumarização, além de realizar essa diminuição, proporciona um ganho, mantendo as sentenças com maior grau de informatividade, ou seja, as palavras mais importantes para representar a ideia do texto, uma característica inerente da sumarização, essencial para o processo de agrupamento. Essa proposta é importante para atenuar a sobregarga de informação. Os sumarizadores usados e suas compressões serão discutidos em detalhes no capítulo 4, seção 4.2.

3.2 MODELO CASSIOPEIA - PROCESSAMENTO

O agrupamento de textos, por similaridade, é usado na etapa de processamento, e acontece quando não se conhecem os elementos do domínio disponível, procurando-se, assim, separar, automaticamente, os elementos em agrupamentos por algum critério de afinidade ou similaridade (RIZZI *et al.*, 2000) e (LOH, 2001). Como os agrupamentos, segundo Alsumait e Domeniconi (2007), não são previamente definidos, o processo não é interativo.

Devido à sumarização de textos proposta neste trabalho, foi possível criar uma solução diferente para o modelo Cassiopeia realizar a etapa de processamento. Na literatura, utiliza-se, conforme mostra a Figura 2, o corte Luhn (superior e inferior). O modelo Cassiopeia define um novo método baseado no corte de Luhn, em que propõe um corte médio na distribuição da frequência da palavras (Figura 4). Para viabilizar essa variação no corte de Luhn, foram utilizados centroides, como forma de representação do esparço amostral, e para organização dos textos nos agrupamentos, o método hierárquico aglomerativo e o algoritmo *Cliques,* para garantir a similaridade entre os textos agrupados.

3.2.1 IDENTIFICAÇÃO DOS ATRIBUTOS

O modelo Cassiopeia identifica as características das palavras no documento, utilizando a frequência relativa, que define a importância de um termo, de acordo com a frequência com que é encontrado no documento. Quanto mais um termo aparecer em um documento, mais importante é, para aquele documento. A frequência relativa é calculada por meio da equação 20, fórmula que normaliza o resultado da frequência absoluta das palavras, evitando que documentos pequenos sejam representados por vetores pequenos e documentos grandes, por vetores grandes.

Com a normalização, todos os documentos serão representados por vetores de mesmo tamanho, como mostra a **Equação 20:**

$$F_{\mathrm{r}} X = \frac{F_{\mathrm{abs}} X}{\mathrm{N}} \qquad (20)$$

Onde $F_{\mathbf{r}} X$ é igual à frequência relativa de *X*, $F_{\mathbf{abs}} X$ 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. Considerado um espaço-vetorial, cada palavra representa uma dimensão (existem tantas dimensões quantas palavras diferentes no documento).

3.2.2 SELEÇÃO DOS ATRIBUTOS

Tendo como base os pesos das palavras, obtidos na frequência relativa, é calculada a média sobre o total de palavras no documento. Nessa etapa, o modelo usa a truncagem, ou seja, um tamanho máximo de 50 posições (WIVES, 2004) para os vetores de palavras, realizando um corte que representa a frequência média das palavras obtidas com os cálculos e, em seguida, realiza a organização dos vetores de palavras (Figura 4). Essa truncagem do vetor de palavras com 50 posições, segundo Wives (2004), é suficiente para estabelecer um vetor com "boas características". De acordo com Wives (2004), o uso de um vetor com mais posições não garante palavras com boas características, mas causa aumento do processamento computacional. O modelo Cassiopeia divide esse vetor de 50 palavras, ordenadas de forma decrescente, com 25 posições à direita e 25 à esquerda da frequência média, calculada para fazer a ordenação do vetor.

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

1. calcular a frequência relativa: quantas vezes cada palavra aparece no documento, dividido pelo número total de palavras do documento;
2. ordenar as palavras em ordem descrescente de frequência (da maior para a menor);
3. achar a frequência média das palavras, somando as frequências relativas e dividindo pelo número total de palavras do documento;
4. encontrar a primeira palavra cuja frequencia mais próxima à média;
5. marcar esta palavra e escolher, incluindo-a, mais as 24 anteriores (esquerda);
6. marcar esta palavra e escolher as 25 posteriores (direita);
7. montar o vetor em ordem decrescente com as 50 palavras escolhidas.

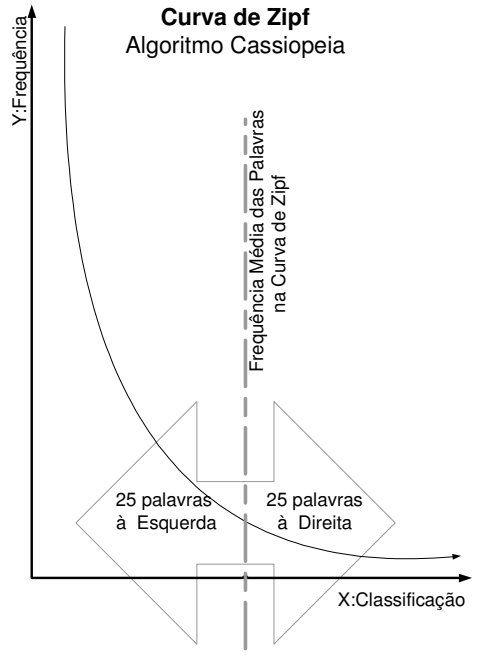


Figura 4: Seleção dos atributos no modelo Cassiopeia.

1 - Algoritmo do método Cassiopeia:

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

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

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

Equação 21:

$$f(P,N) = \frac{\sum_{n=1}^{N} F_r P_n}{N} \qquad (21)$$

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

Em relação à complexidade, sabe-se que m representa a quantidade de palavras e n o número de textos. O espaço necessário para a execução do modelo é dado pela representação de cada texto (ou *cluster*) pelo seu centroide. Como o pior caso é determinado pela geração de um

agrupamento para cada texto, tem-se a complexidade de O(50*n*) = O(*n*), em que 50 é o tamanho dos centroides.

Quanto à complexidade de tempo pode-se dividir o algoritmo em três etapas. (1) gerar os centroides de cada texto, nesse caso as palavras são ordenadas e o método Cassiopeia, apresentado na Figura 4, é aplicado. Considerando o pior caso no qual cada texto contém as *m* palavras, a complexidade de tempo é O($n.m \log_2 m$). (2) o método aglomerativo tem por objetivo gerar os agrupamentos, conforme apresentado na Figura 6, na pior das hipóteses, a cada iteração somente um novo agrupamento será gerado. Tomando como base a existência de *n* agrupamentos (um para cada texto), o pior caso seria obter apenas um novo agrupamento a cada iteração do método, ou seja, após a avaliação de todos os pares de agrupamentos disponíveis. Dessa maneira a complexidade de tempo pode ser definida de acordo Gersting (1995) pela **Equação 22**:

$$f = \sum_{i=1}^{n} i^2 \Rightarrow \frac{n(n+1)(2n+1)}{6} \Rightarrow O(n^3) \qquad (22)$$

(3) o algoritmo *Cliques* determina a similaridade entre os centroides e seu custo é O (1). Sendo assim, a complexidade de tempo do modelo Cassiopeia é O (max ($n.m \log_2 m$, n^3)).

Para efeitos de contraste com a solução matricial de Wives(1999), cabe assinalar que neste caso tem-se a relação de complexidade de espaço O(*mn*) enquanto a complexidade de tempo é de O(*n3*).

3.2.3 USO DO MÉTODO HIERÁRQUICO AGLOMERATIVO E DO ALGORITMO *CLIQUES*

O modelo Cassiopeia, mostrado na Figura 3, utiliza o método hierárquico, para organizar seus textos em agrupamentos, que são particionados, sucessivamente, produzindo uma representação hierárquica, tipo que facilita a visualização dos agrupamentos a cada ciclo de processamento, bem como o grau de similaridade obtidos entre eles com uso do algoritmo *Cliques*. É um método que, de início, não requer definições de número de agrupamentos. A principal vantagem e a característica determinante para escolha do método a ser usado no modelo Cassiopeia é a facilidade de lidar com qualquer medida de similaridade utilizada, ou seja, o algoritmo *Cliques* e a sua consequente aplicabilidade a qualquer tipo de atributo (BERKHIN, 2002).

No método hierárquico aglomerativo (Figura 5), descrito no algoritmo 2, os agrupamentos são recursivamente criados, considerando alguma medida de similaridade. Sendo assim, no início,

os agrupamentos são em número reduzido, com baixo grau de similaridade, mas com o decorrer do processo, eles vão aumentando e tornando-se dissimilares, com alto grau de similaridade entre os documentos de cada agrupamento (SILVA *et al.*, 2005).

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.*

Um exemplo do funcionamento do método hierárquico pode ser visto na Figura 5, com o uso do algoritmo 2, cujos passos são descritos e realizados nessa abordagem. Dessa forma, pode-se interpretar a Figura 6, inicialmente, com cinco agrupamentos, por exemplo [A, B, C, D, E]. Decorridos os passos, forma-se um agrupamento denominado G1, no qual se encontram [A, B]. A similaridade do agrupamento G1, no caso, é medida pela distância D1. O G2 é formado por [D, E], sendo que a medida de similaridade de G2 é igual a D2. No passo seguinte, é formado o agrupamento G3, constituído por [C] e G2. A distância de similaridade de G2 para G3 é a D3. O próximo passo é a formação do agrupamento G4 que, formado por G1 e G3, tem distância de similaridade D4.

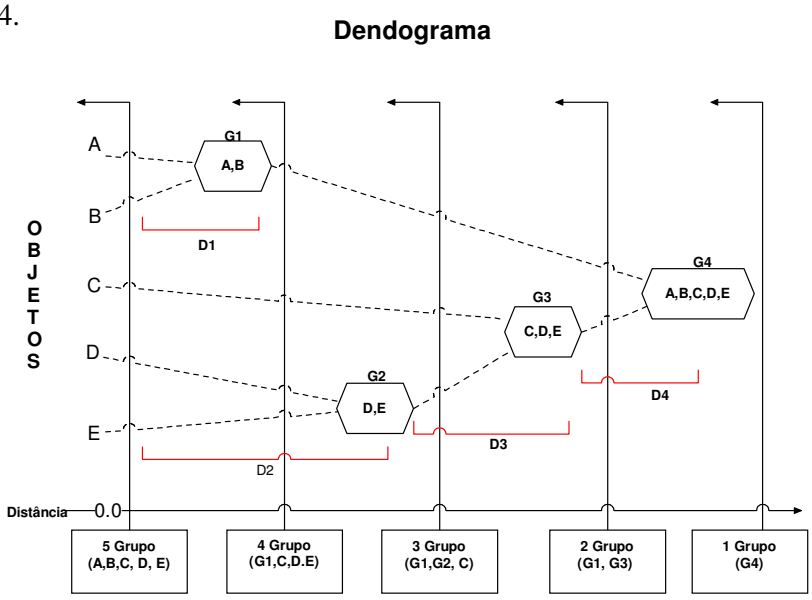


Figura 5: Dendograma do método hierárquico aglomerativo.

No modelo Cassiopeia, os textos são agrupados e o modelo adota o algoritmo *Cliques*, para garantir a similaridade entre documentos e agrupamentos. A justificativa da escolha do algoritmo *Cliques* advém de os agrupamentos tenderem, segundo Wives (2004) e Guelpeli *et al.* (2010), a ser mais coesos e de melhor qualidade, uma vez que os elementos são mais semelhantes

ou próximos. O algoritmo *Cliques* baseia-se em teoremas e axiomas conhecidos da teoria dos grafos, e possuem alta capacidade dedutiva, tendo, portanto, maior fundamentação teórica (ALDENDERFER E BLASHFIELD, 1984). O *Cliques* é um algoritmo da classe *graph-theoretic* (Figura 6).

Devido à sua capacidade de construir agrupamentos mais coesos, o algoritmo *Cliques* é o mais adequado e usado no agrupamento de texto (KOWALSKI, 1997), (WIVES, 2004) e (GUELPELI,2010). Os textos só são adicionados a um agrupamento, caso seu grau de similaridade seja maior do que o limiar definido para todos os textos já presentes nesse agrupamento. O algoritmo 3 descreve os passos do algoritmo *Cliques* (KOWALSKI, 1997), (SALTON e MACGILL, 1983) e (WIVES, 1999).

A adpatação para este trabalho foi definir o grau de similaridade, ou seja, contabilizar o total de palavras comuns entre os textos nos seus vetores e nos agrupamentos em seus centroides[10]. Na primeira fase, a contabilização das palavras ocorre nos vetores dos textos para criar os agrupamentos. Na fase seguinte, todos os textos já estão em agrupamentos, cada um desses agrupamentos contém um centroide de palavras obtidos na primeira fase, começa então o reagrupamento. O reagrupamento contabiliza o total de palavras comuns entre centroides dos agrupamentos, pode surgir agrupamentos, subagrupamentos ou até mesmo a fusão destes. A descrição detalhada será feita no item 3.4.

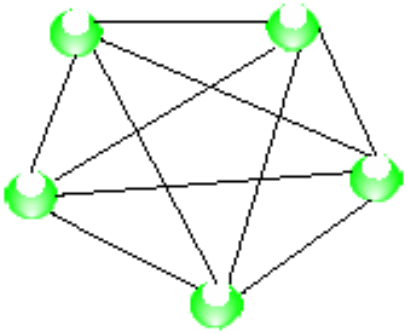


Figura 6: Grafo do algoritmo Cliques.

3 - Algoritmo Cliques:

1. Seleciona 1º elemento e coloca em um novo cluster.

2. Procura o próximo objeto similar.

3. Se 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.

[10] No modelo Cassiopeia o vetor têm 50 posições. Quando na comparação entre textos este vetor é denominado “vetor”, quando é nos agrupamentos ele é denominado “centroide”.

3.3 MODELO CASSIOPEIA - PÓS-PROCESSAMENTO

No pós-processamento, o modelo fornece uma estrutura hierárquica capaz de apresentar bons resultados para RI, justificados pelos que é apresentado no capítulo 5, constatado pelas avaliações obtidas pelas métricas externas e internas. Nessa etapa, o modelo terá como saída os textos agrupados por similaridade e sumarizados.

Este modelo possibilita uma avaliação melhor, em comparação a outros agrupadores de textos, pois estes estão sumarizados, ou seja, têm um número bem menor de sentenças, com alto grau de informatividade, isto é, garantido pela sumarização usada na etapa de pré-processamento.

Com os textos agrupados no pós-processamento, é possível realizar a recuperação de documentos e, a partir da sua análise, pode-se obter outros similares, justificando assim a criação dessa estrutura.

Com essa organização estrutural, uma generalização e/ou especificação de documentos pode ser feita, já que a partir do momento da recuperação, parece ser interessante possibilitar a consulta a outros documentos mais específicos ou mais genéricos. Quando o documento for encontrado pela RI, a estrutura possibilitará ter o texto-fonte e o seu sumário correlato, ou seja, com alto grau de informatividade.

3.4 DESCRIÇÃO DETALHADA DO MODELO CASSIOPEIA

Para um entendimento detalhado do modelo Cassiopeia, nesta seção serão descritos os passos de seu funcionamento. Os procedimentos serão assim representados: {}, usando-se também a divisão em macroetapas, ou seja, pré-processamento, processamento e pós processamento.

Pré-processamento

1. preparar os textos para o processamento;
2. definir um sumarizador;
3. determinar um grau de compressão a ser usado no sumarizador;
4. sumarizar os textos-fonte, criando textos sumarizados.

Processamento

1. {Identificar e selecionar atributos} de cada texto;
2. gerouAgrupamento = verdadeiro;
3. criar agrupamentos de trabalho com base nos textos cujos centróides são cx e cy
4. enquanto gerouAgrupamento faça
5. gerouAgrupamento = falso;
6. para x = 1 até (total de centroides) faça
7. cx = centroide x;
8. {Estabelecer maior grau de similaridade(cx, cy, x)};
9. Se cy está vazio então
10. criar um novo agrupamento contendo cx;
11. gerouAgrupamento = verdadeiro;
12. senão
13. Se cy não está vazio
14. agrupar e criar um centroide respectivo contendo as 25 palavras mais frequentes de cx e de cy, totalizando 50 palavras;
15. gerouAgrupamento = verdadeiro;
16. fimSe
17. fimSe
18. fimPara
19. fimEnquanto
20. Fim.

Pós-Processamento

1. obter textos fontes com seus respectivos sumários em agrupamentos hierarquizados.

{*Identificar e selecionar atributos*}

1. calcular a frequência relativa: quantas vezes cada palavra aparece no documento, dividido pelo número total de palavras do documento;
2. ordenar as palavras em ordem descrescente de frequência (da maior para a menor);
3. achar a frequência média das palavras, somando as frequências relativas e dividindo pelo número total de palavras do documento;
4. encontrar a primeira palavra cuja frequencia mais próxima à média;
5. marcar esta palavra e escolher, incluindo-a, mais as 24 anteriores (esquerda);
6. marcar esta palavra e escolher as 25 posteriores (direita);
7. montar o vetor em ordem decrescente com as 50 palavras escolhidas.

{*Estabelecer maior grau de similaridade*(cx, cy, x)}

1. *scoreMaior = 0;*
2. para y = x+1 até (total de centroides) faça
3. *scoreAtual* = **total de palavras comuns nos centroides x e y // que representa a similaridade entre os centroides**;
4. Se *scoreAtual* > *scoreMaior* então
5. *scoreMaior = scoreAtual;*
6. *cy* = centroide y;
7. FimSe
8. FimPara
9. Fim.

3.5 CONSIDERAÇÕES FINAIS DO MODELO CASSIOPEIA

Com a proposta do uso da sumarização no pré-processamento, o modelo usa, para representação do seu espaço amostral, um vetor de 50 posições denominado centroide, atenuando o problema da alta dimensionalidade e dos dados esparsos e pode-se manter a lista de *stopword*, tornando-o independente do idioma. Isso é importante, pois a interação humana deve ser mínima e, se possível, nenhuma, dado o volume de informações textuais a ser recuperado. Dessa forma, o modelo contribui para atenuar o problema da sobrecarga de informação em RI.

Os bons resultados obtidos em domínios distintos e/ou antagônicos, que serão apresentados e discutidos no capítulo 5, garantem essa avaliação dos agrupadores, mostrando que o modelo Cassiopeia não se restringe a um domínio específico; isto ocorre devido à variação da frequência média feita no corte de Luhn, proposto no modelo.

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 alto grau de informatividade, devido à sumarização.

CAPÍTULO 4 – MÉTODO DE ELABORAÇÃO DOS TESTES

No capítulo 4 serão descritos os métodos de elaboração dos testes realizados neste trabalho. Será explicado o conceito de *corpus* computadorizado e as fases que compõem a escolha de cada um. Os *corpora* serão explicados com base nas fases propostas pela literatura, e mostradas as estatísticas de cada *corpus,* separadamente, contendo, entre várias informações, o número de palavras totais de cada um.

Todos os *corpora* utilizados foram obtidos de bases nacionais e internacionais, nos idiomas português e inglês, nos domínios jornalístico, jurídico e médico. Para viabilizar a elaboração dos testes foram escolhidos sumarizadores nas duas línguas, e os critérios que orientaram as seleções desses sumarizadores serão explicados. Depois de apresentar os sumarizadores, são apresentadas as suas funcionalidades. Os sumarizadores são separados em dois tipos, profissionais e da literatura. Serão apresentadas ainda as funções aleatórias, desenvolvidas com a função também de sumarizar, explicando suas finalidades. Além delas, a definição dos graus de compressão usados nos sumarizado.

4.1 *CORPORA*

Em 1961, foi laçado o primeiro *corpus* linguístico eletrônico denominado *Brown* (KUCERA e FRANCIS, 1961), com 1 milhão de palavras, que impulsionou o desenvolvimento da Linguística de *Corpus* (LC). Foi o pioneiro dos *corpora* eletrônicos, pois, na época, os recursos computacionais eram escassos. Desde o lançamento do *Brown University Standard Corpus of Present-Day American English* ou *corpus Brown*, a área de *corpus* linguístico tem evoluído. Segundo Sardinha (2000), a LC preocupa-se com a coleta e exploração de *corpora*. Não é só nos centros acadêmicos que tem alcançado mais espaço, nas empresas também.

Para Aluísio e Almeida (2006), um *corpus* computadorizado observa um conjunto de requisitos que influencia a validade e confiabilidade da pesquisa baseada em *corpus*. Segundo os autores, autenticidade, representatividade, balanceamento, amostragem, diversidade e tamanho são os fatores a serem observados na coleta da formação de um *corpus*. Para eles, criar um *corpus* é um processo repetitivo, que começa com a seleção dos textos, baseada em algum critério significativo para a pesquisa (critério externo), continua com as investigações empíricas da língua ou variedade linguística sob análise (critério interno), e termina com a revisão de todo o projeto.

A escolha do *corpus* para este trabalho teve duração de dois meses (fevereiro a abril de 2010), da concepção até o momento da compilação.

A primeira fase, denominada por Aluísio e Almeida (2006) critério externo, foi estabelecida com base em critérios representativos para a pesquisa. Para a criação dos *corpus* foram observados os critérios de (1) gratuidade, (2) possibilidade de cópias das próprias bases, (3) classificação das bases para cada um dos domínios determinados para o trabalho (jornalístico, jurídico e médico), (4) resumo do texto original, denominado sumário de referência, elaborado pelo autor.

Na segunda fase, as investigações empíricas recaíram sobre a escolha dos idiomas, inglês e português, conjuntamente com os sumarizadores disponíveis para simulação nos domínios jornalístico, jurídico e médico. Terminada a elaboração dos critérios externos e internos, todo o projeto foi revisado e ajustado a essas necessidades.

Na etapa de manipulação, ocorreram a limpeza e a formatação do *corpus* para o processamento computacional. Na limpeza, foram retiradas as imagens, gráficos, tabelas, números de páginas e todas as anotações que não faziam parte do corpo do texto. Na formatação, todos os textos foram convertidos para o formato ".txt", ou seja, formato compatível para o processamento computacional. Os arquivos foram renomeados, seguindo uma organização sequencial de 1 a 100. Há duas pastas, uma para os textos em português, e outra para os textos em inglês. Devido à necessidade de se utilizar o texto original, e o resumo fornecido pelo autor, foram criadas duas subpastas para esta finalidade. Uma, denominada "Sumários Manuais", na qual estão os resumos fornecidos pelos autores, e outra, denominada "Textos Fontes", na qual estão os textos ".txt". Em cada uma delas existem subpastas, nomeadas de acordo com as categorias obtidas nas bases de origem, sendo assim, cada texto aparece classificado em uma única categoria.

Na Figura 7 é apresentado o diagrama dos *corpora* usados. No total, foram 1000 textos selecionados, com 500 textos-fonte e 500 textos sumários manuais. Todos os textos-fonte e os sumários manuais pertencem a uma única categoria. Não há texto repetido e/ou classificado em mais de uma categoria.

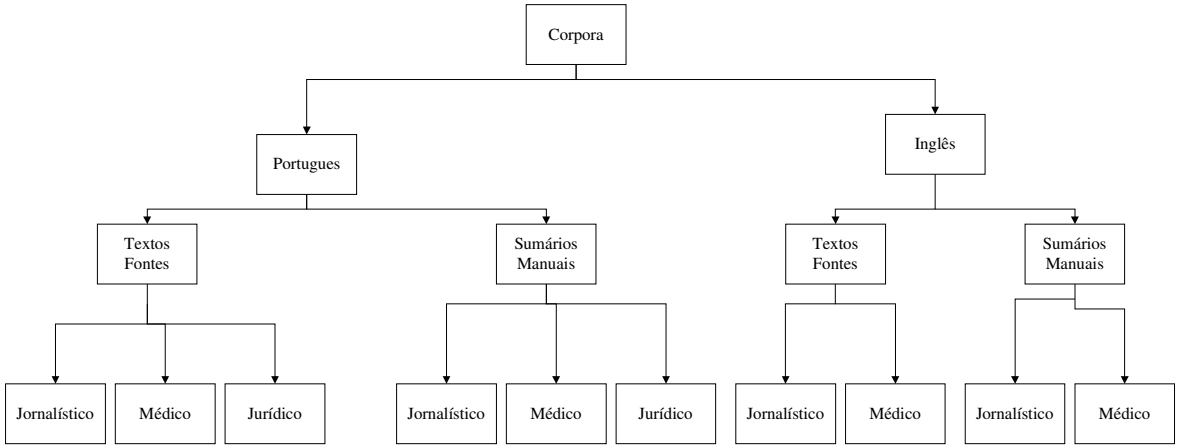


Figura 7: Diagrama dos *corpora* usados neste trabalho.

As estatísticas dos *corpora* serão apresentadas e todos os valores obtidos nas Tabelas 1,2,3,4 e 5 foram calculados com a utilização do software *Get FineCount 2.6*, cuja última versão é de 10 de setembro de 2010.

4.1.1 *CORPORA* EM PORTUGUÊS

Para formação dos *corpora* em português foram escolhidos três domínios, jornalístico, jurídico e médico, totalizando 600 textos, descritos a seguir:

(1) Domínio Jurídico: 100 textos-fonte, entre eles textos científicos retirados de bases especializadas em artigos jurídicos, nos *sites*: http://www.r2learning.com.br/_site/artigos/ e http://www.direitonet.com.br/artigos/, acessados entre 09-02-2010 e 14-02-2010. Os textos foram retirados de nove categorias, classificadas pelo *site* em áreas como ambiental, civil, constitucional, consumidor, família, penal, previdenciário, processual e trabalhista. Nas áreas consumidor, família e trabalhista há quinze textos em cada categoria. Nas áreas ambiental, civil, constitucional, penal e previdenciário, dez para cada categoria, enquanto na categoria processual, apenas cinco. Como resumos de referência humano foram usados os 100 resumos criados pelos próprios autores dos artigos científicos. A Tabela 2 mostra a estatística do *corpus* da categoria jurídica relacionada aos textos-fonte. As linhas têm valores mínimos, máximos, totais e médias relacionadas a cada coluna. Um item que faz parte da coluna, muito importante para este trabalho, por exemplo, é o número de palavras máximo e mínimo. Este *corpus* tem um texto com 501 palavras, no mínimo, e outro com 14.080, no máximo. No total dos 100 textos, são 236.339 palavras, com média de 2.363,39 por texto.

Tabela 2: Estatística dos 100 textos-fonte no domínio jurídico, compostos por 9 categorias e no idioma português.

Itens	Nº de Palavras	Nº de Palavras + Numeral	Nº (%)	Carac-teres	Caracteres + Espaços	Sen-tenças	Palavra+Numeral por Sentença		Caracteres por Palavra+Numeral	
Mínimos	501	506	0,06	2.625	3.142	21	506	21	2.625	506
Máximos	14.080	14.208	10,29	75.903	90.202	697	14.208	697	75.903	14.208
Totais	236.339	242.332	-	1.295.228	1.544.983	11.421	242.332	11.421	1.295.228	242.332
Médias	2.363,39	2.423,32	2,47	12.952,28	15.449,83	114,21	20,69		5,48	

(2) Domínio Médico: 100 textos-fonte, científicos. Foram extraídos da *Cientific Electronic Library Online - SciELO* Brasil, do endereço, http://www.scielo.br/ entre 15-02-2010 e 20-02-2010. O *site* é especializado em artigos científicos, incluindo diversas áreas da saúde. Os textos foram retirados de dez categorias, classificadas pelo *site* como cardiologia, dermatologia, epidemiologia, geriatria, ginecologia, hematologia, neurologia, oncologia, ortopedia e pediatria, e em cada categoria há dez textos. Como resumos de referência humano foram usados 100 resumos, criados pelos próprios autores dos artigos científicos. A tabela 3 mostra a estatística do *corpus* da categoria médica relacionada aos textos-fonte. As linhas têm valores mínimos, máximos, totais e médias, relacionadas a cada coluna. Um item que faz parte da coluna, muito importante para este trabalho, por exemplo, é o número de palavras máximo e mínimo. Esse *corpus* tem um texto com 537 palavras, no mínimo e outro com 7.428, no máximo. No total dos 100 textos são 283.601 palavras, com média de 2.836,01 por texto.

Tabela 3: Estatística dos 100 textos-fonte no domínio médico, compostos por 10 categorias e no idioma português.

Itens	Nº de Palavras	Nº de Palavras + Numeral	Nº (%)	Carac-teres	Caracteres + Espaços	Sen-tenças	Palavra Palavra+Numeral por Sentença		Caracteres por Palavra+Numeral	
Mínimos	537	577	0,41	3.355	3.896	29	577	29	3.355	577
Máximos	7.428	7.537	13,17	43.121	50.718	346	7.537	346	43.121	7.537
Totais	283.601	296.912	-	1.632.902	1.930.394	12.903	296.912	12.903	1.632.902	296.912
Médias	2.836,01	2.969,12	4,48	16.329,02	19.303,94	129,03	21,98		5,76	

(3) Domínio Jornalístico formado pelo corpus TeMário Rino e Pardo (2003), com 100 textos extraídos de dois jornais nacionais, *Folha de São Paulo* e *Jornal do Brasil*. Os textos são classificados de acordo com as seções dos jornais: Especial, Mundo, Opinião da *Folha de São*

Paulo; Internacional e Política do *Jornal do Brasil*. Em cada seção há vinte textos. Todos já têm resumos de referência humana, criados por especialistas humanos. A Tabela 4 mostra a estatística do *corpus* da categoria jornalística relacionada aos textos-fonte. As linhas têm valores mínimos, máximos, totais e médias relacionadas a cada coluna. Um item que faz parte da coluna, muito importante para este trabalho, por exemplo, é o número de palavras máximo e mínimo. Este *corpus* tem um texto com 424 palavras, no mínimo e outro com 1.310, no máximo, No total dos 100 textos, são 62.014 palavras, com média de 620,14 por texto.

Tabela 4: Estatística dos 100 textos no domínio jornalístico, compostos por 5 categorias e no idioma português.

Itens	Nº de Palavras	Nº de Palavras + Numeral	Nº (%)	Carac-teres	Caracteres + Espaços	Sen-tenças	Palavra Palavra+Numeral por Sentença		Caracteres por Palavra+Numeral	
Mínimos	424	431	0	2.160	2.649	14	424	14	2.160	431
Máximos	1.310	1.350	10,29	6.827	8.155	71	1.310	71	6.827	1.350
Totais	62.014	63.132	-	323.919	391.014	3.212	62.014	321.219,31	323.919	63.132
Médias	620,14	631,32	1,77	3.239,19	3.910,14	32,12	19,31		5,22	

4.1.2 *CORPORA* EM INGLÊS

Para a língua inglesa, houve variação apenas nos domínios, jornalístico e médico. Os textos jurídicos não se enquadraram dentro dos critérios estabelecidos, no momento da criação do *corpus*. Não foram encontradas bases com as características, por exemplo, da gratuidade. Para formação dos *corpora,* em inglês, optou-se também por uma variedade de domínios entre textos jornalísticos e médicos, totalizando 400 textos, descritos a seguir:

(1) Domínio Jornalístico. Os textos foram retirados da Agência de Notícias Reuters (http://www.reuters.com/news/politics), entre 01-03-2010 e 15-03-2010. Foram escolhidos de acordo com as categorias determinadas pela Agência de Notícias. Os 100 textos, extraídos de dez categorias, cada uma com dez, formaram assim as categorias: e*conomy, entertainment, G-20, green business, health, housing market, politics, science, sports,* e t*echnology.* Como resumos de referência humana também foram usados 100 textos, obtidos dos *leads* das notícias, criados pelos próprios autores da notícias.

A Tabela 5 mostra a estatística do *corpus* da categoria jornalística relacionada aos textos-fonte. As linhas têm valores mínimos, máximos, totais e médias relacionadas a cada coluna. Um item que faz parte da coluna, muito importante para este trabalho, por exemplo, é o número de

palavras máximo e mínimo. Este *corpus* tem um texto com 25 palavras, no mínimo e outro com 1.257, no máximo, totalizando, nos 100 textos 39.433 palavras, com média de 394,33 por texto.

Tabela 5: Estatística dos 100 textos fontes no domínio jornalístico, compostos por 10 categorias e no idioma inglês.

Itens	Nº de Palavras	Nº de Palavras + Numeral	Nº (%)	Carac-teres	Caracteres + Espaços	Sen-tenças	Palavra Palavra+Numeral por Sentença		Caracteres por Palavra+Numeral	
Mínimos	25	26	0	125	150	2	25	2	125	25
Máximos	1.257	1.257	10,32	5.800	7.127	81	1.257	81	5.800	1.257
Totais	39.433	40.506	-	199.485	241.030	2.123	38.433	2.123	199.485	39.433
Médias	394,33	405,06	2,65	1.994,85	2.410,30	21,23	18,57		5,06	

(2) Domínio Médico - composto por 100 textos-fonte. Foram usados textos científicos do domínio médico, no idioma inglês, extraídos da *Cientific Electronic Library Online - SciELO* Brasil, no endereço http://www.scielo.br/, entre 16-03-2010 e 25-03-2010. Os textos foram retirados de dez categorias classificadas pelo *site*: *cardiology, dermatology, epidemiology, geriatrics, gynecology, hematology, neurology, oncology, orthopedics and pediatrics*. Como resumos de referência humana, foram usados os 100 *abstracts* dos artigos científicos, criados pelos próprios autores. A Tabela 6 mostra a estatística do *corpus* da categoria médica relacionada aos textos-fonte. As linhas têm valores mínimos, máximos, totais e médias relacionadas a cada coluna. Um item que faz parte da coluna, muito importante para este trabalho, por exemplo, é o número de palavras máximo e mínimo. No *corpus* há um texto com 451 palavras, no mínimo e outro com 8.520, no máximo, totalizando, nos 100 textos, 250.258 palavras, com média de 2.502,58 nos 100 textos.

Tabela 6: Estatística dos 100 textos no domínio médico, compostos por 10 categorias e no idioma inglês.

Itens	Nº de Palavras	Nº de Palavras + Numeral	Nº (%)	Caracteres	Caracteres + Espaços	Sentenças	Palavra Palavra+Numeral por Sentença		Caracteres por Palavra+Numeral	
Mínimos	451	472	0,75	2.353	2.816	27	451	27	2.353	451
Máximos	8.520	8.899	19,24	48.533	57.134	940	8.520	940	48.533	8.520
Totais	250.258	265.143	-	1.422.112	1.685.205	14.952	250.258	14.952	1.422.112	250.258
Médias	2.502,58	2.651,43	5,61	14.221,12	16.852,05	149,52	16,74		5,68	

4.2 SUMARIZADORES AUTOMÁTICOS

Todos os sumarizadores são automáticos, conforme explicado na seção 2.5. A escolha desses sumarizadores, a princípio, é justificada pela possibilidade de eles sumarizarem tanto na língua inglesa quanto na língua portuguesa. Inicialmente, deveriam ser dois sumarizadores profissionais e um da literatura (o melhor em resultados qualitativos, comprovado em experimentos publicados).

Outro critério para escolha dos sumarizadores deste experimento é a possibilidade de definir percentuais de compressão. Foram selecionados os que deveriam possibilitar uma faixa de compressão de 50% a 90%.

A primeira dificuldade foi encontrar sumarizadores que pudessem ser usados nos dois idiomas, português e inglês. Apesar do número significativo de sumarizadores profissionais encontrados, em inglês, eles não atendiam aos requisitos de percentual de compressão.

Devido à não disponibilidade, em português, de sumarizadores profissionais, foram adotados apenas os da literatura. Já na língua inglesa, foi possível manter o critério de dois profissionais e um da literatura, todos com a possibilidade de escolha de percentuais de compressão. Foram ainda desenvolvidas funções *baselines,* descritas em 4.2.3.

4.2.1 SUMARIZADORES EM PORTUGUÊS

Para o processo de sumarização na língua portuguesa, foram utilizados três sumarizadores encontrados na literatura, vistos a seguir, que atenderam aos requisitos, pois além de serem gratuitos, todos tinham a possibilidade de sumarizar através da compressão.

Foram encontrados, alguns sumarizadores profissionais, não desenvolvidos especificamente para a língua portuguesa, que foram considerados para análise, mas descartados,

por terem problemas em relação à questão da gratuidade ou da compressão (não realizavam compressão).

Em um trabalho específico de sumarização em português, Leite (2010), na lista de sumarizadores por ele estudada, não encontrou também nenhum profissional, com as características expostas anteriormente.

4.2.1.1 SUPOR

O SuPor(MÓDOLO, 2003) pode ser estruturado em três grandes níveis: superficial, em cujas abordagens encontram-se tamanho da sentença, palavras mais frequentes, nomes próprios, posição e sintagmas sinalizadores; no nível de entidade, o SuPor trabalha com similaridade e relações com *thesaurus*; o último nível, foi considerado por como o de enredo de tópicos.

Para realizar a SA, quaisquer informações processadas pelos três níveis citados podem ser escolhidas e, quando juntas, proporcionam raciocínios diferentes de identificação e extração das sentenças, para compor o SA. O SuPor na realidade, é uma estrutura que incorpora vários métodos para SA, cada um com suas próprias características. O usuário tem o papel de um especialista (projetista do sistema ou conhecedor da metodologia).

Para Leite e Rino (2006), o SuPor usa sete métodos de SA, adaptados para o português, que serão apresentados, para entendimento do funcionamento do SuPor, juntamente com seus algoritmos.

O primeiro método é o da palavra mais frequente, selecionado para compor o extrato, as sentenças que incluem as palavras mais frequentes do texto-fonte, pois elas representam os conceitos mais importantes do texto. A escolha de cada sentença é feita mediante a classificação de sua representatividade no texto. Para isso, é atribuído um *score* para cada sentença, baseado na soma das frequências de suas palavras em todo o texto. Feito isso, determina-se um *score* de corte (*threshold*), com base em medidas estatísticas e, em seguida, são selecionadas as sentenças com as palavras mais frequentes. Os passos desse método são apresentados no algoritmo 4.

4 – Palavras mais frequentes:

1. *Pré-processamento:*
   a. *Remoção das stopwords;*
   b. *Stemming ou geração dos quadrigramas.*
2. *Cômputo das frequências:*
   a. *Cômputo da frequência dos radicais ou dos quadrigramas em todo texto;*
   b. *Atribuição do score a cada sentença, com base na soma das frequências de suas palavras;*
3. *Seleção das sentenças:*
   a. *Obtenção do score mínimo para uma sentença ser selecionada;*
   b. *Seleção das sentenças com score acima ou igual ao mínimo.*

Permitiu-se que os radicais gerados pelo *stemmer* fossem considerados os quadrigramas das palavras, isto é, sequências sobrepostas, de quatro letras de cada palavra. Essa técnica se desdobra em duas *features* do SuPor, uma que utiliza radicais e outra que utiliza quadrigramas no pré-processamento.

O segundo método é o do tamanho da sentença, de acordo com Leite e Rino (2006). O SuPor tem um número mínimo de palavras (cinco), ou seja, somente sentenças com, no mínimo, cinco palavras são selecionadas para compor o SA, sendo assim, é necessário apenas contabilizar o número de palavras e as sentenças que contêm o número de palavras.

A terceira técnica é a da posição, segundo Leite e Rino (2006). O SuPor seleciona as sentenças que pertencem aos parágrafos do início do texto (10% iniciais) e do final (5% finais). Também são selecionadas as sentenças iniciais e finais de parágrafos com mais de duas sentenças.

A quarta técnica é a de nomes próprios. É atribuído um peso para cada sentença, baseado no somatório dos nomes próprios dessa sentença, em todo o documento. Os passos desse método são apresentados no algoritmo 5.

5 – Nomes próprios:

1. *Cômputo das frequências:*
   a. *Cômputo da frequência dos nomes próprios: os nomes próprios são identificados com base na ocorrência de iniciais maiúsculas nas palavras, desde que elas não figurem no início da sentença;*
   b. *Atribuição do score a cada sentença, com base na soma das frequências de seus nomes próprios;*
2. *Seleção das sentenças:*
   a. *Obtenção do score mínimo para uma sentença ser selecionada;*
   b. *Seleção das sentenças com score acima ou igual ao mínimo.*

O quinto método é representado por cadeias lexicais, técnica que calcula a importância de cada cadeia lexical, e esse cálculo, com três heurísticas, que seleciona as sentenças, resulta em muitas cadeias lexicais possíveis. Os passos desse método são apresentados no algoritmo 6.

6 – Cadeias Lexicais:

1. *Pré-processamento do texto fonte:*
   a. *Segmentar o texto fonte em tópicos, usando o algoritmo TextTiling, ou utilizar os parágrafos como tópicos;*
   b. *Selecionar os substantivos do texto usando um etiquetador morfossintático para o português.*
2. *Construção das cadeias lexicais:*
   a. *Criar, para cada palavra candidata de cada tópico, tantas interpretações quantas sejam seus sentidos no Thesaurus;*
   b. *Computar o escore de cada interpretação pela soma dos pesos das relações na interpretação, sendo que os pesos são definidos da seguinte forma: repetição e sinonímia = 10 e antonímia = 7;*
   c. *Escolher a interpretação com escore mais alto para cada cadeia lexical quando todas as palavras de um tópico já tiverem sido computadas;*
   d. *Unir as cadeias lexicais de tópicos diferentes que contêm uma palavra comum com o mesmo sentido;*
   e. *Calcular os escores das cadeias lexicais multiplicando o número de ocorrências de membros da cadeia lexical (tamanho da cadeia lexical) pelo índice de homogeneidade. O índice de homogeneidade é calculado subtraindo de um o resultado da divisão do número de ocorrências distintas dos membros da cadeia lexical pelo seu tamanho, i. e., pelo número de todos os seus membros;*
   f. *Escolher as cadeias lexicais mais fortes cujos escores sejam maiores que a multiplicação da média dos escores de todas as cadeias lexicais do texto fonte pelo desvio padrão desses escores.*

Depois de construir as cadeias lexicais, são usadas três heurísticas para a escolha da sentenças. Na heurística 1 (primeira ocorrência), são selecionadas as sentenças que contêm a primeira ocorrência de um membro de uma cadeia lexical forte, depois, na heurística 2 (membro representativo), o processo é similar ao anterior, com a diferença de que o membro tem de ser representativo. Um membro é considerado representativo se a sua frequência na cadeia for maior que a frequência média de todos os seus membros. Por fim, na heurística 3 (concentração no tópico), a cadeia lexical é dada pela divisão do número de ocorrências dos membros dessa cadeia lexical, no tópico, pelo número total de substantivos do tópico. Para cada cadeia lexical, é encontrado o tópico onde ela é mais concentrada. Seleciona-se esse tópico na primeira sentença que contiver um membro dessa cadeia.

O sexto método apresenta a importância dos tópicos. Nessa técnica, são usados *stopwords* e um algoritmo *stemmer,* adaptado para o potuguês, para extração de quadrigramas, em cujo processo é possível conseguir separar os radicais das palavras. Os passos dessa técnica são apresentados no algoritmo 7.

7 – Importância dos tópicos:

1. *Pré-processamento do texto-fonte:*
   a. *Remoção das stopwords;*
   b. *Stemming ou geração de quadrigramas.*
2. *Divisão do texto em tópicos: uma versão modificada do algoritmo TextTiling, proposta pelos autores do método, é aplicada ao texto para sua divisão em tópicos (tiles).*
3. *Cálculo da força dos tópicos: a força do tópico é definida como a soma da média dos vetores TF-ISF2 (Trm frequency – Inverse sentence frequency) de suas sentenças. Para todos os tópicos são, normalizados no intervalo [0,1].*
4. *Cálculo do número de sentenças de cada tópico: o cálculo é feito com base na importância dos tópicos, um número proporcional de sentenças a extrair de cada tópico.*
5. *Seleção das sentenças para cada tópico: as sentenças são selecionadas como as que têm maior similaridade com o centroide do tópico. O centroide do tópico é o vetor resultante da média dos vetores TF-ISF das sentenças desse tópico. Para o cômputo da similaridade da sentença com o centroide é usada medida de similaridade dos co-senos.*

O sétimo método, denominado mapa de relacionamentos, mostra onde é construído o grafo com ligações com os parágrafos que irão compor o SA.

Os passos desse método são apresentados no algoritmo 8.

8 – Mapas de relacionamentos:

1. *Pré-processamento do texto-fonte:*
   a. *Remoção das stopwords;*
   b. *Stemming ou geração de quadrigramas.*
2. *Construção do mapa de relacionamentos:*
   a. *Cálculo dos vetores TF-IPF3 (Term frequency – Inverse paragraph frequency) para cada parágrafo do texto;*
   b. *Cálculo da similaridade entre os parágrafos. Essa similaridade é determinada pela aplicação de uma medida semelhante à dos co-senos aos vetores TF-IPF.*
   c. *Construção do mapa de relacionamentos, considerando as ligações entre parágrafos que tenham similaridade acima de um valor mínimo.*
3. *Seleção dos parágrafos, como já mencionado, são três as formas possíveis de selecionar os parágrafos para compor o extrato, que correspondem ao modo como o percurso no grafo se dá.*

Os caminhos a serem percorridos nos grafos resultantes do algoritmo 8 são três:

No caminho 1 (denso), são selecionados os parágrafos que contêm o maior número de ligações, até que a taxa de compressão seja atingida. No caminho 2 (profundo), primeiro é selecionado o parágrafo que atenda ao caminho1 (mais denso). A partir do parágrafo mais denso, é selecionado aquele que tem o maior número de ligações com ele próprio. A partir desse segundo parágrafo, é selecionado o terceiro, que tem mais ligações com esse último, e assim sucessivamente, até atingir a taxa de compressão escolhida. No caminho 3 (segmentado), procura-se selecionar diferentes parágrafos de cada tópico do texto-fonte. Os tópicos são produzidos através de uma simplificação do mapa. Usando essa divisão em tópicos, é selecionado pelo menos um parágrafo de cada tópico, e os outros podem ser selecionados a partir dos tópicos com maior número de parágrafos, até se obter a taxa de compressão desejada.

O SuPor, para Leite e Rino (2006), apresenta o mapeamento entre o resultado de um método e o valor da *feature*, o método indica a sentença, mas o valor dessa *feature* é *True*; caso

contrário, será *False.* Leite e Rino (2006) resumem, na Tabela 6, esse mapeamento. Para Módolo (2003), a Figura 9 mostra o módulo de extração do SuPor.

A Figura 9 representa como o SuPor realiza a extração das sentenças para compor o sumário. Primeiramente, são utilizados as *stopwords*, léxico e de *thesauros* no pré-processamento, para obter um conjunto de sentença dos textos. Com esse conjunto de sentenças separado das demais, passa-se para as opções de processamento, nas quais são realizadas a classificação das sentenças, a sua ordenação e finalmente a seleção de sentenças que vão compor o sumário. São usados os percentuais de compressão. A Tabela 7 mostra o processamento que deve ocorrer para cada uma das *features,* sendo verdadeira, a que o autor denomina opções de processamento, mostradas na Figura 9.

Tabela 7: Mapeamento entre métodos e *Features* (LEITE e RINO, 2006).

Métodos	**Condição para a *Feature* assumir *True***
Palavras mais frequentes	*Score* (soma das frequências de cada palavra) deve ser maior que o mínimo.
Tamanho da sentença	Número de palavras da sentença deve ser maior que cinco.
Posição	Sentença deve figurar nos parágrafos iniciais e finais do texto, ou ser uma sentença inicial ou final de qualquer parágrafo.
Nomes próprios	*Score* (soma das frequências de cada nome) deve ser maior que o mínimo.
Cadeias lexicais	A sentença deve ser recomendada por pelo menos uma das heurísticas do método.
Importância dos tópicos	A sentença deve ser saliente no tópico em que se encontra.
Mapa de relacionamentos	A sentença deve ser recomendada por pelo menos um dos três percursos no mapa.

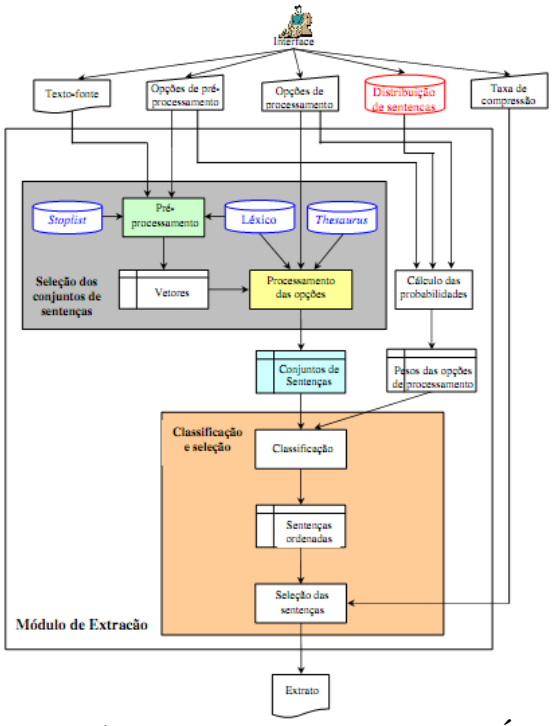


Figura 8: Módulo de extração do SuPor (MÓDOLO, 2003).

4.2.1.2 GIST SUMMARIZER

O GistSumm - GIST SUMMarizer (PARDO,2002) é um sumarizador que usa abordagem superficial, ou seja, estatística. Procura identificar a sentença principal do texto-fonte, que é denominado por Pardo (2002) a "sentença-*gist*" do texto-fonte.

A partir dessa sentença, o sumarizador começa a selecionar as outras que irão compor o sumário automático.

A Figura 10 mostra a arquitetura do GistSumm, com suas funcionalidades. A segmentação sentencial é a primeira etapa do processo do GisSumm, na qual ocorre a delimitação das sentenças através de sinais de pontuação, como ponto final, de exclamação e de interrogação. O ranqueamento de sentenças é a segunda etapa do processo, e nela o usuário escolhe o método que será usado para ranquear as sentenças. A sentença com maior pontuação no *rank* é considerada a sentença-*gist*, a partir da qual as outras s que compõem o sumário são determinadas. Ainda nesta etapa, como mostra a Figura 10, existe a utilização da lista de *stopwords* para o idioma português. De acordo com Pardo (2002), ela é composta de 196 palavras. Ocorre também o uso de um léxico que, de acordo com Pardo *et al.* (2003), foi extraído do Núcleo Interinstitucional de Linguística Computacional- NILC, que assegura ser o maior léxico da língua portuguesa.

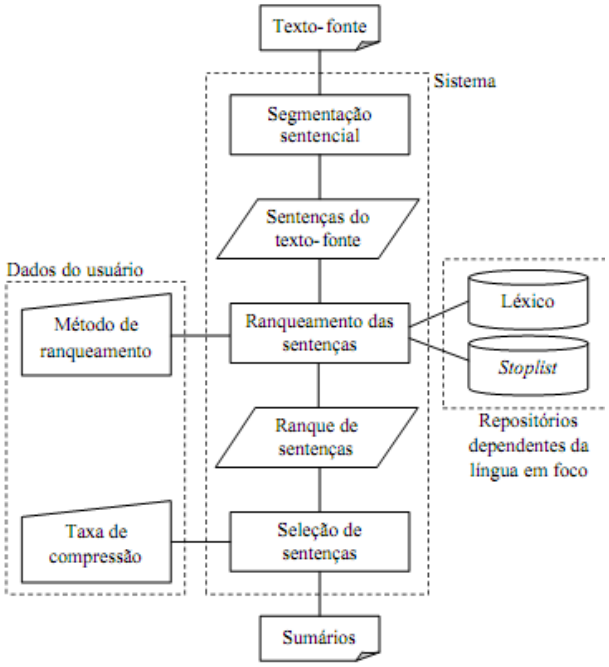


Figura 9: Arquitetura do *GistSumm* (PARDO, 2002).

Para a fase seleção das sentenças, o GistSumm calcula a média da pontuação das sentenças do texto-fonte e assume essa média como um limite para o corte das possíveis sentenças que formarão o sumário. Para Pardo *et al.* (2003), o GistSumm seleciona, com a sentença-*gist,* todas as sentenças do texto-fonte que contenham pelo menos uma palavra canônica[11] da sentença-*gist,* e uma pontuação maior que o limite calculado. Nessa etapa, o usuário seleciona também a compressão. O GistSumm coloca sentenças no sumário, respeitando a ideia da sentença-*gist*, até que o percentual de compressão seja obtido, ou pode acontecer a não obtenção total do percentual de compressão especificado. Assim, o número de sentenças selecionadas para formar o sumário depende da taxa de compressão especificada pelo usuário do sistema. A taxa de compressão é uma medida que determina o tamanho do sumário em relação ao do texto-fonte.

A nova versão do GistSumm de Pardo (2005) foi escolhida para ser usada neste trabalho. Uma das principais extensões implementadas desta versão desenvolvida por Pardo (2002) foi a incorporação da funcionalidade de multidocumentos. Nela vários textos-fonte são inseridos no sistema, que gera um único sumário, diferente da opção monodocumento usada em toda a simulação deste trabalho, que, para cada texto-fonte gera um único sumário.

Pardo (2005) considera outras funcionalidades criadas, como a escolha do ranqueamento de sentença, no qual foram implementados os métodos *average keywords* e intrasentença, que serão

[11] Canônica é uma palavra que significa norma, padrão, regra. São sílabas que seguem uma norma, um padrão , ou seja, uma consoante e uma vogal em cada sílaba.

discutidos mais detalhadamente nos itens 4.2.1.2.1. e 4.2.1.2.2. O método de *Term Frequency – Inverse Sentence Frequency* - TF-ISF não foi usado, pois obteve uma *performace* inadequada, relatada em Pardo (2002).

Afirmam Pardo *et al.* (2006) que o GistSumm foi ampliado, também para tratamento na utilização de texto científico. Esse foi um dos principais motivos para adotar a última versão, já que boa parte dos textos usados neste trabalho são científicos. Um dos problemas relatados por Pardo (2007) seria o de compressões altas, que resultariam em problemas na formação do sumário, mas todos os textos científicos aqui usados são em tamanhos maiores do que os textos usado por Pardo (2007), sendo assim não foi constatado tal problema.

4.2.1.2.1 GIST_AVERAGE_KEYWORD

Pardo (2007) afirma que a pontuação de sentenças pode ocorrer por um dos dois métodos estatísticos simples: *keywords* ou *average keywords.* O método *keywords,* com normalização, é denominado *average keywords,* em função do tamanho das sentenças (medido em número de palavras). A fase seguinte passa para o ranqueamento das sentenças em função da pontuação obtida, sendo que a sentença de maior pontuação é eleita *gist sentence*, isto é, a sentença que melhor representa a ideia principal do texto. As sentenças são selecionadas de modo que (1) contenham pelo menos um *stemmer* comum com a *gistsentence* selecionada na etapa anterior e (2) tenham uma pontuação maior do que um *threshold*, a média das pontuações das sentenças. Em (1), procura-se selecionar sentenças que complementem a ideia principal do texto; em (2), procura-se selecionar somente sentenças relevantes, com base no percentual de compressão determinado pelo usuário.

4.2.1.2.2 GIST_INTRASENTENÇA

Pardo (2007) diz que a sumarização é realizada no interior das sentenças através da exclusão das *stopwords.* Esse método é possível, devido ao desenvolvimento de um segmentador textual, para que possa não só delimitar, de forma mais precisa, as sentenças de um texto, mas também as orações intrasentenciais.

4.2.2 SUMARIZADORES EM INGLÊS

Para o processo de sumarização, em inglês, foram utilizados três sumarizadores, dois profissionais e um da literatura, disponibilizado na web. Esses sumarizadores serão vistos nas próximas subseções. Sumarizadores profissionais na língua inglesa, diferentes dos da língua portuguesa são encontrados em grandes quantidades. Para escolha dos profissionais, foram usados

os critérios de gratuidade e percentuais de compressão. O da literatura foi escolhido por atender aos critérios estabelecidos neste trabalho e, principalmente, por possuir uma versão atualizada e disponível para uso na internet, e também por ter resultados comparativos significativos, apresentados por seu autor, Hassel (2007).

4.2.2.1 *COPERNIC*

O sumarizador profissional *CopernicSummarizer* foi desenvolvido pela *Copernic Inic* e pode ser encontrado no link: http://www.copernic.com/en/products/summarizer/. Em pesquisa pelo *site* e em contato com seu fornecedor, não foi possível obter especificação do seu algoritmo. A versão usada para os testes, neste trabalho, foi a *trial.*

4.2.2.2 *INTELLEXER SUMMARIZER*

O sumarizador profissional *Intellexer Summarizer Pro* foi desenvolvido pela *EffectiveSoft* e pode ser encontrado no link:http://summarizer.intellexer.com/order_summarizer_pro.php. Em busca no *site* e em contato com seu fornecedor, também não foi possível obter especificação do seu algoritmo. A versão usada para os testes neste trabalho, foi a *trial*..

4.2.2.3 *SEWSUM*

O sumarizador *Sewsun* de Hassel (2007), pode ser usado no link http://swesum.nada.kth.se/index-eng.html. Sewsum é um sumarizador da literatura. Para cada idioma, o sistema utiliza um léxico que mapeia as formas flexionadas das palavras de conteúdo para a sua raiz respectiva. O léxico é utilizado para identificação do tema, com base na hipótese de que as sentenças contêm palavras de conteúdo de alta frequência. Observadas as frequências de cada sentença, estas são então modificadas por um conjunto de heurísticas, por exemplo, posição da sentença no texto e sua formatação. *SweSum* exige uma abreviatura para o léxico e um conjunto de heurísticas específicas, para a linguagem correta e *token,* na divisão de frase.

4.2.3 FUNÇÕES *BASELINES*

Foram ainda desenvolvidas quatro funções aleatórias para analisar o comportamento de uma sumarização, com escolha de sentenças que não são escolhidas por algum sumarizador, mas aleatoriamente. Essas funções são denominadas: FA1_S_*Stopword*s, FA2_S_*Stopwords*, FA1_C_*Stopwords* e FA2_C_ *Stopwords*, todas aleatórias, na escolha das sentenças do texto-fonte, tanto para textos em português quanto em inglês. As funções FA1_S_ *Stopwords* e FA2_S_ *Stopwords* retiram as *stopwords* do texto antes da escolha, para que haja uma redução do

número de palavras. Já as FA1_C_*Stopword*s e FA2_C_ *Stopwords* não retiram as *stopwords*. Essas funções adotam ainda a variação de dois métodos, em função do percentual(%) de compressão escolhido. As funções FA2 terminam a sumarização quando atingem o percentual de compressão, independente de onde estejam na frase. Já as funções FA1 vão até o final da frase, não respeitando o percentual de compressão estabelecido. Como são funções aleatórias, para cada uma delas foram realizados três vezes o processo de sumarização. Esse número foi escolhido com base em observações de testes realizados nos *corpora* onde não foi verificado um aumento significativo na média, que justifica a realização de um número maior de sumarização, usando qualquer uma das funções aleatórias.

