UNIVERSIDADE FEDERAL DOS VALES DO JEQUITINHONHA E MUCURI
FACULDADE DE CIÊNCIAS EXATAS
CURSO DE SISTEMAS DE INFORMAÇÃO

MANUTENÇÃO DAS STOPWORDS NO RENDIMENTO DO MODELO CASSIOPEIA

Thiago Gonçalves

DIAMANTINA
2016

UNIVERSIDADE FEDERAL DOS VALES DO JEQUITINHONHA E MUCURI
FACULDADE DE CIÊNCIAS EXATAS
CURSO DE SISTEMAS DE INFORMAÇÃO

MANUTENÇÃO DAS STOPWORDS NO RENDIMENTO DO MODELO CASSIOPEIA

Thiago Gonçalves

Orientador:
Marcus Vinícius Carvalho Guelpeli

Monografia apresentada ao curso de Sistemas de Informação da Universidade Federal dos Vales do Jequitinhonha e Mucuri, com requisito parcial para obtenção do título de Bacharel em Sistemas de Informação.

DIAMANTINA
2016

MANUTENÇÃO DAS STOPWORDS NO RENDIMENTO DO MODELO CASSIOPEIA

Thiago Gonçalves

Orientador:

Marcus Vinícius Carvalho Guelpeli

Monografia apresentada ao curso de Sistemas de Informação da Universidade Federal dos Vales do Jequitinhonha e Mucuri, com requisito parcial para obtenção do título de Bacharel em Sistemas de Informação.

APROVADO em .... / .... / ....

_____

Prof. Ma. Cláudia Beatriz Berti - UFVJM

_____

Prof. Ma. Maria Lúcia Bento Villela - UFVJM

_____

Prof. Dr. Marcus Vinícius Carvalho Guelpeli - UFVJM

Dedico este trabalho à minha mãe Cecília e ao meu pai Manuel, aos amigos e parentes aqui presentes ou mesmo os que estão distantes, mas sempre acreditaram em meu potencial e a cada dia me forneciam forças para continuar, principalmente minha esposa Maiara Cristina, quem esteve a todo tempo presente do meu lado.

AGRADECIMENTOS

Agradeço primeiramente DEUS, por me dar a oportunidade de existir e me tornar quem sou hoje.

Aos meus pais Cecília de Fátima e Manuel de Fátima, pelo amor, incentivo e tudo que me proporcionaram durante toda vida.

À minha esposa Maiara, quem esteve comigo a todo momento do início do curso até a sua conclusão, sempre dando força e incentivando.

Ao meu orientador, Professor Dr. Marcus Vinícius, pela confiança e tempo dedicado na minha orientação.

A todos os professores da UFVJM que, em sua área de conhecimento, contribuíram para meu aprendizado e também para minha formação.

A todos os meus amigos do curso, pelas ideias trocadas e muitas vezes pelas dúvidas sanadas.

A todos que contribuíram, muito obrigado!

RESUMO

Com a grande expansão da internet a quantidade de informação disponível nos meios eletrônicos vem crescendo a cada dia, dificultando cada vez mais a recuperação da informação desejada no momento de uma pesquisa. O modelo Cassiopeia é uma ferramenta que foi desenvolvida para resolver o problema da sobrecarga de informação, usando o processo de clusterização sem retirar as *stopwords* dos textos na etapa de pré-processamento. Como todos os estudos dessa área vistos até o momento, encontrados na literatura, reforçam a grande necessidade de eliminar as *stopwords*, este trabalho tem a finalidade de avaliar a influência da permanência ou não das *stopwords* nos resultados do modelo Cassiopeia. Sendo assim, o método que foi utilizado no modelo Cassiopeia foi a retirada das *stopwords* dos textos a serem *clusterizados*. Os resultados obtidos foram animadores.

Palavras-chave: *Cassiopeia, Clusterização, Recuperação de Informação e Stopwords.*

ABSTRACT

With the great expansion of the Internet the amount of information available in electronic media is growing every day, making it increasingly difficult to recover the desired information at the time of a search. The Cassiopeia model is a tool that has been developed to solve the problem of information overload, using the clustering process without removing the stopwords of texts in the preprocessing step. Like all studies of this area seen so far, in the literature, emphasize the great need to remove stopwords, this study aims to evaluate the influence of the presence or not of stopwords the results of Cassiopeia model. Thus, the method that was used in the Cassiopeia model was the removal of stopwords texts to be clustered. The results were encouraging.