CAPÍTULO 5 – RESULTADOS

No capítulo 5 serão mostrados os experimentos, divididos em duas partes para melhor compreensão da comprovação da hipótese, já apresentada, e serão detalhados nas seções 5.1 e 5.2. Para entendimento da primeira parte do experimento serão apresentadas as análises específicas dos resultados obtidos com as medidas harmônicas *F-Measure* (métrica externa) e coeficiente de Silhouette (métrica interna), com os graus de compressão usados ao longo do experimento, ou seja, 50%, 70%, 80% e 90%. A explicação detalhada, de como se procederam a essas comparações serão detalhadas na subseção 5.1.1, referente à métrica externa e, na subseção 5.1.3, referente à interna, e serão mostrados os respectivos gráficos. Para encerrar a análise dos resultados do primeiro experimento, os gráficos das médias acumuladas serão apresentados.

No segundo experimento serão explicadas e apresentadas as análises específicas dos resultados obtidos com as medidas harmônicas *F-Measure* (métrica externa) e coeficiente de Silhouette (métrica interna), com os graus de compressão usados ao longo do experimento, ou seja, 50%, 70%, 80% e 90%. A explicação detalhada, de como foram realizadas essas comparações será mostrada na subseção 5.2.1, referente à métrica externa, e na subseção 5.2.2, referente à interna e serão analisados os gráficos. Na seção 5.3 será apresentada a comprovação da hipótese; na 5.4. serão expostas as análises dos testes estatísticos, baseados nas tabelas do Apêndice E, geradas pelos softwares estatísticos; na 5.5 serão mostrados os trabalhos correlatos, e na 5.6, a discussão dos resultados.

5.1 PRIMEIRA PARTE DOS EXPERIMENTOS

Na primeira parte foram realizados testes no modelo Cassiopeia, usando os textos-fonte (sem sumarização)[12] e os textos sumarizados[13], obtidos através dos sumarizadores escolhidos e definidos na seção 4.2. O conjunto de cem textos de cada sumarizador e dos textos-fonte (sem sumarização) foi submetido ao modelo Cassiopeia, separadamente, que executou o processo de agrupamento e reagrupamento, com cada conjunto de cem textos individualmente, cem vezes para cada conjunto, obtendo assim uma média aritmética em cada métrica. Esses agrupamentos foram mensurados através das métricas externas ou supervisionadas (*Recall*, *Precision* e *F-Measure*) e

[12] Textos-Fonte (sem sumarização) não sofrem qualquer compressão. Eles são agrupados pelo modelo Cassiopeia gerando agrupamentos de textos que são avaliados pelas métricas externas e internas. Eles são parâmetros comparativos com os textos sumarizados.

[13] Nos textos sumarizados foram usados algoritmos de sumarização e explicados na seção 4.2. Eles são usados no pré-processamento do modelo Cassiopeia para sumarizar os textos-fonte. Foram usadas as compressões de 50%, 70%, 80% e 90%. Os resultados são textos sumarizados, que serão agrupados pelo modelo Cassiopeia e avaliados por métricas externas e internas.

internas ou não supervisionadas (Coesão, Acoplamento e Coeficiente Silhouette*)*, explicadas nas subseções 2.2.5.1 e 2.2.5.2. Foi gerada uma soma acumulada dessas cem médias aritméticas que se encontram nos gráficos dos experimentos, para cada uma das métricas. Esse processo ocorreu separadamente, para cada um dos percentuais de compressão, ou seja, 50%, 70%, 80% e 90% e para cada um dos idiomas, português e inglês.

5.1.1 MÉTRICA EXTERNA: RECALL, PRECISION E F-MEASURE

Para organizar a apresentação dos resultados, será colocada nesta seção apenas a medida harmônica da métrica externa, que é o *F-Measure*. A importância da medida é dada pela maior proximidade dos agrupamentos do valor um, e quanto mais próximos, melhores são os resultados. Os resultados do *Recall* e *Precision* serão colocados no Apêndice A, juntamente com seus comentários e todas as análises. Nesta seção, os textos-fonte (sem sumarização) serão identificados através da linha pontilhada, e as outras linhas representando os demais algoritmos de sumarização usados no experimento. O processo de agrupamento e reagrupamento para esse experimento, no modelo Cassiopeia,, ocorre com cem passos.

Os resultados apresentados foram tirados das médias aritméticas acumuladas de cada algoritmo de sumarização e dos textos-fonte, ao longo dos cem passos. Foi necessário gerar uma soma acumulada dessas cem médias aritméticas, mostradas nos gráficos dos experimentos. Para comparar os resultados, foram usados os resultados dos textos-fonte como referência, que serão apresentados com as compressões de 50%, 70%, 80% e 90%. Acredita-se que com o uso da variação da compressão, possa ser feita uma análise mais uniforme da perda da informatividade nos textos, mediante o aumento do grau de compressão dentro do agrupamento. Essa análise é fundamental, para a qualidade dos agrupamentos gerados, já que pode ser avaliado o grau de compressão que é vantajoso para o agrupamento. Os textos escolhidos para o trabalho pertencem aos domínios jornalístico, jurídico e médico, nos idiomas português e inglês. Como já foi explicado e justificado, no capítulo 4, não existem textos no domínio jurídico, no idioma inglês.

5.1.1.1 USO DA COMPRESSÃO DE 50% NO IDIOMA PORTUGUÊS

Os resultados da Figura 10 demonstram que no domínio jurídico, cinco algoritmos de sumarização aumentaram os valores, em comparação ao *F-Measure* dos textos-fonte, exceto os sumarizadores *Gist Average Keyword* e *Gist Intrasentença*. No domínio jornalístico, todos os algoritmos tiveram seus valores de *F-Measure* inferiores aos dos textos-fonte. O domínio médico teve dois algoritmos que alcançaram valores de *F-Measure,* superiores aos dos textos-fonte, que são as duas funções FA1.

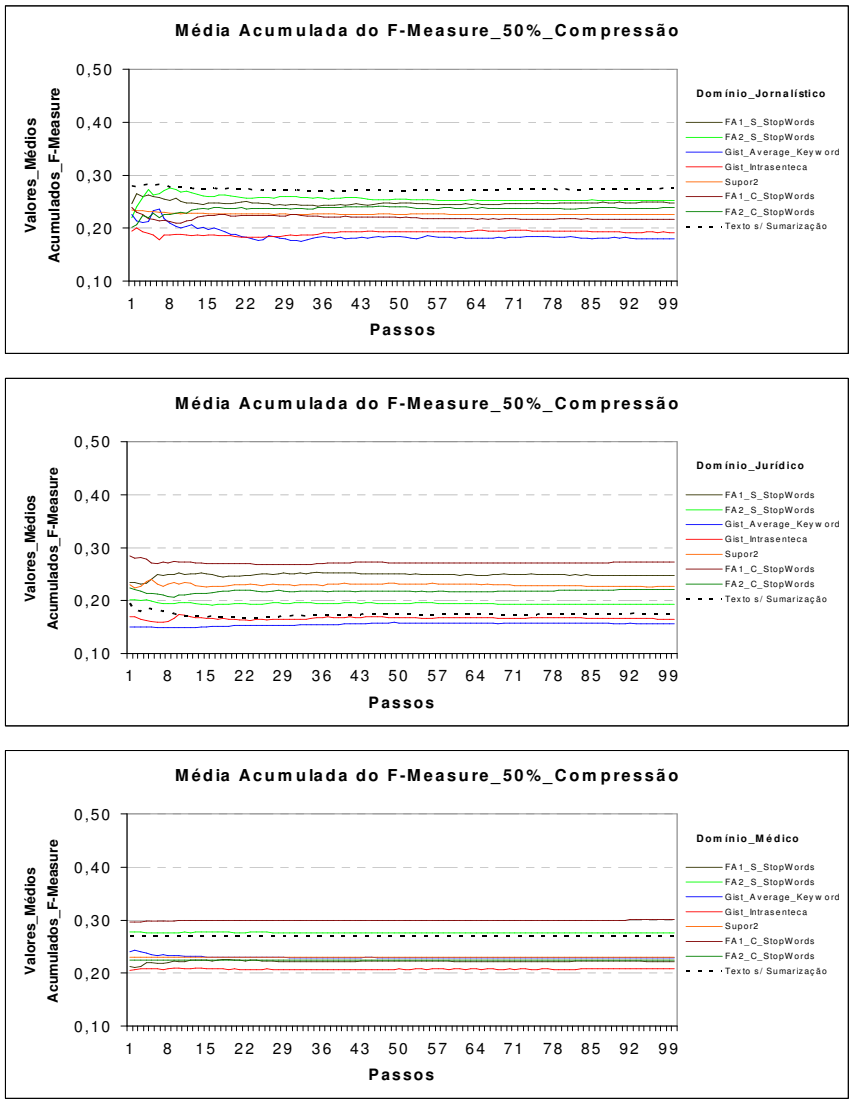


Figura 10: Resultados obtidos pelo modelo Cassiopeia, usando a medida *F-Measure* com 50% compressão no idioma português.

5.1.1.2 USO DA COMPRESSÃO DE 70% NO IDIOMA PORTUGUÊS

Os resultados da Figura 11 demonstram que, no domínio jurídico, cinco algoritmos de sumarização aumentaram os valores, em comparação ao *F-Measure* dos textos-fonte, exceto dois sumarizadores, o *SuPor e* o *Gist Intrasentença*. No domínio jornalístico, apenas a função FA1, sem *stopword,* teve seu valor de *F-Measure* superior aos dos textos-fonte. O domínio médico apresentou dois algoritmos de sumarização que obtiveram valores *F-Measure* superiores aos dos textos-fonte, o *Gist Average Keyword* e a FA2 com *stopwords.*

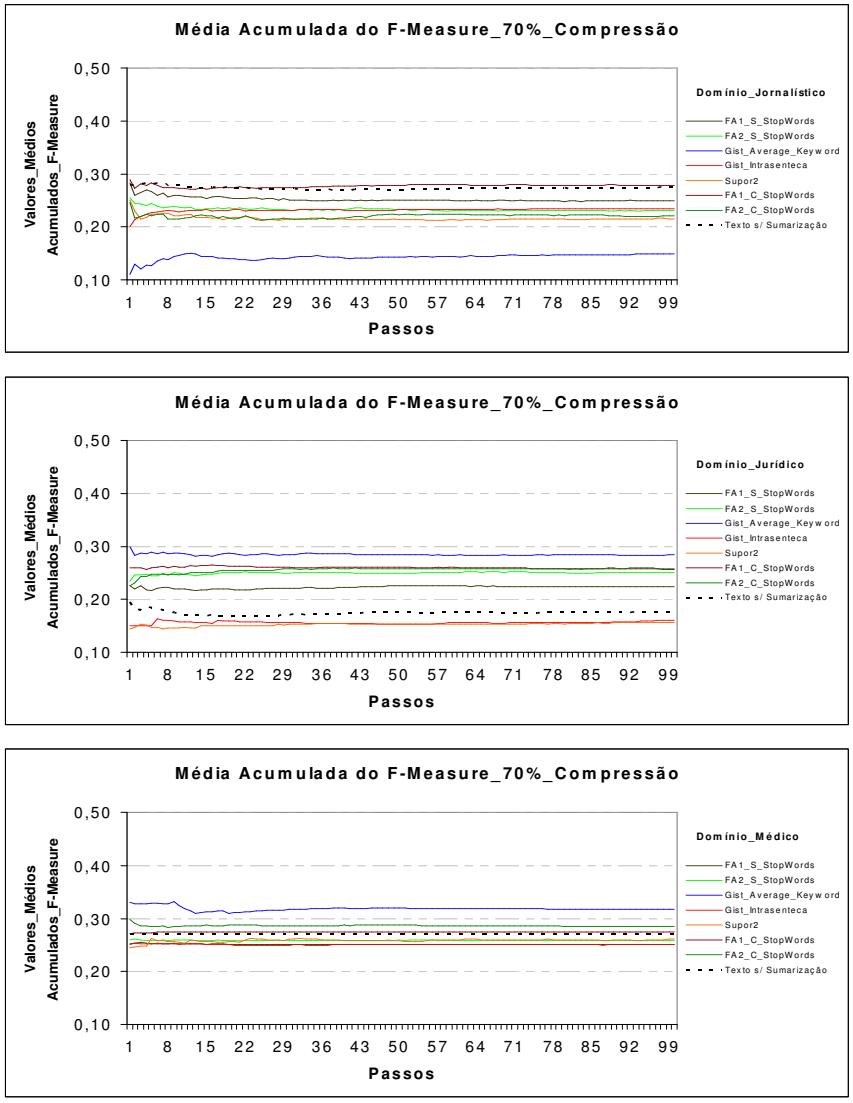


Figura 11: Resultados obtidos pelo modelo Cassiopeia, usando a medida *F-Measure* com 70% compressão no idioma português.

5.1.1.3 USO DA COMPRESSÃO DE 80% NO IDIOMA PORTUGUÊS

Os resultados da Figura 12 demonstram que, no domínio jurídico, todos os algoritmos de sumarização aumentaram os valores *de F-Measure* em comparação aos dos textos-fonte. No domínio jornalístico, nenhum dos algoritmos de sumarização alcançou valores superiores aos da medida *F-Measure* em comparação aos dos textos-fonte. No domínio médico, um algoritmo de sumarização alcançou o valor superior de *F-Measure* dos textos-fonte a função FA1, sem *stopwords*.

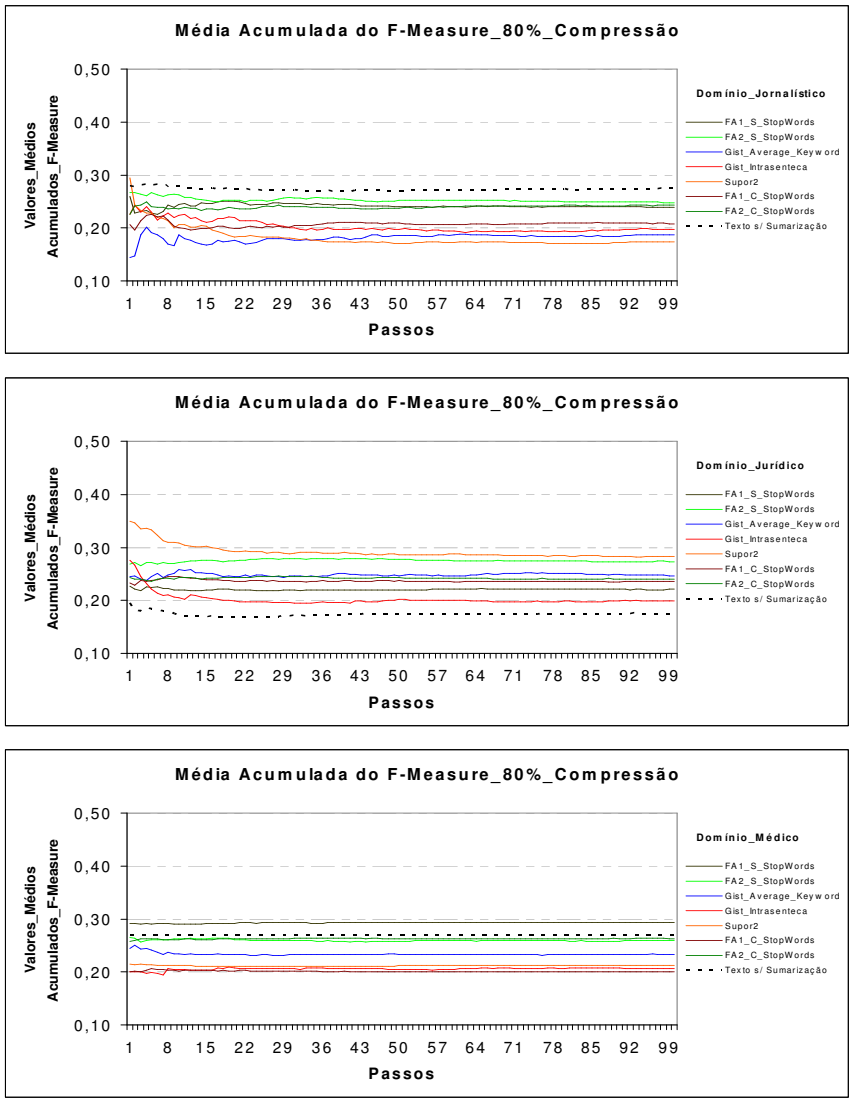


Figura 12: Resultados obtidos pelo modelo Cassiopeia, usando a medida *F-Measure* com 80% compressão no idioma português.

5.1.1.4 USO DA COMPRESSÃO DE 90% NO IDIOMA PORTUGUÊS

Os resultados da Figura 13 demonstram que, no domínio jurídico, todos os algoritmos aumentaram o valor da medida *F-Measure* em comparação aos dos textos-fonte. Para o domínio jornalístico e médico, na medida *F-Measure*, nenhum dos algoritmos teve seus valores superiores aos dos textos-fonte.

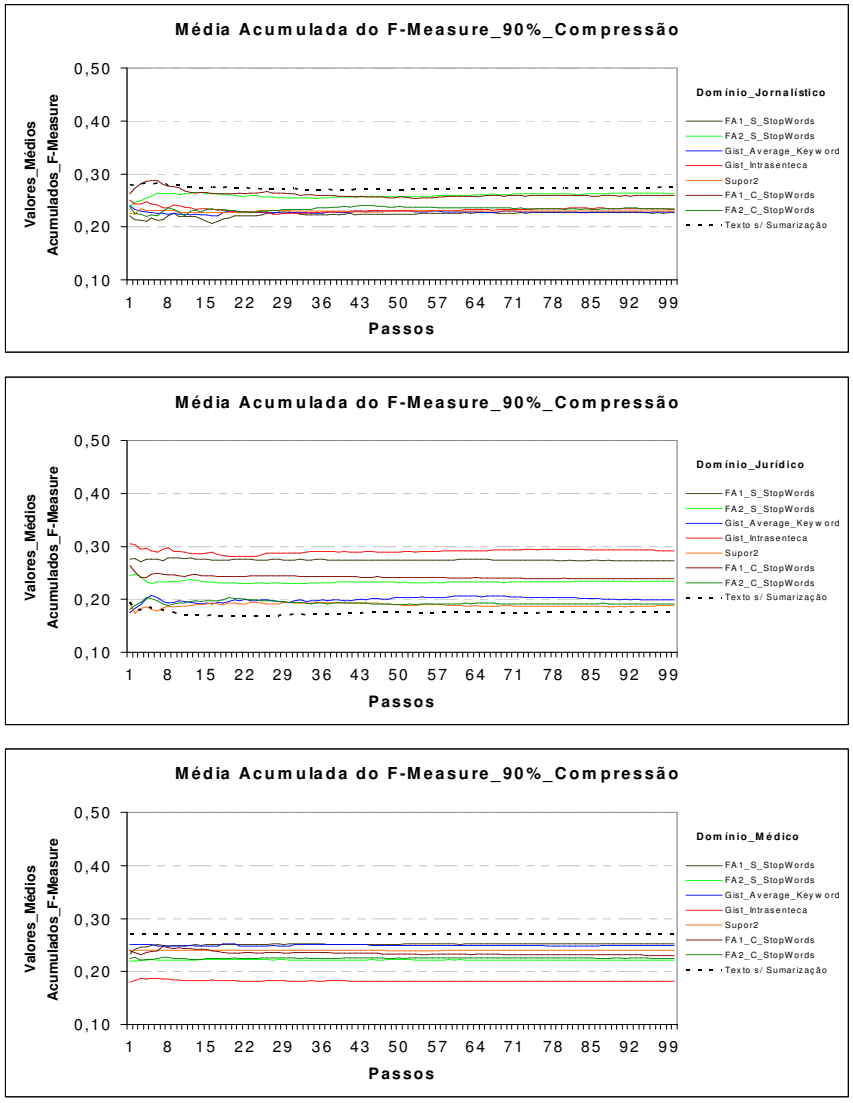


Figura 13: Resultados obtidos pelo modelo Cassiopeia, usando a medida *F-Measure* com 90% compressão no idioma português.

5.1.1.5 USO DA COMPRESSÃO DE 50% NO IDIOMA INGLÊS

Os resultados da Figura 14 demonstram que, no domínio jornalístico, seis algoritmos aumentaram os valores em comparação ao valor de *F-Measure* dos textos-fonte, exceto um algoritmo de sumarização, o *SweSum*. No domínio médico, apenas um algoritmo teve seu valor de *F-Measure* superior ao valor dos textos-fonte, o algoritmo *Intellexer Summarizer*.

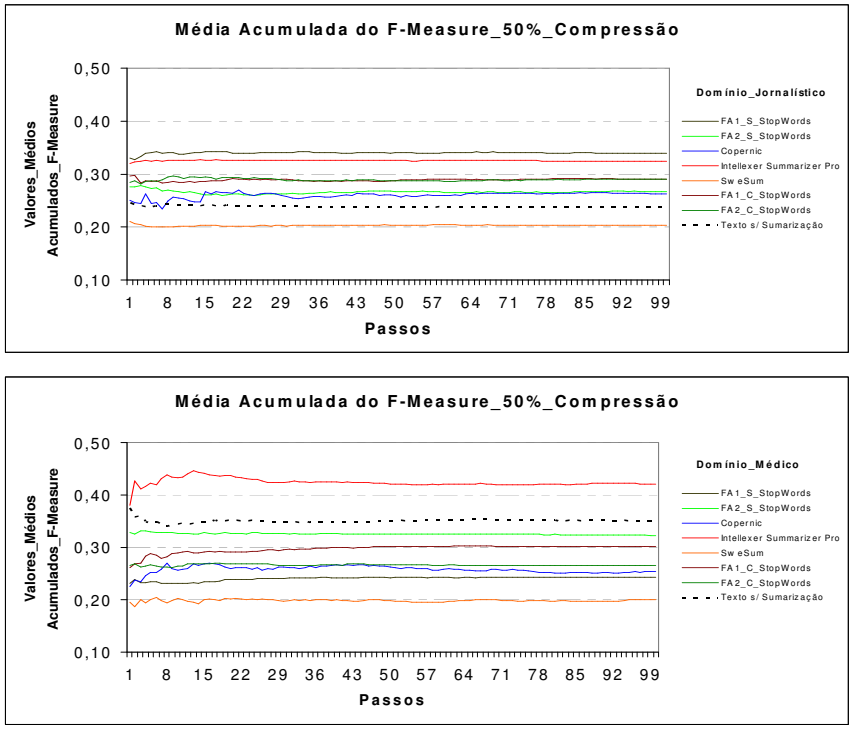


Figura 14: Resultados obtidos pelo modelo Cassiopeia, usando a medida *F-Measure* com 50% compressão no idioma inglês.

5.1.1.6 USO DA COMPRESSÃO DE 70% NO IDIOMA INGLÊS

Os resultados da Figura 15 demonstram que, no domínio jornalístico, dois algoritmos tiveram seus valores de *F-Measure* abaixo dos valores dos textos-fonte, os algoritmos *Intellexer Summarizer* e *SweSum*. No domínio médico, apenas um algoritmo teve seu valor de *F-Measure* superior ao valor dos textos-fonte, o algoritmo *Intellexer Summarizer*.

.

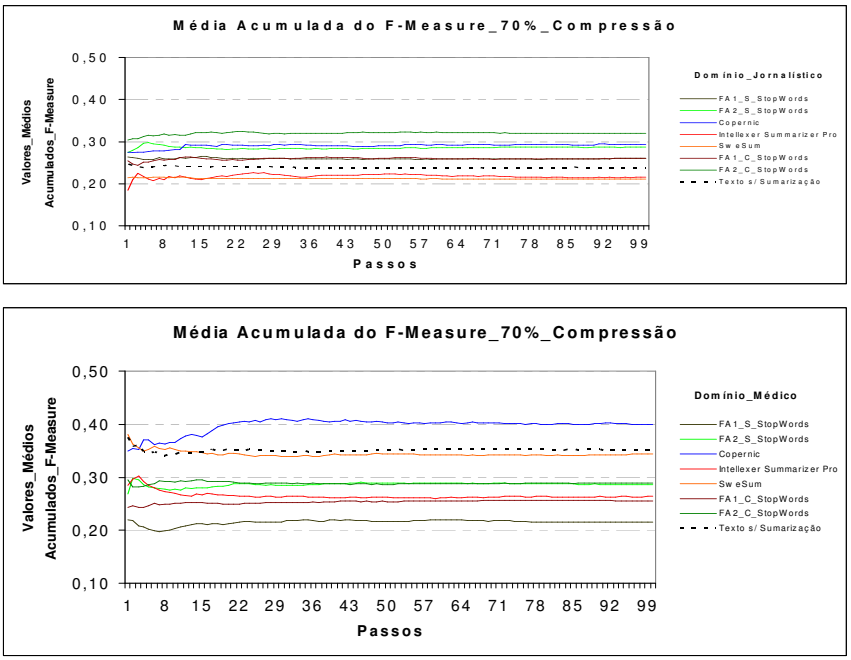


Figura 15: Resultados obtidos pelo modelo Cassiopeia, usando a medida *F-Measure* com 70% compressão no idioma inglês.

5.1.1.7 USO DA COMPRESSÃO DE 80% NO IDIOMA INGLÊS

Os resultados da Figura 17 demonstram que no domínio jornalístico três algoritmos tiveram seus valores de *F-Measure* abaixo dos valores dos textos-fontes, foram os algoritmos FA2 com *stopwords, Intellexer Summarizer e SweSum.* No domínio médico apenas um algoritmo teve seu valor de *F-Measure* superior ao valor dos textos fontes, foi o algoritmo *SweSum.*

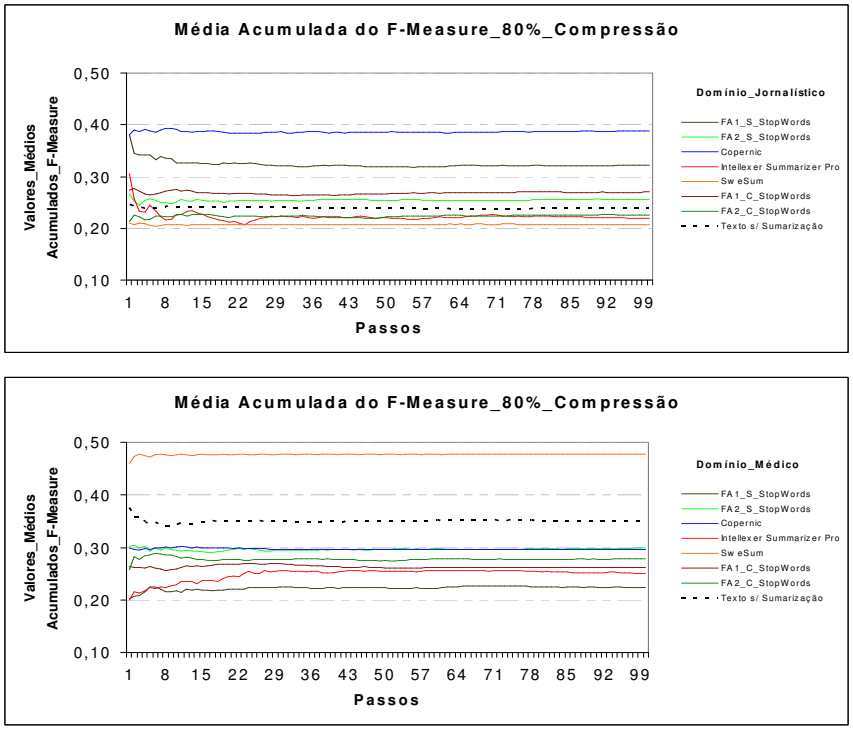


Figura 16: Resultados obtidos pelo modelo Cassiopeia, usando a medida *F-Measure* com 80% compressão no idioma inglês.

5.1.1.8 USO DA COMPRESSÃO DE 90% NO IDIOMA INGLÊS

Os resultados da Figura 17 demonstram que, no domínio jornalístico, todos os algoritmos aumentaram os valores em comparação ao valor de *F-Measure* dos textos-fonte. No domínio médico, nenhum algoritmo teve seu valor de *F-Measure* superior ao valor dos textos-fonte.

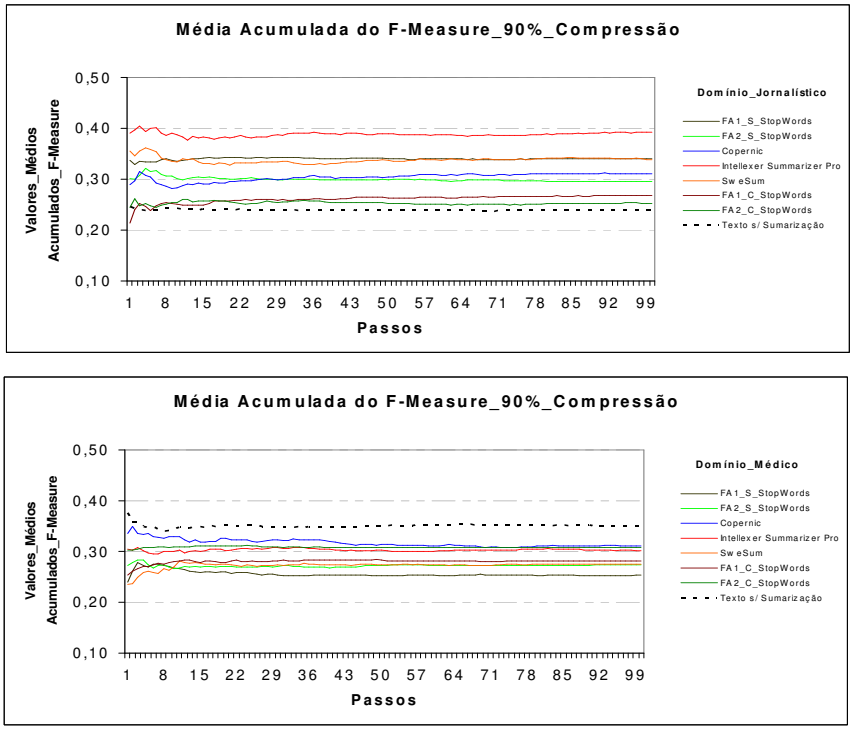


Figura 17: Resultados obtidos pelo modelo Cassiopeia, usando a medida *F-Measure* com 90% de compressão no idioma inglês.

5.1.2 MÉTRICA INTERNA: COESÃO, ACOPLAMENTO COEFICIENTE SILHOUETTE

Para organizar a apresentação dos resultados, será colocada, nesta seção, a medida do Coeficiente Silhoutte da métrica interna. A importância dessa medida é que quanto mais próximo do valor um, melhores são os resultados dos agrupamentos de textos. Os resultados da Coesão e Acoplamento serão colocados no Apêndice B, com seus comentários e análises. Nesta seção, os textos-fonte (sem sumarização) serão identificados através da linha pontilhada, e as outras linhas representarão os demais algoritmos de sumarização usados no experimento. O processo de agrupamento e reagrupamento, no modelo Cassiopeia para esse experimento, ocorreu com cem passos.

Os resultados apresentados no trabalho mostrarão as médias aritméticas acumuladas de cada algoritmo de sumarização e dos textos-fonte, ao longo dos cem passos. Foi necessário gerar uma soma acumulada dessas cem médias aritméticas, que serão apresentadas nos gráficos dos experimentos. Para comparar os resultados, foram usados os dos textos-fonte como referência, e são apresentados com as compressões de 50%, 70%, 80% e 90%. Acredita-se que com uso da variação da compressão, possa ser feita uma análise mais uniforme da perda da informatividade

nos textos, mediante o aumento do grau de compressão dentro do agrupamento de textos. Essa análise é fundamental para a qualidade dos agrupamentos gerados, já que pode ser avaliado qual o grau de compressão vantajoso para o agrupamento. Os textos usados no trabalho pertencem aos domínios jornalístico, jurídico e médico, nos idiomas português e inglês. Como já foi explicado e justificado no capítulo 4, não existem textos no domínio jurídico, no idioma inglês.

5.1.2.1 USO DA COMPRESSÃO DE 50% NO IDIOMA PORTUGUÊS

Os resultados da Figura 18 demonstram que, no domínio jurídico, todos os algoritmos de sumarização tiveram seus valores maiores, em comparação ao Coeficiente Silhouette dos textos-fonte. No domínio médico e no domínio jornalístico, na medida Coeficiente Silhouette, nenhum dos algoritmos teve valores maiores, em comparação ao Coeficiente Silhouette dos textos-fonte.

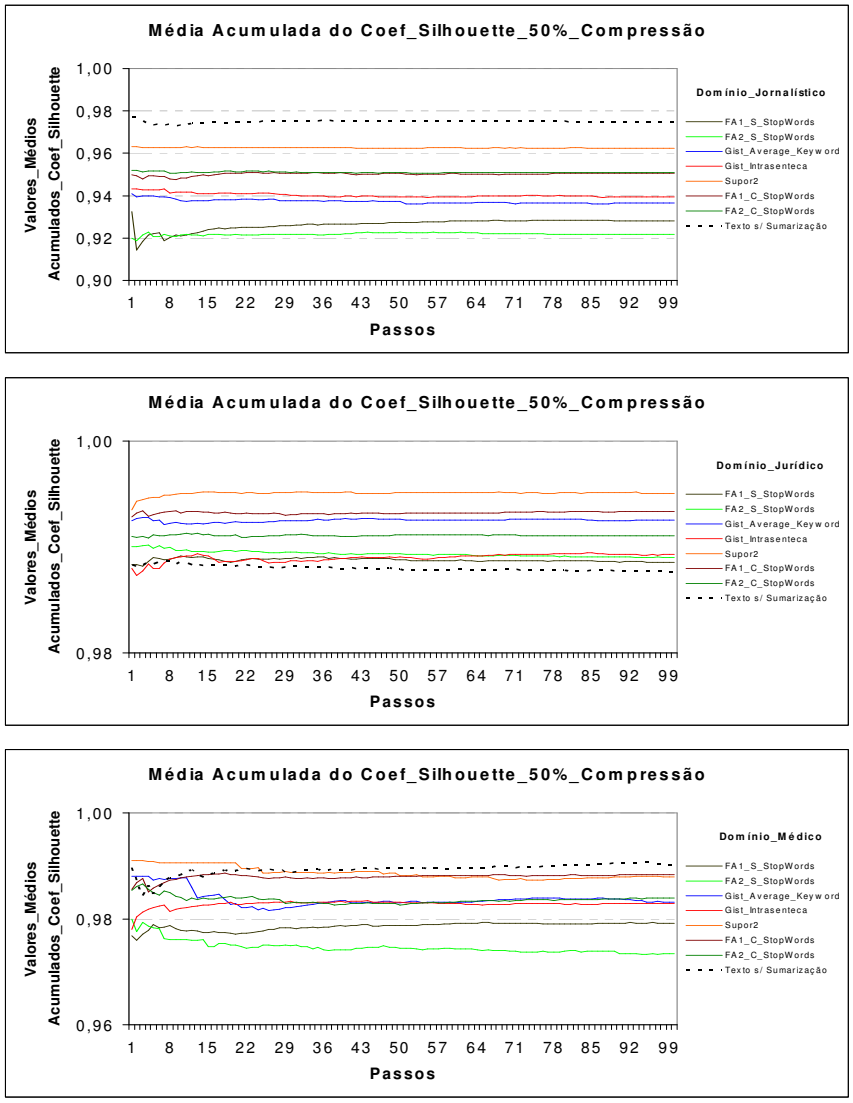


Figura 18: Resultados obtidos pelo modelo Cassiopeia, usando a medida Coeficiente Silhouette com 50% compressão no idioma português.

5.1.2.2 USO DA COMPRESSÃO DE 70% NO IDIOMA PORTUGUÊS

Os resultados da Figura 19 demonstram que, no domínio jurídico, quatro dos algoritmos aumentaram os seus valores em comparação ao valor do Coeficiente Silhouette dos textos, com exceção do algoritmo da literatura *Gist Intrasentença,* e as funções FA1 e FA2, sem *stopwords*. Nos domínios médico e jornalístico, na medida Coeficiente Silhouette, nenhum dos algoritmos ficou com valores acima dos textos-fonte.

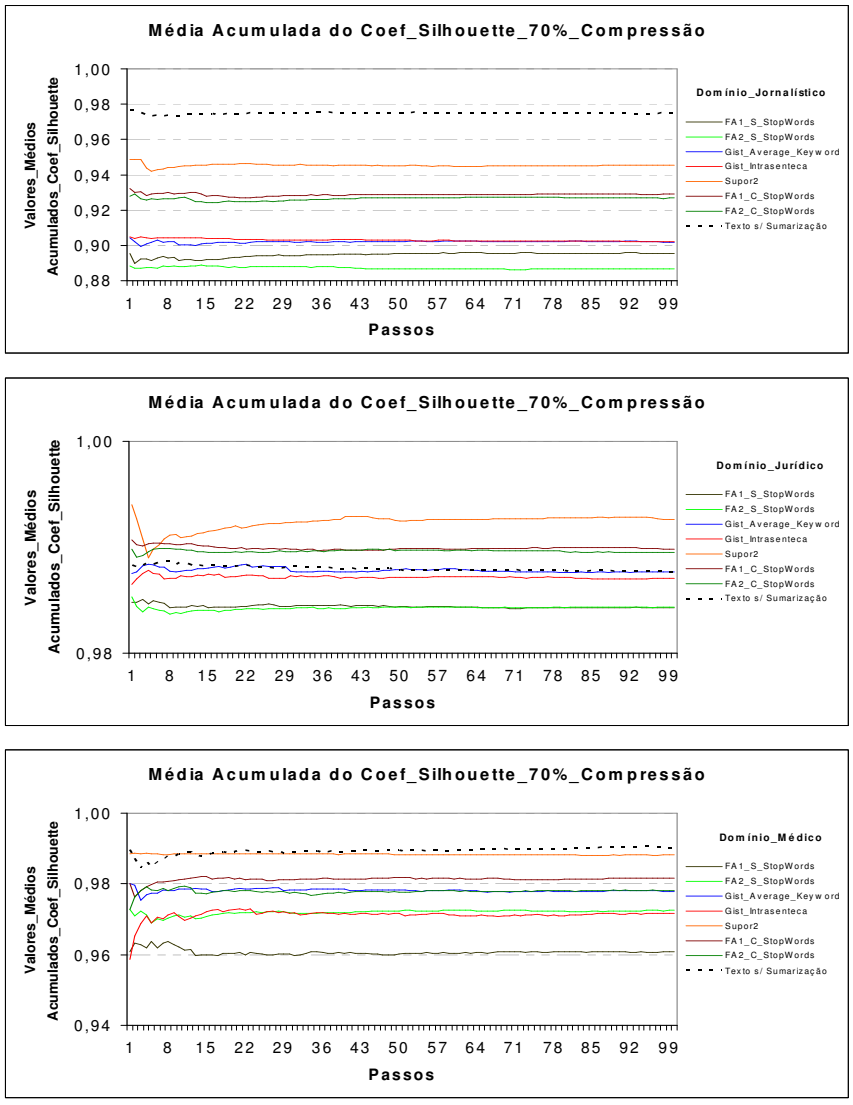


Figura 19: Resultados obtidos pelo modelo Cassiopeia, usando a medida Coeficiente Silhouette com 70% compressão no idioma português.

5.1.2.3 USO DA COMPRESSÃO DE 80% NO IDIOMA PORTUGUÊS

Os resultados da Figura 20 demonstram que o domínio jurídico teve um algoritmo, o *SuPor,* que ficou com seu valor da medida Coeficiente Silhouette acima dos textos-fonte. Nos domínios médico e jornalístico, na medida Coeficiente Silhouette, nenhum dos algoritmos ficou com seus valores acima dos textos-fonte.

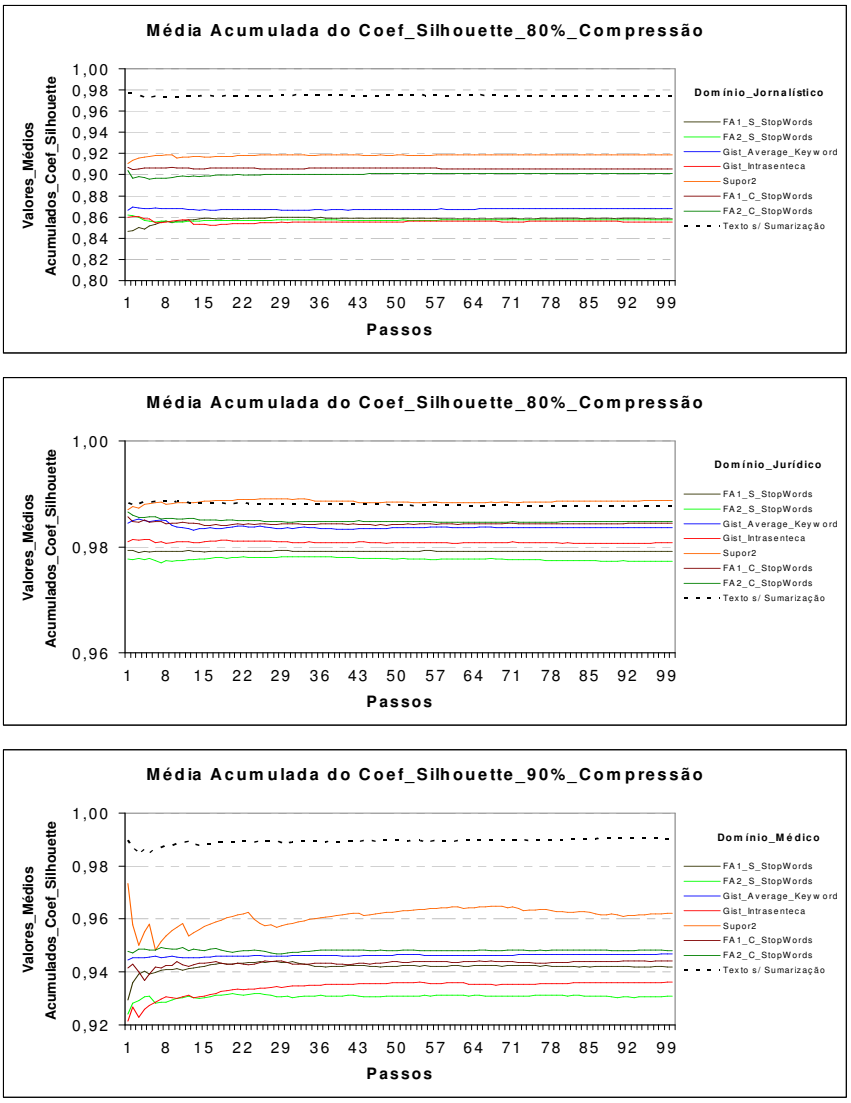


Figura 20: Resultados obtidos pelo modelo Cassiopeia, usando a medida Coeficiente Silhouette com 80% compressão no idioma português.

5.1.2.4 USO DA COMPRESSÃO DE 90% NO IDIOMA PORTUGUÊS

Os resultados da Figura 21 demonstram que em todos os domínios nenhum dos algoritmos ficou com seus valores na medida Coeficiente Silhouette acima dos textos-fonte.

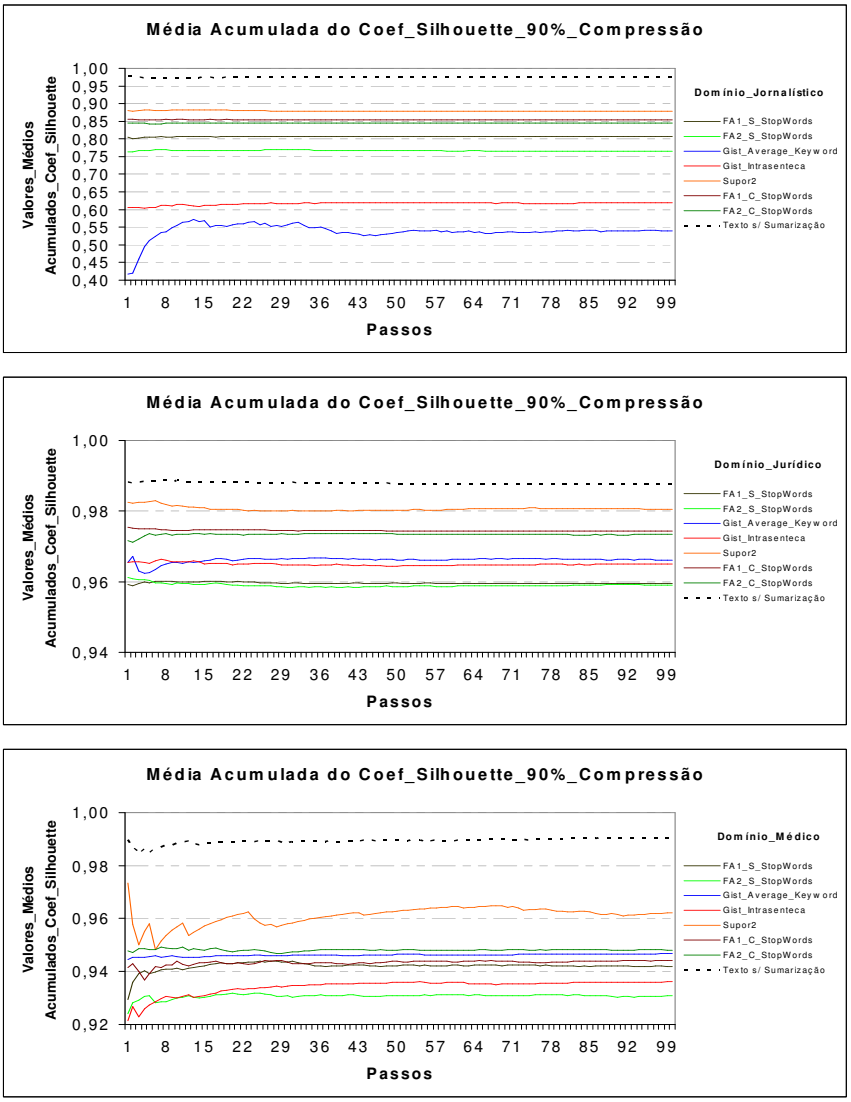


Figura 21: Resultados obtidos pelo modelo Cassiopeia, usando a medida Coeficiente Silhouette com 90% compressão no idioma português.

5.1.2.5 USO DA COMPRESSÃO DE 50% NO IDIOMA INGLÊS

Os resultados da Figura 22 demonstram, que no domínio jornalístico, todos os algoritmos tiveram os seus valores maiores, em comparação ao Coeficiente Silhouette dos textos-fonte. No domínio médico, na medida Coeficiente Silhouette, seis dos algoritmos tiveram os seus valores maiores, em comparação com o valor de Coeficiente Silhouette dos textos-fonte, com exceção do da função aleatória FA2 sem *stopwords*.

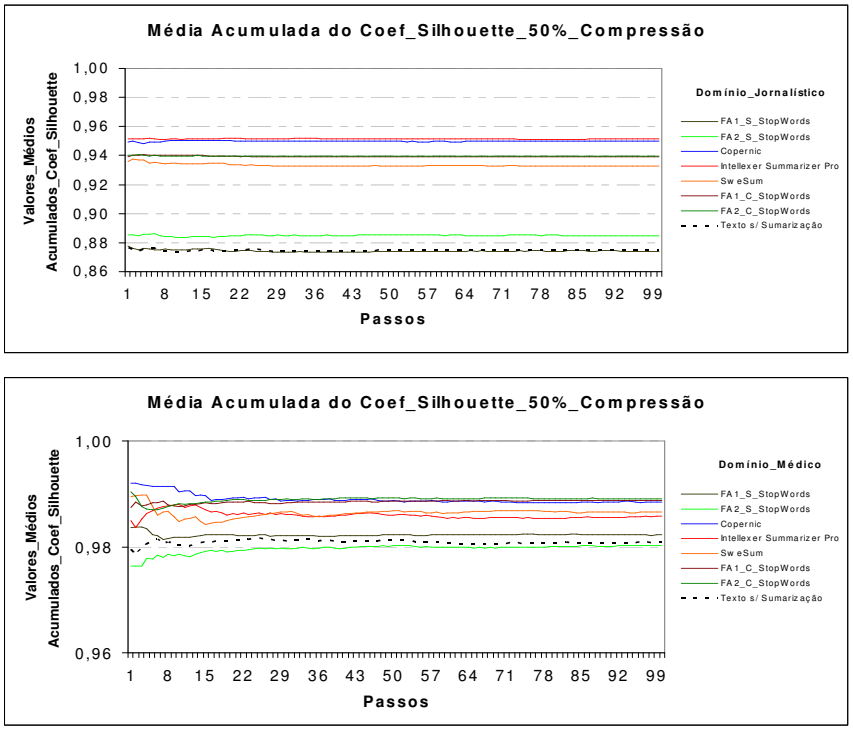


Figura 22: Resultados obtidos pelo modelo Cassiopeia, usando a medida Coeficiente Silhouette com 50% compressão no idioma Inglês.

5.1.2.6 USO DA COMPRESSÃO DE 70% NO IDIOMA INGLÊS

Os resultados da Figura 23 demonstram que, no domínio jornalístico, cinco dos algoritmos tiveram valores maiores em comparação ao Coeficiente Silhouette dos textos-fonte. A exceção são as funções FA1 e FA2 sem *stopwords*. Esses mesmos resultados foram encontrados no domínio médico.

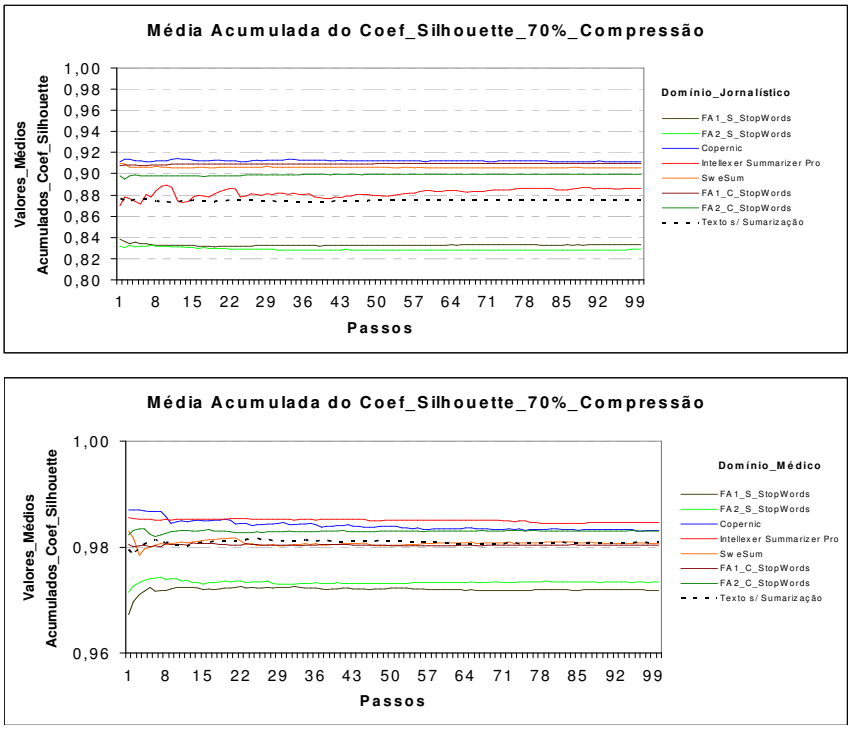


Figura 23: Resultados obtidos pelo modelo Cassiopeia, usando a medida Coeficiente Silhouette com 70% compressão no idioma Inglês.

5.1.2.7 USO DA COMPRESSÃO DE 80% NO IDIOMA INGLÊS

Os resultados da Figura 24 demonstram que, no domínio jornalístico, dois algoritmos tiveram seus valores Coeficiente Silhouette acima dos textos-fonte, os algoritmos profissionais *Intellexer Pro* e *Copernic*. No domínio médico nenhum dos algoritmos ficou com seus valores na medida Coeficiente Silhouette acima dos textos-fonte.

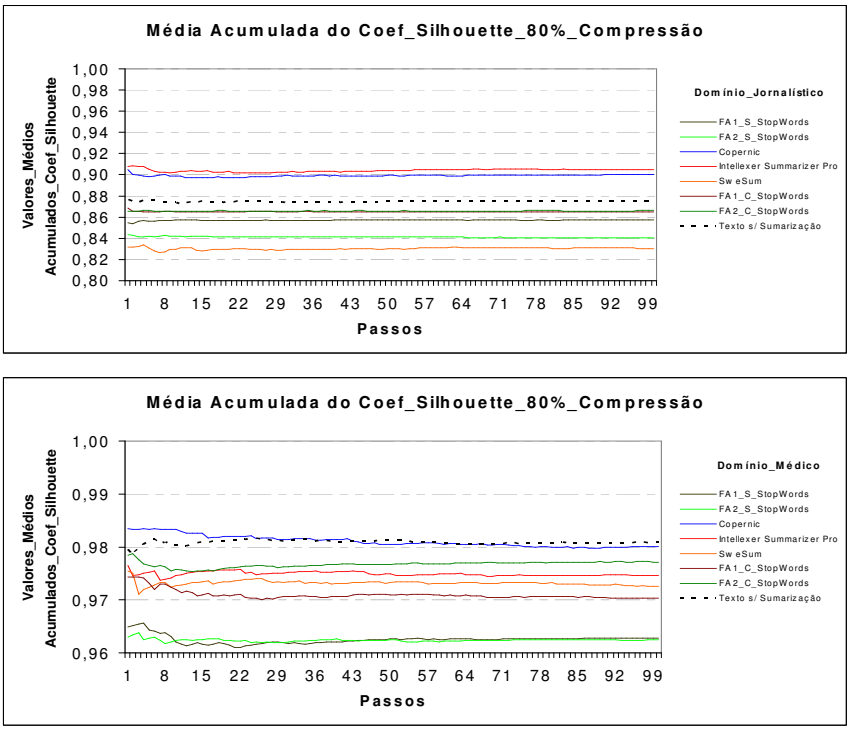


Figura 24: Resultados obtidos pelo modelo Cassiopeia, usando a medida Coeficiente *Silhouette* com 80% compressão no idioma Inglês.

5.1.2.8 USO DA COMPRESSÃO DE 90% NO IDIOMA INGLÊS

Os resultados da Figura 26 demonstram que, no domínio jornalístico, existe apenas um algoritmo, cujo valor está acima da medida do Coeficiente Silhouette dos textos-fonte, o algoritmo da literatura *SweSum.* Já no domínio médico, nenhum dos algoritmos ficou com seus valores na medida Coeficiente Silhouette acima dos textos-fonte.

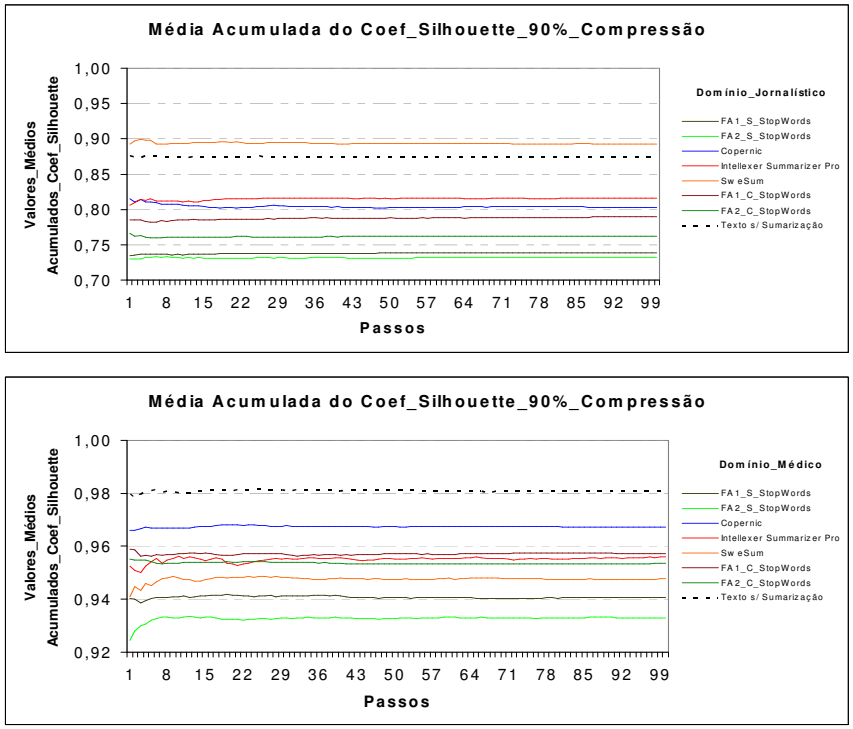
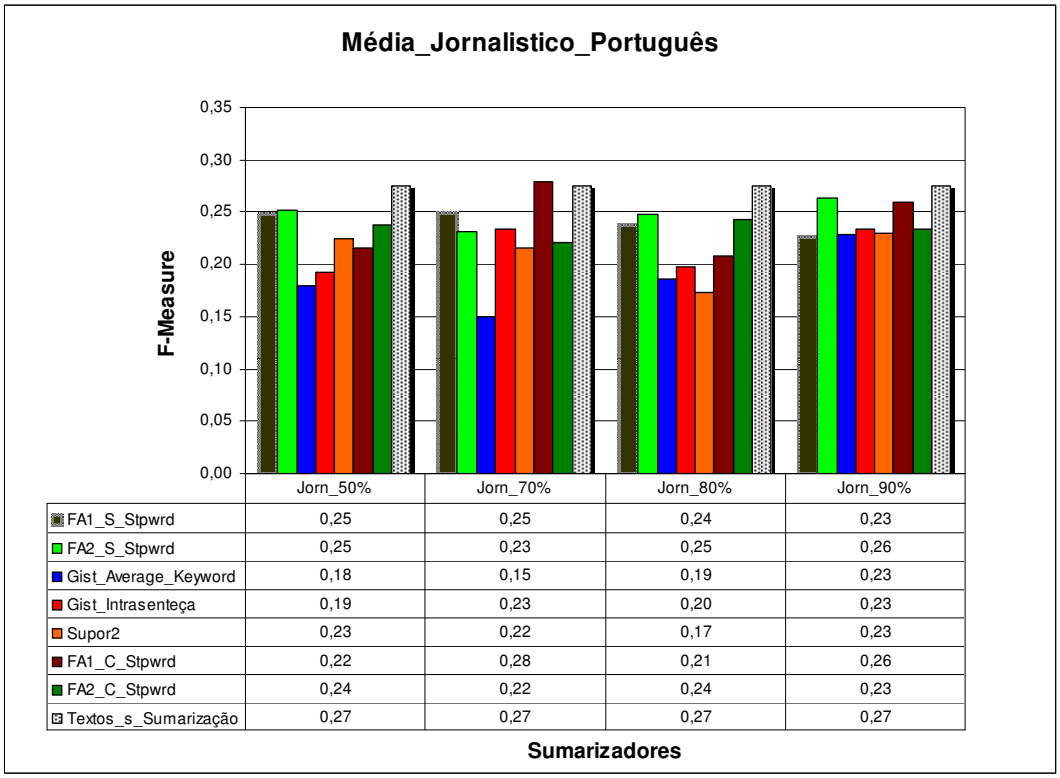


Figura 25: Resultados obtidos pelo modelo Cassiopeia, usando a medida Coeficiente Silhouette com 90% compressão no idioma Inglês.