Keywords: Cassiopeia, clusterização, Information Retrieval e Stopwords.

LISTA DE FIGURAS

LISTA DE TABELAS

Sumário

Capítulo 1 - INTRODUÇÃO

Com a grande quantidade de informações disponíveis atualmente, principalmente pelas facilidades de gerá-las e armazená-las em meios eletrônicos, surgiu o que é denominado na literatura como sobrecarga de informação (NASSIF, 2011), dificultando o processo de recuperação da informação. Segundo Guelpeli (2012), diante de tanta informação disponível, selecionar apenas as que são de interesse de quem a procura facilita o processamento e a recuperação dessas informações.

O problema da sobrecarga de informação tem despertado muito interesse em pesquisas para o desenvolvimento de boas ferramentas de busca e recuperação de informação. Segundo Oliveira *et al*. (2007), o processo manual de organização de documentos é lento e requer a presença constante de um especialista humano, nem sempre disponível.

Para o desenvolvimento de algoritmos para recuperação de informação, é necessário uma padronização dos textos, pois estes estão em um formato não estruturado. Além do mais, tais textos precisam de uma representação para que o processo computacional possa ser capaz de trabalhar com os mesmos. Existem várias maneiras para representação do espaço amostral, sendo o modelo vetorial, abordagem para representação de um texto, o mais utilizado para a recuperação de informação, através de uma matriz que contém os documentos e os termos que representarão o mesmo, como mostra tabela 01 na página 06 (SALAZAR, 2012), (REZENDE et al, 2011), (ALMEIDA, 2007), (CORRÊA et al, 2012), (LOPES, 2004).

Com a representação através do espaço vetorial, onde cada palavra de um documento representa uma dimensão, tem-se tantas dimensões quanto existe palavras, tornando assim uma solução com alta dimensionalidade acarretando um grande custo computacional. Para tratar o problema da alta dimensionalidade, a maioria dos estudos da literatura baseia-se no corte de Luhn, o qual primeiro elimina as palavras com a freqüência mais alta como artigos, preposições, dentre outras que são também chamadas de *stopwords* e logo em seguida elimina as que aparecem isoladamente em um certo texto (REZENDE *et al*, 2011), (SÁ, 2008), (WIVES *et al*, 1998), (WIVES, 1999) e (WIVES, 2004).

O modelo Cassiopeia proposto por Guelpeli (2012), é resultado de uma pesquisa, partindo de uma ideia totalmente oposta à dos pesquisadores vistos na literatura, para elaboração deste trabalho. O modelo resolve o grande problema de alta dimensionalidade da área de Mineração de Texto (*Text Mining*) – MT, com o processo de sumarização em sua etapa de pré-processamento, propiciando assim a manutenção das *stopwords*.

Com a permanência das *stopwords* nos textos, o modelo Cassiopeia se torna independente do idioma, pois não será necessário que o algoritmo possua uma lista que contenha estas para que as mesmas sejam retiradas dos textos. Por outro lado, com a permanência das *stopwords*, poderão aparecer algumas para representar um determinado

texto, sendo que, de acordo com a literatura, as *stopwords* influenciam o desempenho de algoritmos de Recuperação de Informação.

Neste trabalho, o foco principal é avaliar se as *stopwords* selecionadas para compor o vetor de palavras influenciarão os resultados obtidos pelo modelo Cassiopeia na formação dos agrupamentos, pois, as *stopwords* são palavras que podem acarretar perda de desempenho em algoritmos de *clusterização* além de influenciar nos resultados (PRIOR, 2010), (REZENDE *et al*, 2011).

1.1Motivação

A motivação deste trabalho é avaliar se as stopwords presentes nos textos influenciam ou não o rendimento do modelo Cassiopeia, pois, quase toda literatura trabalha com a remoção destas palavras dos textos, principalmente para resolver o problema da alta dimensionalidade e também por estas estarem presentes em todos os domínios de textos.

1.2 Problema

O Cassiopeia não descarta as *stopwords* em seu pré-processamento, podendo assim haver a existência de algumas representando os textos no processo de agrupamento.

1.3 Hipótese

Retirando as *stopwords* de todos os textos, não influenciará os resultados do Modelo Cassiopeia.

1.4 Contribuição

Mostrar que o modelo Cassiopeia, diferentemente da maioria dos modelos vistos na literatura, independe da presença ou ausência de *stopwords*, conseqüentemente provar que o Cassiopeia é independente de idioma.

1.5 Estrutura do Trabalho

Capítulo 2 – Estado da arte

Neste capítulo, serão apresentados conceitos do que são as *stopwords* e suas influências nos documentos de textos, o que são agrupadores aprofundando um pouco em agrupadores de textos e o modelo Cassiopeia. Será abordado também as métricas para avaliação dos resultados obtidos por agrupadores e por fim serão apresentados os testes estatísticos usados no trabalho.

Capítulo 3 – Metodologia

Neste capítulo, será feita uma apresentação da metodologia adotada para o desenvolvimento deste trabalho. Será descrito a estrutura do corpora utilizado para a pesquisa, do software que foi desenvolvido para a retirada das *stopwords* de todos os textos e da utilização do modelo Cassiopeia para o processo de agrupamento. Será realizada a comparação dos resultados obtidos com e sem *stopwords*, e sobre a utilização dos testes estatísticos para validação dos mesmos.

Capítulo 4 – Resultados

Serão apresentados os gráficos com os resultados dos experimentos juntamente com uma análise crítica deles. Será apresentado também, a hipótese do trabalho, sendo apresentado em seguida uma análise dos testes estatísticos e ao final do capítulo, uma discussão dos valores obtidos.

Capítulo 5 – Conclusões

Neste capítulo, serão feitas considerações observando os resultados obtidos e a hipótese deste trabalho, bem como as limitações encontradas e os trabalhos que podem ser realizados no futuro.

Capítulo 2 - ESTADO DA ARTE

Neste capítulo, serão abordados os principais conceitos que fundamentam este trabalho. Inicialmente, uma explicação contextualizada do que são as *stopwords* e quais as suas influências no processo de agrupamento de texto, afinal a influência destas palavras é o foco deste trabalho. Sobre agrupamento, será apresentado um breve conceito, seguido de uma explicação do que é e como funciona o processo de agrupamento de texto observando suas fases. Por fim, fecha-se o item do Capítulo com uma apresentação referente ao modelo Cassiopeia, incluindo o passo a passo de seu funcionamento. As medidas para avaliação dos agrupamentos textuais aqui empregadas, como as métricas externas, composta das medidas *Recall*, *Precision* e *F-Measure* e as métricas internas, composta das medidas Coesão, Acoplamento e *Coeficiente Silhouette*, serão apresentadas e formalizadas. Por fim, será apresentado o conceito dos testes estatísticos, utilizados para avaliar a hipótese deste trabalho.

2.1 *Stopwords*

As *stopwords*, também denominadas como palavras negativas, são palavras que não possuem significados para um domínio específico. Dentre estas palavras tem-se artigos, preposições, conjunções e diversas outras utilizadas na construção sintática das orações. Segundo Loh (2008), o conjunto de palavras muito frequentes e com pouco significado, ou seja, o conjunto de *stopwords*, formam uma *stoplist*. As palavras que compõem uma *stoplist*, facilmente serão encontradas em qualquer tipo de texto.

Como forma de enumerar as *stopwords*, o anexo B, apresenta uma *stoplist* do idioma português, contendo 258 *stopwords*, bem como o anexo C, apresenta uma *stoplist* do idioma inglês, a qual possui 183 *stopwords*. Ambas as *stoplists*, foram propostas por Loh (2008), sendo estas criadas com base em análise estatísticas de diversos autores.