Com objetivo de sintetizar os resultados até então apresentados, e agrupar em gráficos os resultados das medidas *FMeasure* e do Coeficiente Silhouette, serão analisadas as Figuras 26, 27, 28, 29, 30, 31, 32, 33, 34 e 35, que mostram os resultados finais das médias acumuladas, ou seja, o último valor obtido em todas as compressões, de 50%, 70%, 80% e 90%, no idioma português, nos domínios jornalístico, médico e jurídico e no idioma inglês, nos domínios jornalístico e médico.

A Figura 26 indica o resultado do aferimento dos agrupamentos de textos obtidos pelo modelo Cassiopeia, usando a medida *F-Measure*, no idioma português, no domínio jornalístico. São apresentadas também as médias acumuladas de todas as compressões (50%, 70%, 80% e 90%) e os sete algoritmos de sumarização e os textos-fonte.

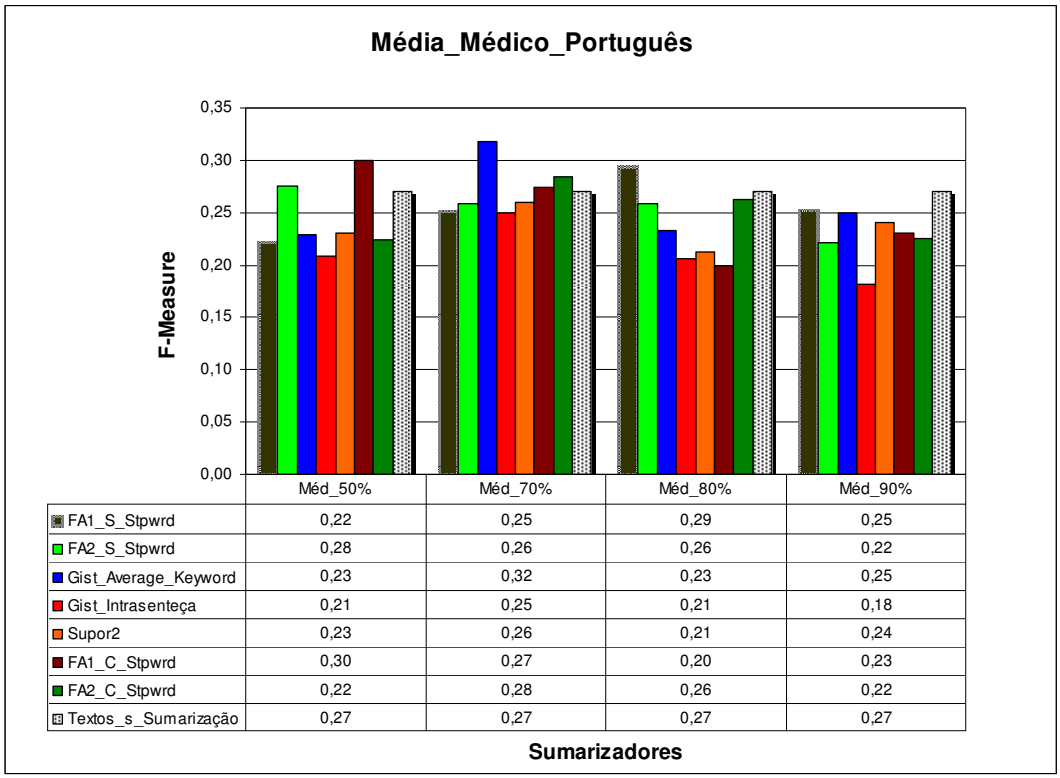
A Figura 26 mostra que os algoritmos de sumarização, em sua grande maioria, não conseguiram aumentar o valor de *F-Measure* nos agrupamentos obtidos pelo modelo Cassiopeia, em comparação com os agrupamentos obtidos com os textos-fonte. A única exceção foi a função FA1_com_*Stopword,* com a compressão de 70%, tendo aumentado o valor dos agrupamentos na medida *FMeasure*.



	Jorn_50%	Jorn_70%	Jorn_80%	Jorn_90%
FA1_S_Stpwrd	0,25	0,25	0,24	0,23
FA2_S_Stpwrd	0,25	0,23	0,25	0,26
Gist_Average_Keyword	0,18	0,15	0,19	0,23
Gist_Intrasenteça	0,19	0,23	0,20	0,23
Supor2	0,23	0,22	0,17	0,23
FA1_C_Stpwrd	0,22	0,28	0,21	0,26
FA2_C_Stpwrd	0,24	0,22	0,24	0,23
Textos_s_Sumarização	0,27	0,27	0,27	0,27

Figura 26: Resultados das médias acumuladas obtidos pelo modelo Cassiopeia, usando a medida *FMeasure* com 50%, 70%, 80% e 90% de compressão no idioma português, no domínio jornalístico.

No domínio médico, Figura 27, com 50% de compressão, dois algoritmos melhoraram o desempenho do agrupamento gerado pelo modelo Cassiopeia, na medida *FMeasure,* em relação aos agrupamentos dos textos-fonte. Com 70%, apenas um, e com 80% e 90% nenhum dos algoritmos de sumarização conseguiu aumentar os valores de *F-Measure* dos agrupamentos obtidos pelo modelo Cassiopeia.

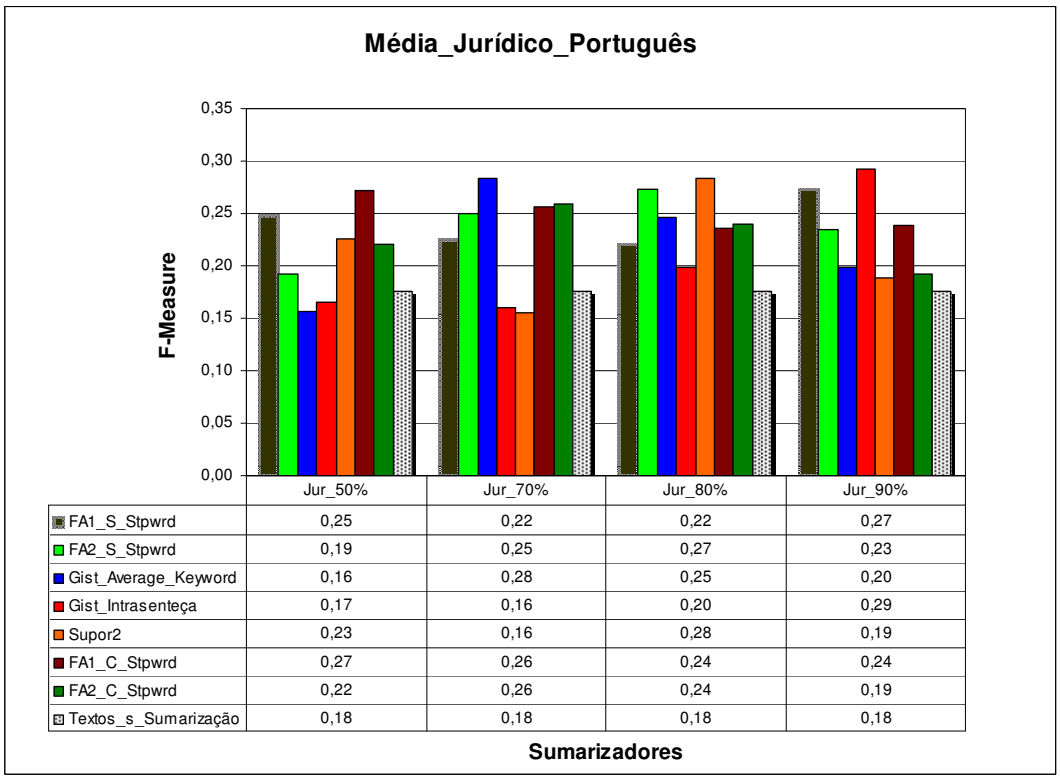


	Méd_50%	Méd_70%	Méd_80%	Méd_90%
FA1_S_Stpwrd	0,22	0,25	0,29	0,25
FA2_S_Stpwrd	0,28	0,26	0,26	0,22
Gist_Average_Keyword	0,23	0,32	0,23	0,25
Gist_Intrasenteça	0,21	0,25	0,21	0,18
Supor2	0,23	0,26	0,21	0,24
FA1_C_Stpwrd	0,30	0,27	0,20	0,23
FA2_C_Stpwrd	0,22	0,28	0,26	0,22
Textos_s_Sumarização	0,27	0,27	0,27	0,27

Figura 27: Resultados das médias acumuladas obtidos pelo modelo Cassiopeia, usando a medida *FMeasure* com 50%, 70%, 80% e 90% de compressão no idioma português, no domínio médico.

O melhor desempenho obtido pelo modelo Cassiopeia, usando a medida *FMeasure*, foi no domínio jurídico, no idioma português. Houve melhora considerável de todos os algoritmos de sumarização, comparados com os dos textos-fonte.

Para a medida *F-Measure*, na Figura 28, com 50 % de compressão, apenas dois algoritmos não tiveram os agrupamentos gerados pelo modelo Cassiopeia com valores de *FMeasure* aumentado. Com 70% são dois algoritmos, e com 80% e 90% todos os algoritmos de sumarização conseguiram aumentar os valores de *F-Measure* dos agrupamentos obtidos.

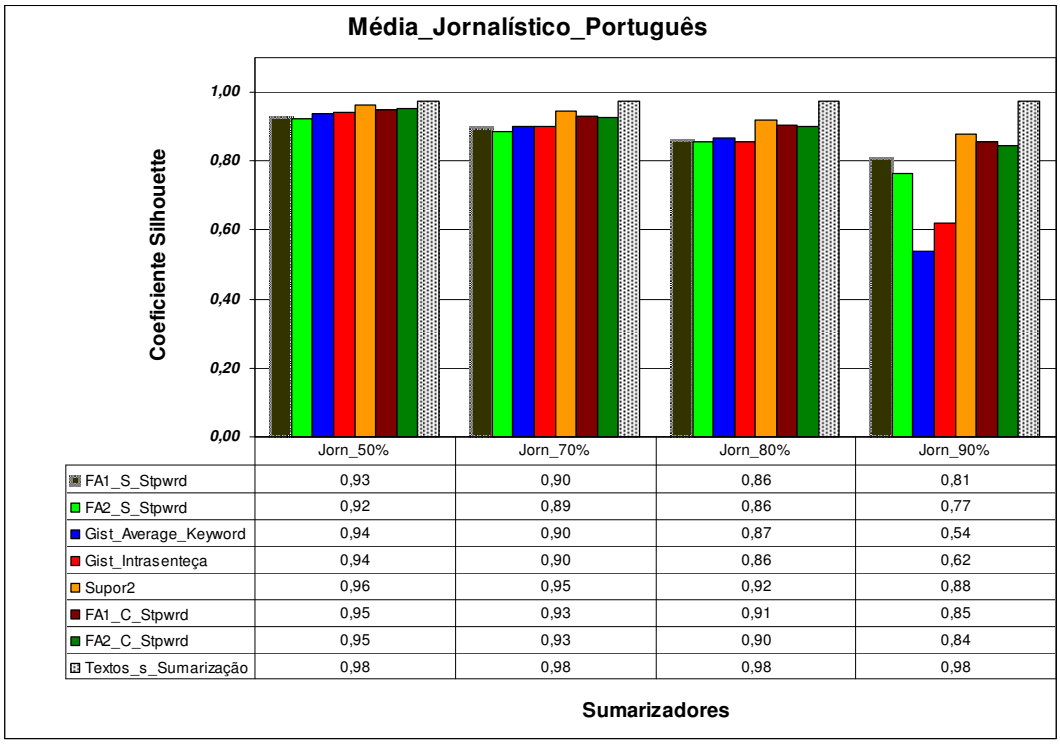


	Jur_50%	Jur_70%	Jur_80%	Jur_90%
FA1_S_Stpwrd	0,25	0,22	0,22	0,27
FA2_S_Stpwrd	0,19	0,25	0,27	0,23
Gist_Average_Keyword	0,16	0,28	0,25	0,20
Gist_Intrasenteça	0,17	0,16	0,20	0,29
Supor2	0,23	0,16	0,28	0,19
FA1_C_Stpwrd	0,27	0,26	0,24	0,24
FA2_C_Stpwrd	0,22	0,26	0,24	0,19
Textos_s_Sumarização	0,18	0,18	0,18	0,18

Figura 28: Resultados das médias acumuladas obtidos pelo modelo Cassiopeia, usando a medida *FMeasure* com 50%, 70%, 80% e 90% de compressão no idioma português, no domínio jurídico.

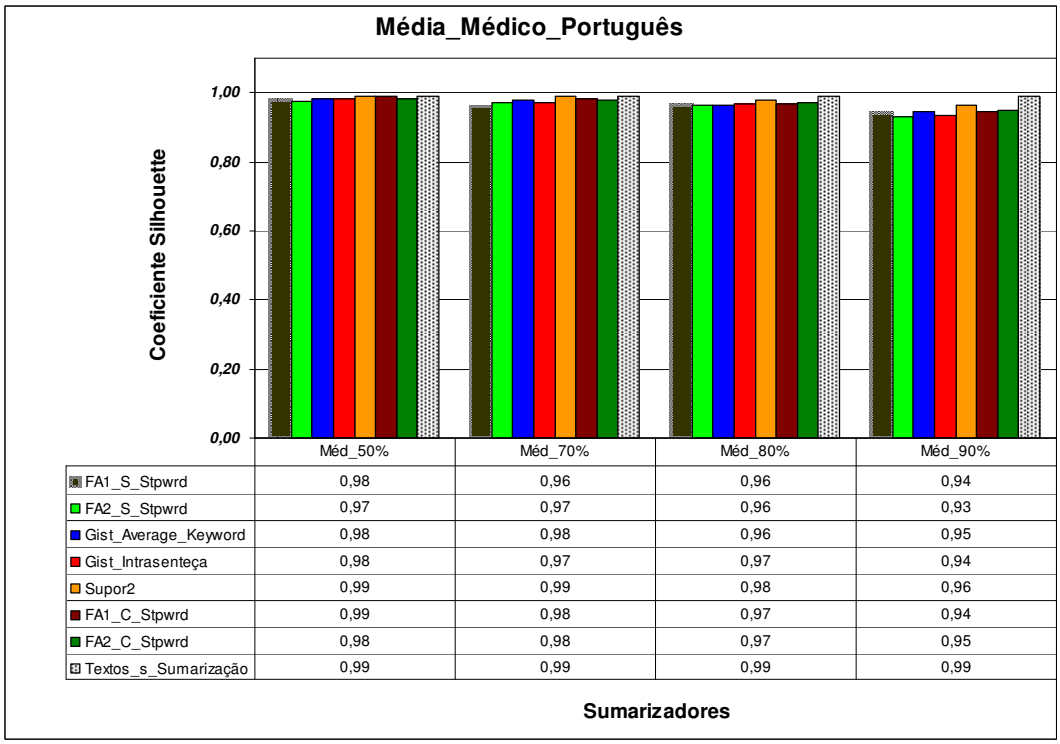
Observa-se, no domínio jornalístico, Figura 29, que os algoritmos de sumarização não conseguiram aumentar o valor dos agrupamentos gerados pelo modelo Cassiopeia, aferidos pela medida Coeficiente Silhouette em relação ao texto-fonte.

No domínio médico, apresentado na Figura 30, com 50% de compressão, dois algoritmos melhoraram o desempenho dos agrupamentos gerados pelo modelo, aferidos pela medida Coeficiente Silhouette, comparados com os textos-fonte. Com 70% apenas um, e com 80% e 90% nenhum dos algoritmos conseguiu aumentar os valores de *F-Measure* dos agrupamentos obtidos.



	Jorn_50%	Jorn_70%	Jorn_80%	Jorn_90%
FA1_S_Stpwrd	0,93	0,90	0,86	0,81
FA2_S_Stpwrd	0,92	0,89	0,86	0,77
Gist_Average_Keyword	0,94	0,90	0,87	0,54
Gist_Intrasenteça	0,94	0,90	0,86	0,62
Supor2	0,96	0,95	0,92	0,88
FA1_C_Stpwrd	0,95	0,93	0,91	0,85
FA2_C_Stpwrd	0,95	0,93	0,90	0,84
Textos_s_Sumarização	0,98	0,98	0,98	0,98

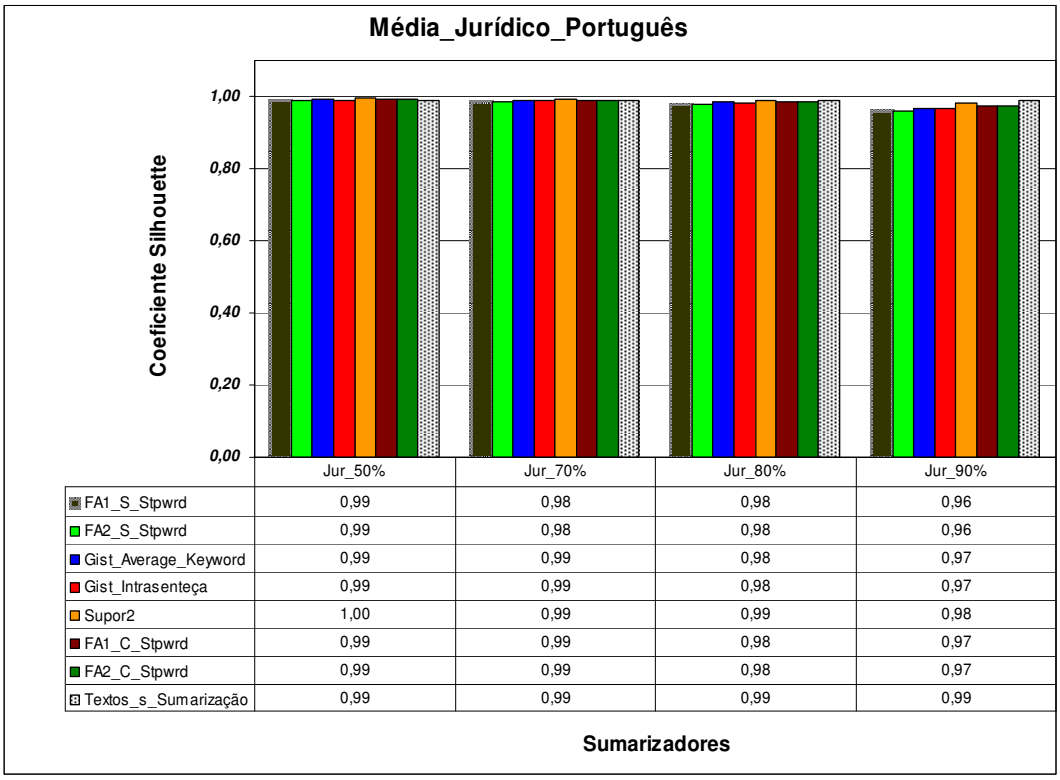
Figura 29: Resultados das médias acumuladas obtidos pelo modelo Cassiopeia, usando a medida Coeficiente Silhouette com 50%, 70%, 80% e 90% de compressão no idioma português, no domínio jornalístico.



	Méd_50%	Méd_70%	Méd_80%	Méd_90%
FA1_S_Stpwrd	0,98	0,96	0,96	0,94
FA2_S_Stpwrd	0,97	0,97	0,96	0,93
Gist_Average_Keyword	0,98	0,98	0,96	0,95
Gist_Intrasenteça	0,98	0,97	0,97	0,94
Supor2	0,99	0,99	0,98	0,96
FA1_C_Stpwrd	0,99	0,98	0,97	0,94
FA2_C_Stpwrd	0,98	0,98	0,97	0,95
Textos_s_Sumarização	0,99	0,99	0,99	0,99

Figura 30: Resultados das médias acumuladas, obtidos pelo modelo Cassiopeia, usando a medida *Coeficiente Silhouette* com 50%, 70%, 80% e 90% de compressão, no idioma português, no domínio médico.

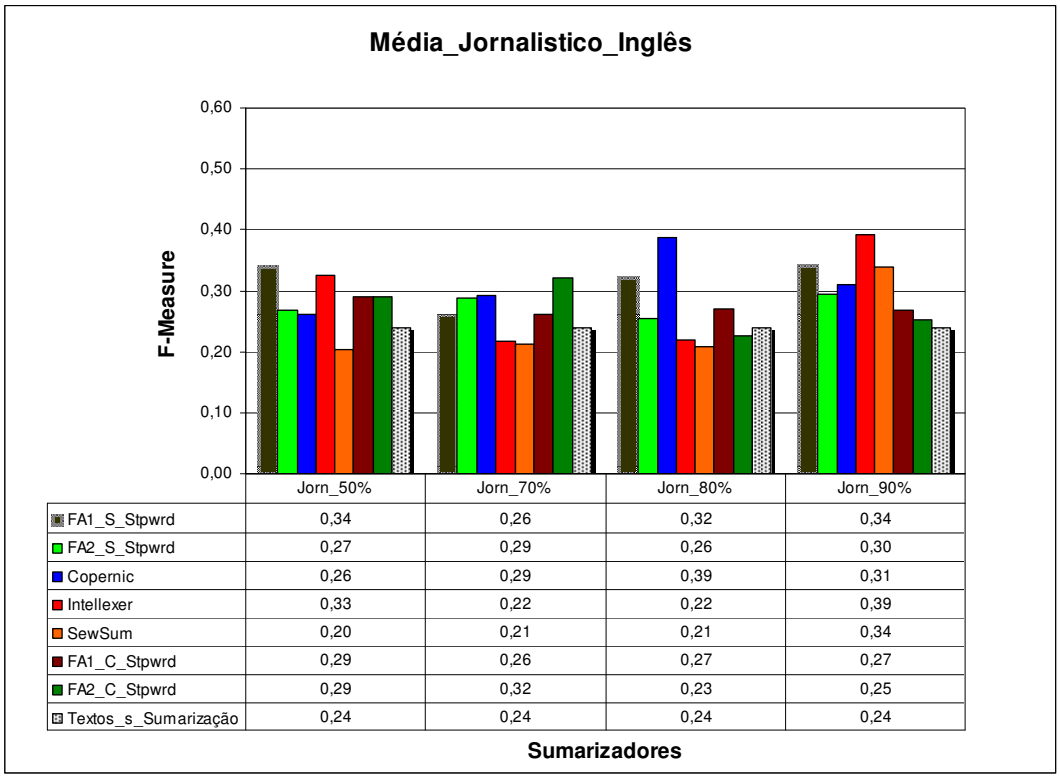
Na medida Coeficiente Silhouette, Figura 31, com 50% de compressão, os seis algoritmos de sumarização fizeram com que os agrupamentos gerados pelo modelo Cassiopeia mantivessem os valores, e um algoritmo conseguiu aumentar o valor dos agrupamentos, em comparação com os textos-fonte. Com 70% de compressão, dois algoritmos diminuíram o valor dos agrupamentos obtidos pelo modelo Cassiopeia, na medida Coeficiente Silhouette. Com 80% de compressão, apenas um algoritmo influenciou o modelo a melhorar o desempenho dos agrupamentos, isso comparado aos gerados pelo modelo Cassiopeia, usando os textos-fonte. Com 90% de compressão, não houve nenhum algoritmo de sumarização que conseguisse fazer com que o modelo Cassiopeia aumentasse o aferimento dos seus agrupamentos.



	Jur_50%	Jur_70%	Jur_80%	Jur_90%
FA1_S_Stpwrd	0,99	0,98	0,98	0,96
FA2_S_Stpwrd	0,99	0,98	0,98	0,96
Gist_Average_Keyword	0,99	0,99	0,98	0,97
Gist_Intrasenteça	0,99	0,99	0,98	0,97
Supor2	1,00	0,99	0,99	0,98
FA1_C_Stpwrd	0,99	0,99	0,98	0,97
FA2_C_Stpwrd	0,99	0,99	0,98	0,97
Textos_s_Sumarização	0,99	0,99	0,99	0,99

Figura 31: Resultados das médias acumuladas, obtidos pelo modelo Cassiopeia usando a medida Coeficiente Silhouette com 50%, 70%, 80% e 90% de compressão no idioma português no domínio jurídico.

Nas Figuras 32, 33, 34 e 35 serão apresentados os resultados dos aferimentos dos agrupamentos de textos obtidos pelo modelo Cassiopeia, usando as medidas *F-Measure* e *Coeficiente Silhouette*, no idioma inglês, nos domínios jornalístico e médico, usando todas as compressões de 50%, 70%, 80% e 90% e os sete algoritmos de sumarização, além dos texto-fonte.



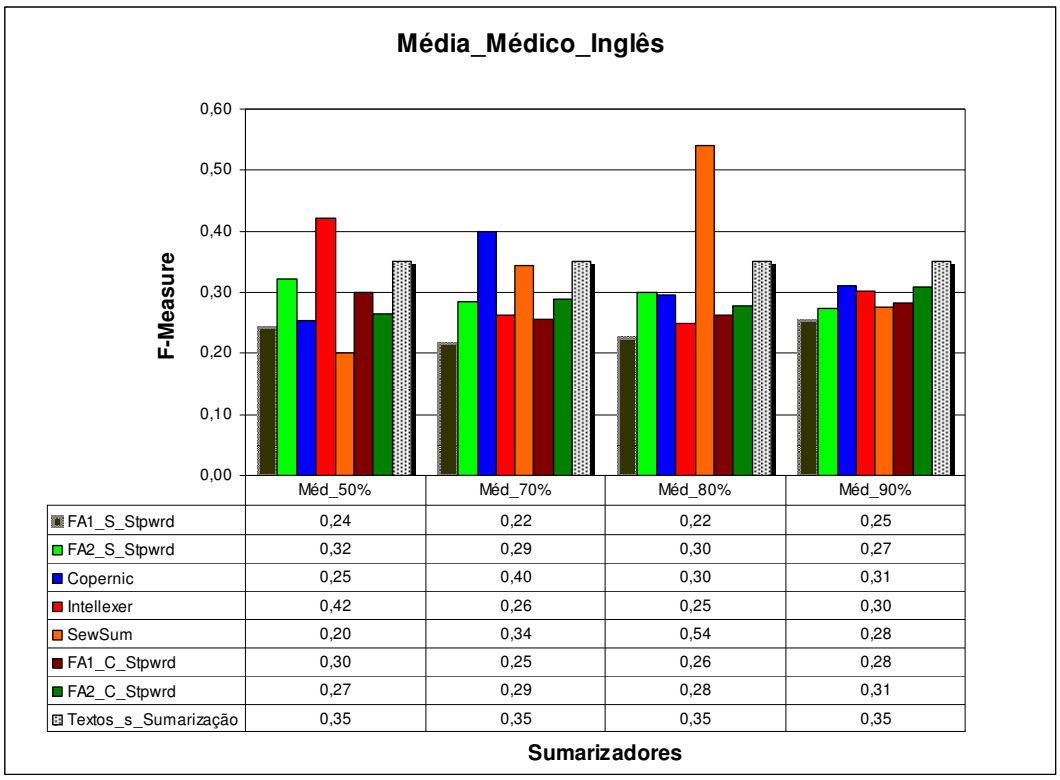
	Jorn_50%	Jorn_70%	Jorn_80%	Jorn_90%
FA1_S_Stpwrd	0,34	0,26	0,32	0,34
FA2_S_Stpwrd	0,27	0,29	0,26	0,30
Copernic	0,26	0,29	0,39	0,31
Intellexer	0,33	0,22	0,22	0,39
SewSum	0,20	0,21	0,21	0,34
FA1_C_Stpwrd	0,29	0,26	0,27	0,27
FA2_C_Stpwrd	0,29	0,32	0,23	0,25
Textos_s_Sumarização	0,24	0,24	0,24	0,24

Figura 32: Resultados das médias acumuladas obtidos pelo modelo Cassiopeia, usando a medida *F-Measure* com 50%, 70%, 80% e 90% de compressão no idioma inglês, no domínio jornalístico.

Pode-se observar, para o domínio jornalístico, mostrado na Figura 32, que, em grande parte, os algoritmos de sumarização possibilitarão ao modelo Cassiopeia gerar agrupamentos de textos com valores de *F-Measure*, maiores do que os agrupamentos dos textos-fonte.

Com a compressão de 50%, a exceção foi *SewSum*; com a de 70%, *SewSum* e *Intellexer*; com a de 80%, *SewSum*, *Intellexer* e *FA2_com_Stopword,* e com 90%, todos os algoritmos de sumarização possibilitaram ao modelo obter agrupamentos com valores de *F-Measure* maiores do que os agrupamentos dos textos-fonte.

No domínio médico, observa-se, na Figura 33, com as compressões de 50%, 70% e 80%, que apenas um algoritmo de sumarização, em cada compressão, possibilitou ao modelo gerar agrupamentos com valor de *F-Measure* maior que os textos-fonte. Com 90%, nenhum dos algoritmos possibilitou gerar agrupamentos que obtivessem valores de *F-Measure* maiores do que os dos textos sem sumarização.



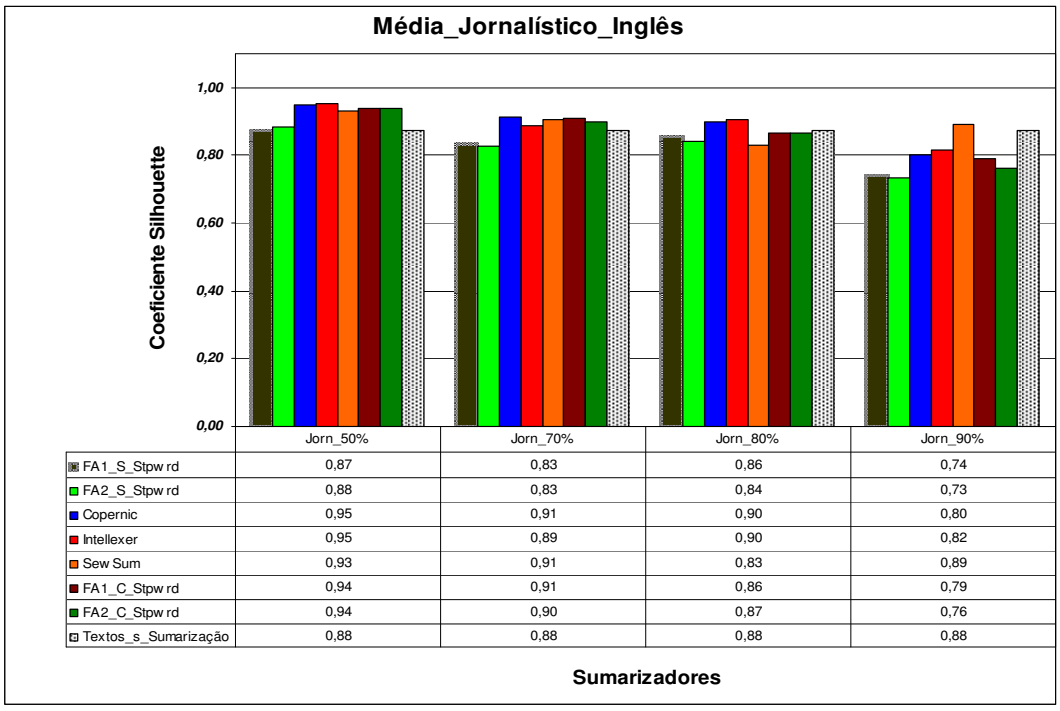
	Méd_50%	Méd_70%	Méd_80%	Méd_90%
FA1_S_Stpwrd	0,24	0,22	0,22	0,25
FA2_S_Stpwrd	0,32	0,29	0,30	0,27
Copernic	0,25	0,40	0,30	0,31
Intellexer	0,42	0,26	0,25	0,30
SewSum	0,20	0,34	0,54	0,28
FA1_C_Stpwrd	0,30	0,25	0,26	0,28
FA2_C_Stpwrd	0,27	0,29	0,28	0,31
Textos_s_Sumarização	0,35	0,35	0,35	0,35

Figura 33: Resultados das médias acumuladas obtidos pelo modelo Cassiopeia, usando a medida *F-Measure* com 50%, 70%, 80% e 90% de compressão no idioma inglês, no domínio médico.

Observa-se, no domínio jornalístico, na Figura 34, que os algoritmos de sumarização com 50% e 70% de compressão melhoraram os desempenhos dos agrupamentos de textos gerados no modelo Cassiopeia, com valores de Coeficiente Silhouette maiores do que agrupamentos dos textos-fonte.

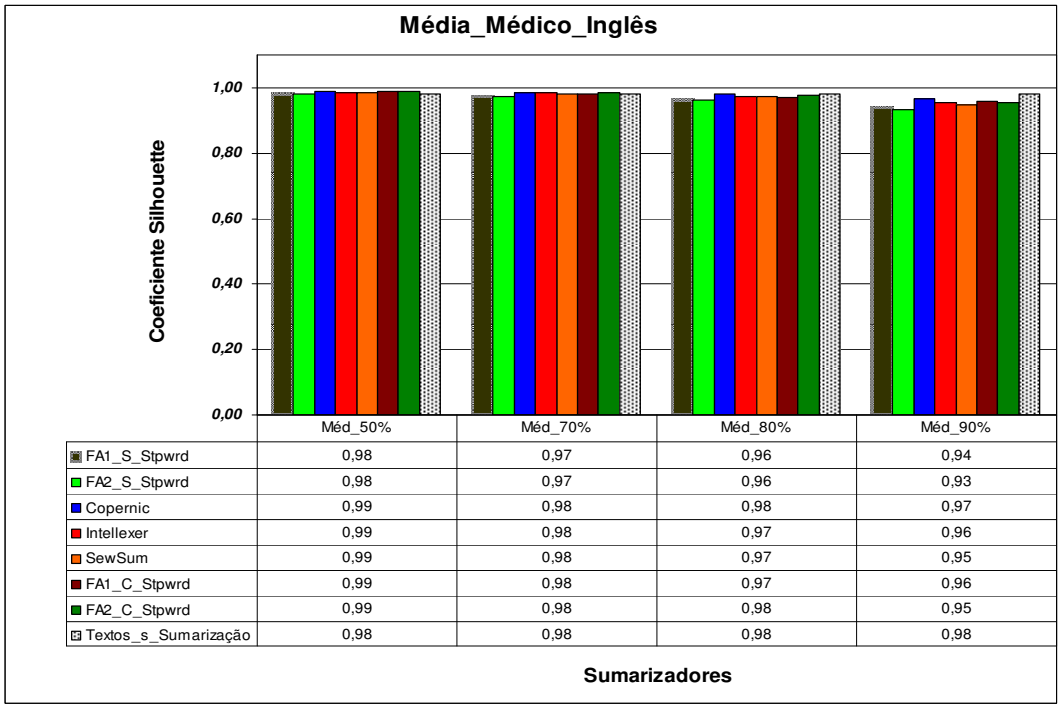
As exceções ficaram com 50% de compressão, apresentadas pela função *FA1_sem_Stopword* e com 70% de compressão pelas funções *FA1 e FA2 sem Stopword*. Com 80% de compressão, apenas dois algoritmos melhoraram os desempenhos dos agrupamentos gerados pelo modelo, sendo assim, aumentaram seus valores de Coeficiente Silhouette. Com 90% de compressão, apenas um aumentou.

No domínio médico, mostrado na Figura 35, o aumento da compressão dos algoritmos de sumarização, foi obtendo agrupamentos com seus valores de Coeficiente Silhouette diminuídos. Com 50% de compressão, apenas um algoritmo teve seus agrupamentos na medida de Coeficiente Silhouette menor que os agrupamentos dos textos-fonte. Com 70% foram dois, com 80% foram cinco e com 90% foram todos.



	Jorn_50%	Jorn_70%	Jorn_80%	Jorn_90%
FA1_S_Stpw rd	0,87	0,83	0,86	0,74
FA2_S_Stpw rd	0,88	0,83	0,84	0,73
Copernic	0,95	0,91	0,90	0,80
Intellexer	0,95	0,89	0,90	0,82
Sew Sum	0,93	0,91	0,83	0,89
FA1_C_Stpw rd	0,94	0,91	0,86	0,79
FA2_C_Stpw rd	0,94	0,90	0,87	0,76
Textos_s_Sumarização	0,88	0,88	0,88	0,88

Figura 34: Resultados das médias acumuladas obtidos pelo modelo Cassiopeia, usando a medida Coeficiente Silhouette com 50%, 70%, 80% e 90% de compressão no idioma inglês, no domínio jornalístico.



	Méd_50%	Méd_70%	Méd_80%	Méd_90%
FA1_S_Stpwrd	0,98	0,97	0,96	0,94
FA2_S_Stpwrd	0,98	0,97	0,96	0,93
Copernic	0,99	0,98	0,98	0,97
Intellexer	0,99	0,98	0,97	0,96
SewSum	0,99	0,98	0,97	0,95
FA1_C_Stpwrd	0,99	0,98	0,97	0,96
FA2_C_Stpwrd	0,99	0,98	0,98	0,95
Textos_s_Sumarização	0,98	0,98	0,98	0,98

Figura 35: Resultados das médias acumuladas obtidos pelo modelo Cassiopeia, usando a medida Coeficiente Silhouette com 50%, 70%, 80% e 90% de compressão no idioma inglês, no domínio médico.

5.2 SEGUNDA PARTE DOS EXPERIMENTOS

Na segunda parte, serão apresentados os experimentos realizados no modelo Cassiopeia, usando os textos-fonte (sem sumarização) e os textos sumarizados, obtidos através dos sumarizadores escolhidos e definidos na seção 4.2.. O conjunto de cem textos de cada sumarizador e do texto-fonte (sem sumarização) foi submetidos ao modelo Cassiopeia, separadamente. O modelo executou o processo de agrupamento e reagrupamento, com cada conjunto de cem textos, individualmente, ocorrendo, assim, cem vezes para cada conjunto, resultando em uma média aritmética de cada métrica. Esses agrupamentos foram mensurados através das métricas externas ou supervisionada (*Recall*, *Precision* e *F-Measure*), e internas ou não supervisionada (*Coesão*, *Acoplamento* e *Coeficiente Silhouette)*, explicadas nas subseções 2.2.5.1 e 2.2.5.2. Os cem resultados obtidos com a média aritmética de cada uma das métricas e para cada agrupamento. Foi também gerada uma soma acumulada desses cem valores, que serão mostrados nos gráficos dos experimentos para cada uma das métricas. Esse processo ocorreu, separadamente, para cada um dos percentuais de compressão, ou seja, 50%, 70%, 80% e 90% e para cada um dos idiomas, português e inglês.

A diferença da primeira parte em relação à segunda parte dos experimentos, surgiu na seleção dos atributos, que foi comparada com métodos tradicionais na literatura, como: RDF, RTF e TFIDF, explicados nas subseções 2.3.1, 2.3.2 e 2.3.3. Segundo Loh (2001), Wives (2004) e Nogueira (2009), a seleção do atributo é a parte mais significativa para o processo de agrupamento. De acordo com os autores, é o fator principal para bons resultados nos agrupamentos. Nogueira (2009) faz um estudo criterioso sobre o assunto, pois trabalha com a seleção de atributos não supervisionados, um critério de escolha para o modelo Cassiopeia, já que se pretende que este seja, em todas as suas etapas, não supervisionado. O autor afirma que o novo conjunto gerado é de dimensão menor do que o original. Como o objeto de estudo deste trabalho são as bases textuais, é fundamental que os resultados finais sejam compreensíveis para o usuário, ou seja, que o subconjunto obtido tenha relação forte com o conjunto original.

Dessa forma, justifica-se a segunda etapa do experimento com o modelo Cassiopeia, para mensurar a qualidade dos agrupamentos, usando como critério a seleção dos atributos e, em seguida mensurado com as métricas internas e externas, usadas também na primeira parte desse experimento, e como todos os *copora* usados para validar o modelo.

5.2.1 MÉTRICA EXTERNA: *RECALL, PRECISION* E *F-MEASURE*

Para organização da apresentação dos resultados, serão colocados no Apêndice C todos os gráficos referentes à média acumulada dos resultados do *Recall*, *Precision e FMeasure*, obtida ao longo das 100 iterações no modelo Cassiopeia. Os resultados das médias finais acumuladas das medidas *Recall* e *Precision* serão também apresentados no Apêndice C.

A Figura 36 mostra o gráfico dos métodos RDF, RTF , TFIDF, e o modelo Cassiopeia com *Stopword*, e o sem *Stopword*, usados para identificar e selecionar os atributos (palavras) nos textos sumarizados a serem agrupados. Para esse experimento foram usados os sumarizadores *Gist_Keyword*, *Gist_Intrasentença* e *SuPor,* com as respectivas compressões de 50%, 70%, 80% e 90% nos textos-fonte, no domínio jornalístico e no idioma português. Os resultados que aparecem na Figura 36 representam as médias finais acumuladas da medida *F-Measure*, obtidas nos agrupamentos de textos ao longo das 100 interações para cada método testado.

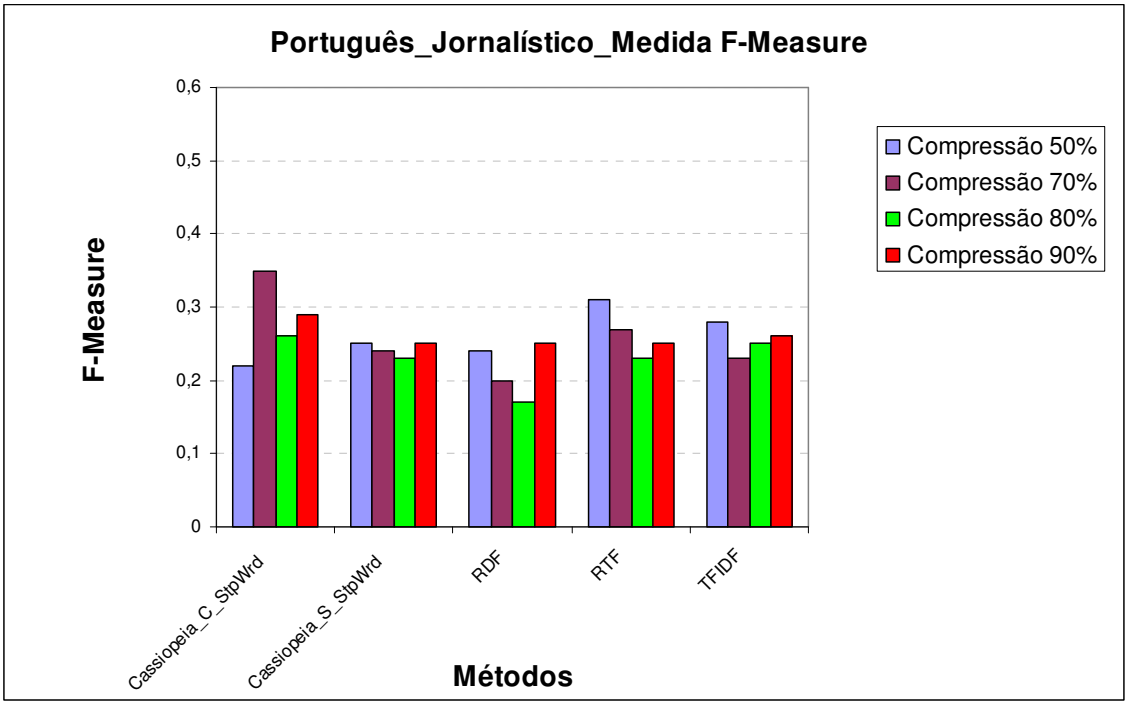


Figura 36: Resultados das médias finais acumuladas da medida *F-Measure* para os agrupamentos obtidos através dos métodos RDF, RTF e TFIDF e do modelo Cassiopeia, com e sem *Stopword*, no domínio jornalistico e no idioma português.

Analisando a compressão de 50%, o melhor desempenho de *F-Measure* dos agrupamentos foi obtido pelo método RTF, com 0,31. Com 70%, 80% e 90% de compressão nos textos-fonte, o modelo Cassiopeia_com_*Stopword* foi o que obteve, em seus agrupamentos, os melhores valores, de 0,35, 0,26 e 0,29 de *F-Measure,* respectivamente. Assim, pode-se afirmar que o modelo Cassiopeia com *Stopword* obteve os melhores desempenhos de *F-Measure,* em 75% de toda a

amostra para o idioma português, no domínio jornalístico, sendo o restante dos 25% da amostra obtido pelo método RTF, com o maior valor de *F-Measure*.

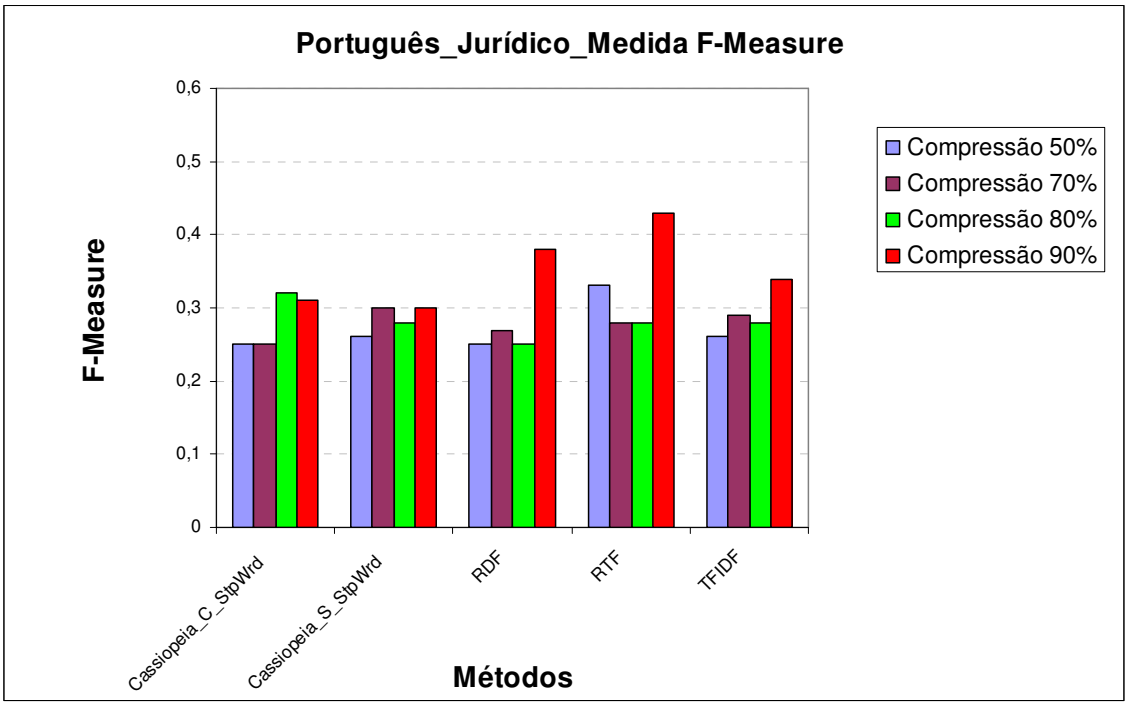


Figura 37: Resultados das médias finais acumuladas da medida *F-Measure* para os agrupamentos obtidos através dos métodos RDF, RTF e TFIDF e do modelo Cassiopeia, com e sem *Stopword*, no domínio jurídico e no idioma português.

Na Figura 37, os resultados mostram as médias finais acumuladas da medida *F-Measure*, obtidas nos agrupamentos de textos ao longo das 100 interações para cada método testado e para o modelo Cassiopeia. Visualizam-se os métodos RDF, RTF, TFIDF e o modelo Cassiopeia com *Stopword*, e o sem *Stopword*, os quais identificam e selecionam os atributos (palavras) nos textos sumarizados a serem agrupados. Para esse experimento foram usados os sumarizadores *Gist_Keyword*, *Gist_Intrasentença* e *SuPor,* com respectivas compressões de 50%, 70%, 80% e 90% nos textos-fonte, no domínio jurídico e no idioma português.

Na compressão de 50%, o melhor método foi o RTF, com um valor de *F-Measure* para os agrupamentos obtidos de 0,33. Com 70% de compressão nos textos-fonte, o modelo *Cassiopeia_sem_Stopword* foi o que obteve, em seus agrupamentos, o melhor valor de *F-Measure* de 0,30. Já para 80% foi o modelo Cassiopeia_com_*Stopword* que obteve o melhor valor de *F-Measure* com 0,32. Com 90% foi o método RTF, que obteve o melhor valor de *F-Measure* para os agrupamentos formados com resultados de 0,43. Sendo assim, pode-se afirmar que o modelo Cassiopeia apresentou os melhores desempenhos de *F-Measure* em 50% de toda a amostra, com e sem uso das *stopword* para o idioma português e no domínio jurídico. Cabe ressaltar que o método RTF também obteve 50% dos maiores valores de *F-Measure* de toda a amostra.

A Figura 38 mostra o gráfico dos métodos da literatura RDF, RTF , TFIDF e o modelo Cassiopeia com *Stopword* e o sem *Stopword*, usados para identificar e selecionar os atributos (palavras) nos textos sumarizados a serem agrupados.

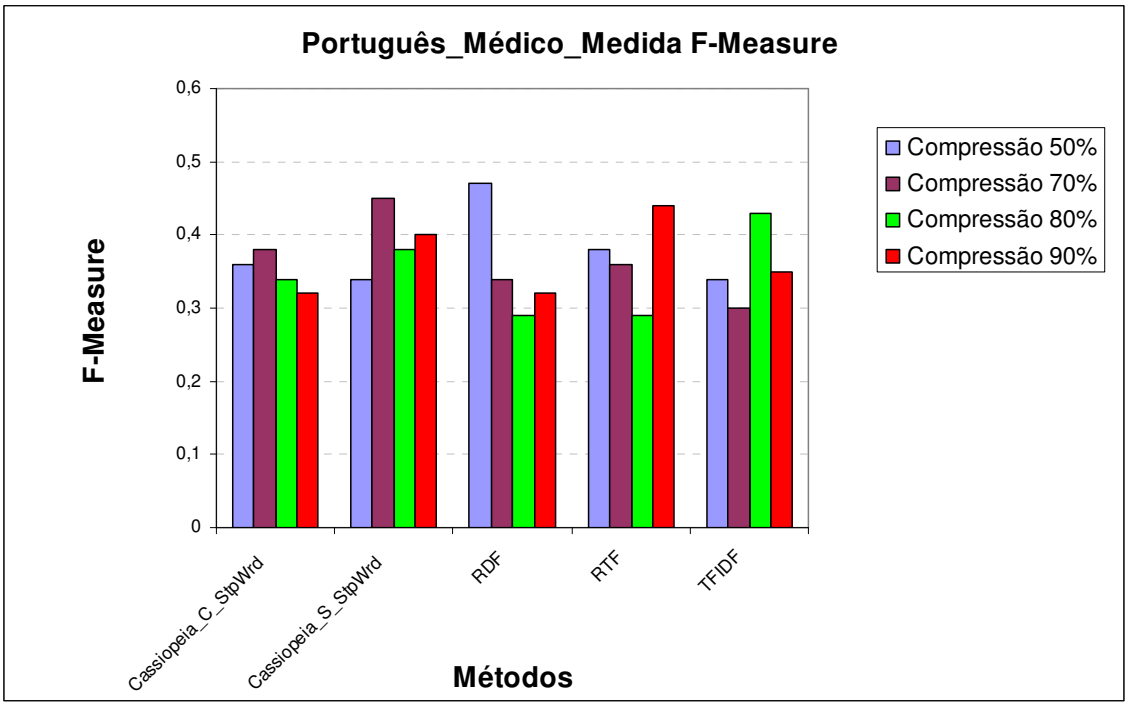


Figura 38: Resultados das médias finais acumuladas da medida *F-Measure* para os agrupamentos obtidos através dos métodos RDF, RTF e TFIDF e do modelo Cassiopeia, com e sem *Stopword*, no domínio médico e no idioma português.

Para esse experimento foram utilizados os sumarizadores *Gist_Keyword*, *Gist_Intrasentença* e *SuPor,* com respectivas compressões de 50%, 70%, 80% e 90% nos textos-fonte, no domínio médico e no idioma português. Os resultados que aparecem na figura representam as médias finais acumuladas da medida *F-Measure*, obtidas nos agrupamentos de textos ao longo das 100 interações para cada método testado.

Com os resultados apresentados na Figura 38, observa-se que, para a compressão de 50%, o melhor valor de *F-Measure* foi no método RDF, cujo valor obtido foi 0,47. Na compressão de 70%, o maior valor de *FMeasure* foi 0,45, obtido no modelo Casiopeia sem *Stopwords*. Para a compressão de 80%, o maior valor da medida *F-Measure* foi 0,43, atingido pelo método TFIDF. A compressão de 90% alcançou o maior valor de *F-Measure* de 0,44, pelo método RTF. Nesse caso, o método RDF obteve 25% da amostra total, o modelo Cassiopeia sem *Stopwords* também obteve 25 % da amostra desse experimento, enquanto o método TFIDF ficou com 25% e o método RTF 25%.

O gráfico representado na Figura 39 mostra o uso dos sumarizadores *Copernic*, *Intellexer* e *SweSum* com as compressões de 50%, 70%, 80% e 90% nos textos-fonte, no domínio jornalístico no idioma inglês. Para o experimento foram empregados os métodos da literatura RDF, RTF , TFIDF e o modelo Cassiopeia com *Stopword* e o sem *Stopword*, a fim de identificar e selecionar os atributos (palavras) nos textos sumarizados a serem agrupados.

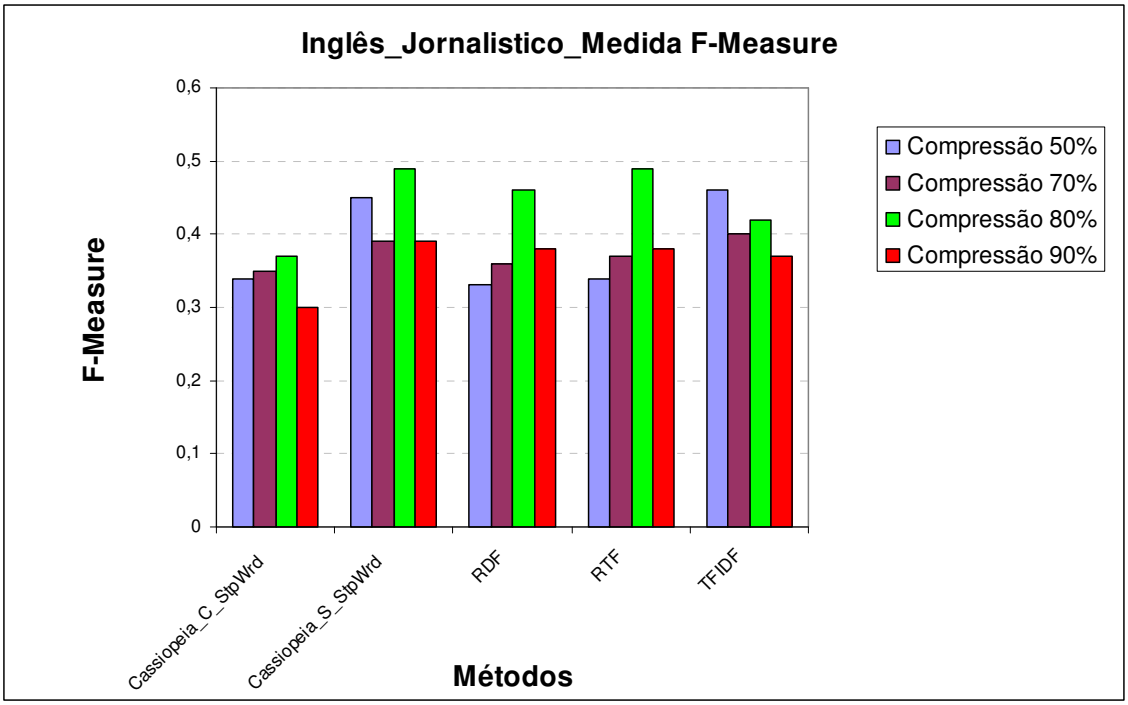


Figura 39: Resultados das médias finais acumuladas da medida *F-Measure* para os agrupamentos obtidos através dos métodos RDF, RTF e TFIDF e do modelo Cassiopeia, com e sem *Stopword*, no domínio jornalístico e no idioma inglês.

Os resultados apresentaram as médias finais acumuladas da medida *F-Measure*, obtidas nos agrupamentos de textos ao longo das 100 interações para cada método testado.

A compressão de 50% , indicou o melhor valor de *F-Measure* dos agrupamentos, atingida pelo pelo método TFIDF, com 0,46. Na compressão de 70%, o maior valor de *FMeasure*, 0,40, foi obtido novamente pelo método TFIDF. Para a compressão de 80%, o maior valor da medida *F-Measure* foi 0,40, alcançado pelo método RDF e pelo modelo Cassiopeia sem *Stopword.* A compressão de 90% atingiu o maior valor de *F-Measure,* 0,39, e foi novamente no modelo Cassiopeia sem *Stopword*. Nesse experimento, o modelo Cassiopeia sem *Stopword* obteve 50% da amostra total, enquanto o método TFIDF, 50% e o método RTF, 25%, pois na compressão de 80% houve um empate com o modelo Cassiopeia sem *Stopword*.

A Figura 40 também mostra a utilização dos sumarizadores *Copernic*, *Intellexer* e *SweSum* com as compressões de 50%, 70%, 80% e 90% nos textos-fonte, só que, nesse caso, os

experimentos ocorreram no domínio médico e no idioma inglês. O experimento usou os métodos da literatura RDF, RTF , TFIDF e o modelo Cassiopeia com *Stopword* e o sem *Stopword*, para identificar e selecionar os atributos (palavras) nos textos sumarizados a serem agrupados.