Diante da enorme quantidade de palavras existentes na composição de um documento, encontrar as que melhor irão representá-lo, se torna uma tarefa demorada e difícil. O problema da alta dimensionalidade, refere-se à enorme quantidade de informação disponível, tornando difícil para um indivíduo encontrar o que ele realmente está procurando. No caso de textos, tem-se uma enorme quantidade de palavras em sua composição, onde grande quantidade destas são as denominadas *stopwords*, aumentando consideravelmente a quantidade de palavras sem trazer informação que irá distinguir um texto do outro. Segundo Zou *et al* (2006), palavras como *stopwords* não exercem nenhuma informação significativa para o documento quando se tratando de recuperação de informação.

Observando a distribuição das palavras de um texto de acordo com a sua ocorrência, conforme figura 01, nota-se a formação de uma curva onde mais à esquerda temos as palavras com um maior número de ocorrências e mais a direita, as palavras com menor número de ocorrências. De acordo com a curva mencionada, pode-se afirmar que a maioria das palavras com os maiores números de ocorrências, são as

stopwords, isso pode ser comprovado com o anexo A, o qual lista todos os termos do texto conforme o seu número de ocorrência no mesmo.

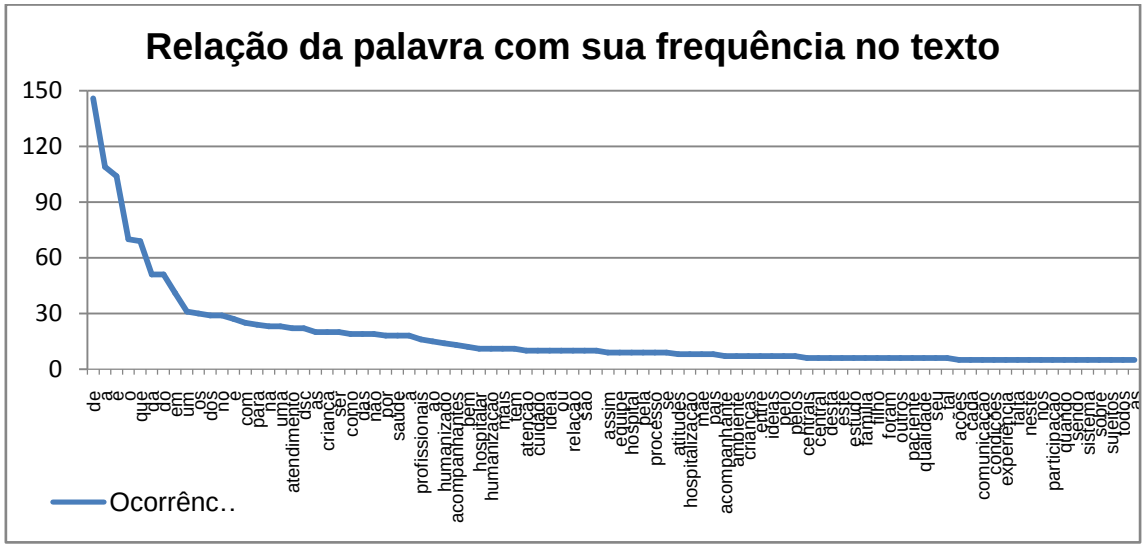


Figura 01: Relação de cada palavra de um texto médico do idioma português com o seu respectivo número de ocorrência.

As *stopwords* estão presentes em todos os textos, independente do domínio. Sendo assim, estas palavras não podem representar ou distinguir textos uns dos outros. Segundo Wives (2004), existem várias maneiras de selecionar as palavras que irão representar cada documento, sendo que a mais simples delas consiste em ignorar as palavras que costumam aparecer em quase todos os documentos, as *stopwords*.

Os trabalhos encontrados na literatura reforçam que as *stopwords* são termos redundantes e desnecessários, além do mais a manutenção destas nos textos irá tornar o processo de recuperação de informação mais lento e também pode chegar a afetar a qualidade dos resultados (WIVES *et al*, 1998), (WIVES, 1999) e (WIVES, 2004), (SÁ, 2008), (REZENDE *et al*, 2011).

Sabendo que palavras como *stopwords* não influenciam para determinar diferença entre textos de domínio distintos, a presença destas palavras irão apenas comprometer o custo computacional do processo de recuperação de informação (LOPES, 2004), (WIVES *et al*, 2004), (GUELPELI, 2012). Porém, se retirá-las dos textos, o algoritmo fica dependente de idioma, pois para a retirada das *stopwords* o algoritmo necessitaria de uma *stoplist* para cada idioma.

Para melhor entender esta questão, as Tabelas 01 e 02 apresentam a quantidade de caracteres e palavras existentes em alguns textos com *stopwords* e sem *stopwords*, respectivamente, sendo os mesmos textos em ambas tabelas.

Tabela 01: Quantidade de palavras em alguns textos com *stopwords*.

Arquivos	Caracteres	Caracteres e espaços	Palavras	Palavras e numerais
Dermatologia	15304	18157	3137	3181
Cardiologia	11101	13091	2137	2138
Ortopedia	6730	7872	1259	1305
Pediatria	16115	19181	3264	3335
Geriatria	8576	10040	1453	1469
TOTAL	57.826	68.341	11.250	11.428

A Tabela 01 apresenta cinco textos para análise, sendo que todos eles são textos normais contendo a presença de *stopwords*. Pode-se observar, avaliando o conjunto dos cinco textos, uma quantidade total de 11.250 palavras, sendo que este valor sobe para 11.428 levando em consideração palavras e numerais. Observando os arquivos que contém tais textos, juntos apresentam um tamanho de 68 *kilobyte*.

Tabela 02: Quantidade de palavras em alguns textos sem as *stopwords*.

Arquivos	Caracteres	Caracteres e espaços	Palavras	Palavras e numerais
Dermatologia	11741	13284	1879	1923
Cardiologia	8526	9525	1216	1217
Ortopedia	5536	6222	838	873
Pediatria	12517	14198	1974	2042
Geriatria	6579	7429	866	871
TOTAL	44.899	50.658	6.773	6.926

Para os resultados da Tabela 02, os textos apresentados na Tabela 01, passaram por um algoritmo que retirou todas as *stopwords* existentes nos textos. Observa-se um total de 6.773 palavras, sendo que este valor sobe para 6.926 levando em considerações palavras e numerais. Observando os novos arquivos gerados com os textos sem *stopwords*, juntos apresentam um tamanho de 52 *kilobyte*.

Avaliando os totais apresentados anteriormente nas duas tabelas, nota-se uma diminuição de 11.250 para 6.773 palavras, resultando em uma queda de 4.477, sendo que avaliando as palavras e numerais a queda é de 11.428 para 6.926, um total de 4.502. Trabalhando com percentual, em ambas as situações os textos apresentaram uma diminuição no número de palavras de mais de 39%. Assim, explica-se a questão de um maior custo computacional quando os textos são tratados com a permanência das *stopwords*, independente do idioma que o texto pertence.

2.2 Agrupamento

O termo agrupamento é bastante utilizado em trabalhos encontrados na literatura como *clusterização*, tradução do termo *clustering*, sendo que os grupos formados pelo processo de *clusterização* são conhecidos como conglomerados ou, do inglês, *clusters*. O processo em questão é também conhecido por diversos outros nomes: *clusterização*, aglomerados ou agrupamento (WIVES, 1999), (WIVES, 2004) e (GUELPELI, 2012).

Segundo Wives (1999), a técnica de agrupamento consiste em organizar uma série desorganizada de objetos em grupos onde os objetos pertencentes ao mesmo grupo possuam muita similaridade e objetos pertencentes a grupos diferentes sejam dissimilares.

Para a formação dos agrupamentos, todas as características que representam um objeto são examinadas e dentre todas, são selecionadas as que juntas melhor representam tal objeto. Assim os objetos são analisados a fim de encontrar características similares entre os mesmos, formando grupos homogêneos com objetos que possuam semelhanças entre si.

2.2.1 Agrupamento de Textos

O processo de agrupamento de textos pode ser definido como: dada uma base de texto P, os elementos da base P devem-se ser agrupados de maneira que os textos mais similares fiquem no mesmo grupo e os menos similares sejam colocados em grupos diferentes. Segundo Arora *et al* (2005), documentos relacionados devem pertencer ao mesmo agrupamento.

Segundo Fan *et al* (2006), o processo de agrupamento de textos é totalmente automático tendo como objetivo repartir uma coleção em grupos de textos de conteúdos similares, a fim de obter mais conhecimentos dos textos e das relações entre eles.