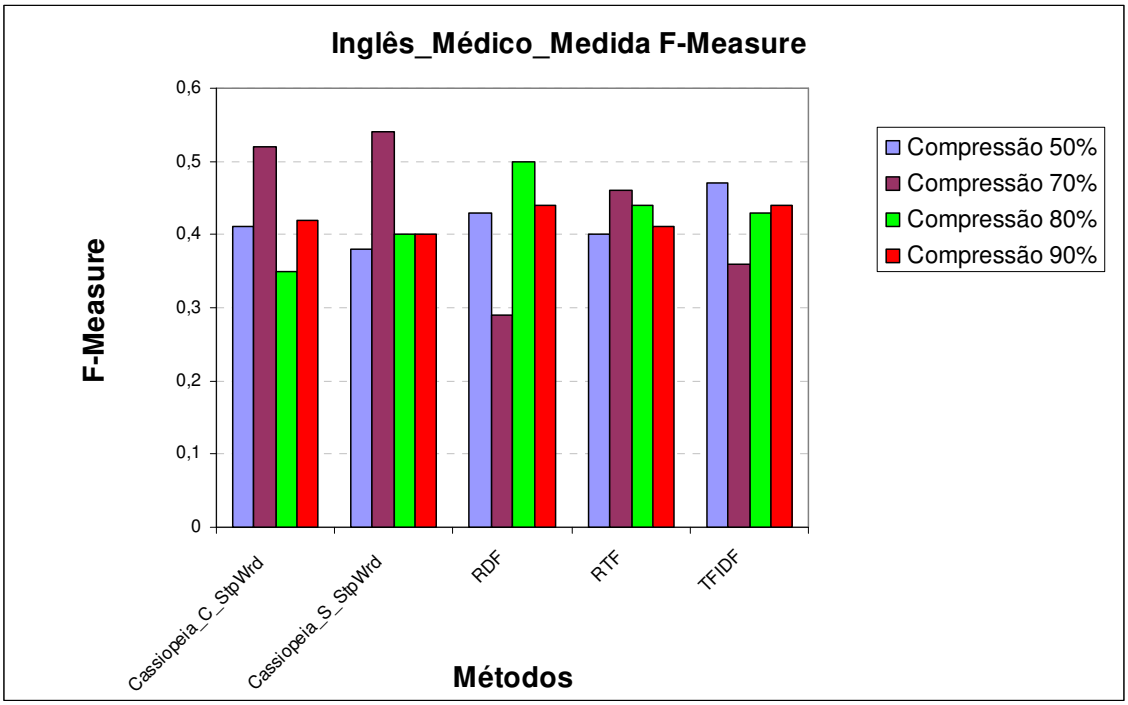


Figura 40: Resultados das médias finais acumuladas da medida *F-Measure* para os agrupamentos obtidos através dos métodos RDF, RTF e TFIDF e do modelo Cassiopeia, com e sem *Stopword*, no domínio médico e no idioma inglês.

Os resultados apresentaram as médias finais acumuladas da medida *F-Measure*, obtidas nos agrupamentos de textos ao longo das 100 interações para cada método testado.

Na compressão de 50%, o melhor método foi o TFIDF, com um valor de *F-Measure* para os agrupamentos obtidos com o valor de 0,47. Com 70% de compressão nos textos-fonte, o modelo Cassiopeia_sem_*Stopword* foi o que obteve em seus agrupamentos o melhor valor de *F-Measure* de 0,54. Já para 80%, o método RDF obteve o melhor valor de *F-Measure* com 0,5. Com 90%, foram os métodos RDF e TFIDF que alcançaram os melhores valores de *F-Measure* para os agrupamentos formados com resultados de 0,44. O método TFIDF apresentou os melhores desempenhos de *F-Measure* em 50% de toda a amostra, já o método RDF obteve 50%. Os métodos RDF e TFIDF apresentaram similaridade em seus valores de *F-Measure* na compressão de 90%, com o valor de 0,44.

5.2.2 MÉTRICA INTERNA: *COESÃO, ACOPLAMENTO* E *COEFICIENTE SILHOUETTE*

Para melhor organização da apresentação dos resultados, serão colocados no Apêndice D todos os gráficos referentes à média acumulada dos resultados da Coesão, Acoplamento e

Coeficiente Silhouette, obtida ao longo das 100 interações no modelo Cassiopeia. Os resultados das médias finais acumuladas das medidas de Coesão e Acoplamento serão também apresentados no Apêndice D.

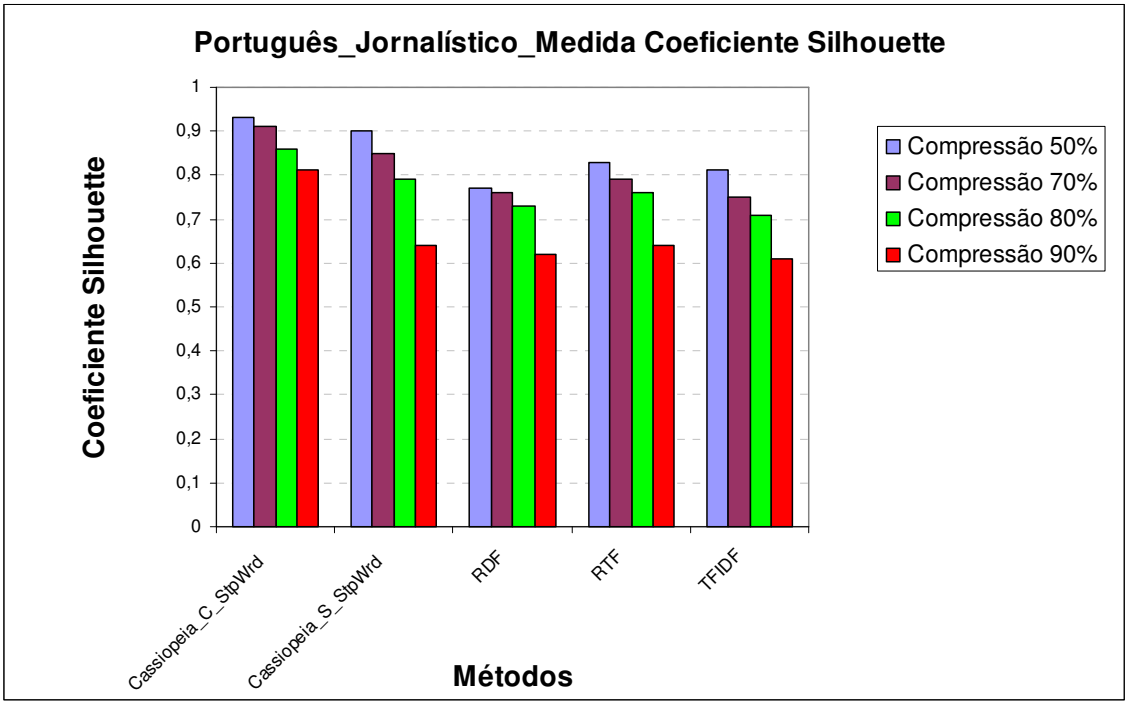


Figura 41: Resultados das médias finais acumuladas da medida *Coeficiente Silhouette* para os agrupamentos obtidos através dos métodos RDF, RTF e TFIDF e do modelo Cassiopeia, com e sem *Stopword*, no domínio jornalístico e no idioma português.

A Figura 41, mostra o gráfico dos métodos RDF, RTF, TFIDF e o modelo Cassiopeia com *Stopword*, e o sem *Stopword*, usados para identificar e selecionar os atributos (palavras) nos textos sumarizados a serem agrupados. Para esse experimento foram utilizados os sumarizadores *Gist_Keyword*, *Gist_Intrasentença* e *SuPor,* com respectivas compressões de 50%, 70%, 80% e 90% nos textos-fonte, no domínio jornalístico no idioma português. Os resultados apresentados na figura indicam médias finais acumuladas da medida Coeficiente Silhouette, obtidas nos agrupamentos de textos ao longo das 100 interações para cada método testado.

Em todas as compressões apresentadas na Figura 41, com 50%, 70%, 80% e 90%, os melhores valores do Coeficiente Silhouette para os agrupamentos foram os do modelo Cassiopeia_com_*Stopword* que obtiveram respectivamente os valores 0,93, 0,91, 0,86 e 0,81. Com esses resultados foi alcançado 100% de toda a amostra para o modelo Cassiopeia_com_*Stopword*.

O gráfico representado na Figura 42 também mostra o uso dos sumarizadores *Gist_Keyword*, *Gist_Intrasentença* e *SuPor* com as compressões de 50%, 70%, 80% e 90% nos

textos-fonte, só que nesse caso os experimentos ocorreram no domínio jurídico, no idioma português.

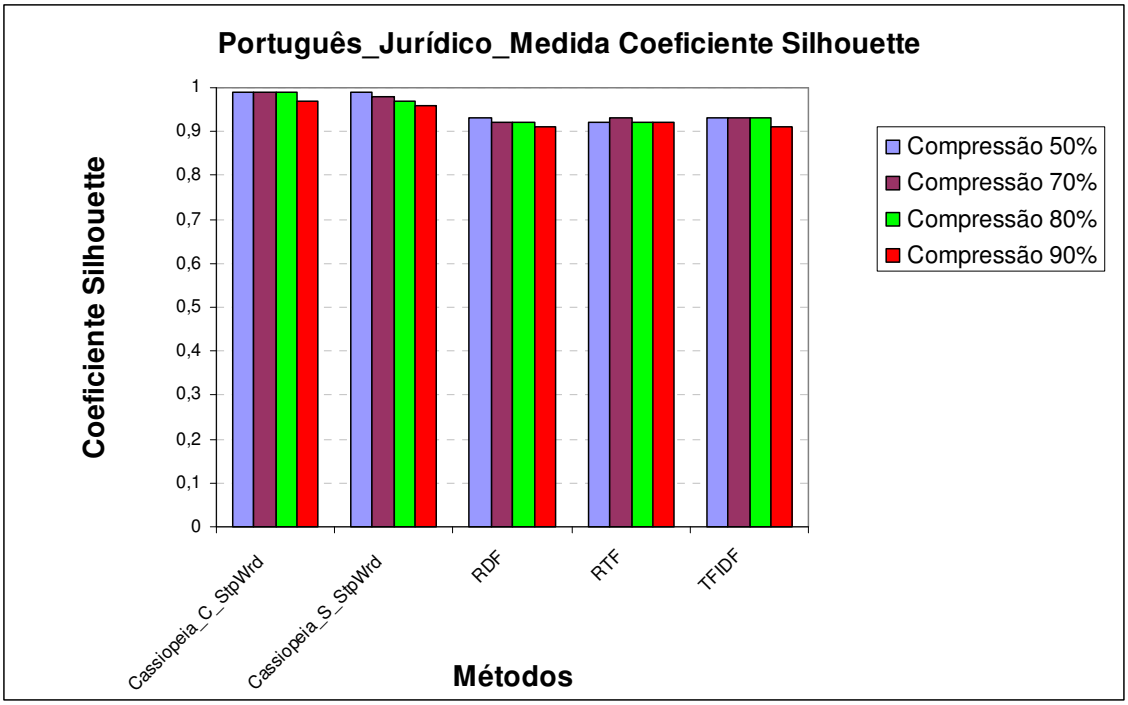


Figura 42: Resultados das médias finais acumuladas da medida *Coeficiente Silhouette* para os agrupamentos obtidos através dos métodos RDF, RTF e TFIDF e do modelo Cassiopeia, com e sem *Stopword*, no domínio jurídico e no idioma português.

O experimento mostra os métodos da literatura RDF, RTF, TFIDF e o modelo Cassiopeia com *Stopword* e o sem *Stopword*, usados para identificar e selecionar os atributos (palavras) nos textos sumarizados a serem agrupados. Os resultados visualizados apresentam as médias finais acumuladas da medida Coeficiente Silhouette, obtidas nos agrupamentos de textos ao longo das 100 interações para cada método testado.

Todas as compressões observadas na Figura 42, de 50%, 70%, 80% e 90%, alcançaram os melhores valores do Coeficiente Silhouette para os agrupamentos pelo modelo Cassiopeia_com_*Stopword* e obtiveram, respectivamente, os valores 0,99, 0,99, 0,99 e 0,97. Com esses resultados, foi atingido 100% de toda a amostra para o modelo Cassiopeia_com_*Stopword*.

O gráfico da Figura 43 apresenta os sumarizadores *Gist_Keyword*, *Gist_Intrasentença* e *SuPor* com as compressões de 50%, 70%, 80% e 90% nos textos-fonte, no domínio médico, no idioma português.

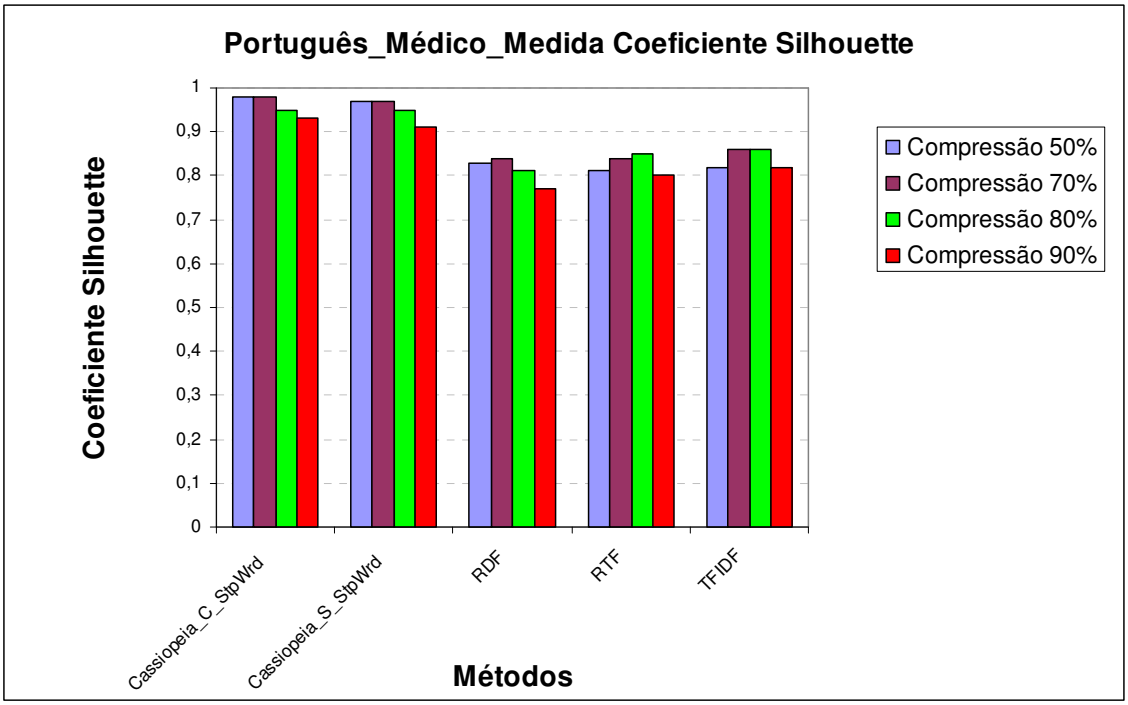


Figura 43: Resultados das médias finais acumuladas da medida *Coeficiente Silhouette* para os agrupamentos obtidos através dos métodos RDF, RTF e TFIDF e do modelo Cassiopeia, com e sem *Stopword*, no domínio médico e no idioma português.

Os resultados mostram as médias finais acumuladas da medida Coeficiente Silhouette, obtidas nos agrupamentos de textos ao longo das 100 interações para cada método testado.

Como pode ser observado na Figura 43 com as compressões 50%, 70%, 80% e 90% os melhores valores do Coeficiente Silhouette obtidos nos agrupamentos foram os do modelo Cassiopeia_com_*Stopword* que atingiram respectivamente 0,98, 0,98, 0,95 e 0,93. Com esses resultados foi alcançado 100% de toda a amostra para o modelo Cassiopeia_com_*Stopword*.

O gráfico representado na Figura 44 mostra os sumarizadores *Copernic*, *Intellexer* e *SweSum,* com as compressões de 50%, 70%, 80% e 90%, nos textos-fonte, no domínio jornalístico, no idioma inglês. Para o experimento foram utilizados os métodos da literatura RDF, RTF, TFIDF e o modelo Cassiopeia com *Stopword* e o sem *Stopword*, a fim de identificar e selecionar os atributos (palavras) nos textos sumarizados a serem agrupados. Os resultados são as médias finais acumuladas da medida *Coeficiente Silhouette*, obtidas nos agrupamentos de textos ao longo das 100 interações para cada método testado.

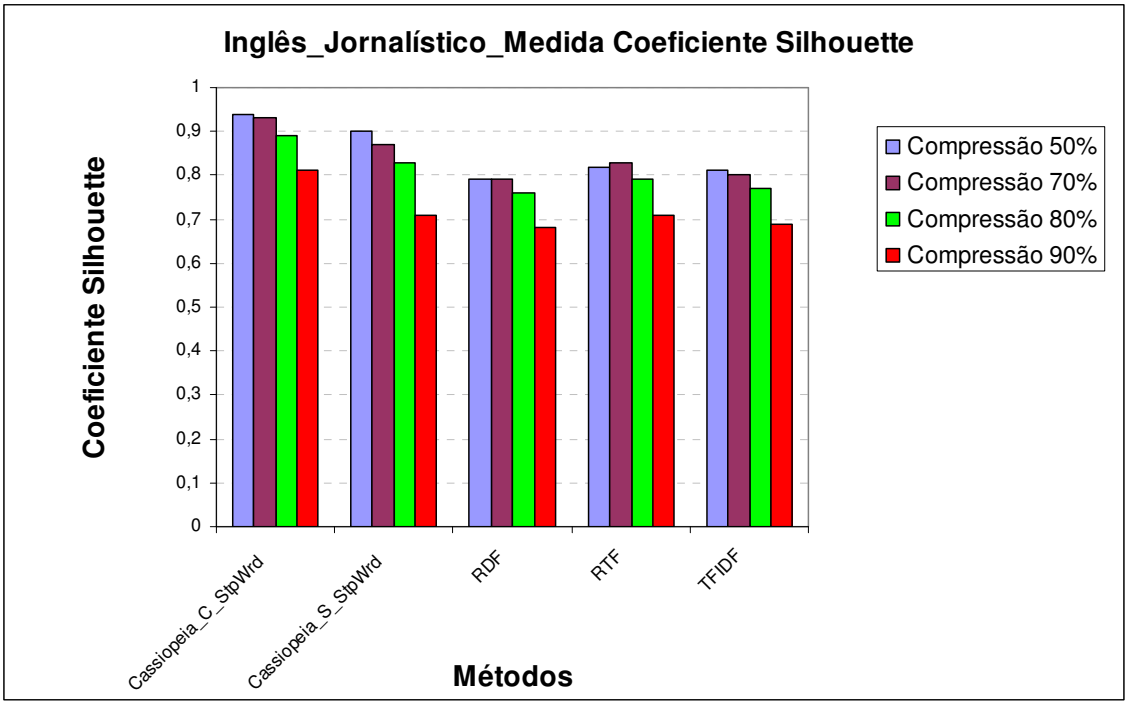


Figura 44: Resultados das médias finais acumuladas da medida *Coeficiente Silhouette* para os agrupamentos obtidos através dos métodos RDF, RTF e TFIDF e do modelo Cassiopeia com e sem *Stopword,* no domínio jornalístico e no idioma inglês.

Na Figura 44, visualiza-se o uso das compressões 50%, 70%, 80% e 90%. Os melhores valores do Coeficiente Silhouette apresentados nos agrupamentos foram os do modelo Cassiopeia_com_*Stopword,* que obtiveram, respectivamente, 0,94, 0,93, 0,89 e 0,81. Com esses resultados, foi alcançado 100% de toda a amostra para o modelo Cassiopeia_com_*Stopword*.

O gráfico representado na Figura 45 também apresenta os sumarizadores *Copernic*, *Intellexer* e *SweSum,* com as compressões de 50%, 70%, 80% e 90% nos textos-fonte. Só que nesse caso, os experimentos ocorreram no domínio médico, no idioma inglês. No experimento foram aplicados os métodos da literatura RDF, RTF , TFIDF e o modelo Cassiopeia com *Stopword*, e o sem *Stopword*, usados para identificar e selecionar os atributos (palavras) nos textos sumarizados a serem agrupados. Os resultados apresentaram as médias finais acumuladas da medida *F-Measure*, obtidas nos agrupamentos de textos ao longo das 100 interações para cada método testado.

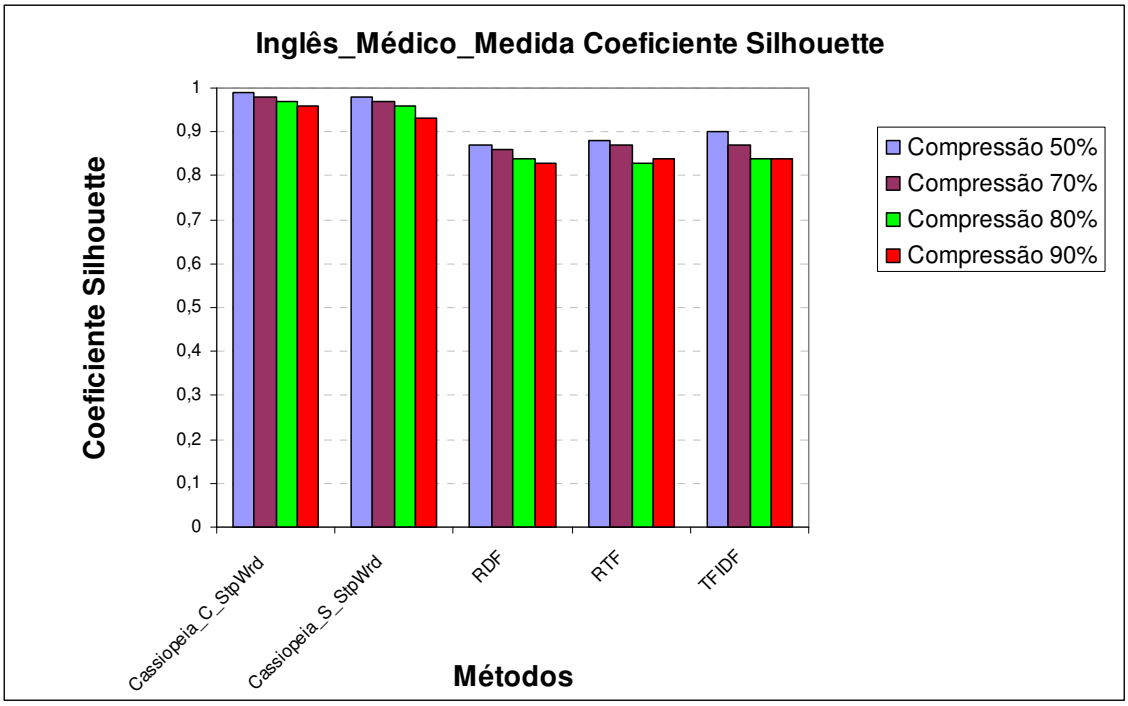


Figura 45: Resultados das médias finais acumuladas da medida Coeficiente Silhouette para os agrupamentos obtidos através dos métodos RDF, RTF e TFIDF e do modelo Cassiopeia, com e sem *Stopword*, no domínio médico e no idioma inglês.

A Figura 45 apresentou os resultados com as compressões 50%, 70%, 80% e 90%. Os melhores valores do Coeficiente Silhouette obtidos nos agrupamentos foram os do modelo Cassiopeia_com_*Stopword,* que mostraram, respectivamente, 0,99, 0,98, 0,97 e 0,96. Com esses resultados foi alcançado 100% de toda a amostra para o modelo Cassiopeia_com_*Stopword*.

5.3 HIPÓTESE

A hipótese nula deste trabalho consiste na afirmação de agrupadores em bases textuais que incluem, na etapa de pré-processamento, a sumarização de texto, e na etapa de processamento, o processo de agrupamento hierárquico de texto, com um novo método para definição do corte de Luhn, não conseguem melhorar suas avaliações nos agrupamentos de textos, tem restrição de domínio, são dependentes do idioma e não conseguem representar seu espaço amostral em uma solução vetorial.

Formalmente pode-se representar essa hipótese nula através da

Equação 23:

$$H_0: k_{\text{outros agrupamento}} = k_{\text{modelo Cassiopeia}} \quad (23)$$

Onde:

H_0 = hipótese nula;
$k_{\text{modelo Cassiopeia}}$ = distribuição das *k* amostras do modelo Cassiopeia;
$k_{\text{outros agrupamento}}$ = distribuição das *k* amostras dos outros agrupamentos;

Se a hipótese nula for considerada falsa, alguma outra afirmativa deve ser verdadeira. Este trabalho propõe a hipótese alternativa H_1, na qual a melhoria do desempenho de agrupadores em bases textuais inclui, na etapa de pré-processamento, a sumarização de textos e, na etapa de processamento, o agrupamento hierárquico de textos, com um novo método para definição do corte de Luhn, com melhoria nas avaliações nos agrupamentos de textos, sem restrição de domínio, independência do idioma e a representação do seu espaço amostral em uma solução vetorial.

A hipótese alternativa está formalmente representada na **Equação 24:**

$$H_1: k_{\text{modelo Cassiopeia}} > k_{\text{outros agrupamento}} \quad (24)$$

A metodologia de teste de hipótese aqui adotada considerou as amostras obtidas nas simulações do modelo Cassiopeia na primeira parte dos experimentos, como processo de agrupamento, usando textos- fonte e textos sumarizados, com diferentes níveis de compressão. com uma distribuição maior, usando as medidas de *Recall*, *Precision*, *F-Measure*, Coesão, Acoplamento e coeficiente de Silhouette. Na segunda parte dos experimentos, como processo de agrupamento, usando diferentes métodos de seleção de atributos encontrados na literatura, comparados com o modelo Cassiopeia com e sem *stopword*, usando as medidas de *Recall*, *Precision*, *F-Measure*, Coesão, Acoplamento e coeficiente de Silhouette.

Assim sendo, 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 *Ranks* estabelecidos.

5.4 ANÁLISE DOS TESTES ESTATÍSTICOS

Para análise dos testes estatísticos, foram geradas as tabelas apresentadas no Apêndice E. Nelas estão contidos os valores gerados para o teste ANOVA de Friedman, e para o de concordância, de Kendall, obtidos com os softwares citados e explicados no mesmo apêndice.. São vinte tabelas, dez para o primeiro experimento, cinco referentes às métricas externas, e cinco referentes às métricas internas; outras dez para o segundo experimento, também composto de cinco das métricas externas e cinco das internas.

As tabelas são compostas de informações, como: compressões (50%, 70%, 80% e 90%); *N* (o número de amostras); *GL* (grau de liberdade); Valor Crítico, α(alfa) com 0,05 nível de significância do teste estatístico; α (qui-quadrado), p-valor (bilateral) <0,05; *SFr p-valor* (bilateral) <0, 0001 que é o cálculo pelo teste ANOVA de Friedman, citado na subseção 2.6.1; Coeficiente de Concordância de Kendall, calculado conforme se explica na subseção 2.6.2, os valores de ordem médio são comparados com os valores do Coeficiente de Concordância de Kendall; os valores de soma de ordens; ordem médio, e média usados para ANOVA de Friedman, e por fim o *desvio padrão*.

Nos testes estatísticos, em todas as tabelas contidas no Apêndice E, observou-se a rejeição da hipótese nula (H_0) e a aceitação da hipótese alternativa (H_1) deste trabalho com um grau de significância de $_{\text{p-valor}}$ (bilateral) <0, 001.

Segundo Callegari e Jacques (2007), o valor de *p-valor* <0,05 significa que se está assumindo uma probabilidade de apenas 5% de que a diferença encontrada no estudo não seja verdadeira. Quanto menor o valor de *p-valor*, menor será a probabilidade de isso acontecer. De uma forma geral, os resultados de um estudo podem variar de “não significante” até “extremamente significante”, como mostra a Tabela 8 a seguir.

Tabela 8: Significância estatística, conforme o *p- valor* (CALLEGARI E JACQUES, 2007).

Valor de P	Significado
>0,05	Não significante
0,01 a 0,05	Significante
0, 001 a 0,01	Muito significante
<0, 001	Extremamente significante

Analisando os valores de todas as tabelas do Apêndice E, obtidas pelos testes estatísticos, o *p-valor* foi menor 0,001, ou seja, segundo a Tabela 7, "extremamente significante". Pode-se concluir que a possibilidade de existir em todos os testes estatísticos realizados neste trabalho um erro do Tipo I é menor, 0,001, lembrando a definição dada no capítulo 2, na seção 2.6 que o erro tipo I consiste em rejeitar H_0 quando a hipótese é verdadeira. Assim, a hipótese H_0 foi rejeitada em todas as amostras dos experimentos, e aceita a hipótese H_1.

5.5 TRABALHOS CORRELATOS

Os trabalhos correlatos na área de agrupamento, comparados neste trabalho, estão relacionados nas Tabelas 9 e 10. As colunas das tabelas correspondem aos itens nos quais os trabalhos correlatos foram analisados, e cada uma de suas linhas corresponde aos trabalhos em ordem cronológica. A Tabela 10 enfatiza os resultados com uso das métricas externas e internas. Observa-se que nenhum dos trabalhos correlatos empregou as métricas internas.

Todos os trabalhos correlatos utilizados nesta tese tiveram seus protótipos procurados para realização dos testes. O único protótipo usado foi Eurekha de Wives (2004), disponibilizado em http://www.inf.ufrgs.br/~wives/wiki/doku.php?id=eurekha. Ressalva-se, entretanto, que o software não possibilitou a saída das medidas *Recall*, *Precision* e *F-Measure*, que tiveram de ser calculadas a partir da análise dos agrupamentos gerados, e do *corpus* TeMário de Rino e Pardo (2003), utilizados para os testes com Eurekha e com Cassiopeia, explicados na subseção 4.1.1.

Os resultados de Maria *et al.* (2008) e Hourdakis *et al.* (2010) foram obtidos dos próprios trabalhos aqui citados. O trabalho de Lopes (2004) forneceu a avaliação dos agrupamentos, através de visualização de dendrograma, mas o autor não usou qualquer tipo de medida, apenas análise dos dendrogramas. O trabalho de Ribeiro (2009) evidenciou somente a avaliação com a medida *Precision,* que obteve uma faixa entre os 0,42 a 0,53, dependendo do método de seleção de atributos usado nos experimentos. Ribeiro (2009) utilizou apenas essa medida, pois agrupou

documentos do resultado de uma consulta ao Google, com etiquetagem e escolha do número de grupos, usando amostragem e o conceito de estabilidade do agrupamento.

Wives (2004) usa agrupamento de textos com lógica *Fuzzy,* e como forma de representação de conhecimento textual utiliza a representação de conceitos. Segundo Wives (2004), conceito é a representação do conteúdo dos documentos, que aparecem como descritores, no processo de identificação de similaridade entre os textos e, consequentemente, para que sejam agrupados em assuntos similares. Wives (2004) consegue obter esses descritores através de vocabulário controlado ou padronizado. Seu trabalho mostra uma forte manipulação na fase de pré-processamento, e uma dependência muito forte do domínio no qual está trabalhando, e também nessa fase, o autor utiliza uma lista de *stopword*. No processamento, Wives (2004) usa a estrutura hierárquica aglomerativa, com algoritmo *Cliques* com operadores *Fuzzy* para o cálculo da similaridade.

Lopes (2004) utiliza uma biblioteca de algoritmo de agrupamento denominada *C-Clustering Library*, que possui os algoritmos *Clustering* Hierárquico, *Clustering K-means* e *Self-Organizing Maps*. Emprega a matriz de similaridade, na qual são realizados os cálculos com *Term Frequency*- TF, *Inverse Document Frequency* -IDF e *Term Frequency * Inverse Document Frequency* -TF-IDF e um processo de pós-processamento, que usa visualização de dendrograma baseado em matriz de cores. Existe forte dependência da fase de pré- processamento, que utiliza *stopwords*, substituição de sinônimos, *stemmer* para português. Nessa fase, Lopes (2004) afirma que garante o uso em outras línguas, entretanto devem ser mudados todos os arquivos. No processamento, é realizado cálculo da matriz de similaridade, e o emprego de algoritmos *clustering* hierárquico, *k - means* e SOM (*Self-Organizing Maps*). No pós-processamento, existe a visualização bastante complexa para identificar padrões e, consequentemente, obter algum conhecimento. O autor relata, ainda, a dependência para uma quantidade de termos a ser descrita para cada *corpus* para a melhora da taxa de acerto.

Maria *et al.* (2008) apresentam, em seu trabalho, uma ferramenta *denominada Clustering Toolkit* - CLUTO, que organiza os agrupamentos e define os conceitos. A autora escolhe um *corpus* e determina um domínio específico (esporte), pois afirma ser mais constante e reduzido. No pré-processamento, emprega a ferramenta FORMA (etiquetagem morfológica e lematização) e duas medidas, TFIDF e C-*Value,* para atribuir valor aos termos e selecioná-los em limiar de similaridade. Maria *et al.* (2008) descrevem alguns problemas na ferramenta CLUTO, que tem de definir o número de agrupamentos para análise dos conceitos, quando ocorre manualmente. Houve vários problemas em relação a agrupamentos gerados devido às medidas de similaridade. A autora afirma haver problemas de semântica em alguns agrupamentos gerados.

Ribeiro (2009) descreve em seu trabalho a utilização de seleção local de características em agrupamento hierárquico de documentos, com o emprego de *bisecting K-means* adaptado. Com isso, propõe o método ZOOM-IN, que seleciona termos a cada passo de divisão do algoritmo hierárquico divisivo. Partindo de uma seleção de grande proporção de termos relevantes a cada divisão, a quantidade de termos a ser selecionada é calculada a partir do tamanho de cada grupo. O autor afirma que o método ZOOM-IN, na escolha do número de características, baseada no tamanho de cada grupo, não mostrou bons resultados. Ribeiro (2009) afirma que o uso do tamanho dos grupos, para calcular a proporção de termos a serem considerados em cada divisão de grupo, demonstrou não funcionar bem para domínios variados. Existe a necessidade de selecionar somente os termos relevantes em cada divisão, para que o método local possa alcançar melhor desempenho que uma seleção global.

Hourdakis *et al.* (2010) adotam a ideia de agrupamento hierárquico para agrupar textos médicos. Utilizam *K-Means,* inicialmente, para separar a coleção em dois agrupamentos. Assim começa a divisão até chegar [SEM SENTIDO!] agrupamentos folhas. Onde não podem mais ocorrer divisões, fica um documento para cada agrupamento. Os autores utilizam o algoritmo *Bayesian Information Criterion* - BIC-Means, e suas variações, para avaliar, sobre algum critério de pontuação, as novas divisões e assim determinar o ponto de parada, evitando chegar ao agrupamento final com apenas um único documento, o que não é muito bom. Usam a pontuação BIC, aplicada localmente, como o critério de divisão de um conjunto de agrupamentos. Dessa forma, consegue-se mensurar a melhoria de um agrupamento, quando ele é dividido. Se a pontuação BIC dos dois novos agrupamentos for inferior à pontuação BIC de seu agrupamento antecessor, o algoritmo não aceita a divisão, pois o critério de coesão é violado, assim, o agrupamento antecessor passa a ser o agrupamento final, na hierarquia, não podendo mais ser dividido.

Os autores usam vetores com centroides para realizar as comparações entre os agrupamentos. Consequentemente, a BIC-*Means* termina quando não há agrupamentos que consigam ser separados, de acordo com BIC. Para o experimento, utilizaram uma coleção denominada *OHSUMED,* retirada da *Medline*, uma base de dados da biblioteca Nacional Americana de Medicina, adotada para comparar os resultados com alguns métodos variantes BIC, para criação de agrupamentos, como *Dynamic BIC-Means*, *BIC-Means and X-Means*. Para avaliação, utilizam um conjunto de 100 perguntas a serem realizadas após os agrupamentos terem sido formados, e avaliaram os conjuntos de textos recuperados, utlizando as medidas *Recall*, *Precision* e *F-Measure*. Os resultados foram satisfatórios, comparados com métodos exaustivos, como *Bisecting Incremental K-Means*. No trabalho, analisam o problema do reagrupamento, e

dizem que poderia melhorar com o método *Dynamic BIC-Means.* Como trabalho futuro, propõem uma variação *BIC-Means* que é o *G-means* algoritmos.

Tabela 9: Tabela comparativa domínio, idioma, complexidade de espaço e interação humana.

Autor (Ano)	***Pré-Processamento***	***Processamento***	***Domínio Idioma***	***Complexidade Espaço***	***Interação humana***
WIVES (2004)	*Stopword* + descritores de vocabulário controlado ou padronizado.	Agrupamento hierárquico aglomerativo por Conceitos.	Dependente	O(mn)	Sim
LOPES (2004)	*Stopword* + substituição de sinônimos e *stemme*r Português.	Biblioteca de agrupamento			
MARIA *et al.,* (2008)	Utilização da ferramenta FORMA, para lematização	Agrupamento + Conceitos			
RIBEIRO (2009)	*Stopword* + *stemmer* de Porter	Agrupamento hierárquico divisivo + seleção de características			
HOURDAKIS *et al.,* (2010)	Não existe descrição	Agrupamento Hierárquico com uso de centroides			
GUELPELI (2012)	Sumarização com níveis de compressão	Novo método de Luhn+ Agrupamento hierárquico aglomerativo+ Cliques+ centroides	Independente	O(n)	Não

Tabela 10: Tabela comparativa das métricas no idioma português e inglês e nos domínios jornalístico e médico

Autor (Ano)	***Idioma***	***Domínio***	**Avaliações Médias**					
			Métricas Externas			**Métricas Internas**		
			Precision	***Recall***	***F-Measure***	***Coesão***	***Acoplamento***	*Coeficiente Silhouette*
WIVES (2004)	Português	Jornalístico	0,35	0,58	0,44	*Não tem*		
LOPES (2004)	Português	Jornalístico	Avaliações de visualização de dendrograma			*Não tem*		
MARIA *et al.,* (2008)	Português	Jornalístico	0,41	0,50	0,45	*Não tem*		
GUELPELI (2012)	**Português**	**Jornalístico**	**0,35**	**0,34**	**0,35**	**0,22**	**0,22**	***0,91***
RIBEIRO (2009)	Inglês	Jornalístico	0,42 a 0,52	Não tem	Não tem	*Não tem*		
GUELPELI (2012)	**Inglês**	**Jornalístico**	**0,53**	**0,36**	**0,37**	**0,29**	**0,34**	***0.93***
HOURDAKIS *et al.* (2010)	Inglês	Médico	0,22	0,35	0,35	*Não tem*		
GUELPELI (2012)	***Inglês***	***Médico***	***0,75***	***0,48***	***0,52***	***0,29***	***0,27***	***0,98***

5.6 DISCUSSÃO DOS RESULTADOS

Esta seção tem como objetivo discutir os resultados como um todo, iniciando-se pela análise do primeiro experimento realizado no modelo Cassiopeia. Foram utilizadas as métricas externas, com as medidas *Recall*, *Precision* e *F-Measure* e as internas, com as medidas de Coesão, Acoplamento e Coeficiente Silhouette das Figuras 26, 27, 28, 29, 30, 31, 32, 33, 34 e 35.

Na medida *F-Measure*, observou-se que os agrupamentos resultantes do modelo Cassiopeia, na comparação com os textos-fonte, e com os demais textos sumarizados, obtiveram nas compressões com 50% e 70%, os resultados mais significativos para o aferimento dos agrupamentos na medida *F-Measure*. Notou-se ainda um decréscimo, a partir do aumento da compressão, o que pareceu ser coerente, poroque à medida que aumentava a compressão dos textos sumarizados, acontecia a perda da informatividade, ou seja, diminuía-se o número de palavras.

A medida Coeficiente de Silhouette teve um resultado idêntico à *F-Measure*. Com o aumento da compressão dos textos sumarizados, surgia uma perda da informatividade, o que se refletia nos agrupamentos resultantes do modelo Cassiopeia. Como no *F-Measure,* na medida Coeficiente Silhouette, observou-se que os agrupamentos resultantes do modelo Cassiopeia, na comparação com os textos-fonte e com os demais textos sumarizados, tinham as compressões com 50% e 70%, e os resultados mais significativos, para o aferimento dos agrupamentos, foram na medida Coeficiente Silhouette.

Em uma análise no nível de domínios, os resultados focalizados na pesquisa pareceram coerentes, uma vez que no domínio jurídico, a presença de palavras raras ou neológicas era mais frequente nos textos; a *performance* foi muito boa, mas no domínio médico não foi tão significativa.

Essa observação não se aplicou ao domínio pobre do léxico, cujas palavras comuns, com alta frequência nos textos, foram retiradas no processo de sumarização, mediante o aumento da compressão, como aconteceu no caso do domínio jornalístico. Isso ficou evidente, porque os textos já eram pequenos em relação à quantidade de palavras, em comparação, por exemplo, com o domínio jurídico ou médico. Ressaltou-se ainda, que os melhores resultados ocorreram no idioma inglês, devido à característica da língua inglesa e/ou à boa qualidade dos sumarizadores em inglês. No domínio médico, no idioma inglês, na medida *F-Measure* houve uma exceção dessa análise.

Os textos-fonte não sofreram, ao longo do experimento, qualquer tipo de redução, enquanto os textos sumarizados, à medida que as compressões iam aumentando, perdiam informatividade, ou seja, as palavras perdiam as sentenças. Constatou-se que, esse fato serviu para todos os sumarizadores, domínios e idiomas. Assim, foi vantajoso usar a sumarização com percentual de 50% a 70%, com ganho de qualidade para os agrupamentos. A partir dessa observação, os ganhos foram pontuais, ou seja, dependeram do idioma, do domínio, da compressão e da qualidade dos sumarizadores.

Por fim, outro fator extremamente relevante referiu-se à concordância entre as métricas externas e internas. Esses resultados foram averiguados entre as medidas *F-Measure* e Coeficiente Silhouette, e se apresentaram bastante positivos para o modelo Cassiopeia, já que tanto a avaliação humana (métricas externas) quanto a avaliação não supervisionada (métricas internas) ficaram próximas em seus resultados, o que garantiu uma boa avaliação do Cassiopeia.

Mediante os resultados obtidos com o modelo Cassiopeia, a contribuição de incluir, na etapa de pré-processamento, a sumarização de texto e um novo método dos limiares do corte de Luhn, no processo de agrupamento, melhorou o desempenho, tal como este é medido pelas métricas externas e internas, em seus agrupamentos

Essa contribuição foi observada, principalmente, no intervalo de 50% até 70%, com exemplos mostrados aqui neste trabalho e discutidos anteriormente. Ressaltou-se também que os resultados foram sensíveis a algumas variáveis, como compressão, domínio, idioma e, principalmente, qualidade dos algoritmos de sumarização, que influenciaram diretamente a *performance* dessas métricas no modelo Cassiopeia.

A qualidade dos sumarizadores constituiu outro ponto importante. Cada algoritmo de sumarização aplicado aos textos-fonte determinou seu desempenho no agrupamento, sendo assim, pôde-se ter ainda o modelo Cassiopeia como uma opção para avaliar a qualidade de cada sumarizador. A avaliação de sumarizadores apresentou um problema crucial, abordado na seção 2.5 do capítulo 2,. A exploração dessa hipótese mostrou-se promissora, já que havia poucas opções na área de sumarização para avaliar resultados de sumarização de forma automática. Como foram expostas, na seção 2.5, as opções se apresentaram extremamente custosas, assim, essa hipótese de avaliação foi colocada como trabalho futuro, na seção 6.3 do capítulo 6.

Observando os resultados do segundo experimento, a discussão apresentada nesta seção abrange o uso das métricas externas, com as medidas *Recall*, *Precision* e *F-Measure* e

das métricas internas, com as medidas de Coesão, Acoplamento e Coeficiente Silhouette, das Figuras 36, 37, 38, 39, 40, 41, 42, 43, 44 e 45.

Como comparação, no segundo experimento, foram utilizados métodos de seleção de atributos bastante usuais na literatura, como RDF, RTF, TFIDF, expostos, na seção 2.3 do capítulo 2. Foram comparados com o método de seleção de atributos do modelo Cassiopeia. O método foi dividido em dois, um com a proposta do modelo Cassiopeia, com uso da *stopwords,* e outro sem uso das *stopwords*. Cabe aqui ressaltar que a retirada das *stopwords* não foi a proposta principal do modelo Cassiopeia, mas foi realizada como uma estratégia, para que o experimento pudesse ser realizado, e será explicada adiante.

Avaliando os resultados das métricas externas, observou-se o desempenho dos métodos do modelo Cassiopeia como bom, no idioma português, obtendo mais de 50% de todas as amostras como o melhor resultado, estabelecendo assim, uma regularidade significativa em todas as amostras, em comparação com os métodos da literatura. No caso do idioma inglês, o resultado foi razoável, já que o método se equiparou aos métodos da literatura.

Nos resultados das métricas internas, nos quais se observou o uso da medida Coeficiente Silhouette, verificou-se um predomínio absoluto do método Cassiopeia, como o maior valor entre todas as amostras. O melhor método do Cassiopeia resultou do uso das *stopwords,* que obteve o maior valor em todas as amostras, ou seja, 100%. O resultado foi tão expressivo que o segundo melhor valor, em todas as amostras, foi do outro método do Cassiopeia, o *sem stopwords*.

Um fator que merece atenção especial é a comparação voltada para os dois métodos do Cassiopeia, com e sem *stopwords*. Houve, nas medidas Coeficiente Silhouette e *F-Measure* uma qualidade muito maior, quando se usou o método com *stopword*. Neste trabalho, em várias oportunidades, foi dito que a retirada das *stopwords* era uma técnica usada no pré-processamento para diminuir a alta dimensionalidade do espaço amostral. Na literatura, sua retirada sempre foi bastante significativa. Os vários autores citados argumentaram que o significado delas tinha pouca expressividade para a recuperação de informação, sendo assim, usaram uma lista para determinar sua exclusão. Ideia diferente de outras áreas que se apóiam no texto ou analisam textos para extrair seus significados e sentidos. Não há, nesses casos, como excluir tais palavras, por exemplo, um artigo “o” ou “a” que podem vir a determinar o gênero do substantivo seguinte, ou até mesmo apontar para algo que já fora citado anteriormente, na superfície textual. Essa questão da relevância das *stopwords* é controversa entre as diversas áreas, mas nesta área da computação já tinha sido levantada por Wives

(2004), citando o trabalho Riloff (1995) "*Little words can make big difference for text classification*".

Deixando de lado exemplos de questão semântica ou gramatical das *stopwords*, acredita-se, na análise feita neste trabalho, que a razão principal da retirada das *stopwords* dos trabalhos da literatura, seja por ser uma solução que pudesse viabilizar todo o processo computacional, sem trazer ganhos significativos para o agrupamento de texto, pois tornaria a técnica dependente da análise humana e, principalmente, criaria uma forte dependência com o idioma.

Os resultados apresentados no segundo experimento foram significativos para compreensão da importância ou não da retirada das *stopwords*. Como o modelo Cassiopeia, com uso da sumarização no pré-processamento, e do novo método para definição do corte de Luhn no processamento, viabilizou a manutenção das *stopwords,* pôde-se assim analisar a qualidade da retirada delas no processo de agrupamento. Pelos resultados obtidos, verificou-se que não houve ganho de qualidade no processo de agrupamento, na retirada das *stopwords*, haja vista que os resultados todos indicaram que o método Cassiopeia com *stopwords* foi melhor em todo o segundo experimento.

Assim, a sumarização usada no pré-processamento diminuiu consideravelmente o número de palavras, mantendo a qualidade delas, característica inerente da sumarização, e ainda viabilizou o processamento, pois reduziu a alta dimensionalidade e os dados esparsos. No processamento, a proposta do novo método do corte de Luhn, no modelo Cassiopeia, além de propiciar o benefício da independência do domínio e do idioma, trouxe qualidade para o processo de agrupamento, e proporcionou ainda um ganho superior, ao contrário da simples retirada das *stopwords*.

Nos testes estatísticos realizados neste trabalho, a hipótese nula H_0 foi rejeitada e assim, a hipótese H_1 foi aceita. Nos valores de todas as tabelas obtidas nos testes estatísticos, o *p-valor* foi menor 0, 001, ou seja, extremamente significante.

Finalizando, como foi mencionado no trabalho, o modelo Cassiopeia é não supervisionado. Os resultados na métrica interna foram bastante significativos, já que houve uma prevalência absoluta da mesma, nos dois experimentos. Conforme na seção 2.3, essa métrica utilizou apenas informações contidas nos grupos gerados para realizar avaliação dos resultados, ou seja, não foram utilizadas informações externas, sendo assim os resultados superiores da métrica interna em relação à métrica externa foram importantes para validar o modelo Cassiopeia como um bom modelo não supervisionado.

CAPÍTULO 6 – CONCLUSÕES

O estudo de informações textuais tem demonstrado ser um importante foco de pesquisas na área acadêmica , já que a internet a cada dia se consolida como principal veículo de distribuição e armazenamento de informações do mundo. Conhecer a fundo essas informções textuais, e criar ferramentas para melhorar a sua manipulação, tem sido importante diferencial para pessoas e empresas. Como esses repositórios de informações textuais crescem muito com a evolução da internet, torna-se necessária a utilização de técnicas que melhorem a recuperação desses textos. A RI, por exemplo, com uso de técnica de agrupamento, permite ao usuário explorar uma grande quantidade de informações, conhecer seu conteúdo, obter inter-relações e ter acesso a informações específicas nestes grandes repositórios.

Baseado em estudos de autores citados nos capítulos 1 e 2 deste trabalho, existe um percentual significativo de documentos textuais, escritos em Linguagem Natural. Esses documentos textuais, em grande parte, são armazenados em repositórios, sem preocupação com estruturas organizadas, que seriam importantes para facilitar a recuperação dessas informações. Dessa forma, esses textos apresentam ambiguidades, e uma série de problemas decorrentes da diferença do vocabulário usado nos textos e dos vários idiomas em que esses documentos são transcritos.

O modelo Cassiopeia, proposto neste trabalho, consiste no uso da sumarização no pré-processamento, que possibilita a redução da quantidade de atributos. Dessa forma, o modelo pode manter a lista de *stopwords,* tornando, assim, o modelo independente do idioma, com um novo método para definição do corte de Luhn, garantindo a independência do domínio. Os resultados obtidos demonstraram um grande avanço na qualidade dos agrupadores, usando o modelo Cassiopeia, em termos de precisão, recuperação de informação, coesão e acoplamento. A análise dos experimentos demonstra que a utilização da técnica de sumarização de texto, agrupamento hierárquico e do novo método do corte de Luhn, apresentam-se como uma boa alternativa para a solução do problema citado, oferecendo resultados quantitativamente semelhantes, ou melhores do que os tradicionais.

A conclusão pode ser discutida em três partes. A primeira diz respeito ao pré-processamento da etapa de RI. Os autores citados ao longo do trabalho confirmam as dificuldades de tratar, em seus trabalhos, esses textos não estruturados, e consideram essa fase crucial para o bom desenvolvimento do processamento. Sendo assim, este trabalho propôs o

uso da sumarização de texto como forma de reduzir, com informatividade, o número alto de atributos existentes nos textos desestruturados, contidos nos repositórios de informação.

Uma das soluções encontradas na literatura, para diminuição dos atributos (palavras) foi a utilização de uma lista de *stopwords*. O grande problema dessa técnica, discutida aqui, na subseção 3.1.1, é a dependência do idioma. Dessa forma, foi criado o modelo Cassiopeia, que usa a sumarização. Além de reduzir boa parte dessas *stopwords*, propicia uma redução de atributos (palavras) pouco significativos para o entendimento da ideia principal do texto original. O modelo Cassiopeia trabalha com algoritmos de sumarização, que realizam a compressão em diferentes percentuais. Isso foi verificado no experimento 1, por meio dos testes realizados, foram averiguados os percentuais ideais de compressão dos algoritmos de sumarização, baseados em seus domínios e no idioma, sem a perda da informatividade nos sumários gerados. Os textos foram substancialmente reduzidos, mas carregaram as palavras mais significativas com a ideia principal do texto-fonte. Foi assim que a fase pré-processamento do modelo Cassiopeia tratou a questão da alta dimensionalidade, um problema muito comum dentro da RI.

A segunda parte desta conclusão começa no processamento, fase que não pode ser discutida isoladamente. Afirmam os autores citados neste trabalho que essa fase é muito dependente do pré-processamento, ou seja, da necessidade de diminuição dos atributos (palavras). Para processar, há necessidade de um modelo para representar o espaço amostral. Uma solução muito comum e de grande consenso na literatura é o uso da matriz de similaridade como estrutura para representar os atributos (palavras). A técnica utilizada em outros trabalhos necessita, obrigatoriamente, do uso da lista de *stopwords*, na fase de pré-processamento, e, sem esse procedimento, torna-se inviável o uso da técnica de representação dos atributos (matriz de similaridade), sem esquecer o problema da matriz esparsa, cujos atributos (palavras) ocorrem apenas uma única vez, e em grande número.

No modelo Cassiopeia, a fase de processamento não utiliza uma matriz de similaridade, isto foi possível pelo uso da sumarização no pré-processamento. Os textos foram reduzidos, pelo menos em 50%, e sendo assim, viabilizou-se o uso de uma representação vetorial para os textos e para os agrupamentos.

Mesmo com a utilização das *stopwords*, o número de atributos continuou muito grande. Para sua representação nos sistemas, os autores utilizaram a identificação e a seleção dos atributos que, segundo vários trabalhos aqui apresentados, influenciaram diretamente na qualidade dos agrupamentos e, consequentemente, na boa *performace* da RI. As propostas identificaram e selecionaram as diversas formas de atributos, visando a atenuar o problema da

alta dimensionalidade e dos dados esparsos (seção 2.4). Para essa fase, o modelo Cassiopeia propôs o uso de uma identificação e seleção nova de atributos, representados na Figura 4, especificados no algoritmo 1, formalizados na equação 21 e discutidos na seção 3.2. O experimento 2 demonstrou a confrontação da seleção de atributos do modelo Cassiopeia, com a seleção de atributos dos outros trabalhos da literatura. Os resultados foram satisfatórios. Indicaram a qualidade nesta seleção, possibilitando assim a independência do domínio.

Na terceira e última parte dessa discussão, chega-se à fase do pós-processamento. Os textos, nas outras propostas, foram agrupados por algum critério de similaridade, já o modelo Cassiopeia apresentou esses textos com boas taxas de mensuração, obtidas com as métricas externas e internas nos experimentos 1 e 2. Acredita-se que a RI, quando utilizar esses agrupamentos, terá bons resultados, já que a métrica externa que usou as medidas como *Recall*, *Precision* e *FMeasure* apresentou bons resultados no agrupamento, e apresentará também na RI.

A partir dessas constatações, na seção seguinte, serão apresentadas as contribuições desta tese e do trabalho realizado durante o desenvolvimento dos experimentos e das análises. Na última seção, serão apontadas algumas possibilidades de trabalhos futuros, que poderão aprofundar ainda mais o tema e aperfeiçoar o estudo aqui apresentado.

6.1 CONTRIBUIÇÕES

A partir dos experimentos e das análises da utilização do modelo Cassiopeia, apresentados neste trabalho, podem-se destacar algumas contribuições importantes para a área de Recuperação de Informação (*Information Retrieval*), mais especificamente, para a área de agrupamento de texto:

. criação de um modelo que possibilite agrupar textos com qualidade, em bases textuais, em domínios distintos e/ou antagônicos;

- criação de um novo método para seleção de atributos;
- criação de um modelo para área de RI que possibilite independência do idioma e, consequentemente, a não interação humana;
- criação de um modelo que melhore o desempenho da coesão e do acoplamento dos agrupadores textuais;
- criação de um modelo que melhore o desempenho da precisão na recuperação de informação;
- redução, com o uso da sumarização, da quantidade de atributos e

- melhoria, com o uso da sumarização, da qualidade da seleção dos atributos

6.2 LIMITAÇÕES

O objetivo de qualquer modelo é ser o mais fiel possível, representando a maior parte do(s) objeto(s), porém os modelos são discretos e abstratos. Por esse motivo, qualquer modelo, mesmo os mais atuais, possuem algumas falhas e não conseguem representar por completo o mundo real. O modelo Cassiopeia não é exceção a essa regra. Da mesma forma, os agrupamentos, no modelo Cassiopeia, nem sempre representam fielmente a realidade e o desejo de cada usuário. Em seus agrupamentos, podem não representar corretamente todos os documentos agrupados. Porém, em se tratando de uma grande quantidade de dados que devem ser recuperados pelo usuário na RI, esses documentos, pelos experimentos realizados neste trabalho, trarão ganhos à área de RI, devido as mensurações obtidas, mas os valores obtidos ainda podem ser questionados, por razão de subjetividade,. Segundo Bovo (2011), o especialista no domínio precisa verificar a compatibilidade dos resultados com o conhecimento disponível do domínio. E, por fim, é o usuário que dará o julgamento final sobre a aplicabilidade dos resultados (BOVO, 2011). Por essa razão, salienta-se a importância de uma análise qualitativa dos resultados, não somente quantitativa, para todos os experimentos aqui realizados.

No trabalho foram usados apenas textos nos idiomas inglês e português e os corpora nos idiomas tinham 100 textos. Acredita-se que este seja um fator limitante, considerando que os textos poderiam ser em número maior e com maior diversificação dos idiomas..

6.3 TRABALHOS FUTUROS

O modelo Cassiopeia poderá, como implementação futura, incluir a fase de aprendizado autônomo. Acredita-se que a inclusão dessa fase venha a obter resultados ainda mais significativos para métricas externas e internas. Uma formalização dessa implementação futura está demostrada na Figura 47. Acredita-se que o aprendizado do percentual de compressão, e o número de palavras que compõem cada centroide dos agrupamentos, que atualmente formam um vetor de tamanho fixo de 50 posições, possa ser aprendido. Esse vetor sofre variações ao longo do processo de agrupamento, não usando todo o espaço reservado. Com aprendizado, consequentemete, levaria a uma diminuição do espaço usado pelo modelo Cassiopeia, causando um impacto na *performance* do algoritmo, no que diz respeito à sua utilização de memória. O aprendizado do percentual de compressão ajudaria a questão da alta dimensionalidade, atenuando ainda mais o problema e, consequentemente, melhorando as

questões de dimensionalidade dos centroides. Outro fator a ser aprendido é o momento de parada do reagrupamento. Atualmente, esses valores são fixos, mas poderiam ser aprendidos, o que possibilitaria uma diminuição no tempo de processamento do algoritmo, já que muitas vezes as palavras que formam os centroides não mais se alterariam no reagrupamento, mas continuariam o reagrupamento, devido aos seus parâmetros fixos. Acredita-se que esse aprendizado do tempo de parada do reagrupamento seria um fator impactante, na questão do processamento do modelo Cassiopeia.

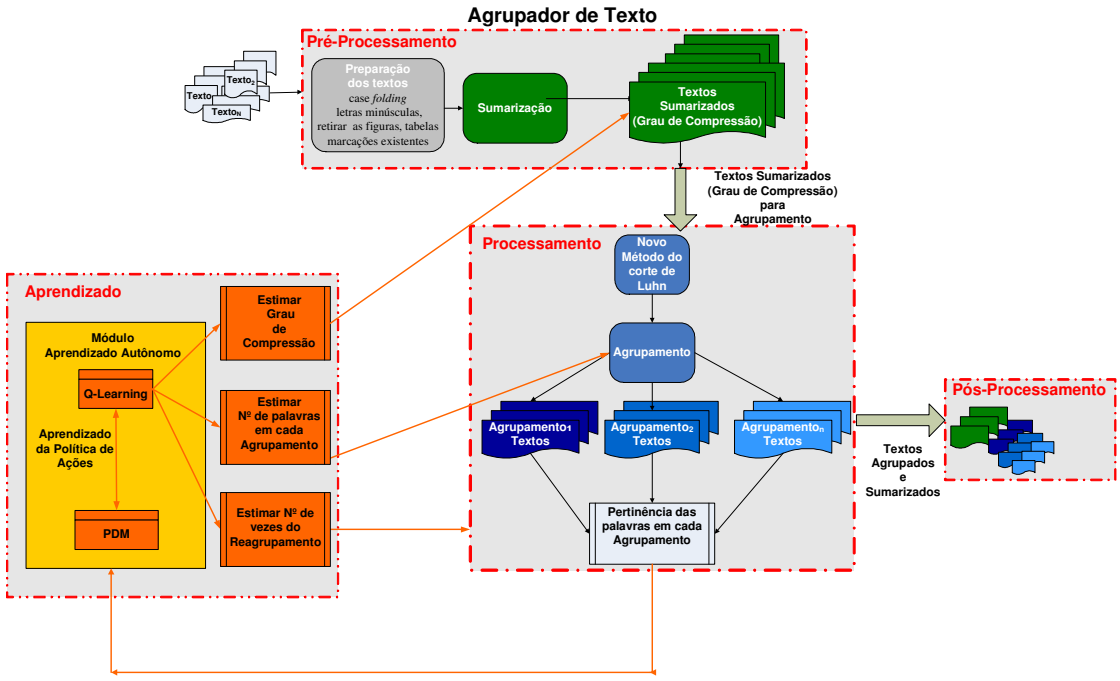


Figura 46: Proposta do modelo Cassiopeia com aprendizado autônomo.

Como foi discutido no experimento 1, e comprovado nos testes estatísticos, o modelo Cassiopeia gerou, como saída, um *rank* de sumários, ou seja, uma classificação ordinal desses algoritmos de sumarização. Atualmente, esse processo de avaliação automática, na área de sumarização, é realizado através da ferramenta *Recall-Oriented Understudy for Gisting Evaluation* – (ROUGE) de Hovy e Lin(1997) e/ou outra ferramenta, *Basic Elements* – (BE) de Tratz e Hovy (2008), uma evolução do ROUGE. O grande problema do ROUGE é a necessidade de que, para cada texto-fonte sumarizado, tem de existir um sumário humano de referência. Esse fator é bastante limitante, impede, muitas vezes, o uso de quantidade de *corpora* volumosos, já que são necessários um sumário humano para cada texto-fonte que compõe os *corpora*. Seria um trabalho humano bastante demorado e oneroso. Uma opção ao ROUGE poderia ser o BE, que tem o problema do *parsing*, que limita a ferramenta a um idioma. O modelo Cassiopeia mostrou uma solução viável de avaliação de sumários

automáticos, apresentando uma boa vantagem: não necessitar de sumários de referência humano e, muito menos, ficar preso a um idioma ou a um domínio.

Outro trabalho futuro que poderá ser realizado pelo modelo Cassiopeia, e que merece atenção, é no pós-processamento. Neste modelo, os índices dos agrupamentos (centroides) estão altamente estimados, ou seja, as palavras indexadas têm uma forte correlação com os textos ali agrupados, então, seria interessante usar alguma técnica para transformar as palavras em categorias. Essa identificação das categorias, através das características presentes em cada centroide de palavra dos seus agrupamentos, na literatura, é denominada *cluster analysis*, ou seja, análise de categoria, definido por Willet (1988). Futuramente, esse modelo poderá usar, no seu pós-processamento, a técnica de análise de categoria, proporcionando descoberta de conhecimento.

O modelo Cassiopeia inicialmente usou o sumarizador denominado Perfil. A proposta desse sumarizador tinha como base a não utilização do percentual de compressão, mas uma redução baseada na análise de cada texto. O Perfil não foi utilizado como um sumarizador para os testes, no modelo Cassiopeia, porque não havia como definir um percentual de compressão no seu algoritmo, e uma mudança desse porte iria descaracterizá-lo, mas esse sumarizador se encontra como parte integrante do modelo, porém não foi utilizado para os testes. Uma análise mais detalhada do Perfil encontra-se nos trabalhos de (GUELPELI e GARCIA, 2007), (GUELPELI *et al.*, 2008), (GUELPELI *et al.*, 2010) e (DELGADO *et al.*, 2010). Acredita-se que com os bons resultados do Perfil, e a capacidade desse sumarizador de definir um grau de compressão automático, com base na análise dos textos, possa fazer parte de um trabalho futuro, uma comparação dele com outros sumarizadores automáticos. É importante ressaltar que o uso do Perfil poderá viabilizar o aprendizado automático no modelo Cassiopeia, mostrado na Figura 66. Acredita-se, com a proposta apresentada na Figura 66, que o modelo Cassiopeia poderá melhorar o seu desempenho.

REFERÊNCIAS

AGGARWAL, C. C. **On the Effects of Dimensionality Reduction on High Dimensional Similarity Search.** In: ACM PODS 2001; 2001 May 21-23, 2001; Santa Barbara, CA: ACM Press; 2001.

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. Edition, Springer, Part I Clustering, pp 87- 108, 2007.

ALUÍSIO, S. M. e ALMEIDA, G. M. B. **O que é e como se constrói um corpus? Lições aprendidas na compilação de vários corpora para pesquisa lingüística.** Calidoscópio Vol. 4, n. 3 , p. 155-177, set/dez 2006 de Unisinos, 2006.

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.

ARMAN, B. K. E AKBARZADEH, M. R. T.**Automatic Text Summarization Using: Hybrid Fuzzy GA-GP** 2006 IEEE International Conference on Fuzzy Systems Sheraton Vancouver Wall Centre Hotel, Vancouver, BC, Canada, 2006.

BAEZA, Y. R. e RIBEIRO, N, B. **Modern information retrieval.** Boston: Addison Wesley,1999 http://people.ischool.berkeley.edu/~hearst/irbook/ Acesso em: 5 SET. 2009.

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. H**ow Much Information? 2010 Report on Enterprise Server Information.** Global Information Industry Center UC San Diego 9500 Gilman Drive, Mail Code 0519La Jolla, CA 92093-0519. http://hmi.ucsd.edu/howmuchinfo.php Acesso em: 12 AGO. 2011.

BOTELHO, G. M. **Seleção de Característica Apoiada por Mineração Visual de Dados.** Dissertação de Mestrado no Instituto de Ciências Matemáticas e de Computação ICMC-USP, São Paulo, Brasil, 2011.

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 Corporation, Tucson, Arizona, 2001.

CRESTANI, F; RIJSBERGEN, C. J. **A model for adaptative information retrieval.** Journal of Intelligent Information Systems, v.8, p.29-56 1997.

CROSS, V. **Fuzzy information retrieval.** Journal of Intelligent Information Systems, Boston, v.3, n.1, p.29-56, 1994.

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.

EDMUNDSON, H.P. **New Methods in Automatic Extracting.** Journal of the ACM. pp. 264-285, 1969.

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

FAN, W., WALLACE, L. RICH, S. and ZHANG, Z., **Tapping into the power of text mining**,Communications of the ACM, vol. 49, 2006.

FANEGO, I. C. **Hacia un modelo lingüístico de resumen automático de artículos médicos en español.** Tese de Doutorado Institut Universitari de lingüística aplicada Universitat Pompeu Fabra, Barcelona,2008.

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

FÁVERO, L. L. **Coesão e Coerência Textuais**. 9. ed. São Paulo: Ática, 2000.

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

FERREIRA, A. B. D. H. **Novo Dicionário Aurélio – Século XXI** . [S.l.]: Nova Fronteira, 1999.

FORMAN, G. **An Extensive Empirical Study of Feature Selection Metrics for Text Classification** Journal of Machine Learning Research 3 1289-1305

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.

GERSTING, J. L**. Fundamentos Matemáticos para a Ciência da Computação.** Editora LTC 3ª Edição, Rio de Janeiro,1995.

GET FINECOUNT Sotware produzido pela Tilti Systems versão 2.6 cuja última versão 10 de setembro de 2010.

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

GUELPELI, M.V.C.; Garcia A.C.B.**Automatic Summarizer Based on Pragmatic Profiles,** International Conference WWW/Internet 2007- IADIS- Volume II pg. 149-153- ISBN: 978-972-8924-44-7 - Outubro de 2007- Vila Real-Portugal .

GUELPELI, M. V. C.; BERNARDINI, F. C.; GARCIA, A. C. B. **Todas as Palavras da Sentença como Métrica para um Sumarizador Automático.** In: Tecnologia da Informação e da Linguagem Humana-TIL, WebMedia, 2008. p. 287-291,Vila Velha,Brasil, 2008.

GUELPELI, M.V.C.; GARCIA A.C.B. BERNADINI, F. C. **An Analysis of Constructed Categories for Textual Classification using Fuzzy Similarity and Agglomerative Hierarchical Methods.** Extend an invitation to you to publish a chapter in the upcoming book on "Emergent Web Intelligence" published by Springer Verlag in the series Studies in Computational Intelligence,2010

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

HASSEL, M. **Resource Lean and Portable Automatic Text Summarization,** PhD-Thesis, School of Computer Science and Communication, KTH, ISBN-978-917178-704-0,2007.

HEARST, M. A. **TextTiling: A quantitative approach to discourse segmentation.** Technical Report Sequoia 93/24, Computer Science Division, University of California, Berkeley, 1993.

HEARST, M. A. **TextTiling: Segmenting text into multi–paragraph subtopic passages**. Computational Linguistics, vol. 23, no. 1 pp. 33–64, 1997.

HOTHO, A.; NURNBERGER, A.; PAAß, G.; AIS, F. **A Brief Survey of Text Mining** In: LDV Forum - GLDV Journal for Computational Linguistics and Language Technology, Vol. 20, Nr. 1 (May 2005) , p. 19-62.

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.

HOVY, E.; LIN, C. **Automated Text Summarization in SUMMARIST.** In: I. Mani and M. Maybury (eds.) Intelligent Scalable Text Summarization ACL 1997 Workshop, pp. 39-46. Madrid, Spain, 1997.

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.

HU, X. R. e ATWELL, E. **A survey of machine learning approaches to analysis of large corpora**. School of Computing, University of Leeds, U.K.LS2 9JT,2003.

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

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=pt-BR&lr=&id=TRc2tBJrsd0C&oi=fnd&pg=PR11&dq=Readings+in+Information+Retrieval&ots=dg7ubtpJkh&sig=cz3HhQes5ryTqS6-IqT-rasxzdU#v=onepage&q&f=false.Acesso em: 15 JUN. 2010.

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.

KOWALSKI, G. **Information Retrieval Systems: Theory and Implementation** Boston: Kluwer Academic, 1997. p.282, 1997.

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

KUCERA, H and FRANCIS W. N. **Brown University Standard Corpus of Present-Day American Englis**h (or just Brown Corpus) as a general corpus (text collection) in the field of corpus linguistics. It contains 500 samples of English-language text, totalling roughly one million words, compiled from works published in the United States in 1961.

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.

LAROCCA, J. N., SANTOS,A. D. S, KAESTNER ,C. A.A. e FREITAS A. A. (2000). **Generating Text Summaries through the Relative Importance of Topics**. Lecture Notes in Computer Science Springer Berlin / Heidelberg Volume 1952/2000, ISSN0302-9743 (Print) 1611-3349 (Online) pp 300, 2000, Brazil. Acessado em 25-11-2011 http://wenku.baidu.com/view/da703be1524de518964b7da8.html?from=related

LAROSE, D. T. **Discovering Knowledge in Data: An Introduction to Data Mining**, 1st ed. Wiley-Interscience, 2004

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.

LEITE, D. S. **Um estudo comparativo de modelos baseados em estatísticas textuais, grafos e aprendizado de máquina para sumarização automática de textos em português.** Dissertação apresentada ao Curso de Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de São Carlos, SP, Brasil, 2010.

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

LIN, C.Y. e HOVY, E.H. **Automatic Evaluation of Summaries Using N-gram Cooccurrence Statistics**. In the Proceedings of the Language Technology Conference. Edmonton, Canada, 2003.

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

LOH, S.; WIVES, L. K.; AMARAL, L. A.; OLIVEIRA J. P. M. **Descoberta de Conhecimento em Textos através da Análise de Sequências Temporais.** Workshop em Algoritmos e Aplicações de Mineração de Dados, WAAMD, II; SBBD, 2006, Florianópolis, ISBN 85-7669-088-8. Florianópolis: Sociedade Brasileira de Computação, 2 p. 49-56, 2006.

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.

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.

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

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: 12 ABR. 2009.

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

MARCU, D. **From Discourse Structures to Text Summaries.** In I. Mani and M. Maybury (eds.), Proc. of the Intelligent Scalable Text Summarization Workshop, pp. 82-88. ACL/EACL'97 Joint Conference. Madrid, Spain, 1997.

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 para um Sumarizador Automático. Vila Velha : WebMedia, 2008. p. 359-363.

MAYBURY, M. **Automated Event Summarization Techniques.** In: Seminar Report of Summarizing Text for Intelligent Communication Seminar. Dagstuhl, Germany, 1993.

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.

MITTAL,V. O.; KANTROWITZ,M.; GOLDSTEIN,J.; CARBONELL,J. G. Selecting Text Spans For Document Summaries: Heuristics And Metrics. In Aaai/Iaai (1999), Pp. 467-473, 1999.

MÓDOLO, M. **SuPor: um Ambiente para a Exploração de Métodos Extrativos para a Sumarização Automática de Textos em Português** Dissertação de Mestrado. Departamento de Computação, UFSCar. São Carlos - SP, 2003.

MOSTAFA, K.; RAZAVIAN, N. S.; OROUMCHIAN, F. RAZI, H. S. **Document Representation and Quality of Text: An Analysis** Book survey of text mining: clustering, classification, and retrieval Second . Editors BERRY, M. E CASTELLANO, M. Edition, Springer, Part I Clustering, pp 219- 231, 2007.

MOURA, M. F. e REZENDE, S.O. Proposta e Experimentação de Modelos de Rotulação para Agrupamentos Hierárquicos de Documentos. LABIC - Laboratory of Computational Intelligence Technical Reports nº RT_302, ICMC-USP, São Carlos - SP, 2007.

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.

NOGUEIRA, B. M. e REZENDE, S. O. **Avaliação de métodos não-supervisionados de seleção de atributos para Mineração de Textos** In: VII Concurso de Teses e Dissertações em Inteligência Artificial (CTDIA 2010) - São Bernardo do Campo, SP. v. 1. p. 1-10, 2010.

OLIVEIRA, H. M. **Seleção de entes complexos usando lógica difusa.** Dissertação (Mestrado em Ciência da Computação) – Instituto de Informática, PUC-RS, Porto Alegre, 1996.

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.; ESPINA, A.P.; RINO, L.H.M.; MARTINS, C.B. **Introdução à Sumarização Automática.** Tech. Report RT-DC 002/2001, Departamento de Computação, Universidade Federal de São Carlos. Abril. 38p , 2001.

PARDO, T.A.S.; RINO, L.H.M.; NUNES, M. G. V. **GistSumm: a summarization tool based on a new extractive method**. PROPOR'03 Proceedings of the 6th international conference on Computational processing of the Portuguese language, Faro, Portugal, 2003.

PARDO, T.A.S. E RINO, L.H.M. **TeMário: Um Corpus para Sumarização Automática de Textos** Série de Relatórios do NILC. NILC-TR -03-09, 2003.

PARDO, T.A.S.; RINO, L.H.M.; **A Coleção TeMário e a Avaliação de Sumarização Automática.** Série de Relatórios Série de Relatórios do NILC. NILC-TR-06-04, 2006.

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. **GistSumm – GIST SUMMarizer: Extensões e Novas Funcionalidades**

Série de Relatórios do NILC. NILC-TR-05-05, 2005.

PARDO, T.A.S.; NUNES, M.G.V.; FILHO, P.P.B.; UZÊDA,V.R. **Estrutura textual e multiplicidade de tópicos na sumarização automática: o caso do sistema GistSumm** .Série de Relatórios do NILC. NILC-TR-10-06,2006.

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.

PAICE, C.D., JONES, P.A. (1993). **The identification of important concepts in highly structure technical papers.** In R. Korfaghe, E. Rasmussen, and P. Willett (eds.), Proc. of the 16th ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 69- 78. ACM Press, June, 1993.

POLLOCK, J.J.; ZAMORA, A. **Automatic Abstracting Research at Chemical Abstracts Service.** Journal of Chemical Information and Compute Sciences 15(4): 226-232, 1975.

PYLE, D.**Data Preparation for Data Mining.** Morgan Kaufmann, San Francisco, CA, 1999.

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.

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.

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

RILOFF, E. **Little words can make big difference for text classification.** In: ANNUAL International ACM-SIGIR Conference on Research and Development in Information Retrieval, SIGIR, 1995. Proceedings New York: ACM Press, 1995. p.130-136.

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.

REZENDE, S. O. , MARCACINI, R. M. e 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) p. 7-21, 2011.

ROSELL, M.: **Text Clustering Exploration – Swedish Text Representation and Clustering Results Unraveled.** PhD thesis, School of Computer Science and Communication, Royal Institute of Technology, Stockholm, Sweden (2009).

SALTON, G. e MACGILL, J. M. **Introduction to Modern Information Retrieval**. New York: McGraw-Hill, 1983.

SALTON, G., WONG, A., e YANG, C. S. **A Vector Space Model for Automatic Indexing.** in Readings in Information Retrieval, K.Sparck Jones and P.Willet, eds.,Morgan Kaufmann Publishers, Inc., San Francisco,1997.

SARDINHA, T. B. **Lingüística de *corpus*: histórico e problemática.** D.E.L.T.A., Vol. 16, N.º 2, 2000 (323-367), 2000.

SAYÃO, M. **Verificação e Validação em Requisitos: Processamento da Linguagem Natural e Agentes.** Tese de Doutorado, PUC, Rio de Janeiro, Abril 2007.

SHAW, B. **Buildng a Better Folksonomy: Web-based Aggregation of Metadata.** Technical Report, 2005.

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. U**ser Modeling and User-Adapted Interaction, Springer ISSN 0924-1868- DOI 10.1007/s11257-004-5270-4.v. 14, n. 5, p. 383-423, 2004.

SPARCK J. K. **Automatic Summarizing: factors and directions.** In I. Mani and M. Maybury (eds.), Advances in automatic text summarization, The MIT Press, pp. 1-12, 1999.

SUBASIC, P. e HUETTNER, A. **Affect Analysis of Text Using Fuzzy Semantic Typing** IEEE Transactions on Fuzzy Systems, Special Issue, 2001.

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

TRATZ, S., Hovy, E.H. . **Summarization Evaluation Using Transformed Basic Elements**. Proceedings of Text Analytics Conference *(TAC-08).* NIST, Gaithersburg, MD, 2008.

TEXT ANALYSIS CONFERENCE **Relatório 2005**. Estados Unidos da América, USA. 8p.

TEXT ANALYSIS CONFERENCE **Relatório 2006**. Estados Unidos da América, USA. 10p.

TRAINA, A. J. .M. e SILVA,C. Y. V. W. **Recuperação de Imagens Médicas por Conteúdo Utilizando Wavelets e PCA.** XII Congresso Brasileiro de Informática em Saúde - CBIS, Porto de Galinhas, Pernambuco, 2010.

VAN RIJSBERGEN, C. J. **Information Retrieval**, Book 2nd ed. London: Butterworths, 1979.

VAN RIJSBERGEN, C. J. **Probabilistic retrieval revisited**. The Computer Journal, Journal: The Computer Journal - CJ , vol. 35, no. 3, pp. 291-298, 1992.

VENTURA, J. M. J. **Extracção de Unigramas Relevantes. D**issertação apresentada na Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa para obtenção do grau de Mestre em Engenharia Informática,Lisboa,Portugal, 2008.

VIANNA, D. S. **Heurísticas híbridas para o problema da filogenia.** Tese de doutorado, Pontifícia Universidade Católica - PUC, Rio de Janeiro, 2004.

WARTIK S. P. **Boolean Operations**. In Collection of Information Retrieval: Data Structures & Algorithms, p.264-292, 1992.

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

WITTEN I. H., FRANK, E. **Data Mining: Practical Machine Learning Tools and Techniques**, 2nd ed. San Francisco, CA: Morgan Kaufmann, 2005

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.

ZIPF, G. K. **Human Behavior and the Principle of Least Effort**. Oxford, England: Addison-Wesley Press. (1949). xi, 585 pp, 1949.

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

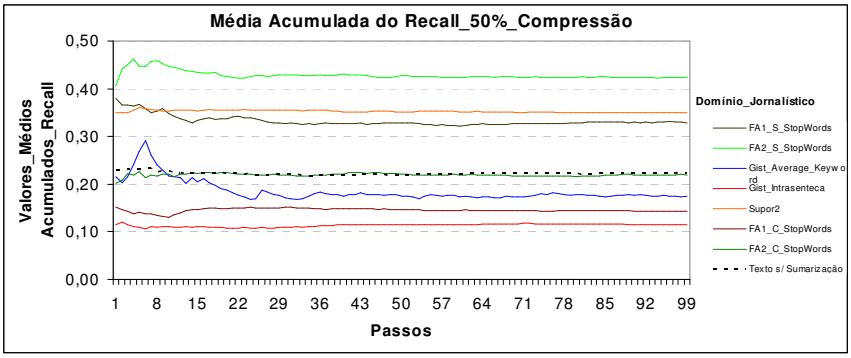
MÉTRICA EXTERNA COM AS MEDIDAS: *RECALL* E *PRECISION*

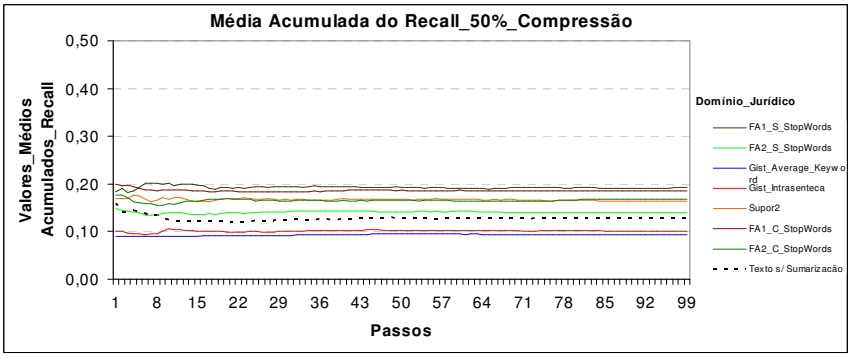
O Apêndice A mostra a continuidade dos resultados obtidos na primeira parte dos experimentos descritos na subseção 5.1.1, onde foi apresentada a medida *F-Measure*. Com as medidas *Recall* e *Precision*, fazem parte do conjunto com *F-Measure* (que é medida harmônica do *Recall* e do *Precision*) da métrica externa. Como forma de organização, no Apêndice A, foram realizados as mesmas comparações descritas na subseção 5.1.1.1. e os resultados foram apresentados com as compressões de 50%, 70%, 80% e 90%. Os textos escolhidos pertencem aos domínios, jornalístico, jurídico e médico nos idiomas português e inglês.

As figuras seguem a mesma numeração estabelecida para a medida *F-Measure*. O diferencial aparece com a letra "a" depois da numeração que representa a figura que mostra a medida *Recall* e a letra "*b*" para representar a medida *Precision*.

USO DA COMPRESSÃO DE 50% NO IDIOMA PORTUGUÊS

O melhor resultado da medida *Recall*, mostrado na figura 10a, foi no domínio jurídico cuja maioria dos algoritmos aumentou o valor em comparação com o valor de *Recall* dos textos fontes, com exceção dos algoritmos da literatura *Gist Intrasentença e Gist Average Keyword*. No domínio jornalístico, aparecem quatro algoritmos abaixo dos textos fontes. Dois algoritmos da literatura, o *Gist Intrasentença e* o *Gist Average Keyword*, e as duas funções aleatórias, FA1 e FA2, ambas mantendo as *stopwords*. O domínio médico foi o que obteve o pior resultado nesta medida, apenas o algoritmo *Gist Average Keyword* da literatura está com seu valor de *Recall* superior ao dos textos fontes.





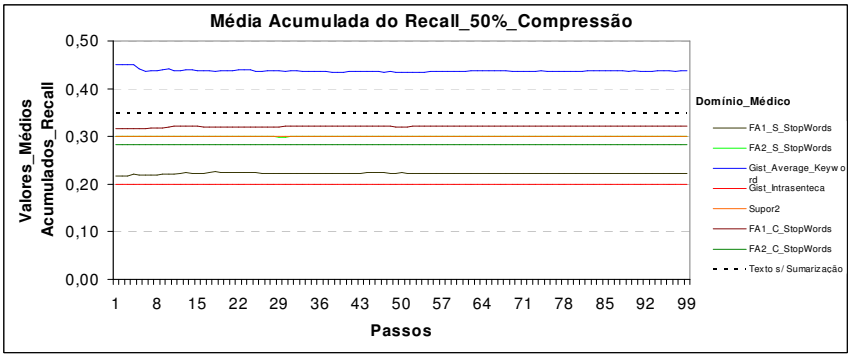


Figura 10a: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Recall* com 50%compressão, no idioma português.

O melhor resultado da medida *Precision*, mostrado na Figura 10b, foi no domínio jurídico cujos algoritmos aumentaram o valor em comparação ao valor de P*recision* dos textos fontes. No domínio jornalístico existem quatro algoritmos que estão abaixo dos textos fontes. Dois algoritmos da literatura o *Gist Average Keyword* e o *SuPor* e as duas funções aleatórias FA2. O domínio médico apresenta quatro algoritmos abaixo dos textos fontes. Os algoritmos da literatura, *Gist Intrasentença, Gist Average Keyword e SuPor*, e a função aleatória FA2 com *stopwords*.

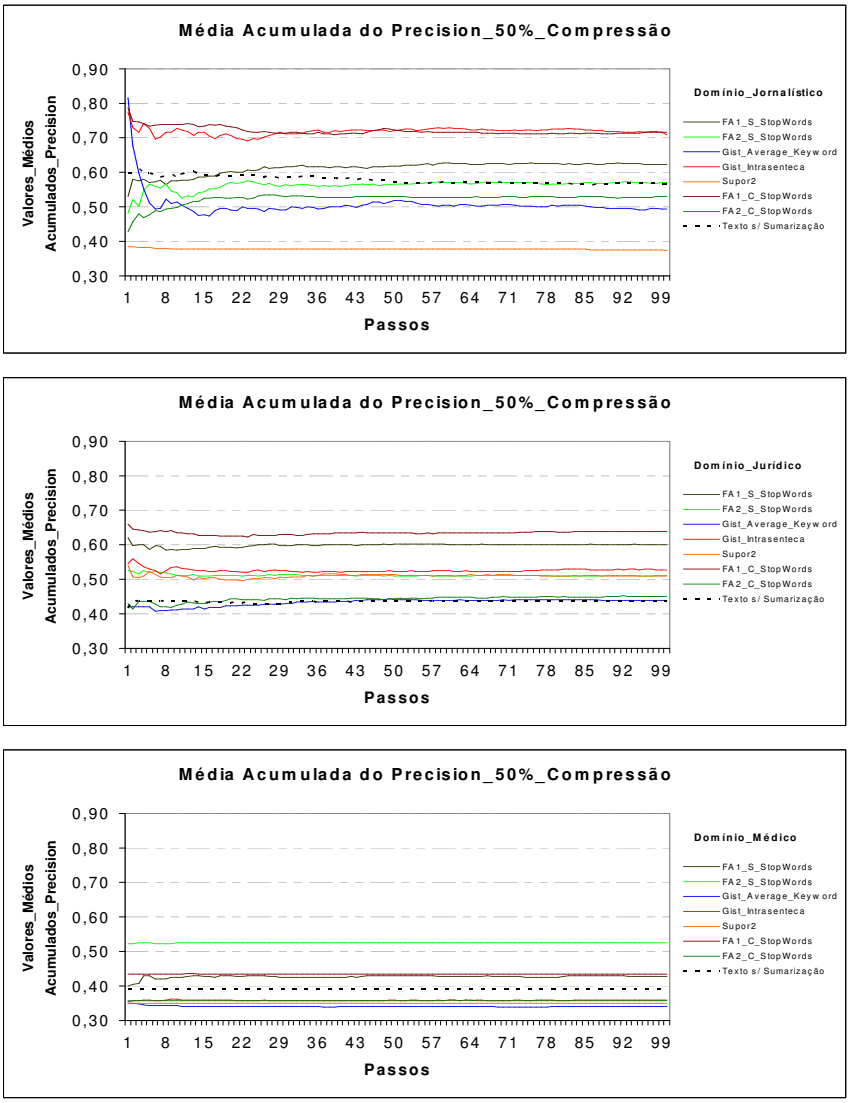


Figura 10b: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Precision* com 50%compressão, no idioma português.

USO DA COMPRESSÃO DE 70% NO IDIOMA PORTUGUÊS

O melhor valor obtido pela medida *Recall*, apresentado na figura 11a, foi no domínio jurídico cuja maioria dos algoritmos aumentou o valor em comparação com o valor de *Recall* dos textos fontes, com exceção dos algoritmos da literatura, o *Gist Intrasentença e* o *SuPor*. No domínio jornalístico existem cinco algoritmos abaixo dos textos fontes. Os três algoritmos da literatura, o *Gist Intrasentença,* o *Gist Average Keyword* e o *SuPor*; as duas funções aleatórias, FA1 e FA2, ambas sem as *stopwords*. No domínio médico aparecem quatro algoritmos abaixo dos textos fontes. Dois algoritmos da literatura, o *SuPor e* o *Gist Average Keyword*, e as duas funções aleatórias, FA1 e FA2, sem as *stopwords*.

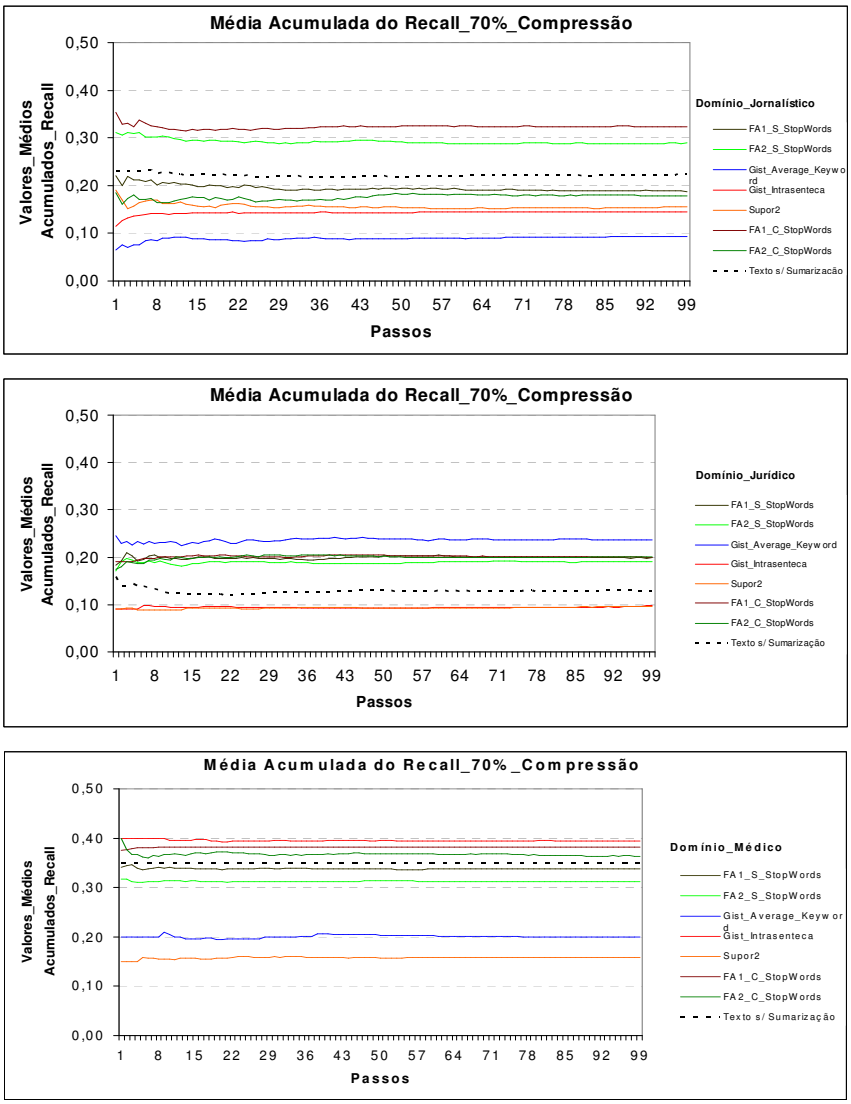


Figura 11a: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Recall* com 70%compressão, no idioma português.

A Figura 11b mostra que o melhor resultado da medida *Precision* foi referente ao domínio jurídico, cuja maioria dos algoritmos aumentou seus valores em comparação com o valor de *Precision* dos textos fontes, com exceção do algoritmo FA1 sem as *stopwords*. Para o domínio jornalístico e médico, dentro dos resultados da medida *Precision*, os resultados foram parecidos, ou seja, apenas um algoritmo ficou abaixo dos textos fontes em cada domínio, respectivamente o *Gist Average Keyword* e a FA1, sem as *stopwords.*

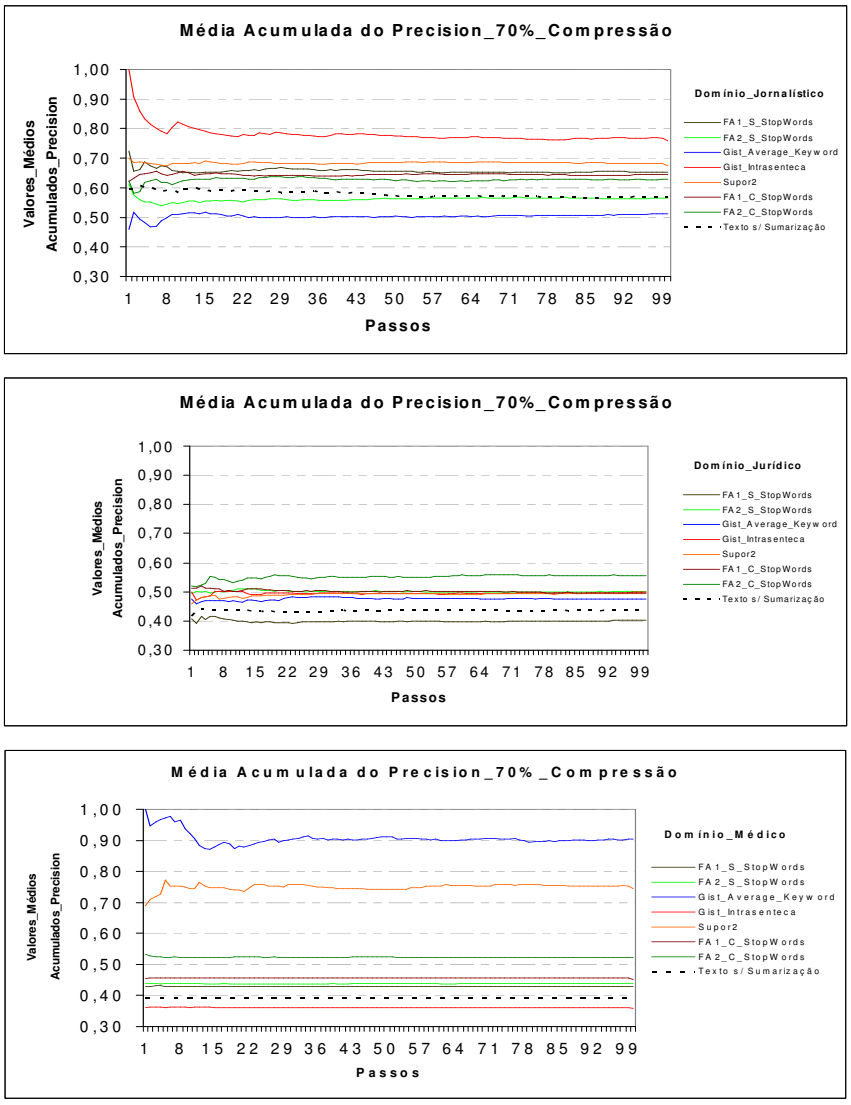


Figura 11b: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Precision* com 70%compressão, no idioma português.

USO DA COMPRESSÃO DE 80% NO IDIOMA PORTUGUÊS

Na Figura 12a, o melhor resultado da medida P*recision* foi no domínio jurídico, cuja maioria dos algoritmos aumentou seus valores em comparação com o valor de *Precision* dos textos fontes, com exceção do algoritmo da literatura, *Gist Intrasentença.* No domínio jornalístico são vistos três algoritmos, abaixo do valor dos textos fontes. Os dois algoritmos da literatura, o *Gist Average Keyword* e o *SuPor*, e uma função aleatória, a FA1 com *stopwords*. No domínio médico, existem três algoritmos abaixo do valor dos textos fontes. Um algoritmo da literatura, o *SuPor* e as duas funções aleatórias, FA1.

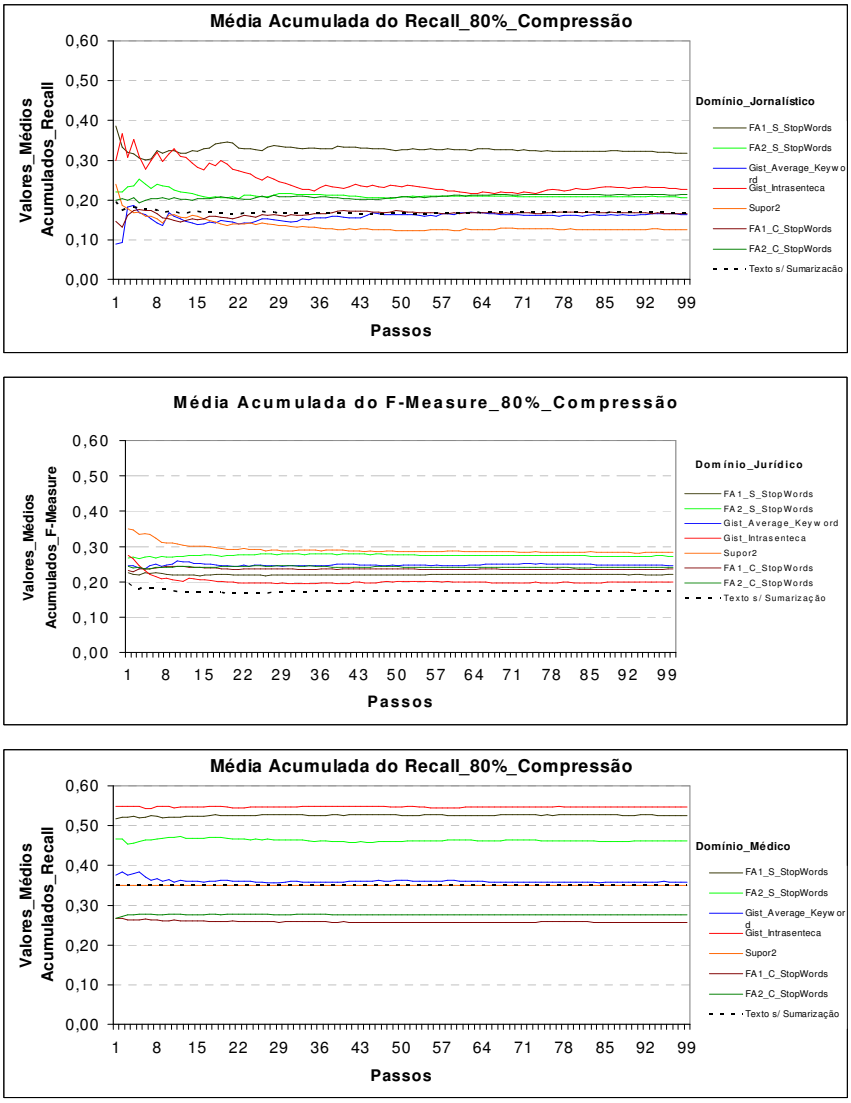


Figura 12a: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Recall* com 80%compressão, no idioma português.

A medida Precision mostrada na Figura 12b teve, no domínio jurídico, o melhor resultado, cuja maioria dos algoritmos aumentou seus valores em comparação com o valor de *Precision* dos textos fontes, com exceção do algoritmo da função aleatória FA2 com *stopwords.* No domínio jornalístico aparecem dois algoritmos abaixo do valor dos textos fontes, os dois algoritmos da literatura, o *Gist Intrasentença e* o *Gist Average Keyword*. No domínio médico, existem dois algoritmos que estão acima do valor dos textos fontes, as duas funções aleatórias FA1 e FA2, sem *stopwords*.

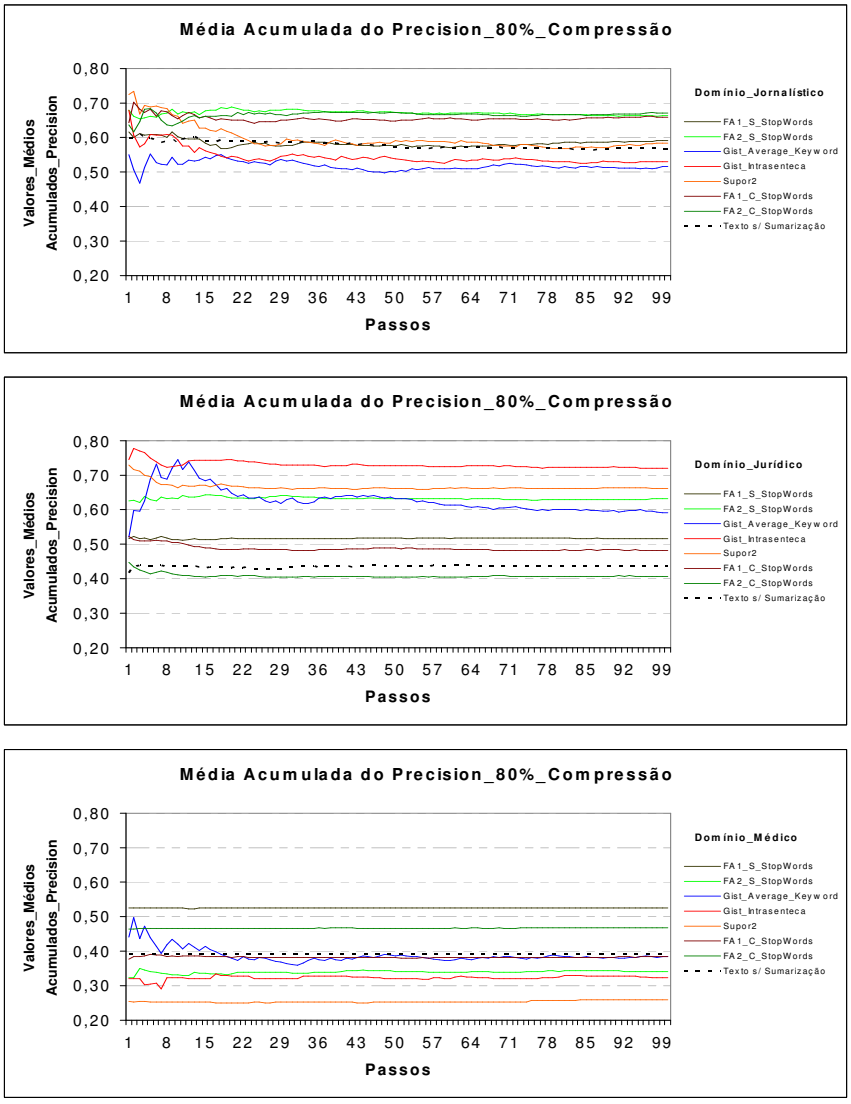


Figura 12b: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Precision* com 80%compressão, no idioma português.

USO DA COMPRESSÃO DE 90% NO IDIOMA PORTUGUÊS

Na Figura 13a, o melhor resultado da medida *Recall* foi no domínio jurídico cujos algoritmos aumentaram os valores em comparação ao valor de *Recall* dos textos fontes. No domínio jornalístico, existem dois algoritmos abaixo do valor dos textos fontes. São os algoritmos das duas funções aleatórias FA1 e FA2, ambas sem as *stopwords*. O domínio médico foi o que obteve o pior resultado nessa medida, apenas dois algoritmos *Gist Intrasentença* da literatura e a função FA2 sem *stopwords*, estão com seus valores de *Recall* superiores ao dos textos fontes.

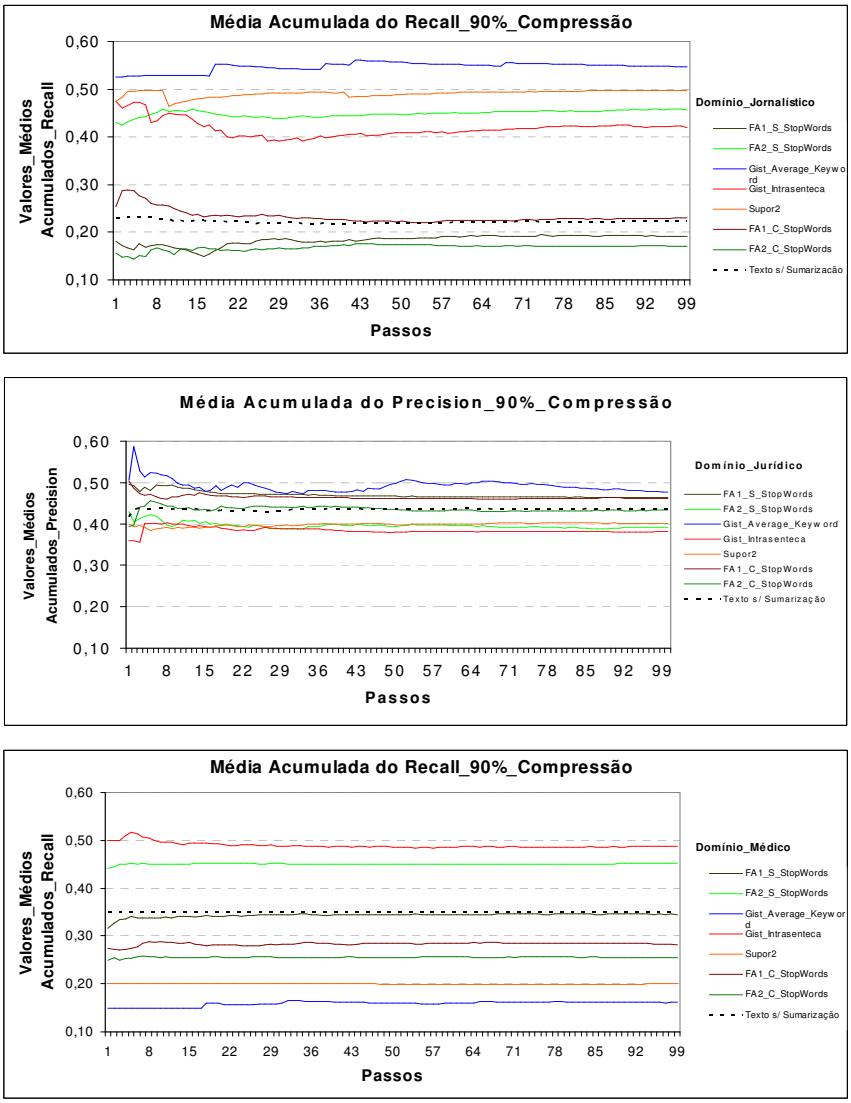


Figura 13a: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Recall* com 90%compressão, no idioma português.

Na Figura 13b a medida *Precision* teve, no domínio jornalístico, três algoritmos abaixo dos valores comparados com o valor de *Precision* dos textos fontes. Esses algoritmos são os três pertencentes à literatura, o *Gist Intrasentença,* o *Gist Average Keyword e* o *SuPor.* No domínio jurídico, existem três algoritmos que ficaram abaixo do valor dos textos fontes. São os algoritmos das duas funções aleatórias a FA1 e o *Gist Average Keyword*, algoritmo da literatura.

O melhor comportamento da medida *Precision* foi no domínio médico. Nesse domínio aparecem dois algoritmos abaixo do valor dos textos fontes. São os algoritmos da função aleatória FA2 sem *stopwords* e algoritmo da literatura, o *Gist Intrasentença.*

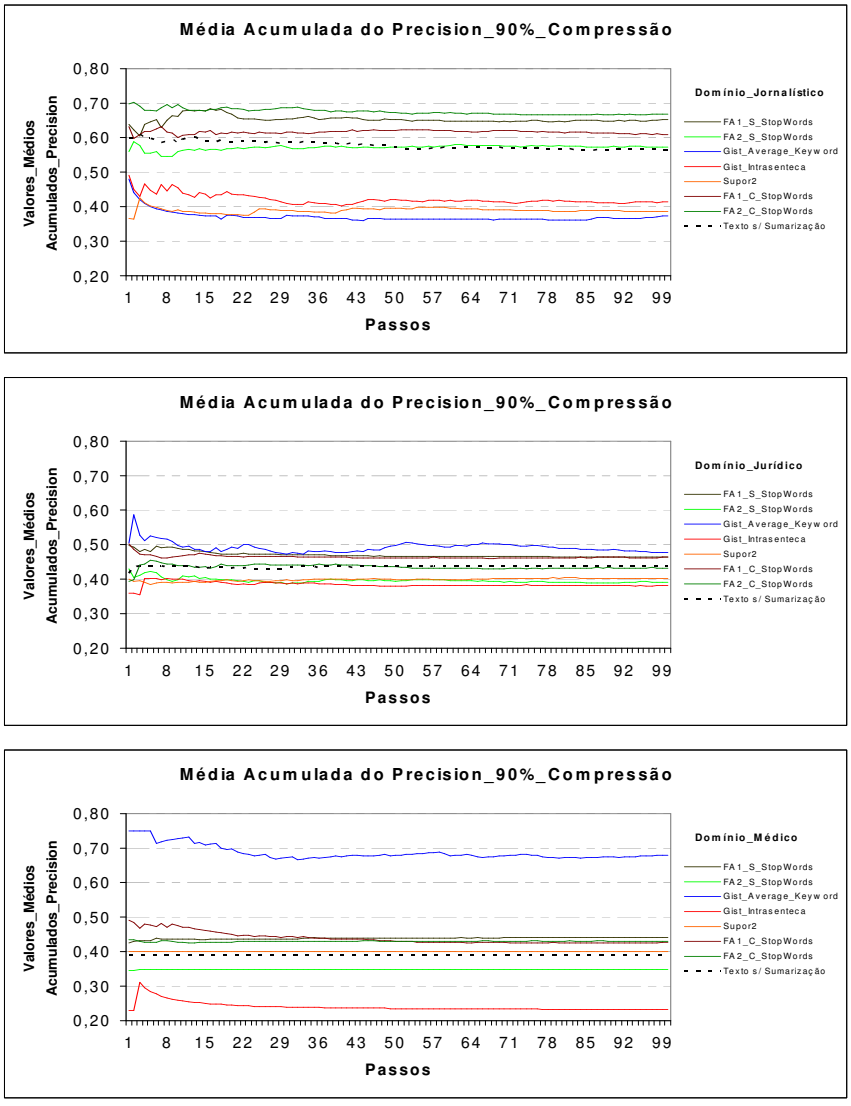


Figura 13b: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Precision* com 90%compressão, no idioma português.

USO DA COMPRESSÃO DE 50% NO IDIOMA INGLÊS

O melhor resultado da medida *Recall*, mostrado na Figura 14a, foi no domínio jornalístico cujos algoritmos aumentaram os valores em comparação ao valor de *Recall* dos textos fontes, exceto um algoritmo, o *Copernic*. No domínio médico, apenas um algoritmo teve seu valor de *Recall* superior aos textos fontes, foi o algoritmo *Intellexer Summerizer*.

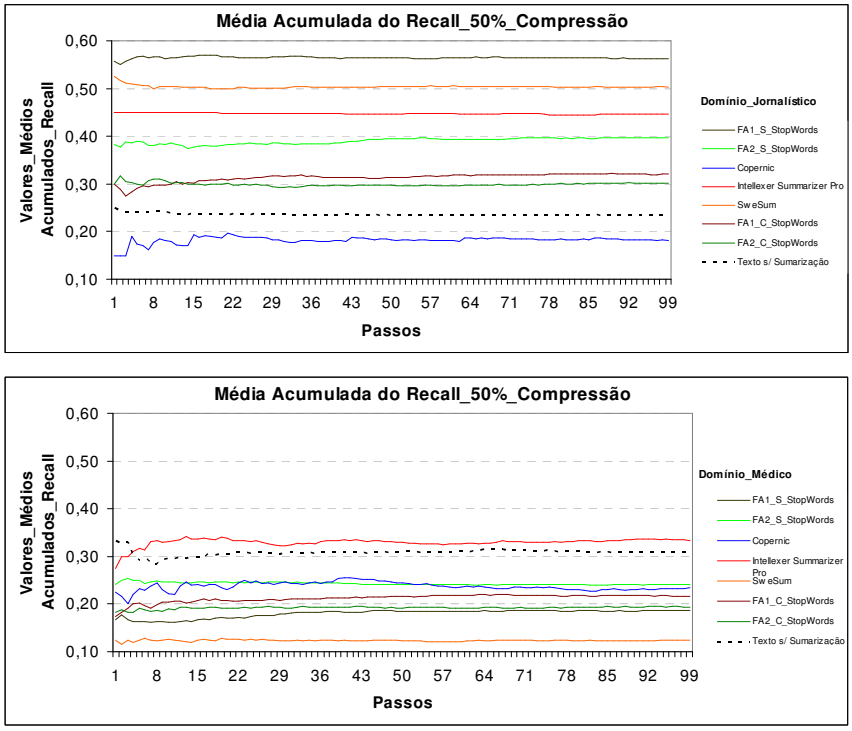


Figura 14a: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Recall* com 50%compressão, no idioma inglês.

Na Figura 14b, o melhor resultado da medida *Precision* foi no domínio jornalístico cujos algoritmos aumentaram seus valores em comparação ao valor de *Precision* dos textos fontes, exceto um algoritmo, *SweSum.* Para o domínio médico, nenhum dos algoritmos de sumarização tiveram seus valores superiores ao dos textos fontes.

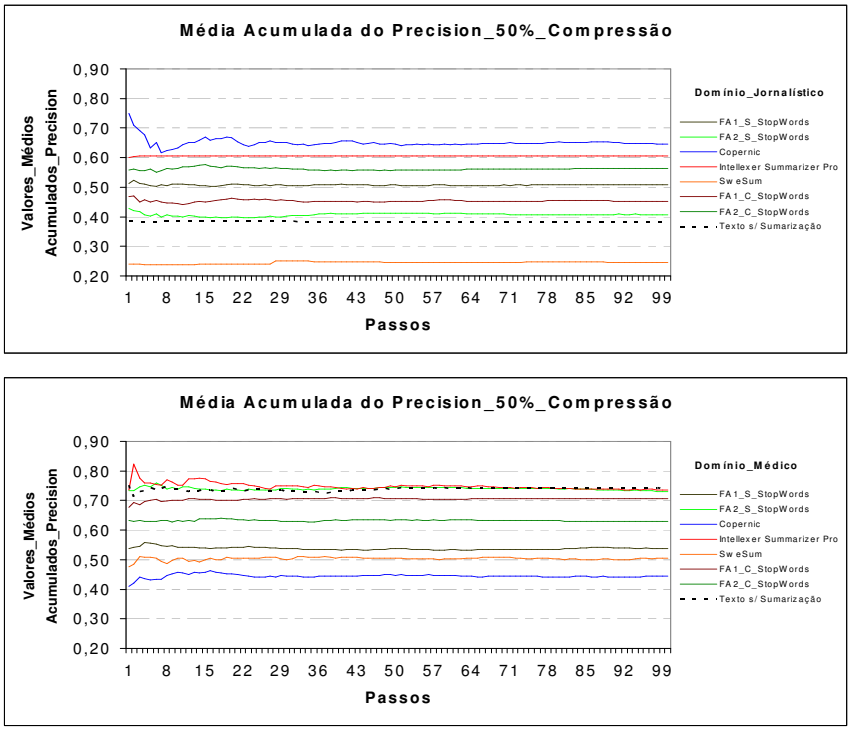


Figura 14b: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Precision* com 50%compressão, no idioma inglês.

USO DA COMPRESSÃO DE 70% NO IDIOMA INGLÊS

O melhor resultado da medida *Recall*, mostrado na figura 15a, foi no domínio jornalístico, cujos algoritmos aumentaram os valores em comparação ao valor de *Recall* dos textos fontes. No domínio médico, apenas um algoritmo teve seu valor de *Recall* superior ao valor dos textos, o algoritmo o *Copernic*.

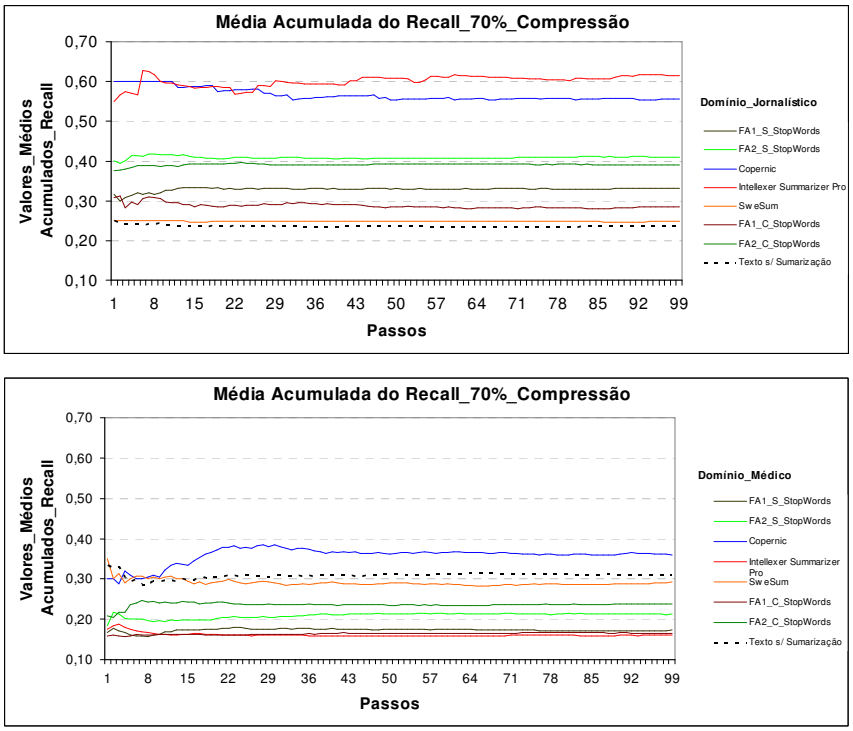


Figura 15a: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Recall* com 70% compressão, no idioma inglês.