O processo de agrupamento de textos possui três fases: pré-processamento, processamento e pós-processamento.

Na etapa de pré-processamento, é realizado uma padronização nos documentos de textos para que eles possam ser tratados pelos algoritmos de agrupamento.

Na segunda etapa, a de processamento, são selecionadas as palavras que irão representar os textos. Segundo Wives (2004), a seleção de características consiste no processo de identificar o melhor conjunto de palavras para representar um documento, melhorando a qualidade da informação que representa um documento. Para a etapa em questão, utiliza-se um vetor para o armazenamento dos termos (características) que irão representar o documento. O modelo vetorial é uma das abordagens mais usadas na representação de documentos de textos de uma coleção, onde vetores em um espaço multidimensional representam os documentos e os termos selecionados para representá-

los, como mostra Tabela 03, em que d_i representa os documentos e t_i os termos (SALAZAR, 2012), (REZENDE et al, 2011), (ALMEIDA, 2007), (CORRÊA et al, 2012), (LOPES, 2004).

Tabela 03: Tabela documento – termo (MANNING *et al*, 2008).

	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}

Por fim, na etapa de pós-processamento são usadas medidas de validação dos agrupamentos formados. Segundo Rezende *et al* (2012), a avaliação dos agrupamentos formados pode ser realizada de forma subjetiva, utilizando um conhecimento de um especialista de domínio, ou de forma objetiva por meio de índices estatísticos que indicam a qualidade dos resultados.

Um grande problema encontrado com os estudos referentes a agrupamentos de textos é o problema da alta dimensionalidade. Segundo Galho (2003), é evidente a necessidade do uso da categorização automática de documentos de textos para melhorar o processo de recuperação de informações textuais.

Dentre as técnicas de redução da alta dimensionalidade para tratamento de textos, a mais usual na literatura de acordo com (REZENDE, et al, 2011), (CORRÊA, et al, 2012), (SÁ, 2008) é a seleção de termos através do corte de Luhn, que se baseia na lei de Zipf, também conhecida como princípio do menor esforço. Com a curva de Zipf, Figura 02, pode-se observar a representação da distribuição da frequência de ocorrências de palavras em um texto, de maneira que, em suas ordenadas, tem se a frequência da palavra, e nas abscissas, a posição da palavra em relação às outras palavras do texto com base na frequência de cada uma.

Luhn (1958), propôs que pode ser definido limites de corte para a curva de Zipf, um superior e um inferior. Segundo Guelpeli (2012), limites estes também chamados de limiares de corte de Luhn. O corte superior tem por finalidade retirar as palavras com maiores frequências em um texto, as *stopwords*, enquanto com o corte inferior, elimina-se as palavras que são muito específicas, encontradas uma única vez em um determinado texto. Assim Luhn propôs que os termos mais relevantes estão na região do pico de uma curva imaginária conforme mostra figura 01, curva esta localizada entre os dois cortes por ele proposto.

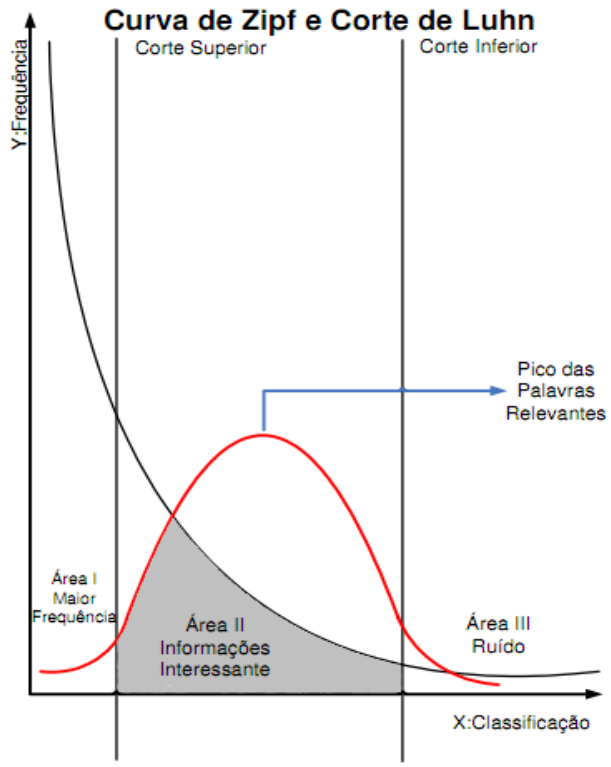


Figura 02: Curva de Zipf com os cortes superior e inferior de Luhn.