Na Figura 15b, a medida *Precision* teve no domínio jornalístico dois algoritmos abaixo dos valores dos textos fontes. Esses algoritmos são o *Intellexer Summarizer e* o *SweSum*. No domínio médico, apenas um algoritmo teve seu valor de *Precision* superior ao valor dos textos fontes, o algoritmo *Intellexer Summarizer.*

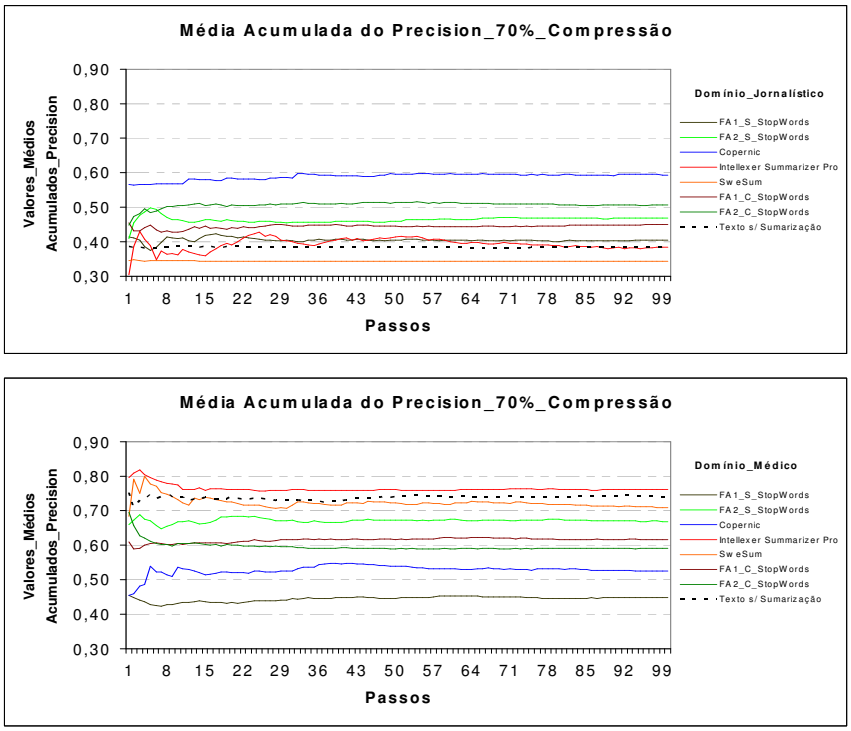


Figura 15b: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Precision* com 70%compressão, no idioma inglês.

USO DA COMPRESSÃO DE 80% NO IDIOMA INGLÊS

Na Figura 16a, a medida R*ecall* teve no domínio jornalístico um algoritmo abaixo do valor comparado com o valor de *Recall* dos textos fontes. Esse algoritmo é o *Intellexer Summarizer*. No domínio médico, apenas um algoritmo atingiu seu valor de *Recall* superior ao dos textos fontes, o algoritmo *SweSum.*

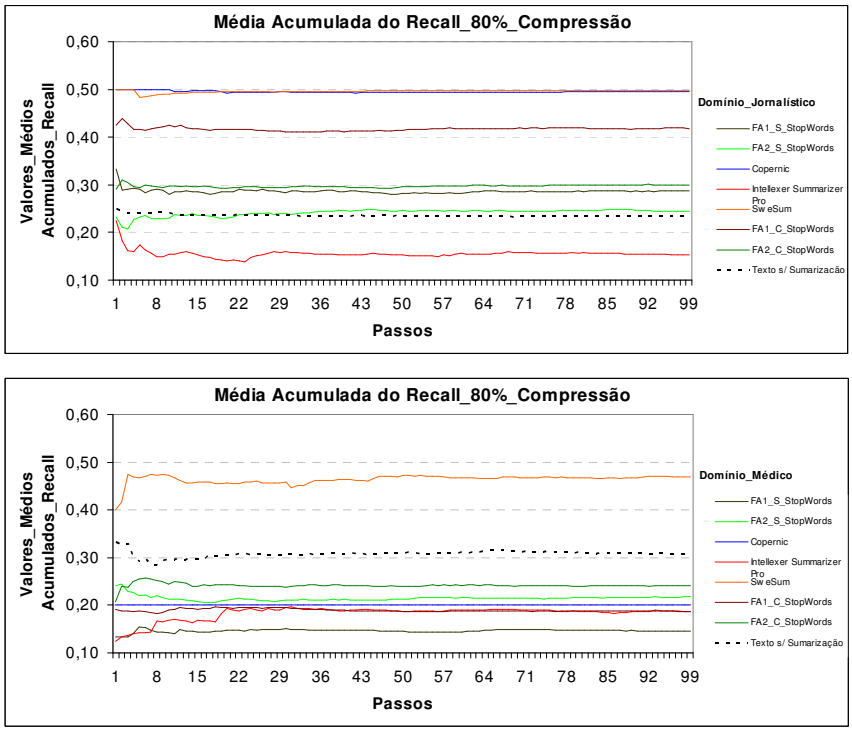


Figura 16a: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Recall* com 80%compressão, no idioma inglês.

A medida *Precision* da Figura 16b teve, no domínio jornalístico, um algoritmo abaixo do valor comparado com o valor de *Recall* dos textos fontes. Esse algoritmo é *SweSum.* No domínio médico, apenas um algoritmo teve seu valor de *Recall* superior ao valor do texto sem sumarização, representado pela linha pontilhada, o algoritmo *SweSum.*

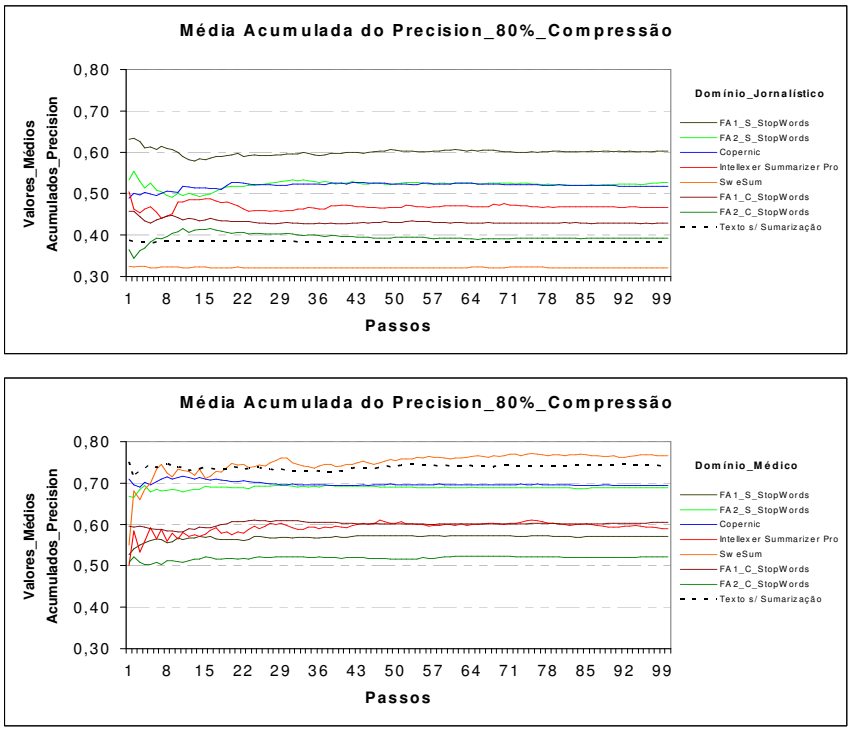


Figura 16b: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Precision* com 80%compressão, no idioma inglês.

USO DA COMPRESSÃO DE 90% NO IDIOMA INGLÊS

O melhor resultado da medida *Recall*, na Figura 17a, foi no domínio jornalístico cujos algoritmos aumentaram os valores em comparação ao valor de *Recall* dos textos fontes. No domínio médico, nenhum algoritmo teve seu valor de *Recall* superior ao valor do texto sem sumarização, representado pela linha pontilhada.

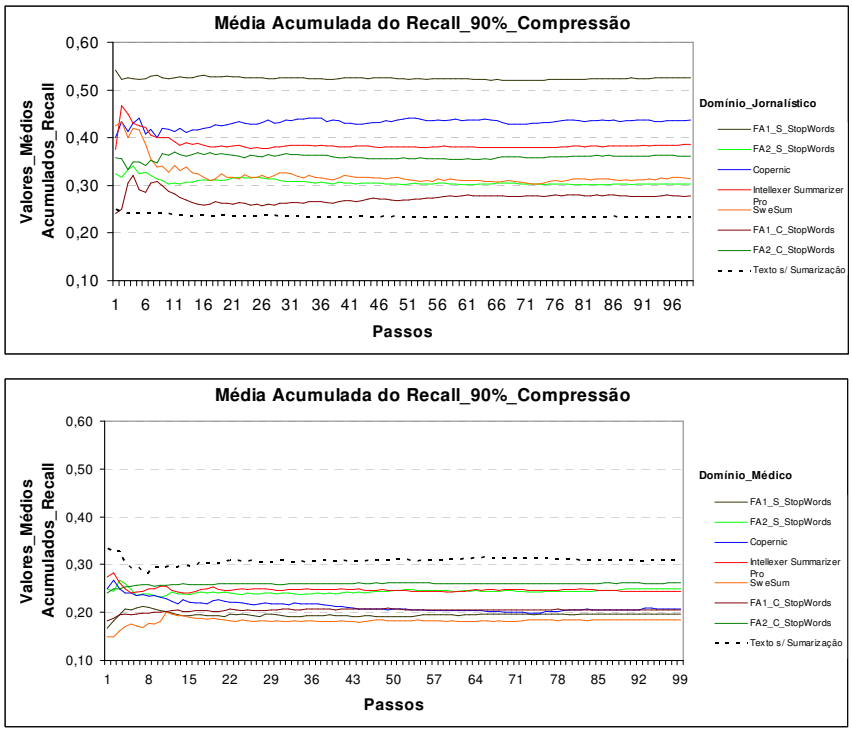


Figura 17a: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Recall* com 90%compressão, no idioma inglês.

A medida *Precision*, apresentada na Figura 17b, teve no domínio jornalístico um algoritmo abaixo do valor comparado com o valor de *Precision* dos textos fontes. No domínio médico, apenas um algoritmo teve seu valor de *Precision* superior ao valor dos textos fontes, o algoritmo *Copernic*.

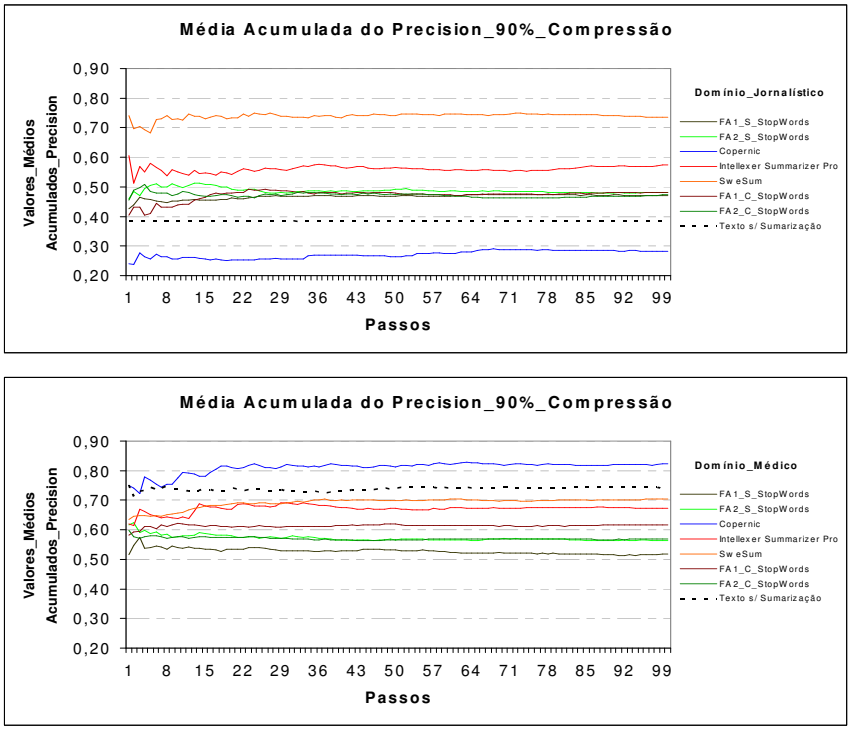


Figura 17b: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Precision* com 90%compressão, no idioma inglês.

APÊNDICE B

MÉTRICA EXTERNA COM AS MEDIDAS: COESÃO E ACOPLAMENTO

O Apêndice B mostra a continuidade dos resultados obtidos da primeira parte dos experimentos descritos na subseção 5.1.3, onde foi apresentada a medida Coeficiente Silhouette. Com as medidas Coesão e Acoplamento, que fazem parte do conjunto com Coeficiente Silhouette (que é medida harmônica da Coesão e do Acoplamento) da métrica interna. Como forma de organização, no Apêndice B, foram realizados as mesmas comparações descritas na subseção 5.1.3. e os resultados foram apresentados com as compressões de 50%, 70%, 80% e 90%. Os textos escolhidos pertecem aos domínios, jornalístico, jurídico e médico nos idiomas português e inglês.

As figuras seguem a mesma numeração estabelecida para a medida Coeficiente Silhouette. O diferencial, aparece com a letra "a" depois da numeração que representa a figura que mostra a medida Coesão e a letra "*b*" para representar a medida Acoplamento..

USO DA COMPRESSÃO DE 50% NO IDIOMA PORTUGUÊS

O melhor resultado da medida de Coesão, mostrado na Figura 18a, aparece no domínio jurídico cujos algoritmos tiveram seus valores maiores em comparação com o valor da Coesão dos textos fontes, com exceção dos algoritmos da literatura, o *Gist Intrasentença,* o *SuPor*, *e* a FA1, sem as *stopwords.*

No domínio jornalístico apenas um algoritmo teve seu valor maior em comparação com o valor da Coesão dos textos fontes, foi o algoritmo da literatura *Gist Average Keyword.* O domínio médico foi o que obteve o pior resultado na medida de Coesão, nenhum dos algoritmos ficaram acima do valor dos textos fontes..

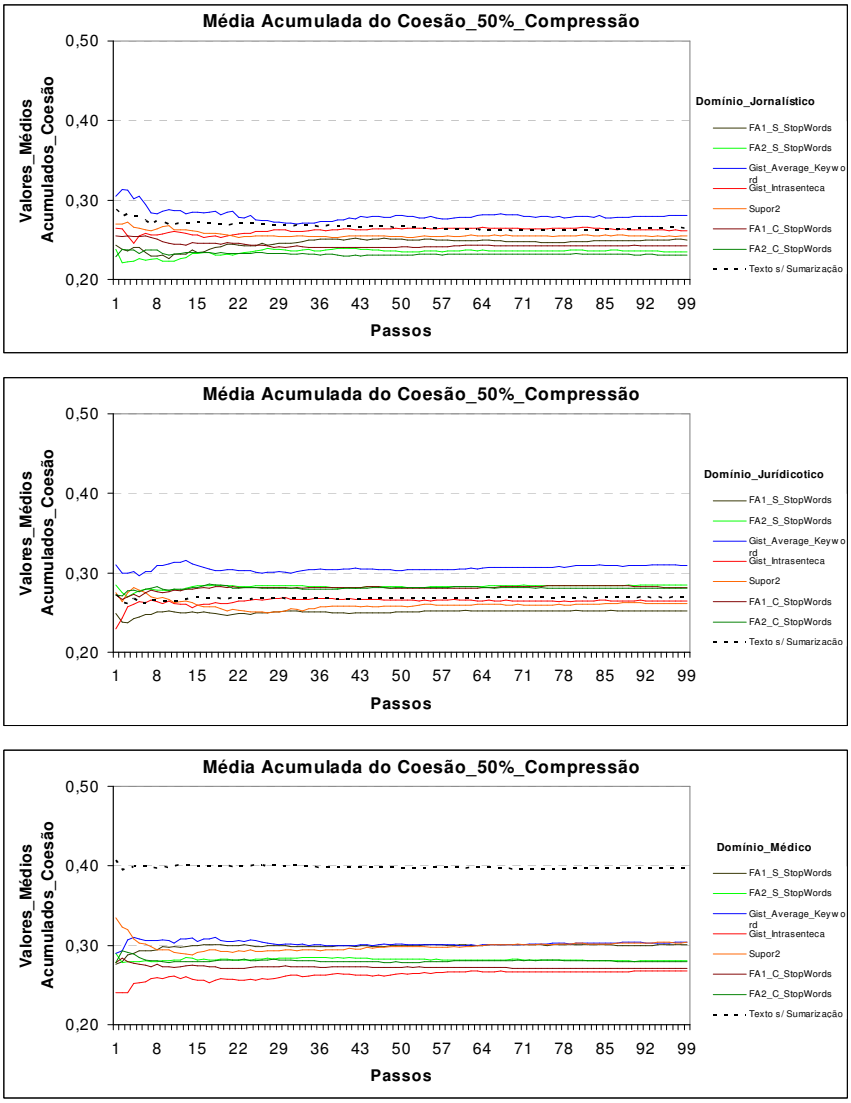


Figura 18a: Resultados obtidos pelo modelo Cassiopeia, usando a medida Coesão com 50%compressão, no idioma português.

Na Figura 18b, os piores resultados da medida Acoplamento foram no domínio jurídico e no domínio médico, cujos algoritmos ficaram com seus valores abaixo do valor dos textos fontes. No domínio jornalístico, apenas um algoritmo ficou abaixo da medida de Acoplamento, em comparação com os textos fontes, o algoritmo da literatura *Gist Intrasentença.*

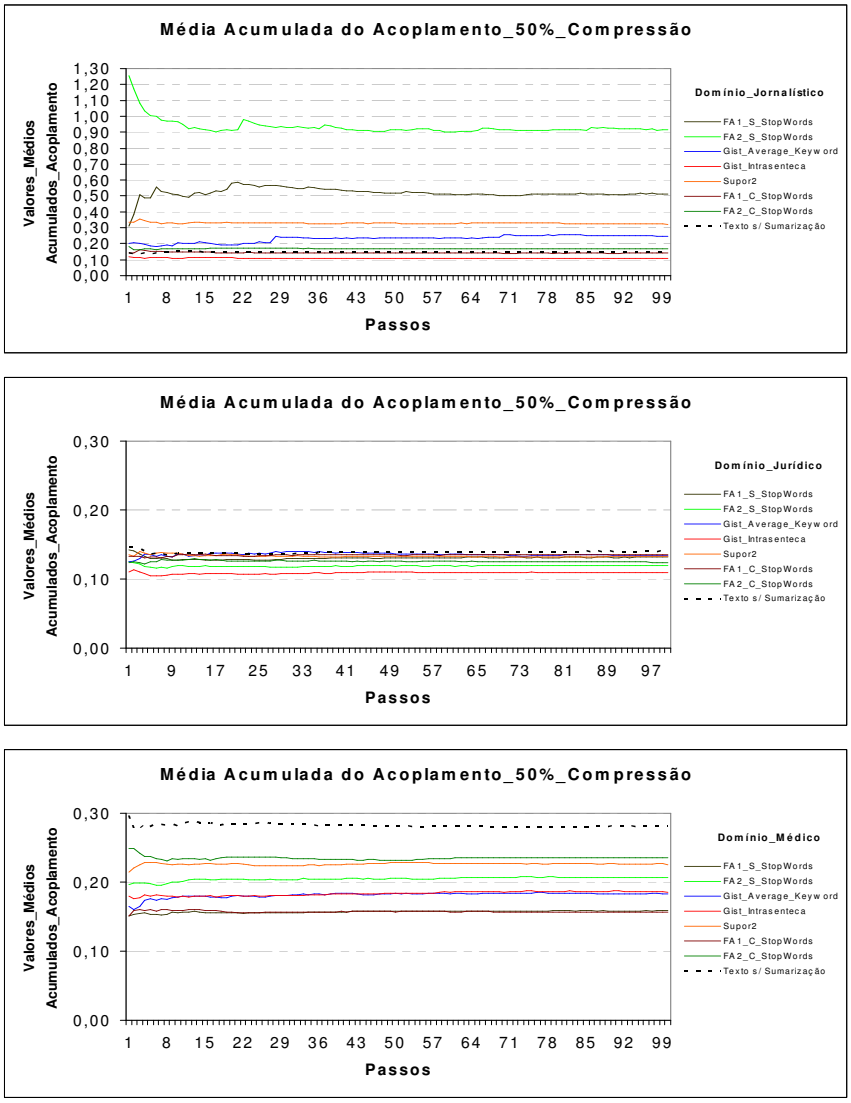


Figura 18b: Resultados obtidos pelo modelo Cassiopeia, usando a medida Acoplamento com 50%compressão, no idioma português.

USO DA COMPRESSÃO DE 70% NO IDIOMA PORTUGUÊS

O melhor resultado da medida Coesão, apresentado na figura 19a, foi no domínio jurídico cujos algoritmos tiveram seus valores aumentados em comparação com o valor de Coesão dos textos fontes, com exceção das funções aleatórias FA2. No domínio jornalístico, existem quatro algoritmos com os valores abaixo do valor dos textos fontes na medida de Coesão. O algoritmo da literatura, o *Gist Intrasentença,* temas duas funções aleatórias FA1 e FA2 com as *stopwords*. No domínio médico, na medida de Coesão, nenhum dos algoritmos alcançaram valores acima dos textos fontes.

.

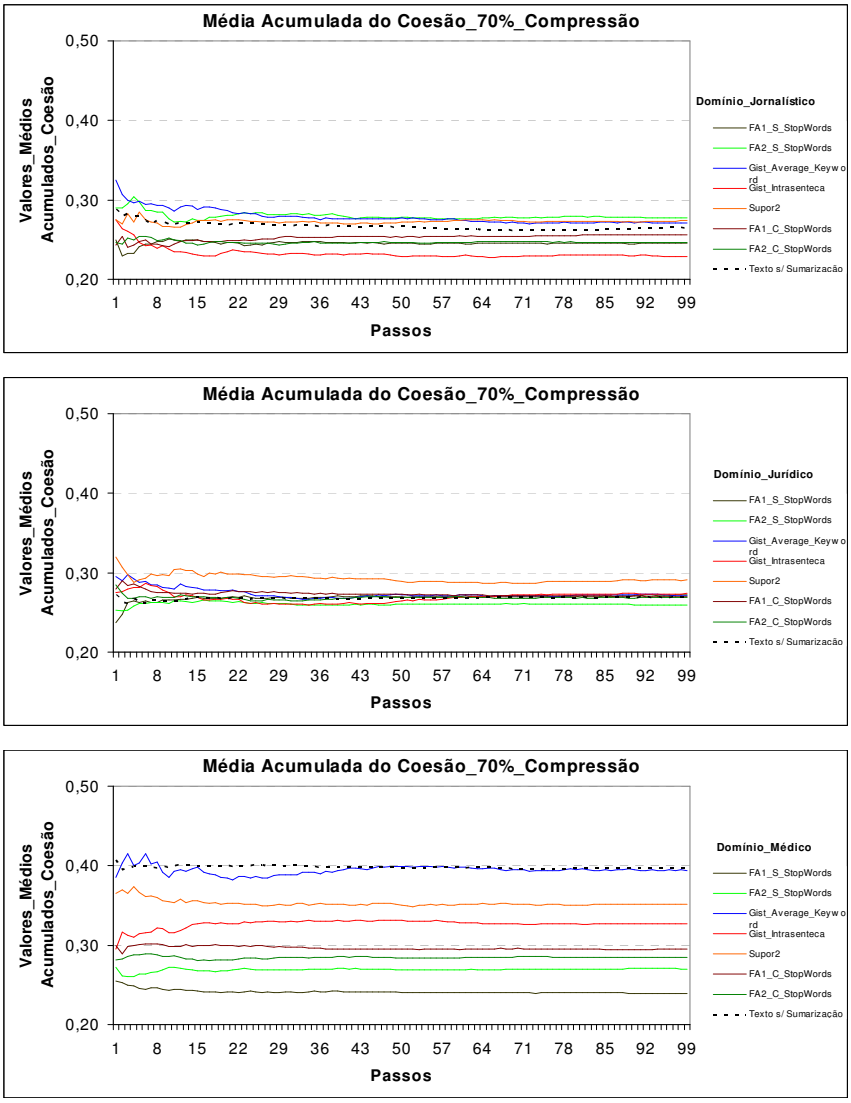


Figura 19a: Resultados obtidos pelo modelo Cassiopeia, usando a medida Coesão com 70%compressão, no idioma português.

No domínio jornalístico, mostrada na Figura 19b, a medida Acoplamento teve o melhor resultado, apenas um algoritmo ficou com seu valor abaixo dos textos fontes. Foi o algoritmo da literatura, o *Gist Intrasentença*. No domínio jurídico, dois algoritmos ficaram com seus valores acima da medida de Acoplamento, em comparação com os textos fontes. Os algoritmos foram da literatura *Gist Intrasentença* e FA2 com *stopwords*. O pior resultado da medida Acoplamento, apresentado na Figura 19b, foi no domínio médico, cujos algoritmos ficaram com seus valores abaixo do valor dos textos fontes.

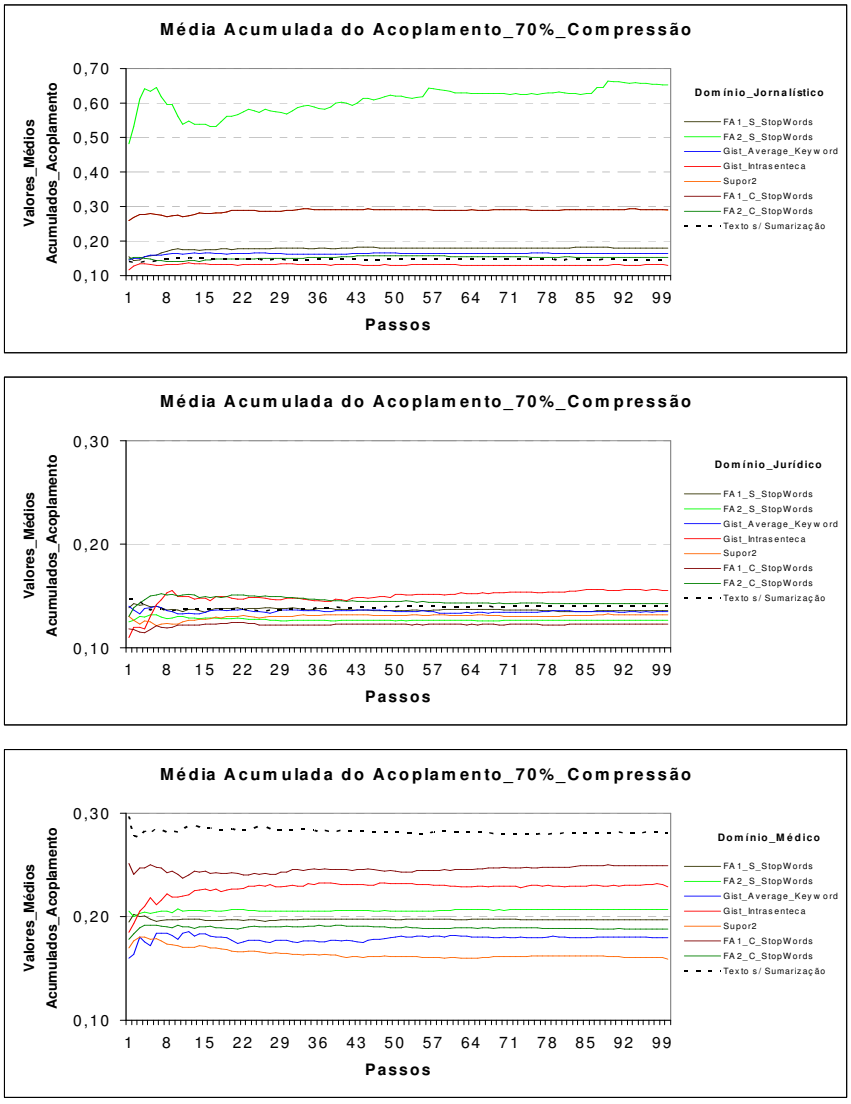


Figura 19b: Resultados obtidos pelo modelo Cassiopeia, usando a medida Acoplamento com 70%compressão, no idioma português.

USO DA COMPRESSÃO DE 80% NO IDIOMA PORTUGUÊS

O melhor resultado da medida Coesão, apresentado na figura 20a, está no domínio jurídico, cuja boa parte dos algoritmos aumentou os valores em comparação com o valor de Coesão dos textos fontes, com exceção do algoritmo da literatura, *Gist Intrasentença* e FA2 sem *stopwords*. No domínio jornalístico, existem dois algoritmos com valores acima do valor dos textos fontes, os algoritmos da literatura, o *Gist Average Keywords* e o *SuPor*. No domínio médico, nenhum dos algoritmos ficou com seus valores acima dos textos fontes.

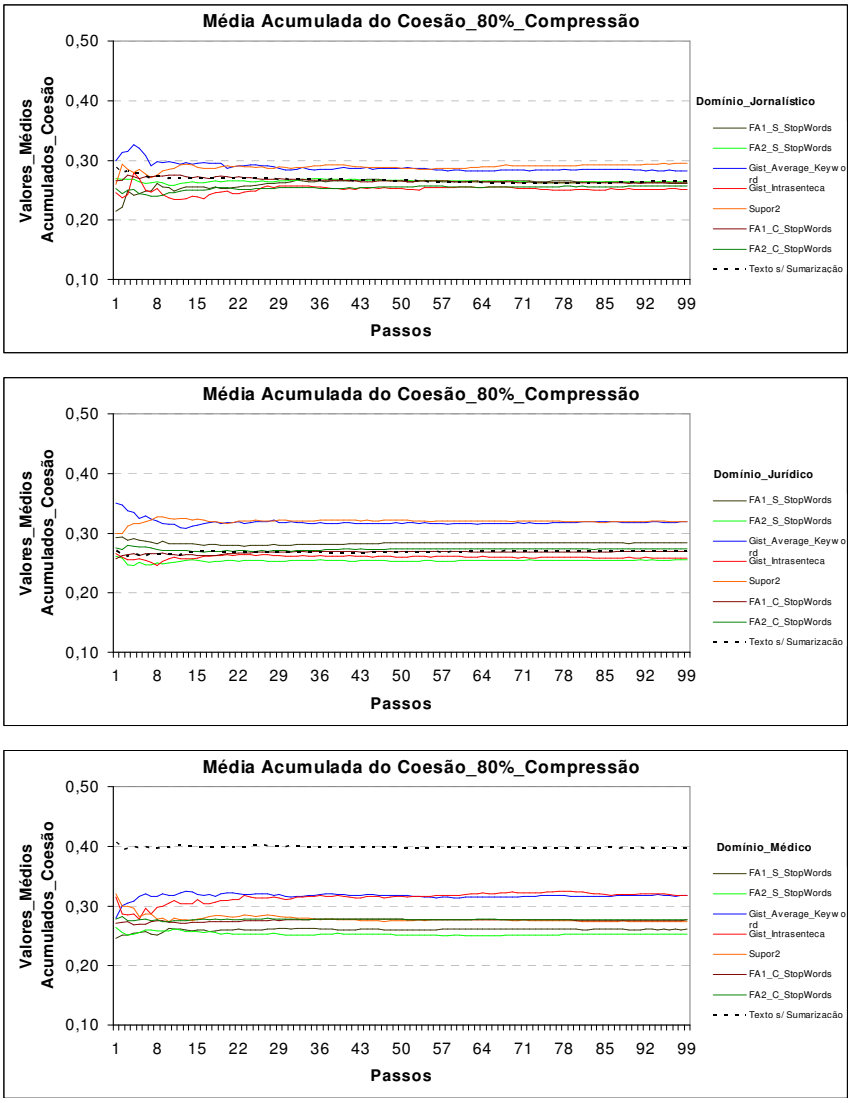


Figura 20a: Resultados obtidos pelo modelo Cassiopeia, usando a medida Coesão com 80%compressão, no idioma português.

O melhor resultado da medida Acoplamento, apresentada na Figura 20b, foi no domínio jornalístico cujos algoritmos ficaram com seus valores acima dos textos fontes. No jurídico, o algoritmo da função aleatória FA1 sem *stopwords* ficou com seu valor acima do valor dos textos fontes. O domínio médico teve três algoritmos que ficaram com seus valores acima do valor dos textos fontes, foram os algoritmos da literatura, *Gist Intrasentença* e as funções aleatórias FA2.

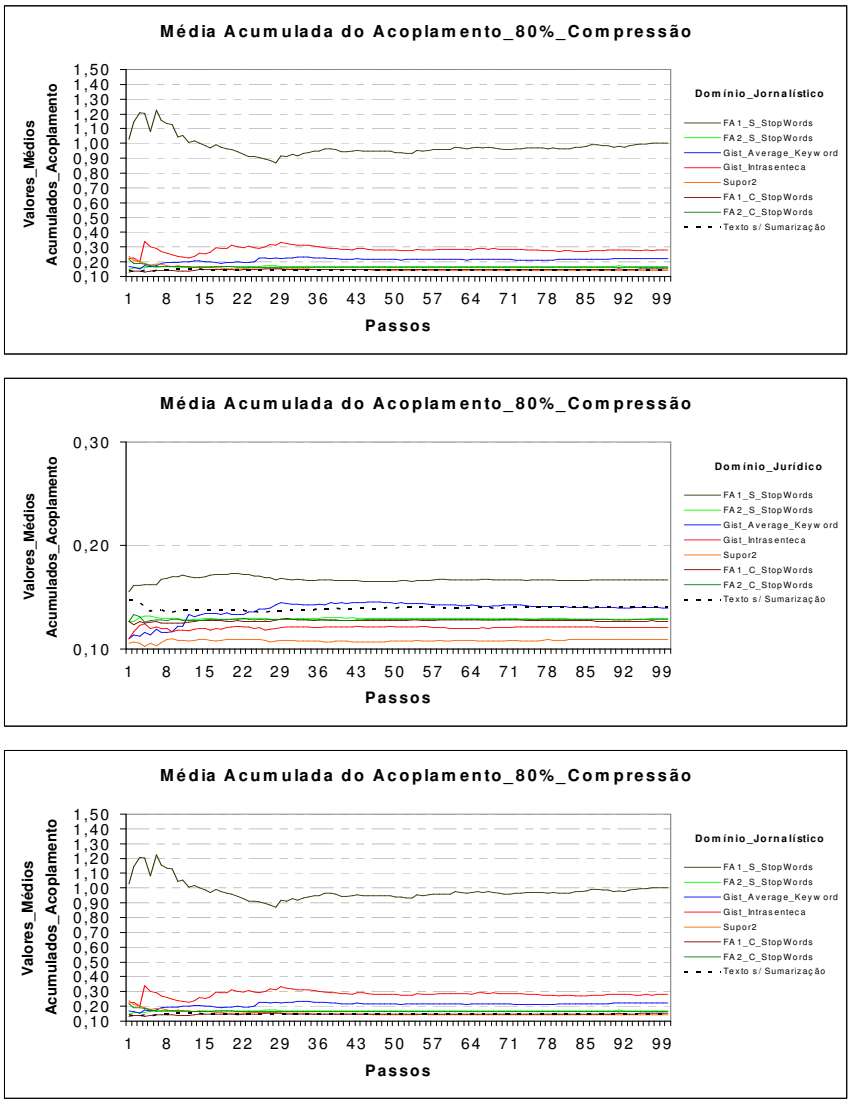


Figura 20b: Resultados obtidos pelo modelo Cassiopeia, usando a medida Acoplamento com 80%compressão, no idioma português.

USO DA COMPRESSÃO DE 90% NO IDIOMA PORTUGUÊS

O melhor resultado da medida Coesão, mostrado na Figura 21a, foi no domínio jornalístico cujos algoritmos aumentaram os seus valores em comparação com o valor de Coesão dos textos fontes. No domínio jurídico, existem três algoritmos que ficaram com valores acima dos textos fontes. Os algoritmos da literatura, *Gist Average Keyword* e *SuPor*, e a função aleatória FA2 com as *stopwords*. No domínio médico, na medida *Coesão*, nenhum dos algoritmos teve seus valores acima dos textos fontes.

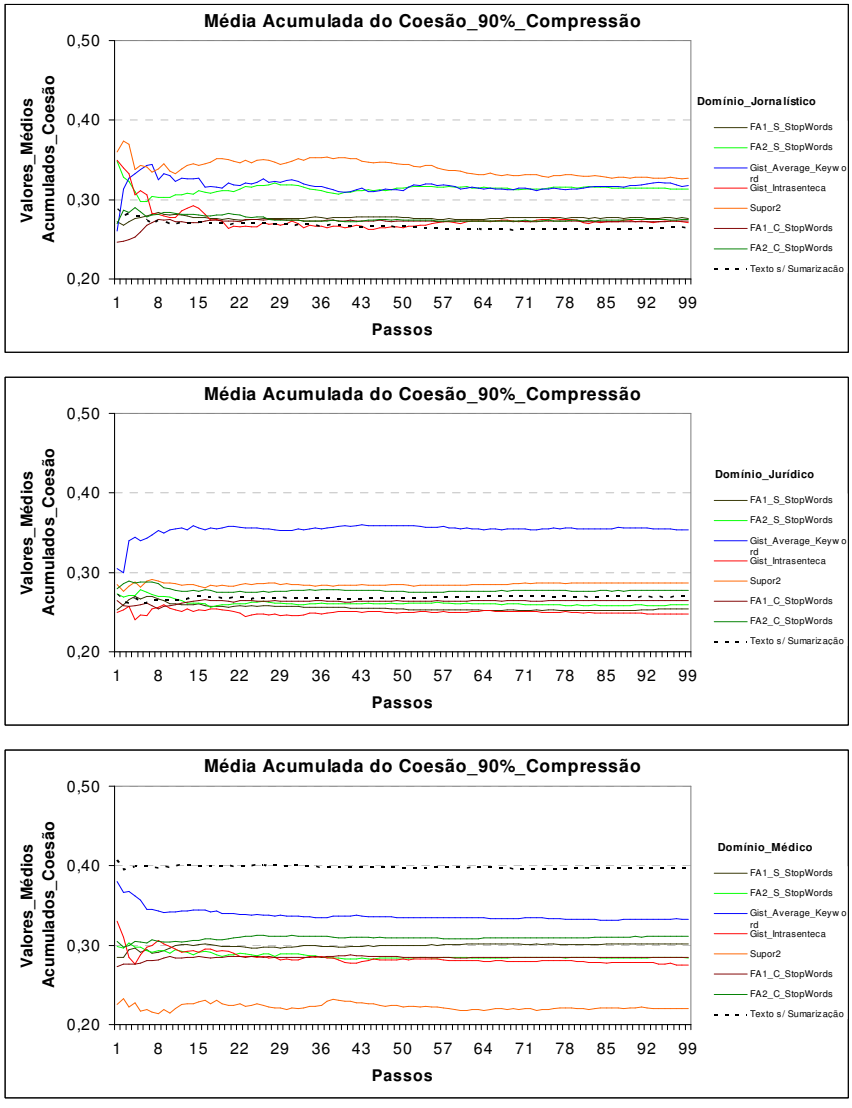


Figura 21a: Resultados obtidos pelo modelo Cassiopeia, usando a medida Coesão com 90%compressão, no idioma português.

O melhor resultado da medida Acoplamento, apresentada na Figura 21b, foi no domínio jornalístico cujos algoritmos ficaram com seus valores acima dos textos fontes. No domínio jurídico, apenas um algoritmo, o *SuPor*, teve seu valor de Acoplamento abaixo dos textos fontes. No domínio médico, três algoritmos ficaram com seus valores abaixo do valor dos textos fontes. Foram os algoritmos FA1 e FA2 com stopwords e o algoritmo da literatura, *Gist Average Keyword.*

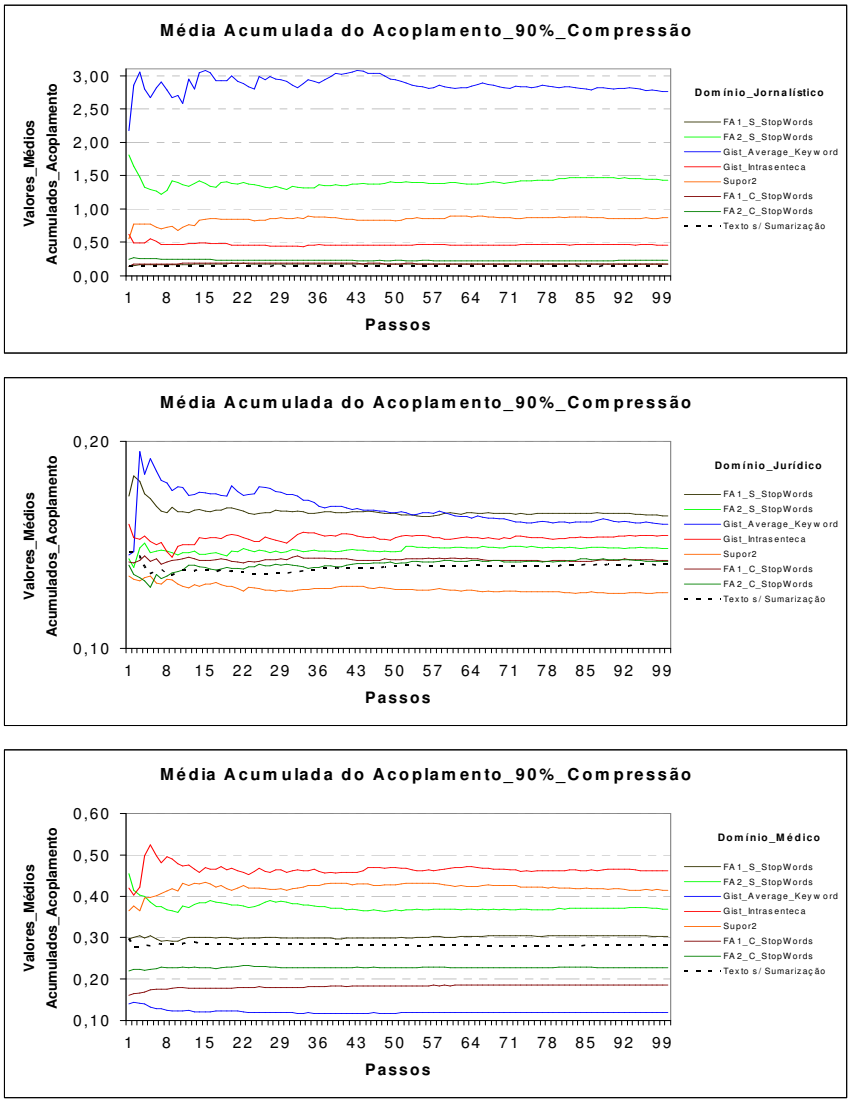


Figura 21b: Resultados obtidos pelo modelo Cassiopeia, usando a medida Acoplamento com 90%compressão, no idioma português.

USO DA COMPRESSÃO DE 50% NO IDIOMA INGLÊS

No domínio jornalístico, na medida de Coesão, mostrado na figura 22a, aparecem três algoritmos que ficaram com seus valores acima dos textos fontes, o algoritmo da literatura *SweSum*, algoritmo profissional *Intellexer Pro* e a função aleatória FA1 sem *stopwords*. Já no domínio médico houve o melhor resultado da medida Acoplamento, todos os algoritmos ficaram com seus valores acima dos textos fontes.

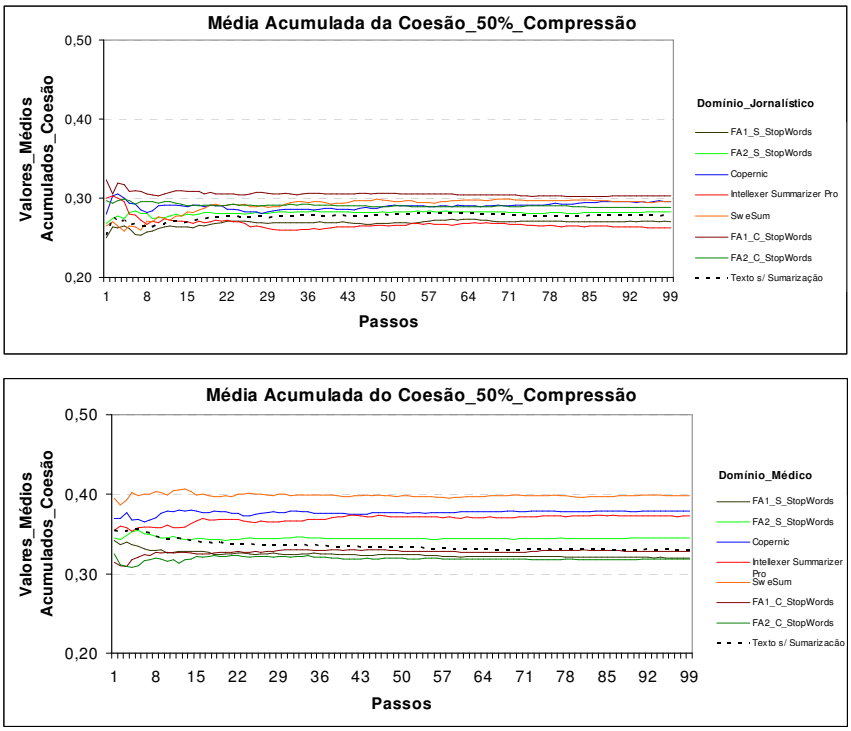


Figura 22a: Resultados obtidos pelo modelo Cassiopeia, usando a medida Coesão com 50%compressão, no idioma inglês.

No domínio jornalístico na medida Acoplamento, apresentada na Figura 22b, aparecem dois algoritmos que ficaram com seus valores abaixo dos textos fontes, o algoritmo da literatura FA1 sem *stopword* e o algoritmo profissional *Intellexer Pro*. Já no domínio médico aparecem três algoritmos que ficaram com seus valores acima dos textos fontes, o algoritmo da literatura *SweSum* e o algoritmo profissional *Intellexer Pro* e *Copernic*.

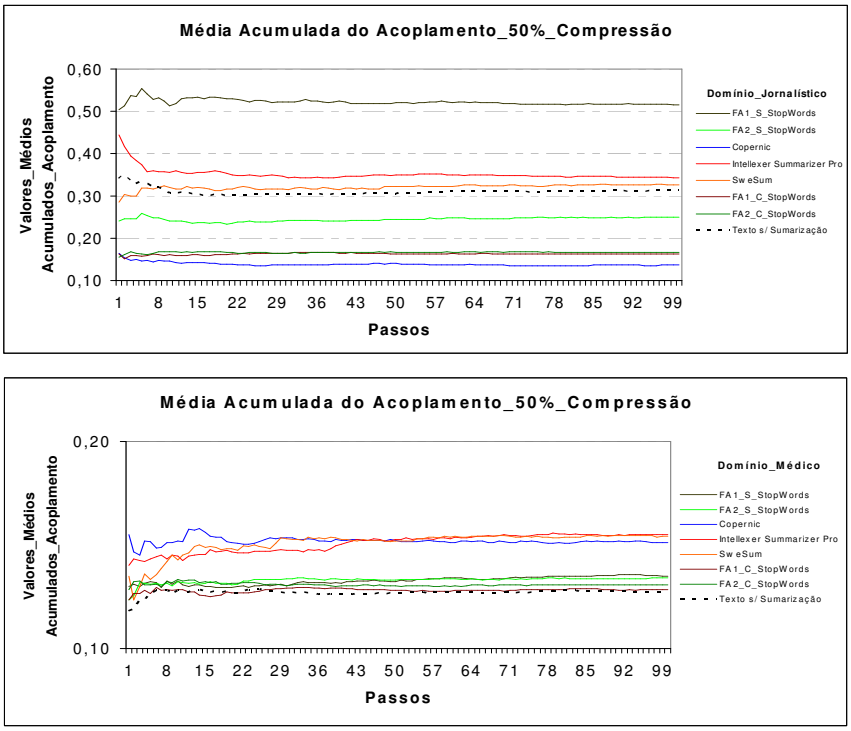


Figura 22b: Resultados obtidos pelo modelo Cassiopeia, usando a medida Acoplamento com 50%compressão, no idioma inglês.

USO DA COMPRESSÃO DE 70% NO IDIOMA INGLÊS

No domínio jornalístico, apresentado na figura 23a, existem três algoritmos com os valores acima dos textos fontes. Os algoritmos profissionais, *Intellexer Pro, Copernic* e o algoritmo da literatura, *SweSum*. No domínio médico, existem três algoritmos com os valores acima dos textos fontes, os algoritmos da literatura *SweSum* e as funções aleatórias FA1 e FA2 com *stopwords*.

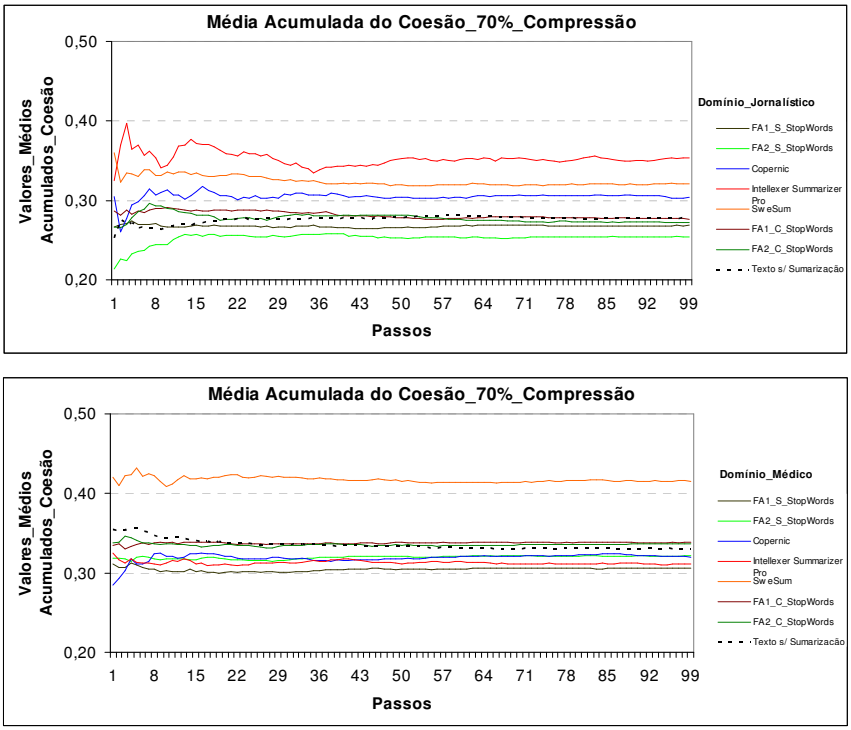


Figura 23a: Resultados obtidos pelo modelo Cassiopeia, usando a medida Coesão com 70%compressão, no idioma inglês.

No domínio jornalístico, apresentado na Figura 23b, existem três algoritmos com os valores acima dos textos fontes. Os algoritmos profissionais *Intellexer Pro, Copernic* e a função aleatória FA2 sem *stopwords*. No domínio médico, existem dois algoritmos com valores abaixo dos textos fontes, o algoritmo profissional *Intellexer Pro* e a função aleatória FA1 sem *stopwords*.

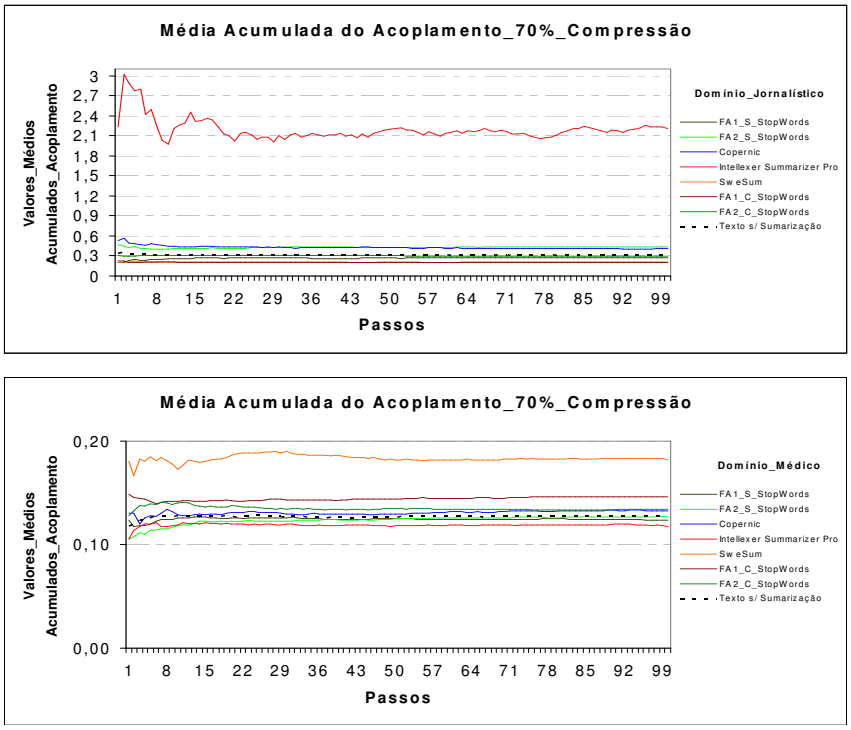


Figura 23b: Resultados obtidos pelo modelo Cassiopeia, usando a medida Acoplamento com 70%compressão no idioma inglês.

USO DA COMPRESSÃO DE 80% NO IDIOMA INGLÊS

No domínio jornalístico, apresentado na figura 24a, existem três algoritmos que tiveram seus valores maiores do que os textos fontes, os algoritmos profissionais *Intellexer Pro, Copernic* e a função aleatória FA1 sem *stopwords*. No domínio médico, existem dois algoritmos com valores acima dos textos fontes, o algoritmo da literatura *SweSum* e algoritmo profissional *Intellexer Pro*.

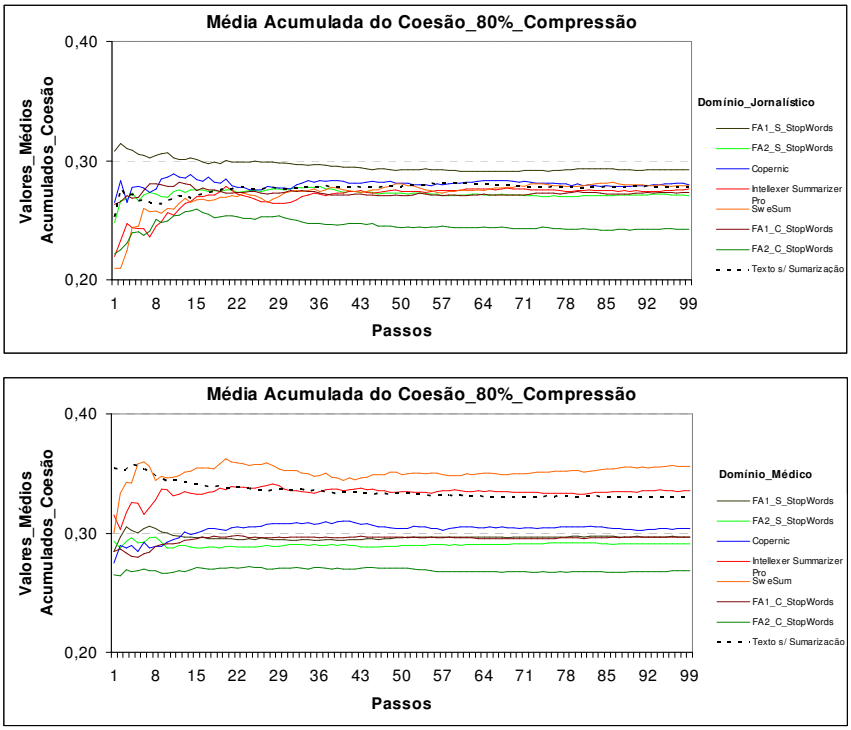


Figura 24a: Resultados obtidos pelo modelo Cassiopeia, usando a medida Coesão com 80%compressão, no idioma inglês.

A Figura 24b mostra os resultados no domínio jornalístico, cujos dois algoritmos têm valores acima dos textos fontes, o algoritmo da literatura *SweSum* e a função aleatória FA1 com *stopwords*. O domínio médico teve o pior resultado na medida Acoplamento, nenhum dos algoritmos ficaram com valores abaixo da linha pontilhada que representa o texto sem sumarização.

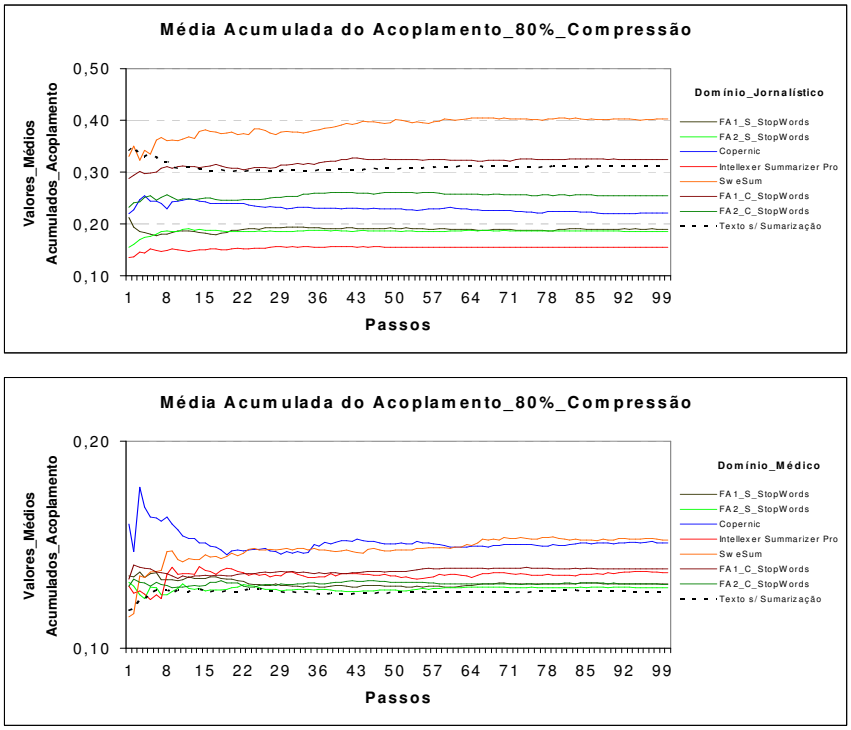


Figura 24b: Mostra os resultados obtidos pelo modelo Cassiopeia, usando a medida Acoplamento com 80%compressão, no idioma inglês.

USO DA COMPRESSÃO DE 90% NO IDIOMA INGLÊS

A Figura 25a mostra os resultados no domínio jornalístico, na medida Coesão cujos algoritmos não tiveram valores acima dos textos fontes. No domínio médico, existe apenas um algoritmo cujo valor está acima dos textos fontes, o algoritmo profissional *Copernic*.

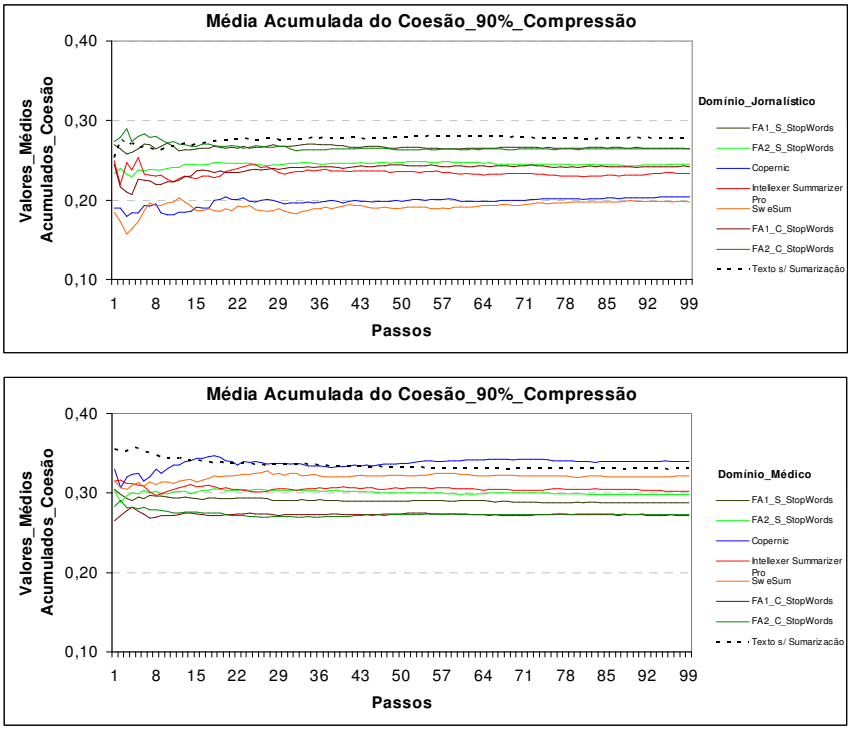


Figura 25a: Resultados obtidos pelo modelo Cassiopeia, usando a medida Coesão com 90%compressão, no idioma inglês.

A Figura 25b mostra os resultados no domínio jornalístico, na medida Acoplamento, cujos três algoritmos aparecem com valores acima dos textos fontes, as funções aleatórias FA1 e FA2 com *stopwords*. No domínio médico, na medida Acoplamento, todos os algoritmos tiveram seus valores acima dos textos fontes.

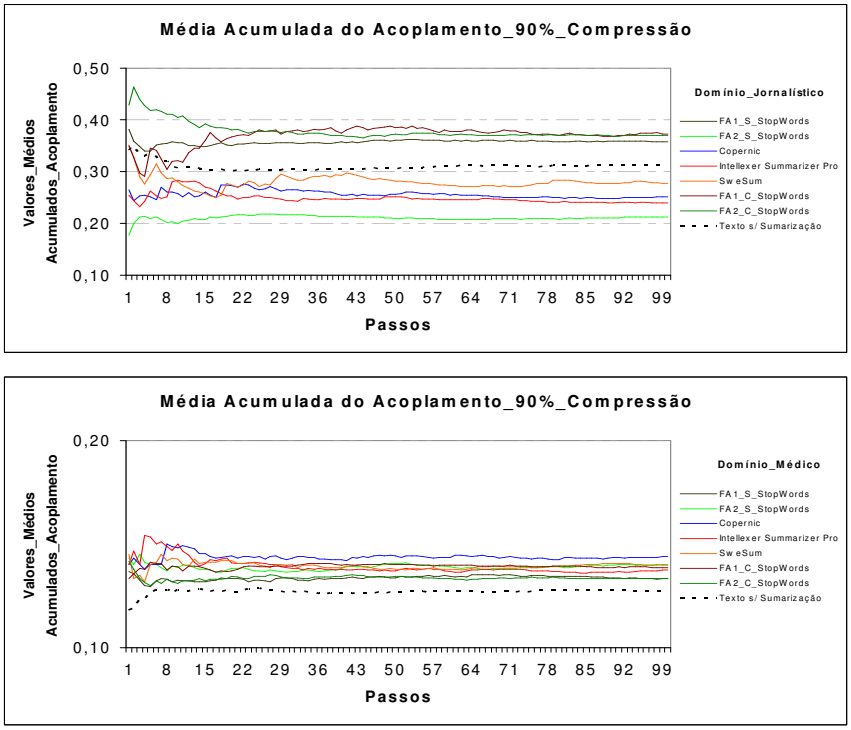


Figura 25b: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Acoplamento* com 90% compressão, no idioma inglês.

APÊNDICE C

MÉTRICA EXTERNA COM AS MEDIDAS: *RECALL* E *PRECISION* REFERENTES ÀS MÉDIAS FINAIS ACUMULADAS

As Figuras 36, 37, 38, 39 e 40 com as suas subdivisões em a,b,c,d mostram como se comportam os métodos da literatura RDF, RTF e TFIDF e o modelo Cassiopeia sem ou com *stopword* referente à identificação e seleção de atributos em bases textuais, usadas na segunda parte do experimento ao longo das 100 interações. As Figuras 36, 37, 38, 39 e 40 e suas subdivisões a e b apontam para o comportamento dos métodos com suas devidas compressões de 50%, 70%, 80% e 90%, seus domínios jornalístico, jurídico (apenas no idioma português) e médico e nos dois idiomas português e inglês, cuja última média final acumulada obtida de cada um dos sumarizadores ao longo da iteração, é somada as três médias obtidas do *Gist_Keyword, Gist_Intra* e *SuPor*. Esse cálculo é realizado para cada um dos métodos da literatura RDF, RTF e TFIDF e para o modelo Cassiopeia sem ou com *stopwords*. Os resultados são mostrados nas Figuras 36, 37, 38, 39 e 40 com as medidas *Recall* e *Precision*.

A necessidade da apresentação das Figuras 36, 37, 38, 39 e 40 advém da importância da complementação da análise da medida *F-Measure* que é uma medida harmônica do *Recall* e *Precision*, discutida na subseção 5.1.1 deste trabalho.

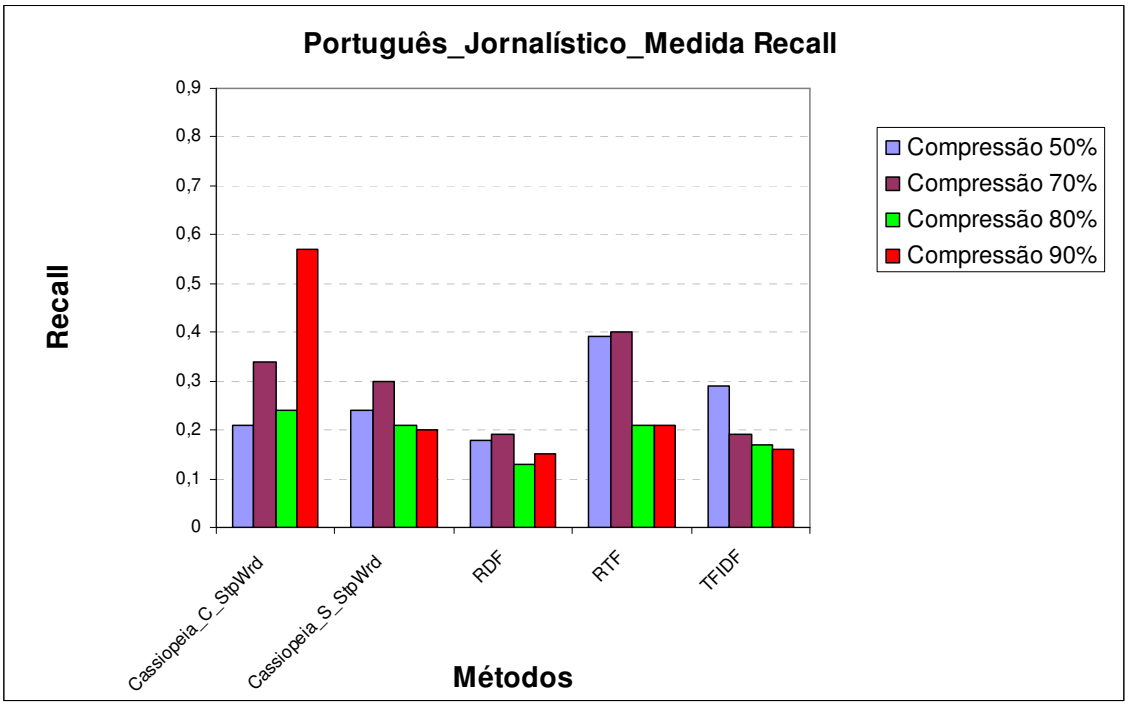


Figura 36a: Resultados das médias finais acumuladas da medida *Recall* para os agrupamentos obtidos através dos métodos RDF, RTF e TFIDF e do modelo Cassiopeia com e sem *Stopwords,* no domínio jornalístico e no idioma português.

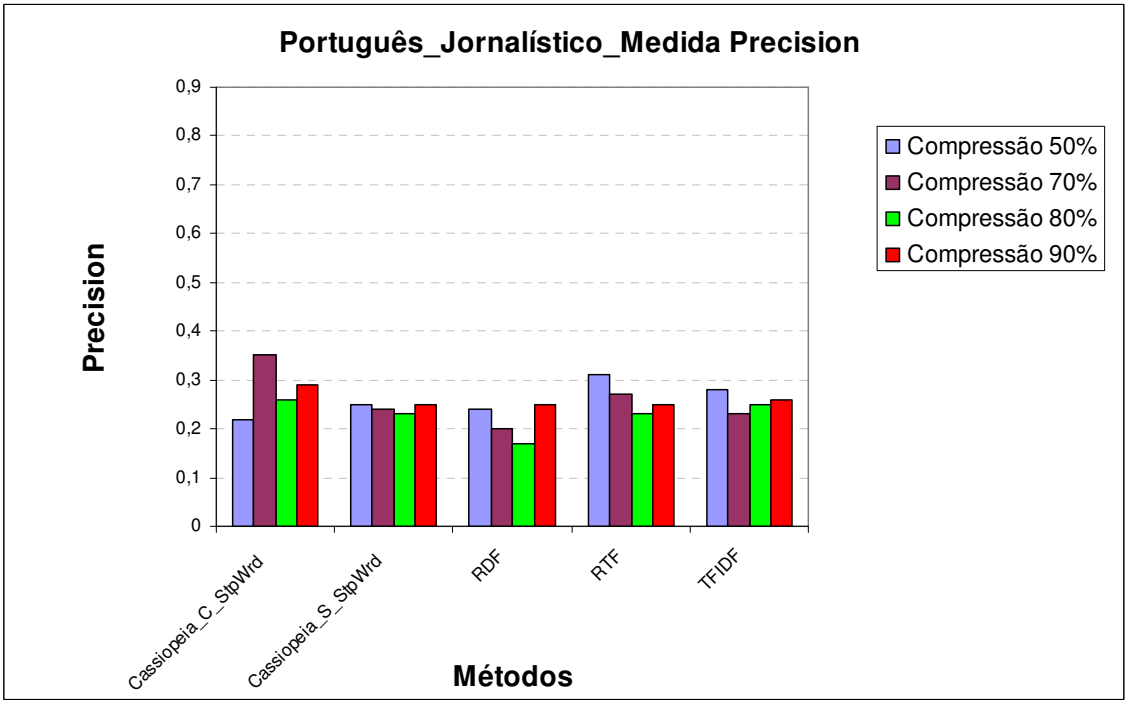


Figura 36b: Resultados das médias finais acumuladas da medida *Precision* para os agrupamentos obtidos através dos métodos RDF, RTF e TFIDF e do modelo Cassiopeia com e sem *Stopwords,* no domínio jornalístico e no idioma português.

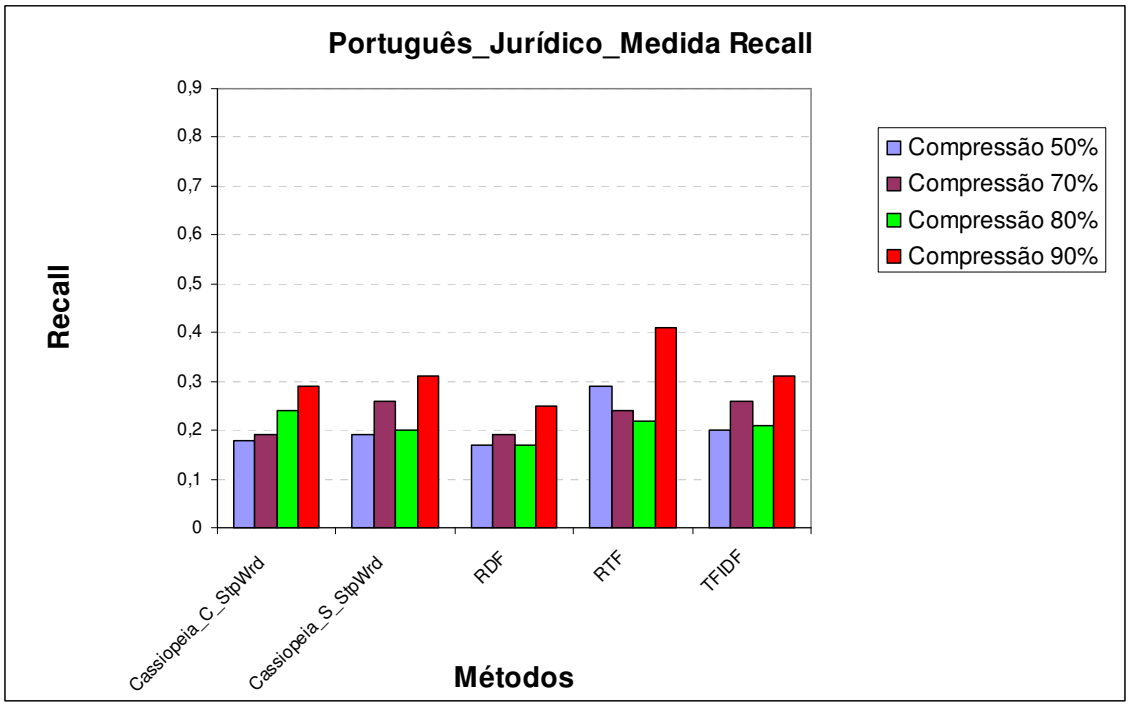


Figura 37a: Resultados das médias finais acumuladas da medida *Recall* para os agrupamentos obtidos através dos métodos RDF, RTF e TFIDF e do modelo Cassiopeia com e sem *Stopwords*, no domínio jurídico e no idioma português.

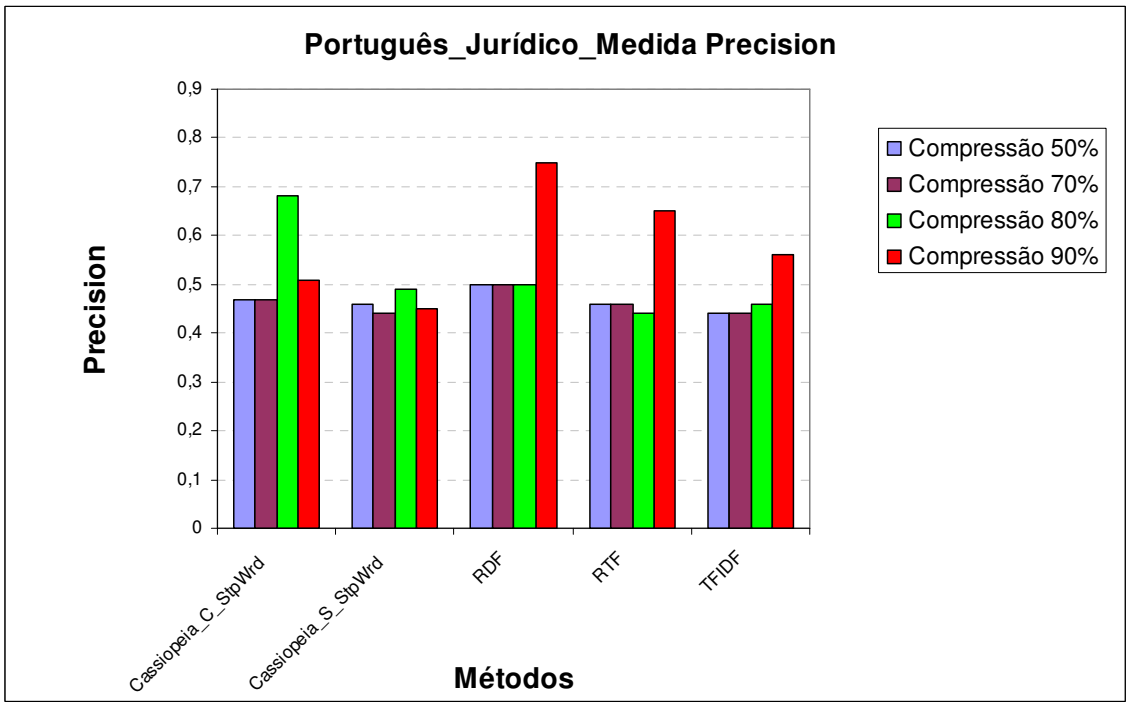


Figura 37b: Resultados das médias finais acumuladas da medida *Precision* para os agrupamentos obtidos através dos métodos RDF, RTF e TFIDF e do modelo Cassiopeia com e sem *Stopwords*, no domínio jurídico e no idioma português.

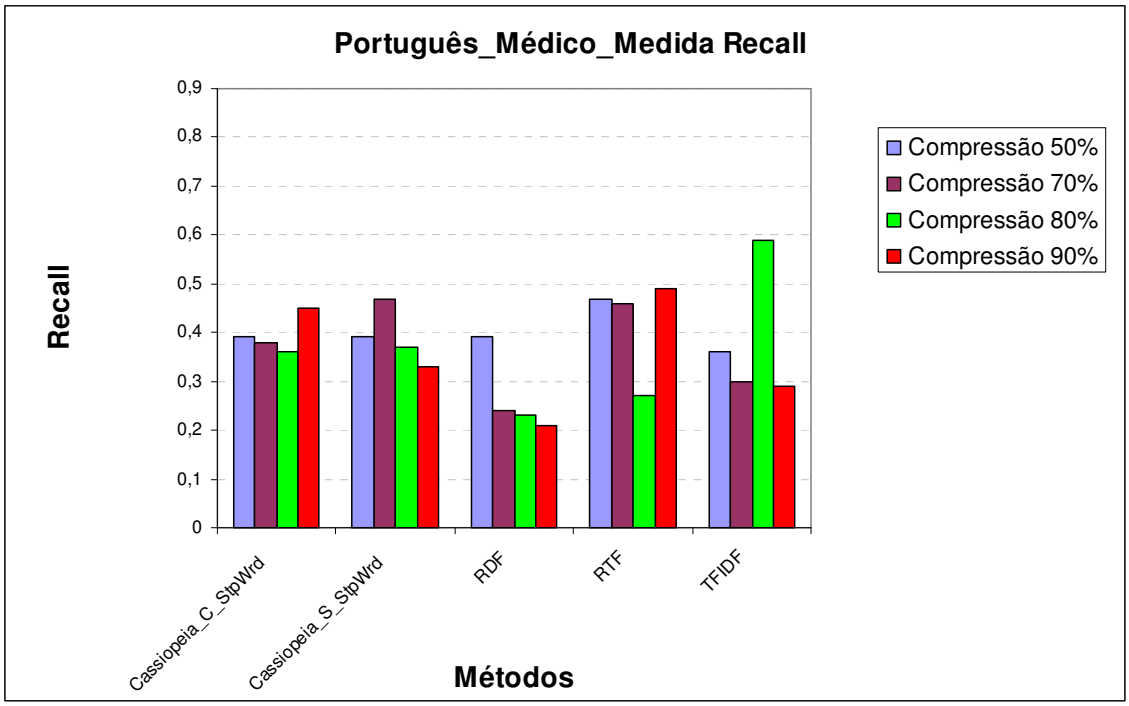


Figura 38a: Resultados das médias finais acumuladas da medida *Recall* para os agrupamentos obtidos através dos métodos RDF, RTF e TFIDF e do modelo Cassiopeia com e sem *Stopwords*, no domínio médico e no idioma português.

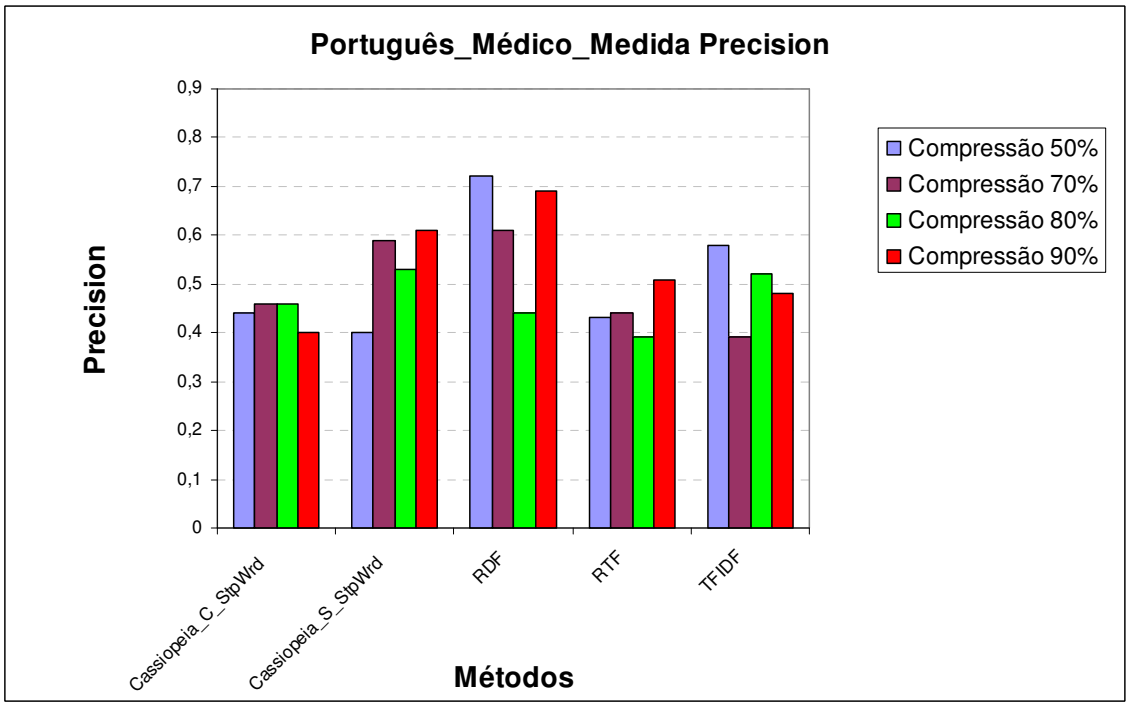


Figura 38b: Resultados das médias finais acumuladas da medida *Precision* para os agrupamentos obtidos através dos métodos RDF, RTF e TFIDF e do modelo Cassiopeia com e sem *Stopwords*, no domínio médico e no idioma português.

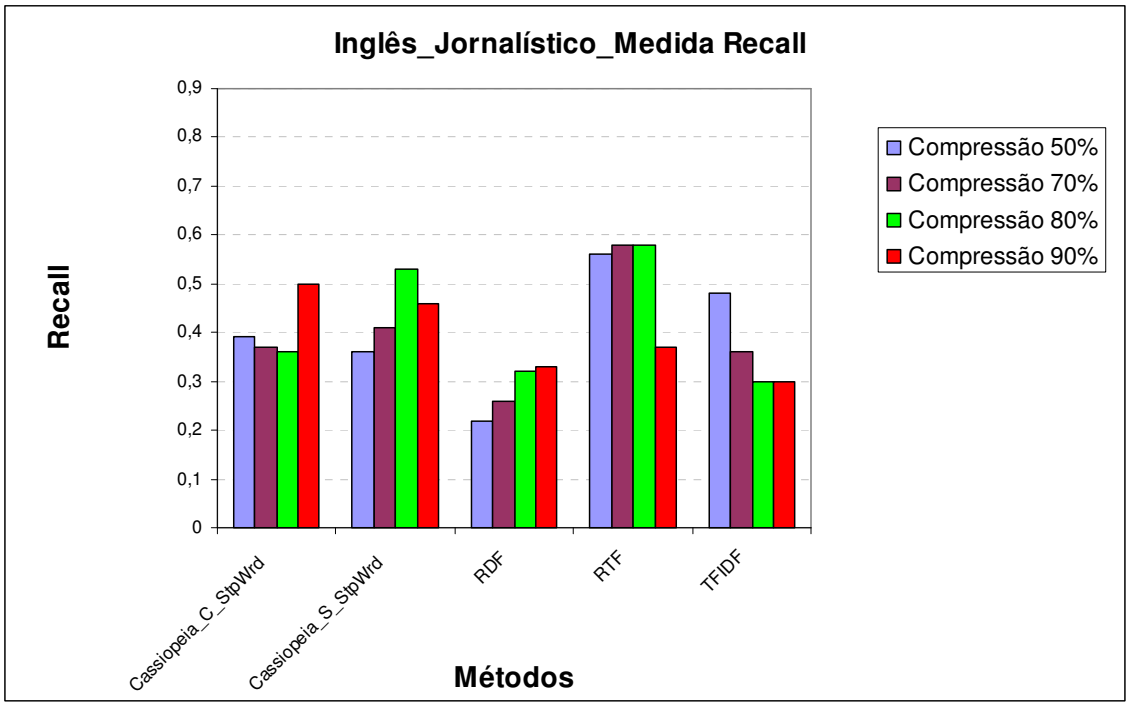


Figura 39a: Resultados das médias finais acumuladas da medida *Recall* para os agrupamentos obtidos através dos métodos RDF, RTF e TFIDF e do modelo Cassiopeia com e sem *Stopwords*, no domínio jornalístico e no idioma inglês.

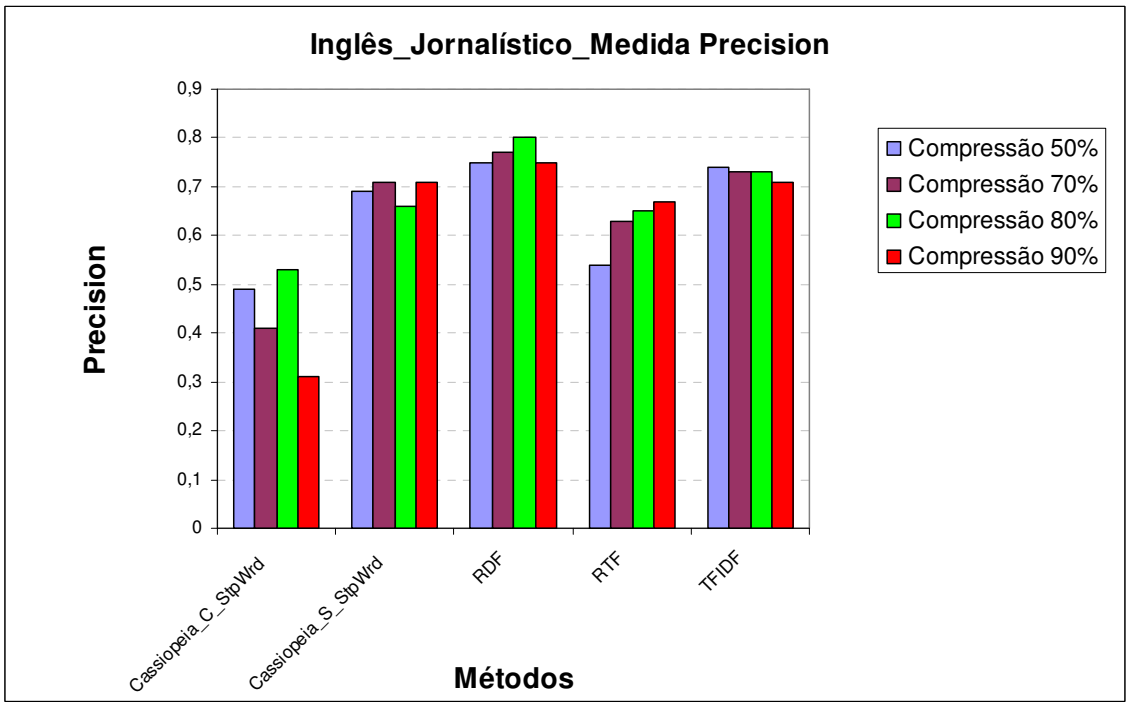


Figura 39b: Resultados das médias finais acumuladas da medida *Precision* para os agrupamentos obtidos através dos métodos RDF, RTF e TFIDF e do modelo Cassiopeia com e sem *Stopwords,* no domínio jornalístico e no idioma inglês.

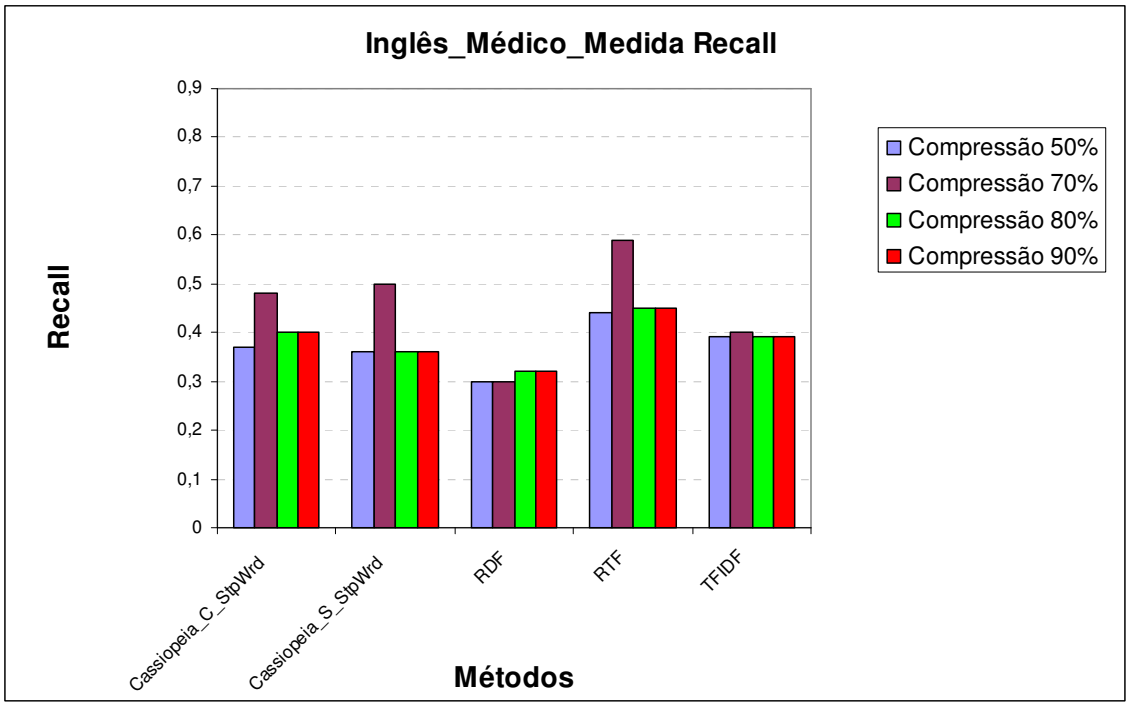


Figura 40a: Resultados das médias finais acumuladas da medida *Recall* para os agrupamentos obtidos através dos métodos RDF, RTF e TFIDF e do modelo Cassiopeia com e sem *Stopwords*, no domínio médico e no idioma inglês.

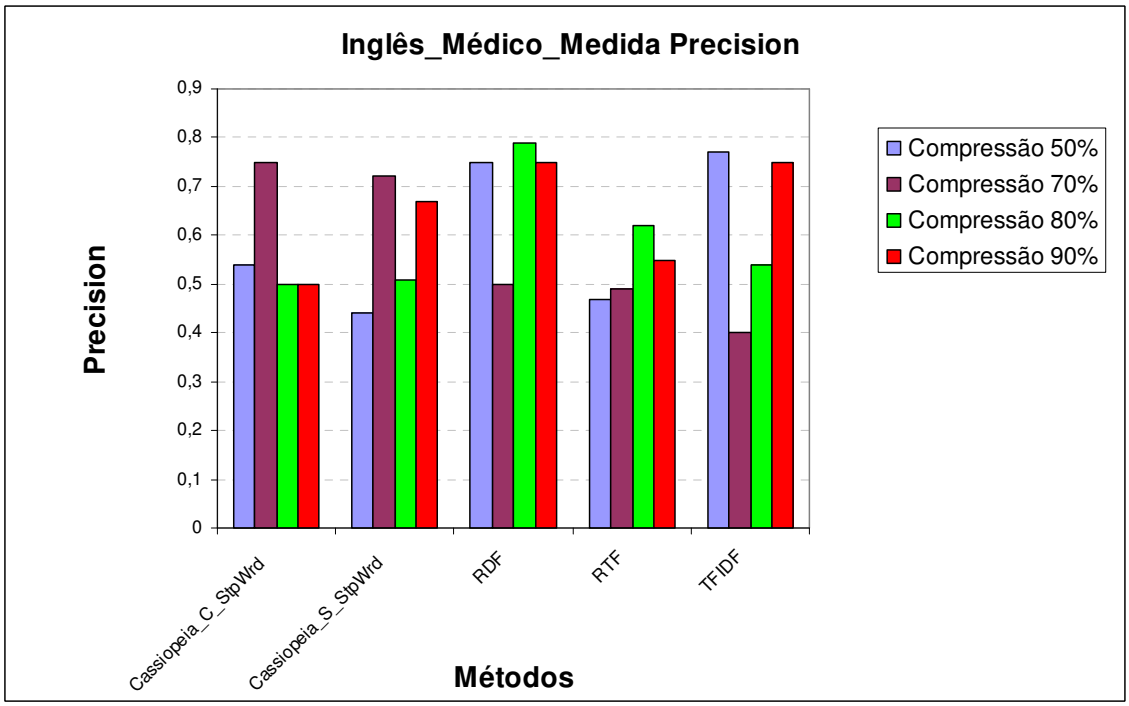


Figura 40b: Resultados das médias finais acumuladas da medida *Precision* para os agrupamentos obtidos através dos métodos RDF, RTF e TFIDF e do modelo Cassiopeia com e sem *Stopwords*, no domínio médico e no idioma inglês.

APÊNDICE D

MÉTRICA INTERNA COM AS MEDIDAS: *COESÃO E ACOPLAMENTO* REFERENTES ÀS MÉDIAS FINAIS ACUMULADAS

As Figuras 41, 42, 43, 44 e 45 com as suas subdivisões em a,b,c,d mostram como se comportam os métodos da literatura RDF, RTF e TFIDF e o modelo Cassiopeia sem ou com *stopword* referente à identificação e seleção de atributos em bases textuais, usadas na segunda parte do experimento ao longo das 100 interações. As Figuras 41, 42, 43, 44 e 45 e suas subdivisões a e b apontam para o comportamento dos métodos com suas devidas compressões de 50%, 70%, 80% e 90%, seus domínios jornalístico, jurídico (apenas no idioma português) e médico e nos dois idiomas português e inglês, cuja última média final acumulada obtida de cada um dos sumarizadores ao longo da iteração, é somada as três médias obtidas do *Gist_Keyword, Gist_Intra* e *SuPor*. Esse cálculo é realizado para cada um dos métodos da literatura RDF, RTF e TFIDF e para o modelo Cassiopeia sem ou com *stopwords*. Os resultados são mostrados nas Figuras 41, 42, 43, 44 e 45 com as medidas Coesão e Acoplamento .

A necessidade da apresentação das Figuras 41, 42, 43, 44 e 45 advém da importância da complementação da análise da medida *F-Measure* é uma medida harmônica do Coesão e Acoplamento discutida no item 5.1.2.2 deste trabalho.

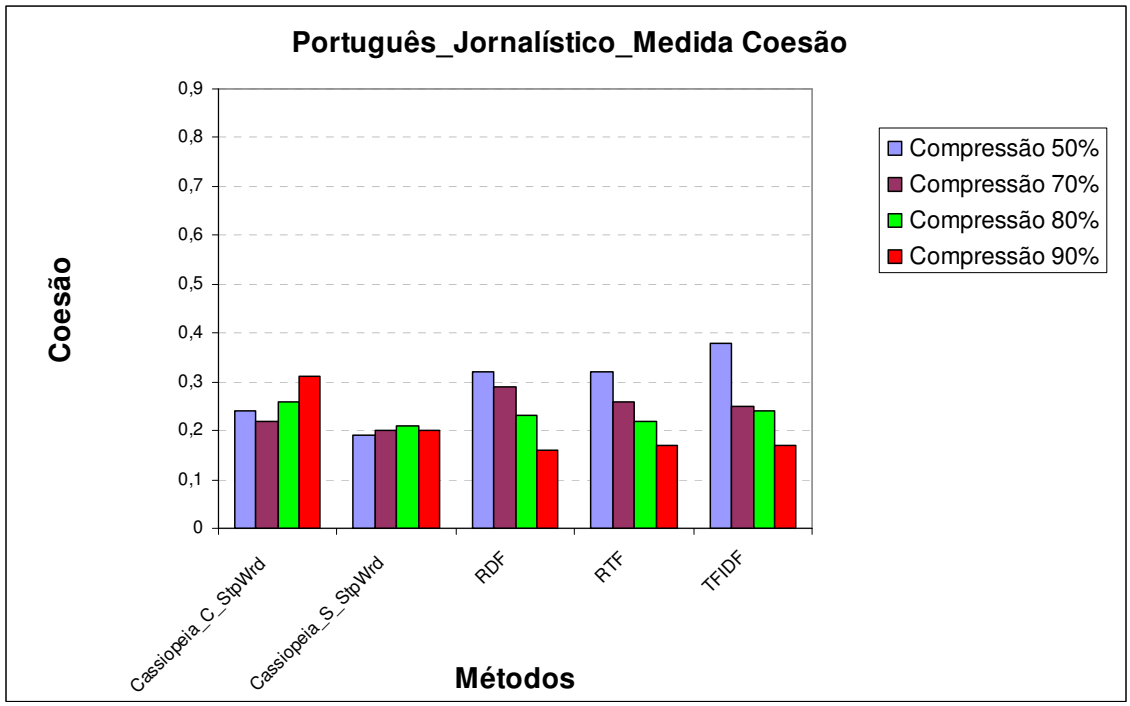


Figura 41a: Resultados das médias finais acumuladas da medida Coesão para os agrupamentos obtidos através dos métodos RDF, RTF e TFIDF e do modelo Cassiopeia com e sem *Stopwords,* no domínio jornalístico e no idioma português.

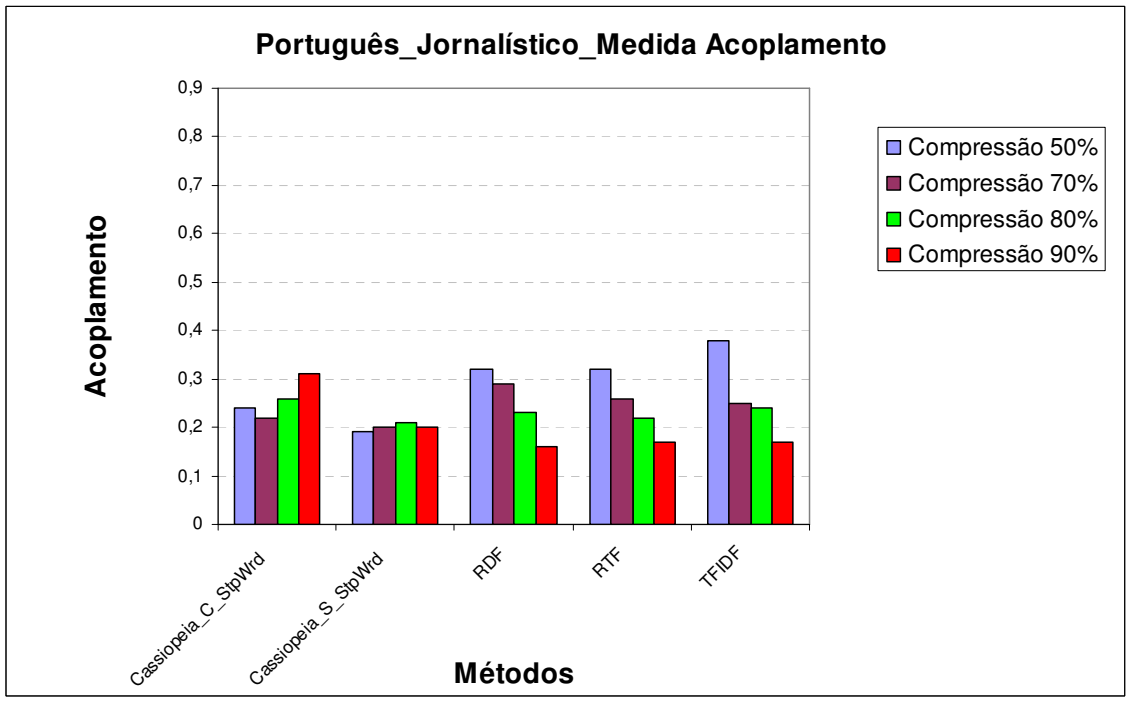


Figura 41b: Resultados das médias finais acumuladas da medida Acoplamento para os agrupamentos obtidos através dos métodos RDF, RTF e TFIDF e do modelo Cassiopeia com e sem *Stopwords,* no domínio jornalístico e no idioma português.

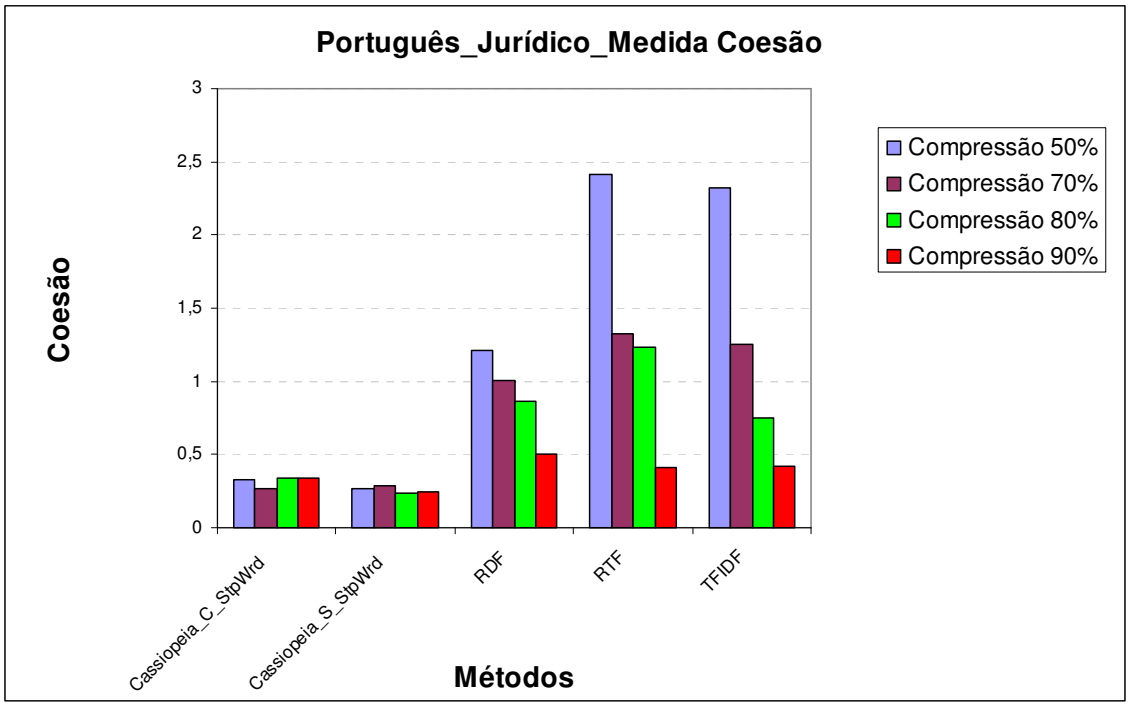


Figura 42a: Resultados das médias finais acumuladas da medida Coesão para os agrupamentos obtidos através dos métodos RDF, RTF e TFIDF e do modelo Cassiopeia com e sem *Stopwords*, no domínio jurídico e no idioma português.

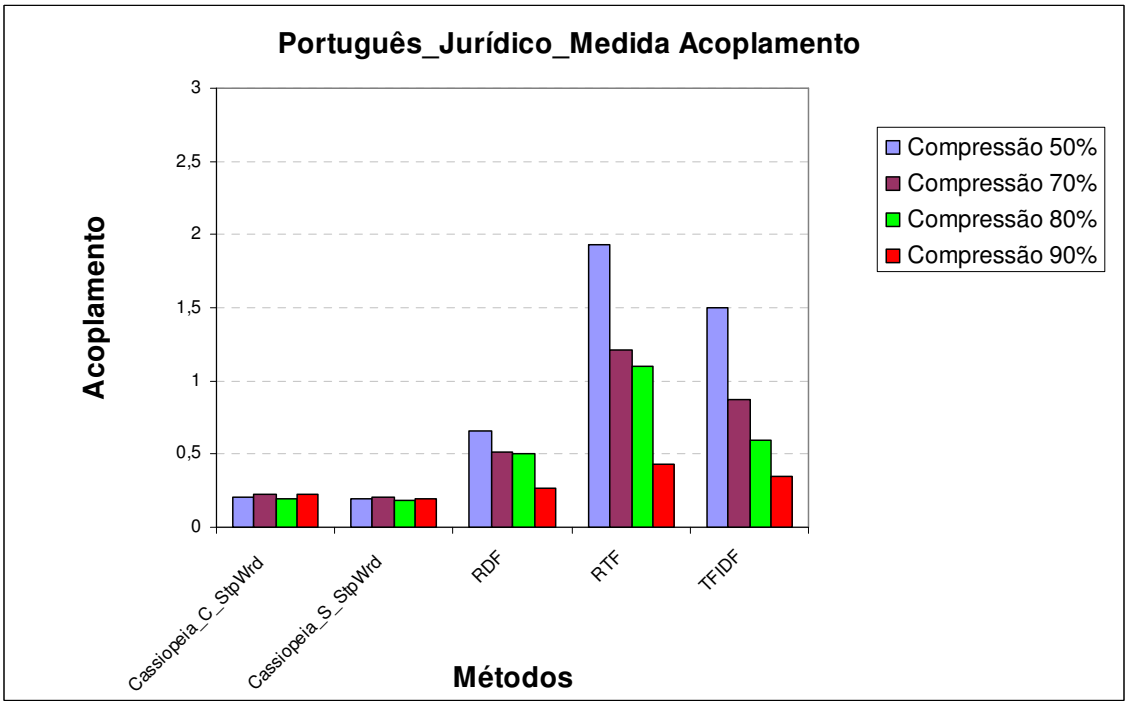


Figura 42b: Resultados das médias finais acumuladas da medida Acoplamento para os agrupamentos obtidos através dos métodos RDF, RTF e TFIDF e do modelo Cassiopeia com e sem *Stopwords*, no domínio jurídico e no idioma português.

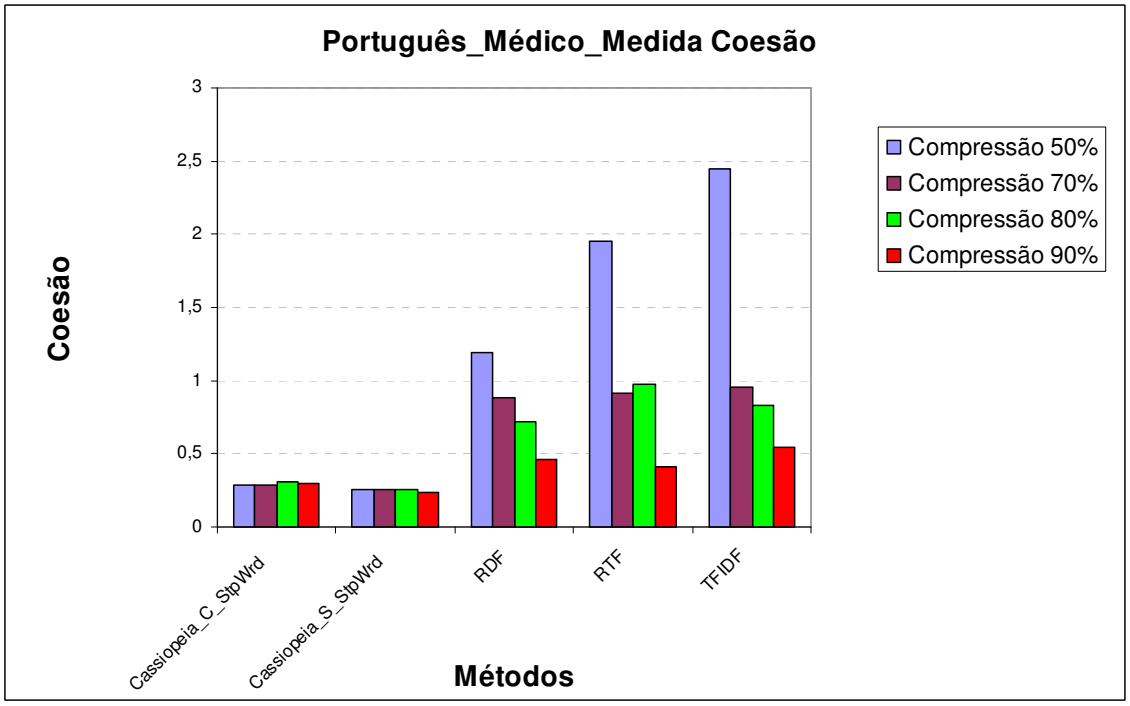


Figura 43a: Resultados das médias finais acumuladas da medida Coesão para os agrupamentos obtidos através dos métodos RDF, RTF e TFIDF e do modelo Cassiopeia com e sem *Stopwords,* no domínio médico e no idioma português.

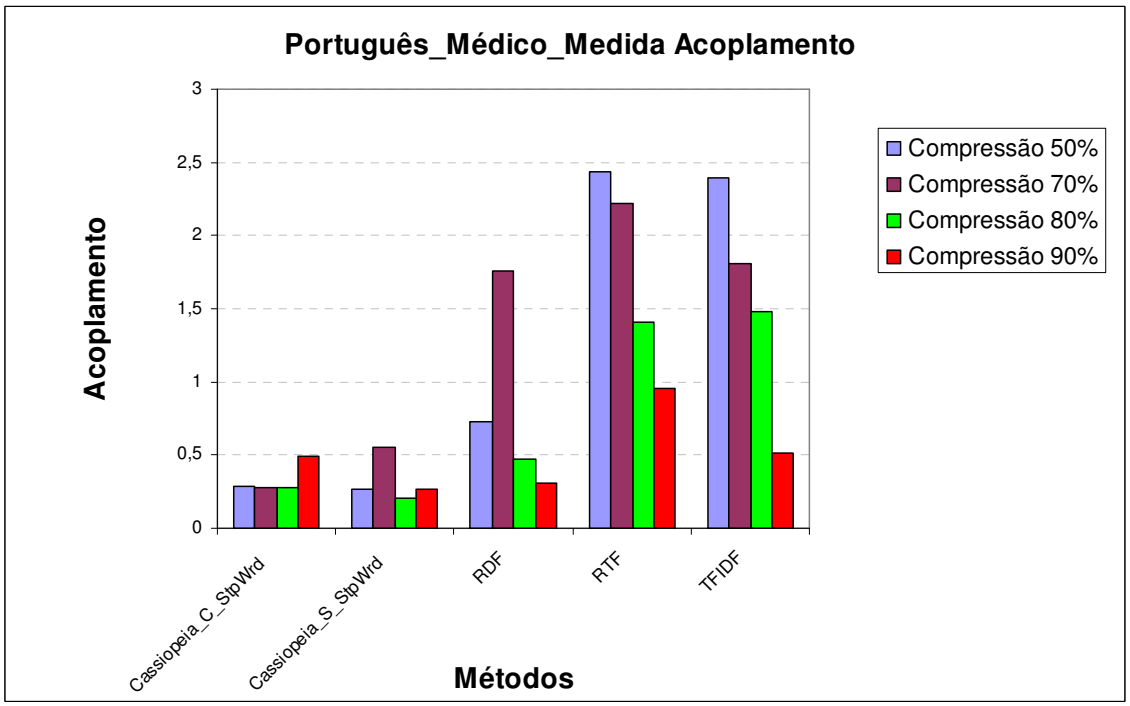


Figura 43b: Resultados das médias finais acumuladas da medida Acoplamento para os agrupamentos obtidos através dos métodos RDF, RTF e TFIDF e do modelo Cassiopeia com e sem *Stopwords,* no domínio médico e no idioma português.

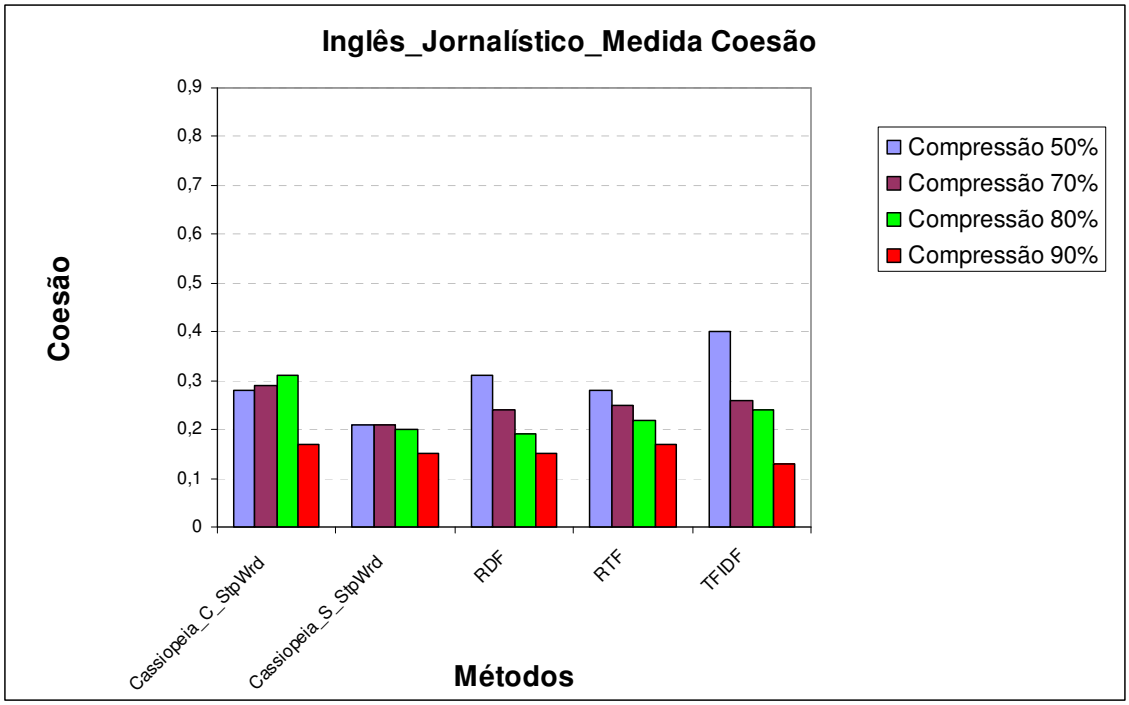


Figura 44a: Resultados das médias finais acumuladas da medida Coesão para os agrupamentos obtidos através dos métodos RDF, RTF e TFIDF e do modelo Cassiopeia com e sem *Stopwords,* no domínio jornalístico e no idioma inglês.

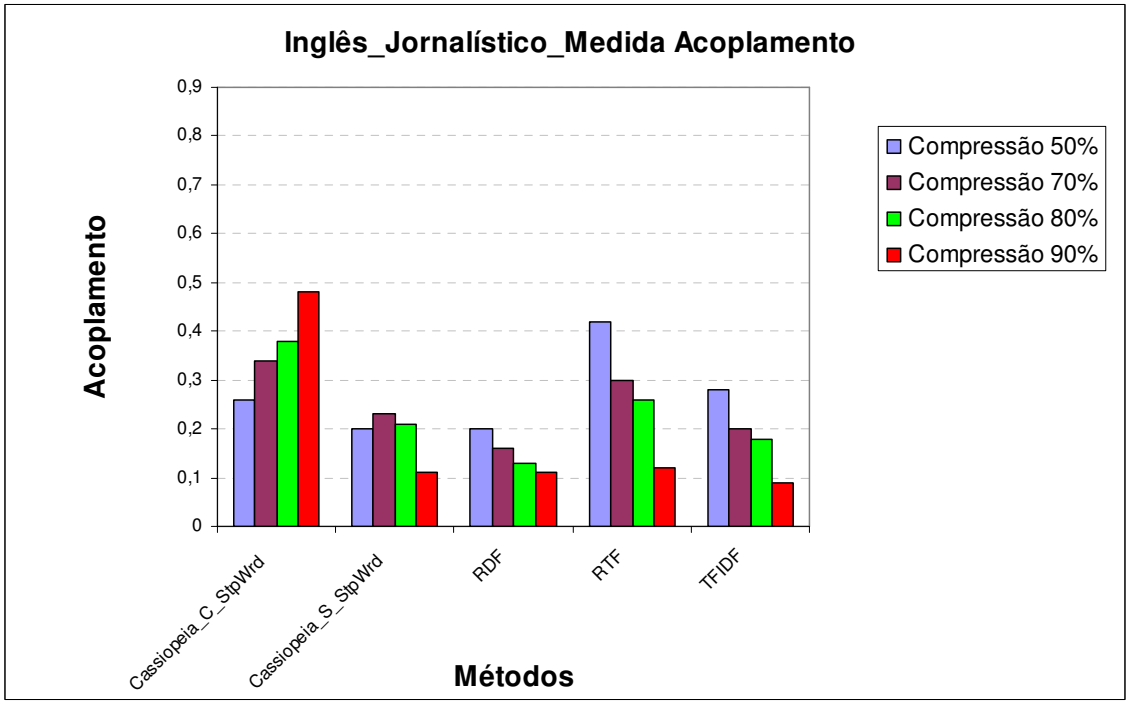


Figura 44b: Resultados das médias finais acumuladas da medida Acoplamento para os agrupamentos obtidos através dos métodos RDF, RTF e TFIDF e do modelo Cassiopeia com e sem *Stopwords,* no domínio jornalístico e no idioma inglês.