2.2.2 Modelo Cassiopeia

Inicialmente, para um melhor entendimento, a Figura 03 apresenta uma visão geral do funcionamento do modelo Cassiopeia, sendo descrito a seguir um detalhamento de suas etapas.

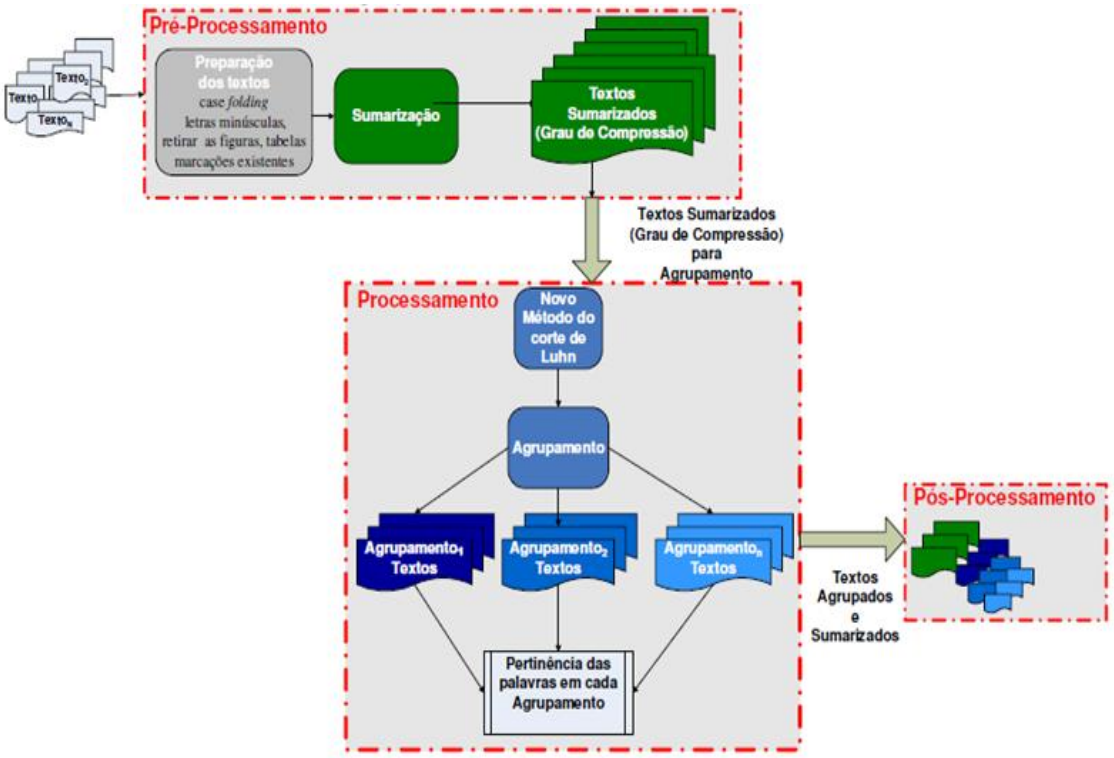


Figura 03: Modelo Cassiopeia.

O modelo Cassiopeia, proposto por Guelpeli (2012), assim como demais agrupadores começa o processo de agrupamento com a preparação dos textos. Fase esta, de pré-processamento, a qual os textos são preparados para o processo computacional. Segundo Guelpeli (2012), nessa fase o modelo Cassiopeia utiliza a técnica *case folding* para colocar todas as letras em minúsculas, além de descartar as figuras, tabelas e outras marcações existentes. Para que a técnica de agrupamento seja aplicada em uma coleção, os documentos devem ser preparados para o processamento (ALMEIDA, 2007), (SALAZAR et al 2011), (CORRÊA et al, 2012), (GUELPELI, 2012), (SALAZAR, 2012).