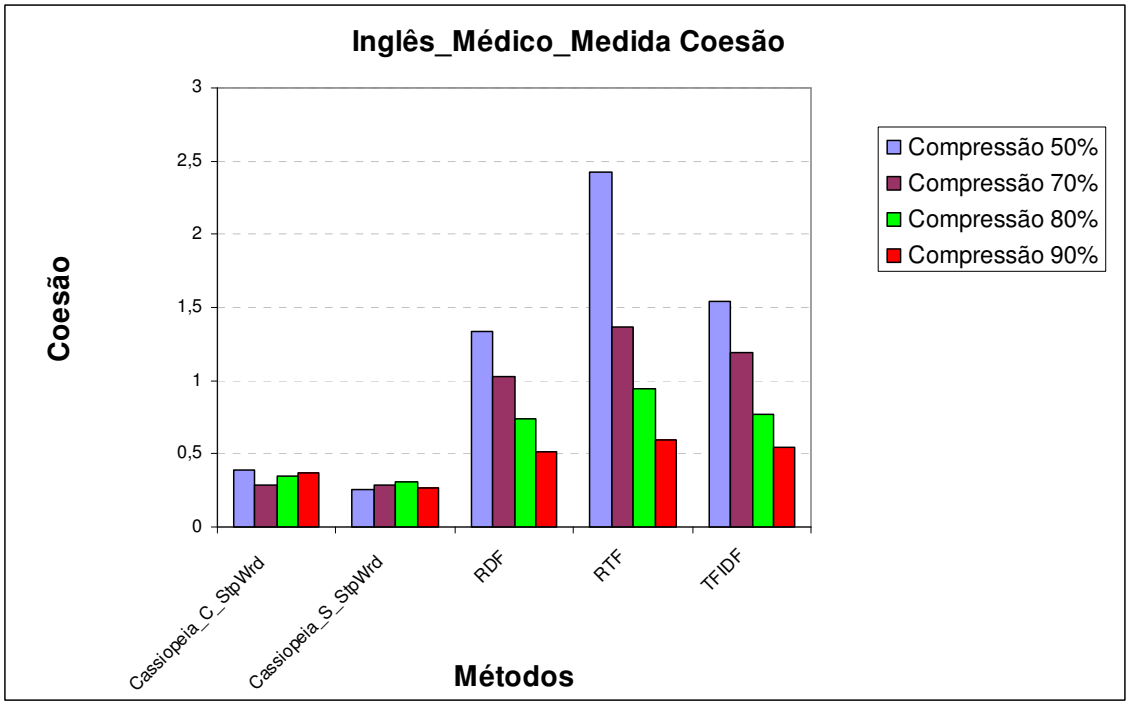


Figura 45a: Resultados das médias finais acumuladas da medida Coesão para os agrupamentos obtidos através dos métodos RDF, RTF e TFIDF e do modelo Cassiopeia com e sem *Stopword*s, no domínio médico e no idioma inglês.

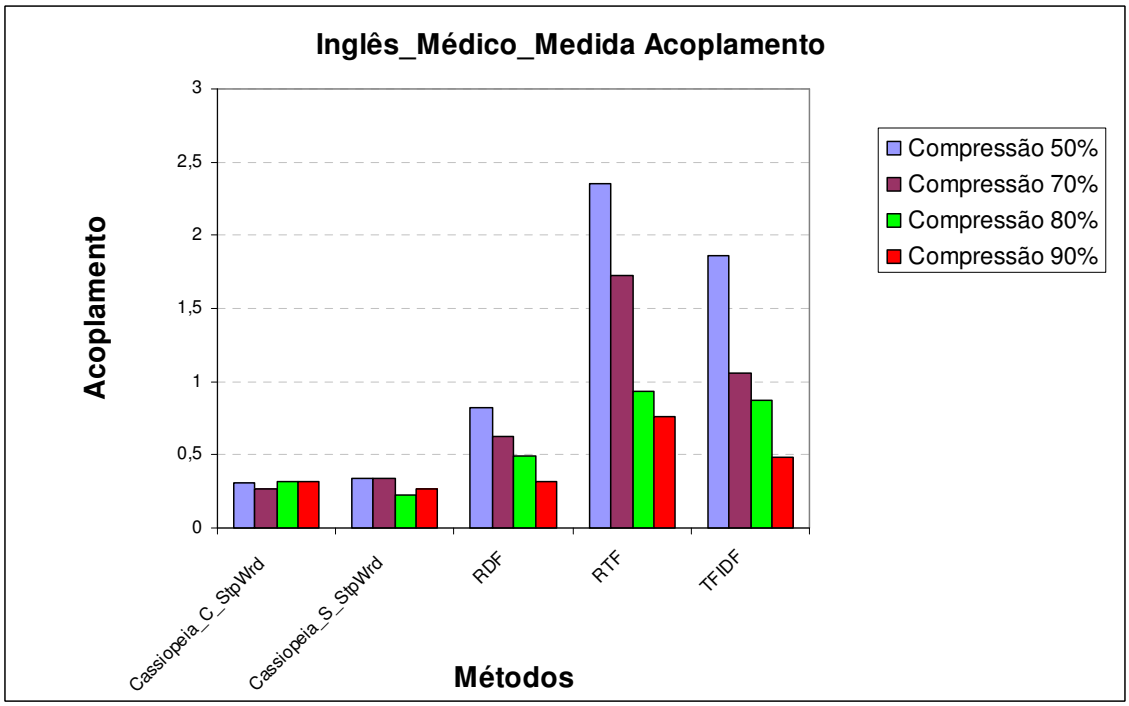


Figura 45b: Resultados das médias finais acumuladas da medida Acoplamento para os agrupamentos obtidos através dos métodos RDF, RTF e TFIDF e do modelo Cassiopeia com e sem *Stopwords,* no domínio médico e no idioma inglês.

APÊNDICE E

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. Aqui neste trabalho foram usados para realizar os testes estatísticos dos experimentos e comprovação da hipótese os seguintes softwares *StatPlus*® (http://www.analystsoft.com/en/products/statplus/) uma versão *Trial*, *InfoStat* (http://www.infostat.com.ar/) uma versão livre e o *Xlstat* (http://www.xlstat.com/en/products-solutions.html) uma versão para estudantes. Esses softwares foram escolhidos porque tinham os testes estatísticos ANOVA de Friedman e o coeficiente de concordância de Kendall adotado neste trabalho.

Tabela 1: Teste Estatístico das amostras das Medidas Externas do Domínio Jornalístico do idioma Inglês referente ao experimento 1.

Teste Estatístico **ANOVA de Friedman e Coeficiente de Concordância de Kendall** **Comparando amostras múltiplas relacionadas** **Jornalístico-Inglês** **Estatísticas Descritivas**																
Compressões	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**
	ANOVA de Friedman								**Coeficiente de Concordância de Kendall**							
N=100; **α**=0,05	**α(qui-quadrado) p-valor (bilateral)<**0,05				**SFr p-valor (bilateral)<**0,0001				**Coef. de Concordância de Kendall**				**Ordem médio**			
GL=7**;ValorCr=**14,06714	696,3119	692,9482	690,5437	687,9212	18.691,26	9.728,23	7.229,45	5.638,3	0,9947	0,9899	0,9865	0,9827	0,9947	0,9898	0,9864	0,9826
Algoritmos Sumarização	**Ordem médio**				**Soma de ordens**				**Média**				**Desvio Padrão**			
FA1_S_*StopWords*	7,995	4,525	7,005	6,625	799,5	452,5	700,5	662,5	0,3397	0,26	0,3231	0,3396	0,002	0,000	0,008	0,002
FA2_S_*StopWords*	3,865	6,365	5,015	4,325	386,5	636,5	501,5	432,5	0,268	0,2863	0,2534	0,3009	0,005	0,005	0,005	0,004
Copernic	3,135	6,635	7,995	4,675	313,5	663,5	799,5	467,5	0,2594	0,2888	0,388	0,3038	0,005	0,004	0,004	0,008
Intellexer Summarizer	7,005	1,855	2,395	8,	700,5	185,5	239,5	800,	0,3278	0,2183	0,2221	0,3889	0,004	0,005	0,011	0,005
SweSum	1,	1,145	1,035	6,375	100,	114,5	103,5	637,5	0,2003	0,2112	0,2099	0,3379	0,002	0,003	0,001	0,007
FA1_C_*StopWords*	5,485	4,47	5,955	2,785	548,5	447,	595,5	278,5	0,2898	0,2594	0,2686	0,2612	0,002	0,003	0,004	0,009
FA2_C_*StopWords*	5,505	8,	2,63	2,19	550,5	800,	263,	219,	0,29	0,3194	0,2235	0,2521	0,001	0,003	0,005	0,004
Texto s/ Sumarização	2,01	3,005	3,97	1,025	201,	300,5	397,	102,5	0,2401	0,2401	0,2401	0,2401	0,001	0,001	0,001	0,001

Tabela 2: Teste Estatístico das amostras das Medidas Externas do Domínio Médico do idioma Inglês referente ao experimento 1.

Teste Estatístico **ANOVA de Friedman e Coeficiente de Concordância de Kendall** **Comparando amostras múltiplas relacionadas** **Médico-Inglês** **Estatísticas Descritivas**																
Compressões	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**
	ANOVA de Friedman								**Coeficiente de Concordância de Kendall**							
N=100; **α**=0,05	**α(qui-quadrado) p-valor (bilateral)**<0,05				**SFr p-valor (bilateral)**<0,0001				**Coef. de Concordância de Kendall**				**Ordem médio**			
GL=7**;ValorCr=**14,06714	697,7558	687,1595	695,9952	675,7841	30780,45	5297,45	17205,11	2762,76	0,9968	0,9817	0,9943	0,9654	0,9968	0,9815	0,9942	0,9651
Algoritmos Sumarização	**Ordem médio**				**Soma de ordens**				**Média**				**Desvio Padrão**			
FA1_S_*StopWords*	2,01	1,	1,035	1,175	201,	100,	103,5	117,5	0,2384	0,2158	0,2212	0,2538	0,004	0,006	0,005	0,007
FA2_S_*StopWords*	6,	4,37	5,405	2,37	600,	437,	540,5	237,	0,3276	0,2881	0,2982	0,2703	0,004	0,005	0,004	0,002
Copernic	3,09	7,935	5,59	6,69	309,	793,5	559,	669,	0,258	0,3968	0,3	0,3161	0,008	0,014	0,000	0,009
Intellexer Summarizer	7,995	2,905	2,095	5,14	799,5	290,5	209,5	514,	0,4235	0,2634	0,2486	0,3021	0,008	0,008	0,012	0,004
SweSum	1,	6,135	8,	2,715	100,	613,5	800,	271,5	0,1995	0,3422	0,5347	0,2721	0,002	0,006	0,011	0,008
FA1_C_*StopWords*	4,985	2,175	2,875	3,74	498,5	217,5	287,5	374,	0,2962	0,2532	0,2622	0,2791	0,007	0,005	0,004	0,004
FA2_C_*StopWords*	3,915	4,55	4,	6,17	391,5	455,	400,	617,	0,2692	0,2898	0,2799	0,3097	0,003	0,002	0,004	0,002
Texto s/ Sumarização	7,005	6,93	7,	8,	700,5	693,	700,	800,	0,3501	0,3501	0,3501	0,3501	0,004	0,004	0,004	0,004

Tabela 3: Teste Estatístico das amostras das Medidas Externas do Domínio Jornalístico do idioma Português referente ao experimento 1.

Teste Estatístico **ANOVA de Friedman e Coeficiente de Concordância de Kendall** **Comparando amostras múltiplas relacionadas** **Jornalístico-Português** **Estatísticas Descritivas**																
Compressões	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**
	ANOVA de Friedman								**Coeficiente de Concordância de Kendall**							
N=100; **α**=0,05	**α(qui-quadrado) p-valor (bilateral)<**0,05				**SFr p-valor (bilateral)<**0,0001				**Coef. de Concordância de Kendall**				**Ordem médio**			
GL=7;**ValorCr=**14,06714	696,3119	687,7634	658,5533	633,7081	2.420,9	5.564,32	1.573,03	946,38	0,9947	0,9825	0,9408	0,9053	0,9947	0,9823	0,9402	0,9043
Algoritmos Sumarização	**Ordem médio**				**Soma de ordens**				**Média**				**Desvio Padrão**			
FA1_S_*StopWords*	6,165	6,005	5,55	1,935	616,5	600,5	555,	193,5	0,2485	0,2524	0,2414	0,2242	0,006	0,006	0,005	0,007
FA2_S_*StopWords*	6,695	4,59	6,935	6,425	669,5	459,	693,5	642,5	0,2547	0,232	0,2526	0,2593	0,008	0,005	0,005	0,004
Gist_Average_Keyword	1,33	1,	1,775	2,855	133,	100,	177,5	285,5	0,1856	0,1433	0,1822	0,2292	0,012	0,007	0,008	0,003
Gist_Intrasenteca	1,765	4,34	3,215	3,355	176,5	434,	321,5	335,5	0,1903	0,2293	0,2027	0,2319	0,005	0,004	0,012	0,005
SuPor2	4,03	2,305	1,52	3,015	403,	230,5	152,	301,5	0,2295	0,2149	0,1803	0,23	0,003	0,007	0,020	0,001
FA1_C_*StopWords*	3,065	7,795	3,605	6,625	306,5	779,5	360,5	662,5	0,22	0,2774	0,208	0,2619	0,004	0,005	0,005	0,007
FA2_C_ *StopWords*	4,955	2,765	5,41	3,86	495,5	276,5	541,	386,	0,2379	0,2196	0,2398	0,2337	0,007	0,004	0,002	0,006
Texto s/ Sumarização	7,995	7,2	7,99	7,93	799,5	720,	799,	793,	0,2715	0,2715	0,2715	0,2715	0,004	0,004	0,004	0,004

Tabela 4: Teste Estatístico das amostras das Medidas Externas do Domínio Jurídico do idioma Português referente ao experimento 1.

Teste Estatístico **ANOVA de Friedman e Coeficiente de Concordância de Kendall** **Comparando amostras múltiplas relacionadas** **Jurídico-Português** **Estatísticas Descritivas**																
Compressões	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**
	ANOVA de Friedman								**Coeficiente de Concordância de Kendall**							
N=100; **α**=0,05	**α(qui-quadrado) p-valor (bilateral)<**0,05				**SFr p-valor (bilateral)<**0,0001				**Coef. de Concordância de Kendall**				**Ordem médio**			
GL=7**;ValorCr=**14,06714	697,0755	692,3145	683,7627	686,6847	6.373,12	4.296,21	5.975,36	7.101,32	0,9958	0,989	0,9768	0,981	0,9958	0,9889	0,9766	0,9808
Algoritmos Sumarização	**Ordem médio**				**Soma de ordens**				**Média**				**Desvio Padrão**			
FA1_S_ *StopWords*	6,98	4,005	2,965	7,02	698,	400,5	296,5	702,	0,2492	0,2215	0,2202	0,2727	0,004	0,004	0,001	0,004
FA2_S_ *StopWords*	3,995	5,12	6,985	5,015	399,5	512,	698,5	501,5	0,1934	0,2496	0,2742	0,2308	0,005	0,002	0,005	0,003
Gist_Average_Keyword	1,005	8,	5,87	3,74	100,5	800,	587,	374,	0,1568	0,2829	0,2495	0,1999	0,005	0,005	0,003	0,006
Gist_Intrasenteca	2,25	1,775	2,145	7,98	225,	177,5	214,5	798,	0,1686	0,1572	0,2029	0,2898	0,003	0,005	0,012	0,004
SuPor2	6,015	1,225	8,	2,41	601,5	122,5	800,	241,	0,2302	0,1517	0,2914	0,1896	0,002	0,004	0,014	0,003
FA1_C_ *StopWords*	8,	6,615	4,45	5,985	800,	661,5	445,	598,5	0,2705	0,2601	0,2393	0,2414	0,003	0,001	0,003	0,004
FA2_C_ *StopWords*	5,005	6,26	4,585	2,785	500,5	626,	458,5	278,5	0,2188	0,2571	0,2404	0,1922	0,003	0,006	0,002	0,004
Texto s/ Sumarização	2,75	3,	1,	1,065	275,	300,	100,	106,5	0,1743	0,1743	0,1743	0,1743	0,006	0,006	0,006	0,006

Tabela 5: Teste Estatístico das amostras das Medidas Externas do Domínio Médico do idioma Português referente ao experimento 1.

Teste Estatístico **ANOVA de Friedman e Coeficiente de Concordância de Kendall** **Comparando amostras múltiplas relacionadas** **Médico-Português** **Estatísticas Descritivas**																
Compressões	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**
	ANOVA de Friedman								**Coeficiente de Concordância de Kendall**							
N=100; **α**=0,05	**α(qui-quadrado) p-valor (bilateral)<**0,05				**SFr p-valor (bilateral)<**0,0001				**Coef. de Concordância de Kendall**				**Ordem médio**			
GL=7**;ValorCr=**14,06714	698,3249	695,7834	695,3441	688,5914	41.271,2	16.335,98	14.785,49	5.975,36	0,9976	0,994	0,9933	0,9837	0,9976	0,9939	0,9933	0,9835
Algoritmos Sumarização	**Ordem médio**				**Soma de ordens**				**Média**				**Desvio Padrão**			
FA1_S_ *StopWords*	2,485	1,55	8,	6,455	248,5	155,	800,	645,5	0,2198	0,2501	0,29	0,2497	0,002	0,001	0,000	0,002
FA2_S_ *StopWords*	7,	3,53	5,51	2,285	700,	353,	551,	228,5	0,28	0,26	0,2601	0,22	0,000	0,000	0,001	0,000
Gist_Average_Keyword	4,52	8,	4,	6,505	452,	800,	400,	650,5	0,2305	0,3194	0,231	0,25	0,002	0,005	0,004	0,000
Gist_Intrasenteca	1,015	1,53	2,305	1,	101,5	153,	230,5	100,	0,21	0,25	0,2081	0,1806	0,000	0,000	0,004	0,002
SuPor2	4,47	3,39	2,59	4,86	447,	339,	259,	486,	0,23	0,2593	0,2102	0,24	0,000	0,003	0,001	0,000
FA1_C_ *StopWords*	8,	5,5	1,105	4,	800,	550,	110,5	400,	0,3	0,27	0,2001	0,2331	0,000	0,000	0,001	0,005
FA2_C_ *StopWords*	2,51	7,	5,495	2,895	251,	700,	549,5	289,5	0,22	0,2884	0,26	0,2243	0,000	0,004	0,000	0,005
Texto s/ Sumarização	6,	5,5	6,995	8,	600,	550,	699,5	800,	0,27	0,27	0,27	0,27	0,000	0,000	0,000	0,000

Tabela 6: Teste Estatístico das amostras das Medidas Internas do Domínio Jornalístico do idioma Inglês referente ao experimento 1.

Teste Estatístico **ANOVA de Friedman e Coeficiente de Concordância de Kendall** **Comparando amostras múltiplas relacionadas** **Jornalístico-Inglês** **Estatísticas Descritivas**																
Compressões	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**
	ANOVA de Friedman								**Coeficiente de Concordância de Kendall**							
N=100; **α**=0,05	**α(qui-quadrado) p-valor (bilateral)<**0,05				**SFr p-valor (bilateral)<**0,0001				**Coef. de Concordância de Kendall**				**Ordem médio**			
GL=7;ValorCr=14,06714	689,2925	697,1593	684,9938	698,6644	6.373,12	24.296,21	4.519,11	51.789,38	0,9847	0,9959	0,9786	0,9981	0,9845	0,9959	0,9783	0,9981
Algoritmos Sumarização	**Ordem médio**				**Soma de ordens**				**Média**				**Desvio Padrão**			
FA1_S_ *StopWords*	1,375	1,515	3,19	1,99	137,5	151,5	319,	199,	0,8718	0,8303	0,8598	0,7398	0,004	0,002	0,001	0,001
FA2_S_ *StopWords*	2,82	1,485	2,	1,01	282,	148,5	200,	101,	0,8851	0,83	0,84	0,73	0,005	0,000	0,000	0,000
Copernic	7,5	7,	7,35	5,07	750,	700,	735,	507,	0,95	0,91	0,9	0,8018	0,000	0,000	0,000	0,004
Intellexer Summarizer Pro	7,5	3,82	7,65	5,93	750,	382,	765,	593,	0,95	0,8824	0,903	0,818	0,000	0,006	0,005	0,004
SweSum	4,08	7,	1,	8,	408,	700,	100,	800,	0,9308	0,91	0,83	0,8911	0,003	0,000	0,000	0,003
FA1_C_ *StopWords*	5,46	7,	4,27	4,	546,	700,	427,	400,	0,94	0,91	0,8662	0,7892	0,000	0,000	0,005	0,003
FA2_C_ *StopWords*	5,46	5,	4,91	3,	546,	500,	491,	300,	0,94	0,9	0,87	0,7601	0,000	0,000	0,000	0,001
Texto s/ Sumarização	1,805	3,18	5,63	7,	180,5	318,	563,	700,	0,8756	0,8756	0,8756	0,8756	0,005	0,005	0,005	0,005

Tabela 7: Teste Estatístico das amostras das Medidas Internas do Domínio Médico do idioma Inglês referente ao experimento 1.

Teste Estatístico **ANOVA de Friedman e Coeficiente de Concordância de Kendall** **Comparando amostras múltiplas relacionadas** **Médico-Inglês** **Estatísticas Descritivas**																
Compressões	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**
	ANOVA de Friedman								**Coeficiente de Concordância de Kendall**							
N=100; **α**=0,05	**α(qui-quadrado) p-valor (bilateral)<**0,05				**SFr p-valor (bilateral)<**0,0001				**Coef. de Concordância de Kendall**				**Ordem médio**			
GL=7**;ValorCr=**14,06714	675,565	644,1726	662,4018	690,3754	2.737,09	1.142,33	1.744,17	7.101,32	0,9651	0,9202	0,9463	0,9863	0,9647	0,9194	0,9457	0,9861
Algoritmos Sumarização	**Ordem médio**				**Soma de ordens**				**Média**				**Desvio Padrão**			
FA1_S_ *StopWords*	2,04	1,5	1,55	2,015	204,	150,	155,	201,5	0,98	0,97	0,9603	0,94	0,000	0,000	0,002	0,000
FA2_S_ *StopWords*	2,04	1,5	1,485	1,	204,	150,	148,5	100,	0,98	0,97	0,96	0,9299	0,000	0,000	0,000	0,001
Copernic	6,04	5,505	6,815	7,	604,	550,5	681,5	700,	0,99	0,9814	0,98	0,97	0,000	0,003	0,000	0,000
Intellexer Summarizer Pro	6,	7,155	4,855	5,195	600,	715,5	485,5	519,5	0,9899	0,9869	0,9735	0,9579	0,001	0,005	0,005	0,004
SweSum	5,76	5,085	3,865	3,56	576,	508,5	386,5	356,	0,9893	0,98	0,9702	0,9497	0,003	0,000	0,001	0,002
FA1_C_ *StopWords*	6,04	5,085	3,8	5,6	604,	508,5	380,	560,	0,99	0,98	0,97	0,96	0,000	0,000	0,000	0,000
FA2_C_ *StopWords*	6,04	5,085	6,815	3,63	604,	508,5	681,5	363,	0,99	0,98	0,98	0,9501	0,000	0,000	0,000	0,001
Texto s/ Sumarização	2,04	5,085	6,815	8,	204,	508,5	681,5	800,	0,98	0,98	0,98	0,98	0,000	0,000	0,000	0,000

Tabela 8: Teste Estatístico das amostras das Medidas Internas do Domínio Jornalístico do idioma Português referente ao experimento 1.

Teste Estatístico																
ANOVA de Friedman e Coeficiente de Concordância de Kendall																
Comparando amostras múltiplas relacionadas																
Jornalístico-Português																
Estatísticas Descritivas																
Compressões	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**
	ANOVA de Friedman								**Coeficiente de Concordância de Kendall**							
N=100; **α**=0,05	**α(qui-quadrado) p-valor (bilateral)<**0,05				**SFr p-valor (bilateral)<**0,0001				**Coef. de Concordância de Kendall**				**Ordem médio**			
GL=7;**ValorCr=**14,06714	698,4094	686,1908	694,0474	699,5297	43.468,46	4.919,39	11.543,05	147.242,49	0,9977	0,9803	0,9915	0,9993	0,9977	0,9801	0,9914	0,9993
Algoritmos Sumarização	**Ordem médio**				**Soma de ordens**				**Média**				**Desvio Padrão**			
FA1_S_*StopWords*	1,895	2,38	2,025	4,	189,5	238,	202,5	400,	0,9279	0,8959	0,8593	0,8097	0,004	0,005	0,003	0,002
FA2_S_*StopWords*	1,105	1,205	2,13	3,	110,5	120,5	213,	300,	0,92	0,89	0,86	0,7695	0,000	0,000	0,000	0,002
Gist_Average_Keyword	3,5	3,2	4,	1,	350,	320,	400,	100,	0,94	0,9	0,87	0,5396	0,000	0,000	0,000	0,023
Gist_Intrasenteca	3,5	3,215	1,845	2,	350,	321,5	184,5	200,	0,94	0,9001	0,8581	0,618	0,000	0,001	0,004	0,004
SuPor2	7,	7,	6,99	7,	700,	700,	699,	700,	0,96	0,948	0,9198	0,88	0,000	0,004	0,001	0,000
FA1_C_*StopWords*	5,5	5,565	6,01	5,97	550,	556,5	601,	597,	0,95	0,93	0,91	0,8507	0,000	0,000	0,000	0,003
FA2_C_*StopWords*	5,5	5,435	5,	5,03	550,	543,5	500,	503,	0,95	0,9287	0,9	0,8408	0,000	0,003	0,000	0,003
Texto s/ Sumarização	8,	8,	8,	8,	800,	800,	800,	800,	0,9775	0,9775	0,9775	0,9775	0,004	0,004	0,004	0,004

Tabela 9: Teste Estatístico das amostras das Medidas Internas do Domínio Jurídico idioma Português referente ao experimento 1.

Teste Estatístico **ANOVA de Friedman e Coeficiente de Concordância de Kendall** **Comparando amostras múltiplas relacionadas** **Jurídico-Português** **Estatísticas Descritivas**																
Compressões	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**
	ANOVA de Friedman								**Coeficiente de Concordância de Kendall**							
N=100; **α**=0,05	**α(qui-quadrado) p-valor (bilateral)<**0,05				**SFr p-valor (bilateral)<**0,0001				**Coef. de Concordância de Kendall**				**Ordem médio**			
GL=7;ValorCr=14,06714	644,	691,8353	619,2261	657,7169	1.138,5	8.388,75	758,95	1.539,95	0,92	0,9883	0,8846	0,9396	0,9192	0,9882	0,8834	0,939
Algoritmos Sumarização	**Ordem médio**				**Soma de ordens**				**Média**				**Desvio Padrão**			
FA1_S_*StopWords*	4,04	1,53	3,365	1,84	404,	153,	336,5	184,	0,99	0,9801	0,98	0,96	0,000	0,001	0,000	0,000
FA2_S_*StopWords*	4,04	1,53	3,365	1,84	404,	153,	336,5	184,	0,99	0,9801	0,98	0,96	0,000	0,001	0,000	0,000
Gist_Average_Keyword	4,04	5,49	3,565	4,68	404,	549,	356,5	468,	0,99	0,99	0,9805	0,9695	0,000	0,000	0,002	0,002
Gist_Intrasenteca	4,04	5,49	3,365	2,93	404,	549,	336,5	293,	0,99	0,99	0,98	0,9637	0,000	0,000	0,000	0,005
SuPor2	7,72	5,49	7,365	6,98	772,	549,	736,5	698,	0,9992	0,99	0,99	0,98	0,003	0,000	0,000	0,000
FA1_C_*StopWords*	4,04	5,49	3,445	4,91	404,	549,	344,5	491,	0,99	0,99	0,9802	0,9704	0,000	0,000	0,001	0,002
FA2_C_*StopWords*	4,04	5,49	4,165	4,82	404,	549,	416,5	482,	0,99	0,99	0,982	0,97	0,000	0,000	0,004	0,000
Texto s/ Sumarização	4,04	5,49	7,365	8,	404,	549,	736,5	800,	0,99	0,99	0,99	0,99	0,000	0,000	0,000	0,000

Tabela 10: Teste Estatístico das amostras das Medidas Internas do Domínio Médico idioma Português referente ao experimento 1.

Teste Estatístico **ANOVA de Friedman e Coeficiente de Concordância de Kendall** **Comparando amostras múltiplas relacionadas** **Médico-Português** **Estatísticas Descritivas**																
Compressões	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**
	ANOVA de Friedman								**Coeficiente de Concordância de Kendall**							
N=100; **α**=0,05	**α(qui-quadrado) p-valor (bilateral)<**0,05				**SFr p-valor (bilateral)<**0,0001				**Coef. de Concordância de Kendall**				**Ordem médio**			
GL=7;**ValorCr=**14,06714	645,7803	697,348	654,2874	685,8398	1.179,13	26.032,31	1.416,99	4.794,99	0,9225	0,9962	0,9347	0,9798	0,9218	0,9962	0,934	0,9796
Algoritmos Sumarização	**Ordem médio**				**Soma de ordens**				**Média**				**Desvio Padrão**			
FA1_S_*StopWords*	3,32	1,005	2,265	3,165	332,	100,5	226,5	316,5	0,98	0,96	0,9611	0,9399	0,000	0,000	0,003	0,001
FA2_S_*StopWords*	1,345	2,51	2,255	1,175	134,5	251,	225,5	117,5	0,9721	0,97	0,961	0,9299	0,004	0,000	0,003	0,001
Copernic	3,8	5,015	1,935	5,52	380,	501,5	193,5	552,	0,9812	0,98	0,96	0,9499	0,003	0,000	0,000	0,001
Intellexer Summarizer Pro	3,32	2,49	4,945	2,485	332,	249,	494,5	248,5	0,98	0,9699	0,9699	0,9364	0,000	0,001	0,001	0,005
SweSum	6,925	7,51	6,86	6,94	692,5	751,	686,	694,	0,99	0,99	0,9792	0,9596	0,000	0,000	0,003	0,003
FA1_C_*StopWords*	6,925	5,015	4,74	3,18	692,5	501,5	474,	318,	0,99	0,98	0,9693	0,94	0,000	0,000	0,003	0,000
FA2_C_*StopWords*	3,52	4,995	5,015	5,535	352,	499,5	501,5	553,5	0,9805	0,9799	0,9702	0,95	0,002	0,001	0,001	0,000
Texto s/ Sumarização	6,845	7,46	7,985	8,	684,5	746,	798,5	800,	0,9898	0,9898	0,9898	0,9898	0,001	0,001	0,001	0,001

Tabela 11: Teste Estatístico das amostras das Medidas Externas do Domínio Jornalístico idioma Inglês referente ao experimento 2.

Teste Estatístico **ANOVA de Friedman e Coeficiente de Concordância de Kendall** **Comparando amostras múltiplas relacionadas** **Jornalistico-Inglês** **Estatísticas Descritivas**																
Compressões	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**
	ANOVA de Friedman								**Coeficiente de Concordância de Kendall**							
N=100; **α**=0,05	**α(qui-quadrado) p-valor (bilateral)<**0,05				**SFr p-valor (bilateral)<**0,0001				**Coef. de Concordância de Kendall**				**Ordem médio**			
GL=4**;ValorCr=**9,488	393,448	400,	381,656	394,8098	5.944,96	1,00E+30	2.059,74	100.820,39	0,9836	1,	0,9541	0,987	0,9835	1,	0,9537	0,9868
Métodos Seleção Atributos	**Ordem médio**				**Soma de ordens**				**Média**				**Desvio Padrão**			
Cassiopeia_C_StpWrd	2,	1,	1,01	3,97	200,	100,	101,	397	0,3439	0,3497	0,3773	0,4097	0,004	0,005	0,008	0,002
Cassiopeia_S_StpWrd	3,09	4,	4,14	5	309,	400,	414,	500	0,4272	0,422	0,483	0,4797	0,004	0,003	0,006	0,002
RDF	1,	2,	2,89	2,445	100,	200,	289,	244,5	0,3277	0,3763	0,4557	0,3898	0,002	0,002	0,011	0,002
RTF	3,91	3,	4,86	1	391,	300,	486,	100	0,4306	0,3898	0,4942	0,3589	0,002	0,003	0,005	0,004
TFIDF	5,	5,	2,1	2,585	500,	500,	210,	258,5	0,47	0,4498	0,4442	0,3912	0,005	0,006	0,003	0,003

Tabela 12: Teste Estatístico das amostras das Medidas Externas do Domínio Médico idioma Inglês referente ao experimento 2.

Teste Estatístico **ANOVA de Friedman e Coeficiente de Concordância de Kendall** **Comparando amostras múltiplas relacionadas** **Médico-Inglês** **Estatísticas Descritivas**																
Compressões	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**
	ANOVA de Friedman								**Coeficiente de Concordância de Kendall**							
N=100; **α**=0,05	**α(qui-quadrado) p-valor (bilateral)<**0,05				**SFr p-valor (bilateral)<**0,0001				**Coef. de Concordância de Kendall**				**Ordem médio**			
GL=4**;ValorCr=**9,488	398,8028	381,9027	387,8084	380,0062	5.944,96	1,00E+30	2.059,74	20.105,49	0,997007	0,954757	0,969521	0,950015	0,996977	0,9543	0,969213	0,94951
Métodos Seleção Atributos	**Ordem médio**				**Soma de ordens**				**Média**				**Desvio Padrão**			
Cassiopeia_C_StpWrd	2,005	4,62	1	1,325	200,5	462	100	132,5	0,3706	0,507	0,3551	0,3952	0,004	0,004	0,005	0,005
Cassiopeia_S_StpWrd	2,995	4,38	2,255	2,995	299,5	438	225,5	299,5	0,3947	0,5052	0,3974	0,4288	0,010	0,004	0,003	0,006
RDF	5	1,54	4,19	4,045	500	154	419	404,5	0,45	0,3813	0,4491	0,44	0,001	0,004	0,005	0,002
RTF	3,995	3	4,81	1,71	399,5	300	481	171	0,4075	0,4415	0,4556	0,3987	0,003	0,004	0,006	0,003
TFIDF	1,005	1,46	2,745	4,925	100,5	146	274,5	492,5	0,3501	0,3805	0,4029	0,4499	0,003	0,002	0,004	0,007

Tabela 13: Teste Estatístico das amostras das Medidas Externas do Domínio Jornalístico idioma Português referente ao experimento 2.

Teste Estatístico																
ANOVA de Friedman e Coeficiente de Concordância de Kendall																
Comparando amostras múltiplas relacionadas																
Jornalístico-Português																
Estatísticas Descritivas																
Compressões	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**
	ANOVA de Friedman								**Coeficiente de Concordância de Kendall**							
N=100; **α**=0,05	**α(qui-quadrado) p-valor (bilateral)<**0,05				**SFr p-valor (bilateral)<**0,0001				**Coef. de Concordância de Kendall**				**Ordem médio**			
GL=4;ValorCr=9,488	399,4011	396,7759	384,8715	319,7408	66.022,4	12.183,42	2.518,58	394,4	0,9985	0,9919	0,9622	0,7994	0,9985	0,9919	0,9618	0,7973
Métodos Seleção Atributos	**Ordem médio**				**Soma de ordens**				**Média**				**Desvio Padrão**			
Cassiopeia_C_StpWrd	1,	5,	4,98	5,	100,	500,	498,	500,	0,1909	0,3512	0,2592	0,2925	0,004	0,004	0,006	0,005
Cassiopeia_S_StpWrd	3,	3,	3,845	2,8	300,	300,	384,5	280,	0,2598	0,2762	0,2482	0,2561	0,002	0,005	0,004	0,006
RDF	2,005	1,93	1,	1,54	200,5	193,	100,	154,	0,2392	0,2497	0,2	0,25	0,004	0,002	0,000	0,000
RTF	4,995	4,	2,51	3,51	499,5	400,	251,	351,	0,3104	0,3001	0,2391	0,2599	0,003	0,001	0,003	0,001
TFIDF	4,	1,07	2,665	2,15	400,	107,	266,5	215,	0,2962	0,2411	0,2403	0,2531	0,006	0,003	0,002	0,005

Tabela 14: Teste Estatístico das amostras das Medidas Externas do Domínio Jurídico idioma Português referente ao experimento 2.

Teste Estatístico **ANOVA de Friedman e Coeficiente de Concordância de Kendall** **Comparando amostras múltiplas relacionadas** **Jurídico-Português** **Estatísticas Descritivas**																
Compressões	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**
	ANOVA de Friedman								**Coeficiente de Concordância de Kendall**							
N=100; **α**=0,05	**α(qui-quadrado) p-valor (bilateral)<**0,05				**SFr p-valor (bilateral)<**0,0001				**Coef. de Concordância de Kendall**				***Ordem médio***			
GL=4;ValorCr=9,488	395,9253	391,7424	375,5744	348,8727	9.619,57	4.696,59	1.522,25	675,54	0,9898	0,9794	0,9389	0,8722	0,9897	0,9791	0,9383	0,8709
Métodos Seleção Atributos	**Ordem médio**				**Soma de ordens**				**Média**				**Desvio Padrão**			
Cassiopeia_C_StpWrd	2,095	3,425	5,	1,565	209,5	342,5	500,	156,5	0,26	0,2098	0,3231	0,3401	0,000	0,001	0,005	0,002
Cassiopeia_S_StpWrd	4,	4,995	3,215	2,22	400,	499,5	321,5	222,	0,2912	0,2299	0,2924	0,3445	0,003	0,002	0,005	0,005
RDF	1,005	1,	1,	4,775	100,5	100,	100,	477,5	0,25	0,1765	0,26	0,39	0,000	0,005	0,000	0,000
RTF	5,	2,1	2,945	4,22	500,	210,	294,5	422,	0,3179	0,2009	0,2907	0,3838	0,004	0,004	0,004	0,006
TFIDF	2,9	3,48	2,84	2,22	290,	348,	284,	222,	0,2681	0,2103	0,2901	0,3464	0,004	0,002	0,001	0,009

Tabela 15: Teste Estatístico das amostras das Medidas Externas do Domínio Médico idioma Português referente ao experimento 2.

Teste Estatístico **ANOVA de Friedman e Coeficiente de Concordância de Kendall** **Comparando amostras múltiplas relacionadas** **Médico-Português** **Estatísticas Descritivas**																
Compressões	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**
	ANOVA de Friedman								**Coeficiente de Concordância de Kendall**							
N=100; **α**=0,05	**α(qui-quadrado) p-valor (bilateral)<**0,05				**SFr p-valor (bilateral)<**0,0001				**Coef. de Concordância de Kendall**				***Ordem médio***			
GL=4**;ValorCr=**9,488	382,4126	400,	395,9087	394,5634	2.152,61	1,00E+30	9.580,	7.184,92	0,956	1,	0,9898	0,9864	0,9556	1,	0,9897	0,9863
Métodos Seleção Atributos	**Ordem médio**				**Soma de ordens**				**Média**				**Desvio Padrão**			
Cassiopeia_C_StpWrd	2,955	4,	3,	1,755	295,5	400,	300,	175,5	0,36	0,3809	0,3386	0,3246	0,000	0,003	0,004	0,005
Cassiopeia_S_StpWrd	1,27	5,	4,	4,	127,	500,	400,	400,	0,3446	0,4507	0,3797	0,399	0,008	0,003	0,002	0,003
RDF	5,	2,	1,635	1,25	500,	200,	163,5	125,	0,4683	0,3401	0,2924	0,319	0,004	0,001	0,005	0,004
RTF	3,98	3,	1,365	5,	398,	300,	136,5	500,	0,3798	0,3604	0,2897	0,4403	0,001	0,002	0,002	0,003
TFIDF	1,795	1,	5,	2,995	179,5	100,	500,	299,5	0,3497	0,2999	0,4304	0,345	0,002	0,001	0,002	0,006

Tabela 16: Teste Estatístico das amostras das Medidas Internas do Domínio Jornalístico idioma Inglês referente ao experimento 2.

Teste Estatístico **ANOVA de Friedman e Coeficiente de Concordância de Kendall** **Comparando amostras múltiplas relacionadas** **Jornalístico-Inglês** **Estatísticas Descritivas**																
Compressões	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**
	ANOVA de Friedman								**Coeficiente de Concordância de Kendall**							
N=100; **α**=0,05	**α(qui-quadrado) p-valor (bilateral)<**0,05				**SFr p-valor (bilateral)<**0,0001				**Coef. de Concordância de Kendall**				**Ordem médio**			
GL=4**;ValorCr=**9,488	400,	396,0042	400,	399,6076	1,00E+30	9.811,29	1,00E+30	100.820,39	1,	0,99	1,	0,999	1,	0,9899	1,	0,999
Métodos Seleção Atributos	**Ordem médio**				**Soma de ordens**				**Média**				**Desvio Padrão**			
Cassiopeia_C_StpWrd	5,	5,	5,	5,	500,	500,	500,	500,	0,94	0,92	0,8796	0,8398	0,000	0,000	0,002	0,001
Cassiopeia_S_StpWrd	4,	4,	4,	4,	400,	400,	400,	400,	0,89	0,87	0,8206	0,7715	0,000	0,000	0,002	0,004
RDF	1,	1,37	1,	1,01	100,	137,	100,	101,	0,79	0,7878	0,73	0,74	0,000	0,004	0,000	0,000
RTF	3,	3,	3,	3,	300,	300,	300,	300,	0,8297	0,8297	0,7686	0,76	0,002	0,002	0,003	0,000
TFIDF	2,	1,63	2,	1,99	200,	163,	200,	199,	0,8	0,7904	0,75	0,7498	0,000	0,002	0,000	0,001

Tabela 17: Teste Estatístico das amostras das Medidas Internas do Domínio Médico idioma Inglês referente ao experimento 2.

Teste Estatístico **ANOVA de Friedman e Coeficiente de Concordância de Kendall** **Comparando amostras múltiplas relacionadas** **Médico-Inglês** **Estatísticas Descritivas**																
Compressões	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**
	ANOVA de Friedman								**Coeficiente de Concordância de Kendall**							
N=100; **α**=0,05	**α(qui-quadrado) p-valor (bilateral)<**0,05				**SFr p-valor (bilateral)<**0,0001				**Coef. de Concordância de Kendall**				**Ordem médio**			
GL=4;ValorCr=9,488	397,9508	388,4555	396,8977	381,6716	19.225,72	3.331,19	12.665,75	2.061,58	0,9949	0,9711	0,9922	0,9542	0,9948	0,9708	0,9922	0,9537
Métodos Seleção Atributos	**Ordem médio**				**Soma de ordens**				**Média**				**Desvio Padrão**			
Cassiopeia_C_StpWrd	5,	4,97	5,	5,	500,	497,	500,	500,	0,99	0,9801	0,97	0,96	0,000	0,001	0,000	0,000
Cassiopeia_S_StpWrd	4,	4,03	4,	4,	400,	403,	400,	400,	0,98	0,9707	0,96	0,9401	0,000	0,003	0,000	0,001
RDF	1,	1,1	1,96	1,735	100,	110,	196,	173,5	0,87	0,8697	0,8498	0,8398	0,000	0,002	0,001	0,001
RTF	2,445	2,615	2,025	1,77	244,5	261,5	202,5	177,	0,89	0,88	0,8504	0,8401	0,000	0,000	0,002	0,001
TFIDF	2,555	2,285	2,015	2,495	255,5	228,5	201,5	249,5	0,8911	0,8777	0,8503	0,8449	0,003	0,004	0,002	0,005

Tabela 18: Teste Estatístico das amostras das Medidas Internas do Domínio Jornalístico idioma Português referente ao experimento 2.

Teste Estatístico																
ANOVA de Friedman e Coeficiente de Concordância de Kendall																
Comparando amostras múltiplas relacionadas																
Jornalístico-Português																
Estatísticas Descritivas																
Compressões	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**
	ANOVA de Friedman								**Coeficiente de Concordância de Kendall**							
N=100; **α**=0,05	**α(qui-quadrado) p-valor (bilateral)<**0,05				**SFr p-valor (bilateral)<**0,0001				**Coef. de Concordância de Kendall**				**Ordem médio**			
GL=4**;ValorCr=**9,488	400	395,7304	395,0607	399,5878	1,00E+30	9.175,94	7.918,4	95.971,41	1,	0,9893	0,9877	0,999	1,	0,9892	0,9875	0,999
Métodos Seleção Atributos	**Ordem médio**				**Soma de ordens**				**Média**				**Desvio Padrão**			
Cassiopeia_C_StpWrd	5	5,	5,	5,	500	500,	500,	500,	0,9303	0,9102	0,8797	0,747	0,002	0,001	0,002	0,006
Cassiopeia_S_StpWrd	4	4,	4,	4,	400	400,	400,	400,	0,9014	0,86	0,81	0,669	0,005	0,000	0,000	0,003
RDF	1	1,645	1,795	1,51	100	164,5	179,5	151,	0,7796	0,76	0,7303	0,6302	0,002	0,000	0,002	0,001
RTF	3	3,	3,	3,	300	300,	300,	300,	0,8302	0,7889	0,75	0,65	0,001	0,003	0,000	0,000
TFIDF	2	1,355	1,205	1,49	200	135,5	120,5	149,	0,8104	0,7569	0,7242	0,63	0,002	0,005	0,005	0,000

Tabela 19: Teste Estatístico das amostras das Medidas Internas do Domínio Jurídico idioma Português referente ao experimento 2.

Teste Estatístico **ANOVA de Friedman e Coeficiente de Concordância de Kendall** **Comparando amostras múltiplas relacionadas** **Jurídico-Português** **Estatísticas Descritivas**																
Compressões	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**
	ANOVA de Friedman								**Coeficiente de Concordância de Kendall**							
N=100; **α**=0,05	**α(qui-quadrado) p-valor (bilateral)<**0,05				**SFr p-valor (bilateral)<**0,0001				**Coef. de Concordância de Kendall**				**Ordem médio**			
GL=4**;ValorCr=**9,488	400,	375,6278	399,3544	398,1242	1,00E+30	1.525,8	61.248,18	21.012,11	1,	0,9391	0,9984	0,9953	1,	0,9385	0,9984	0,9953
Métodos Seleção Atributos	**Ordem médio**				**Soma de ordens**				**Média**				**Desvio Padrão**			
Cassiopeia_C_StpWrd	5,	4,925	4,5	5,	500,	492,5	450,	500,	0,99	0,99	0,98	0,97	0,000	0,000	0,000	0,000
Cassiopeia_S_StpWrd	4,	4,075	4,5	4,	400,	407,5	450,	400,	0,96	0,9815	0,98	0,96	0,000	0,004	0,000	0,000
RDF	2,5	2,455	2,485	1,015	250,	245,5	248,5	101,5	0,93	0,9248	0,92	0,9001	0,000	0,005	0,000	0,001
RTF	1,	1,735	1,	2,485	100,	173,5	100,	248,5	0,92	0,92	0,91	0,9099	0,000	0,000	0,000	0,001
TFIDF	2,5	1,81	2,515	2,5	250,	181,	251,5	250,	0,93	0,9205	0,9203	0,91	0,000	0,002	0,002	0,000

Tabela 20: Teste Estatístico das amostras das Medidas Internas do Domínio Médico idioma Português referente ao experimento 2.

Teste Estatístico **ANOVA de Friedman e Coeficiente de Concordância de Kendall** **Comparando amostras múltiplas relacionadas** **Médico-Português** **Estatísticas Descritivas**																
Compressões	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**	**50%**	**70%**	**80%**	**90%**
	ANOVA de Friedman								**Coeficiente de Concordância de Kendall**							
N=100; **α**=0,05	**α(qui-quadrado) p-valor (bilateral)<**0,05				**SFr p-valor (bilateral)<**0,0001				**Coef. de Concordância de Kendall**				**Ordem médio**			
GL=4;ValorCr=9,488	399,5649	399,4171	394,9429	393,4842	90.920,39	67.840,18	7.731,51	5.978,53	0,9989	0,9985	0,9874	0,9837	0,9989	0,9985	0,9872	0,9835
Métodos Seleção Atributos	**Ordem médio**				**Soma de ordens**				**Média**				**Desvio Padrão**			
Cassiopeia_C_StpWrd	4,51	5,	4,995	5,	451,	500,	499,5	500,	0,98	0,98	0,9599	0,93	0,000	0,000	0,001	0,000
Cassiopeia_S_StpWrd	4,49	4,	4,005	4,	449,	400,	400,5	400,	0,9798	0,9699	0,95	0,9184	0,001	0,001	0,000	0,004
RDF	2,5	1,015	1,	1,14	250,	101,5	100,	114,	0,83	0,8203	0,8106	0,7923	0,000	0,002	0,002	0,004
RTF	1,	1,985	2,195	1,91	100,	198,5	219,5	191,	0,78	0,83	0,83	0,8	0,000	0,000	0,000	0,000
TFIDF	2,5	3,	2,805	2,95	250,	300,	280,5	295,	0,83	0,8511	0,8361	0,8095	0,000	0,003	0,005	0,002

APÊNDICE F

OS RESULTADOS PARCIAIS DE PUBLICAÇÕES DA PESQUISA

Capítulos de livros publicados

1. GUELPELI, M. V. C.; BERNARDINI, F. C.; GARCIA, A. C. B. **An Analysis of Constructed Categories for Textual Classification using Fuzzy Similarity and Agglomerative Hierarchical Methods**. Emergent Web Intelligence: Advanced Semantic Technologies Emergent Web Intelligence: Advanced Semantic Technologies Series: Advanced Information and Knowledge Processing Badr, Y.; Chbeir, R.; Abraham, A.; Hassanien, A.-E. (Eds.). Springer London Dordrecht Heidelberg New York 1st Edition, 2010, XVI, 544 p. 178 illus., Hardcover, ISSN 1610-3947,ISBN 978-1-84996-076-2 e-ISBN 978-1-84996-077-9 e DOI 10.1007/978-1-84996-077-9, 2010.

Artigos completos publicados em periódicos

1. GUELPELI, M. V. C.; GARCIA, A. C. B.; BRANCO H. A. **The Cassiopeia Model: A study with other algorithms for attribute selection in text clusterization**. International Journal of Web Applications, Print ISSN: 0974-7710,Online ISSN: 0974-7729,Volume: 3,Issue: 3 (September 2011), p. 110-121,USA, 2011.

Trabalhos completos publicados em anais de congressos

1. GUELPELI, M. V. C.; GARCIA, A. C.; B.BRANCO H. A. **The process of summarization in the pre-processing stage in order to improve measurement of texts when clustering.** In Proceedings of the The 6th International Conference for Internet Technology and Secured Transactions, IEEE, p ??? - ??? Abu Dhabi,UAE, 2011.
2. .GUELPELI, M. V. C.; GARCIA, A. C.; B.BRANCO H. A. **The Cassiopeia Model: Using summarization and clusterization for semantic knowledge management.** In Proceedings of the ICADIWT 2011-The Fourth International Conference on the Applications of Digital Information and Web Technologies 2011, IEEE, p 97 - 105 Stevens Point, USA, 2011.
3. GUELPELI, M. V. C.; BRANCO H. A.; GARCIA, A. C. B. **CASSIOPEIA: A Model Based on Summarization and Clusterization used for Knowledge Discovery in Textual Bases**. In Proceedings of the IEEE NLP-Ke'2009 - IEEE International

Conference on Natural Language Processing and Knowledge Engineering, Dalian, September 24-27, China, 2009.

4. GUELPELI, M. V. C.; BERNARDINI, F. C.; GARCIA, A. C. B. **Todas as Palavras da Sentença como Métrica para um Sumarizador Automático.** In: Tecnologia da Informação e da Linguagem Humana-TIL, WebMedia, 2008. p. 287-291,Vila Velha,Brasil, 2008.
5. GUELPELI, M. V. C.; GARCIA, A. C. B. **An Analysis of Constructed Categories for Textual Classification using Fuzzy Similarity and Agglomerative Hierarchical Methods.** In: The 2007 IEEE International Conference on Signal-Image Technologies and Internet-Based System, 2007. SITIS '07. Third International IEEE Conference on On page(s): 92 - 99 Location: Shanghai,ISBN: 978-0-7695-3122-9,Digital Object Identifier(DOI): 10.1109/SITIS.2007.109,China, 2007.
6. GUELPELI, M. V. C.; GARCIA, A. C. B. **Automatic Summarizer Based on Pragmatic Profiles.** In: IADIS International Conference WWW/Internet 2007, Vila Real,ISBN: 978-972-8924-44-7, IADIS Press,v.II. p.149 - 153,Portugal, 2007.
7. OLIVEIRA, M. A.; GUELPELI, M. V. C **The Performance of BLMSumm: Distinct Languages with Antagonistic Domains and Varied Compressions**. The Second International Conference on Information Science and Technology (ICIST 2012), Wuhan during March 23-25, China, 2012.
8. OLIVEIRA, M. A.; GUELPELI, M. V. C. **BLMSumm - Métodos de Busca Local e Metaheurísticas na Sumarização de Textos** In: Proceedings of the ENIA - VIII Encontro Nacional de Inteligência Artificial 2011, p. 287 - 298 Natal, Brasil, 2011.
9. DELGADO, C. H.; VIANNA, C. E.; GUELPELI, M. V. C. **Comparando sumários de referência humanos com extratos ideais no processo de avaliação de sumários extrativos**. In: IADIS Ibero-Americana WWW/Internet 2010, p. 293 - 300 Algarve, Portugal, 2010.

PESQUISAS CORRELACIONADAS AO TRABALHO DE TESE

1. **BLMSumm - Métodos de Busca Local e Metaheurísticas na Sumarização de Textos**. Dissertação de Mestrado do aluno Marcelo de Oliveira Arantes do departamento de Ciência e Tecnologia da Computação da Universidade Federal de Itajubá – UNIFEI, onde realizo a função de coorientador.
2. **Estudo do uso do algoritmo genético no problema da sumarização automática**. Trabalho de conclusão dos alunos Rodrigo Rufino Gomes e Rômulo Lindgren de Araujo Oliveira do curso de Ciência da Computação do Centro Universitário de Barra Mansa.Orientação realizada no ano de 2011.
3. **O processo de sumarização e avaliação de sumários extrativos:uma comparação de sumários de referência humanos com extratos ideais.** Trabalho de conclusão dos alunos Carlos Henrique Delgado e Caroline Evangelista Vianna do curso de Ciência da Computação do Centro Universitário de Barra Mansa.Orientação realizada no ano de 2011.

ANEXO

ANEXO A

ESCOLHA DA TÉCNICA TESTE ESTATÍSTICO A PARTIR DO NÚMERO DE AMOSTRAS

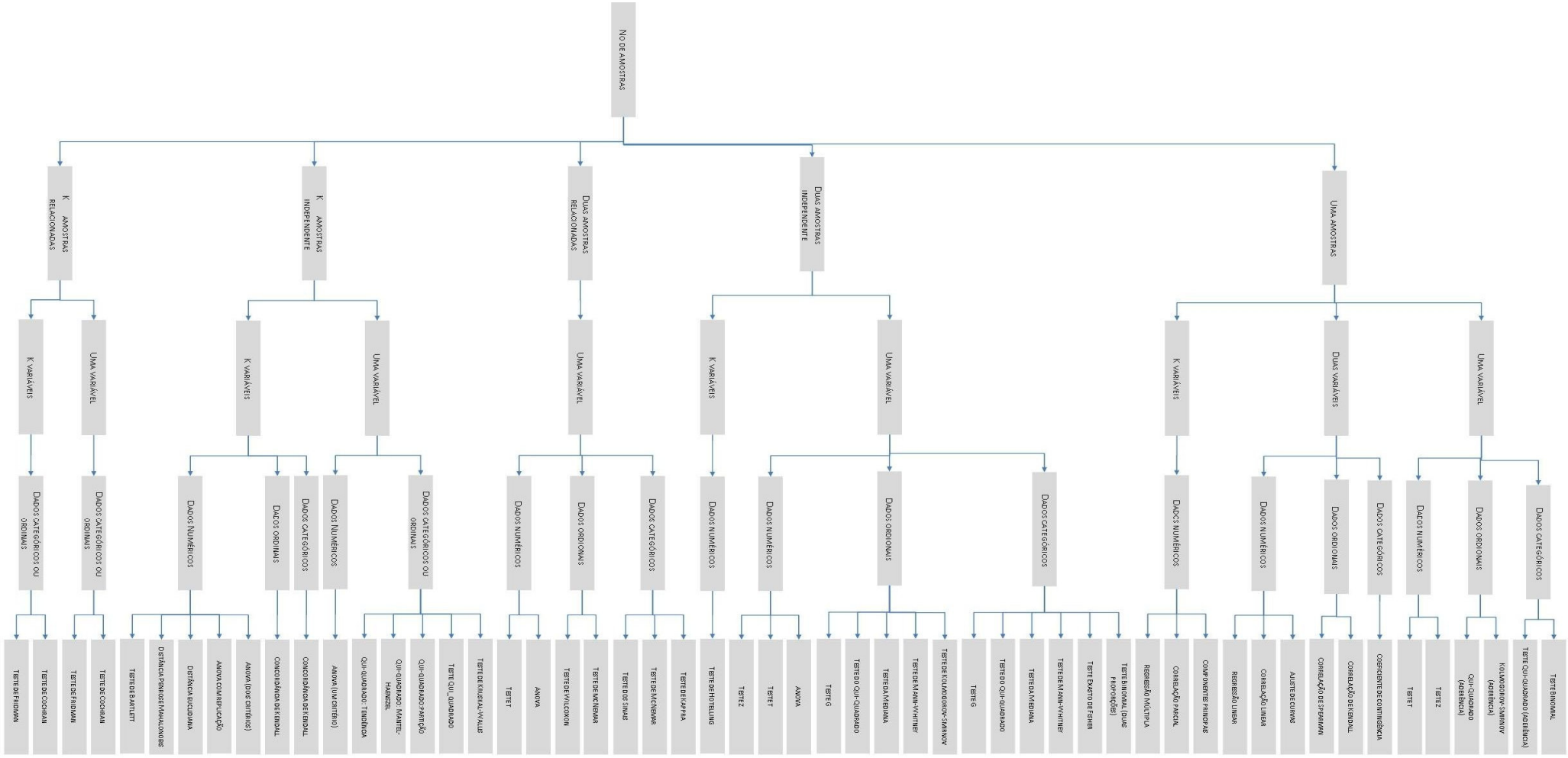


Figura 47: Diagrama para escolha da técnica teste estatístico a partir do número de amostras (CALLEGARI E JACQUES, 2007).