O modelo Cassiopeia resolve o problema da alta dimensionalidade de forma diferente das formas vistas na literatura, as quais simplesmente eliminam as *stopwords* dos textos e demais dados esparsos. Segundo Guelpeli (2012), a sumarização é a proposta do modelo Cassiopeia para a diminuição do número de palavras na etapa de pré-processamento, gerando um ganho qualitativo, quantitativo e viabilizando a manutenção das *stopwords*. Com a sumarização, o modelo Cassiopeia reduz o número de palavras do texto de maneira que este passa a ser composto apenas por palavras que melhor irão representá-lo, resolvendo o problema da alta dimensionalidade sem retirar as *stopwords*, tornando assim o modelo independente de idioma (GUELPELI *et al*, 2009), (GUELPELI *et al*, 2011).

Com a utilização da sumarização na etapa de pré-processamento, onde o modelo teve a possibilidade de manter um bom nível de informatividade ao mesmo tempo que é realizada a diminuição do número de palavras, o modelo Cassiopeia realiza uma solução

diferente na etapa de processamento. Na literatura, a seleção dos termos ocorre com base no corte de Luhn conforme Figura 02. Segundo Guepeli (2012), modelo Cassiopeia propõe um corte médio na distribuição da frequência das palavras (Figura 04).

Na identificação dos termos do documento, o modelo Cassiopeia identifica a importância do termo conforme sua frequência no documento através da frequência relativa de cada termo. Tal frequência normaliza os termos pelo tamanho dos documentos (GUELPELI, 2012), (WIVES, 2004), como mostra **Equação 1**.

$$F_{\mathbf{r}}X = \frac{F_{abs}X}{N} \tag{1}$$

Onde F_rX é a frequência relativa de X, $F_{abs}X$ é a frequência absoluta de X, quantidade de vezes que X aparece no documento e N é o número total de termos do texto.

Em seguida é calculada a média em relação à frequência relativa de todos os termos do documento, definindo assim a localização do corte do modelo Cassiopeia em relação à curva de Zipf, como mostra Figura 04. Segundo Guelpeli (2012), o modelo usa truncagem, ou seja, um tamanho máximo de 50 posições. Para estabelecer um vetor com bons termos, 50 posições é o suficiente, o uso de mais termos não garante melhoras e acarretará aumento no processo computacional (WIVES, 2004). Assim o modelo Cassiopeia trabalha com os 25 termos da direita e os 25 termos da esquerda da frequência média para representação do documento.

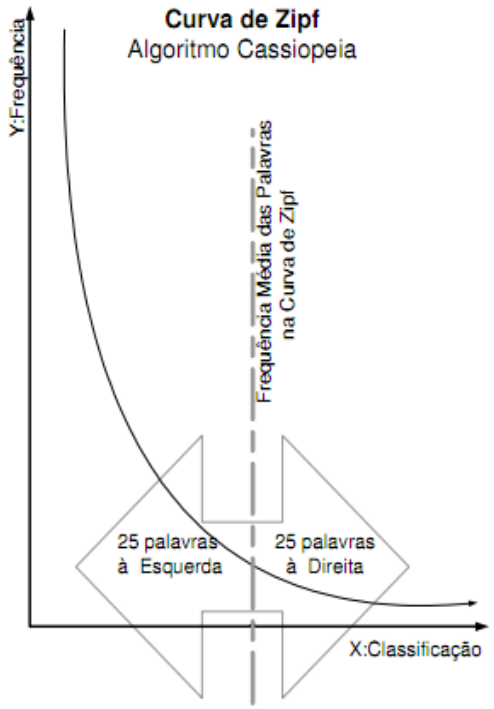


Figura 04: Seleção de termos no modelo Cassiopeia.

O modelo Cassiopeia, utiliza o método hierárquico aglomerativo para organizar seus textos em agrupamentos, produzindo uma representação hierárquica, facilitando a visualização dos agrupamentos, bem como o grau de similaridade obtido entre eles com o uso do algoritmo cliques (GUELPELI, 2012).

No pós-processamento, o modelo oferece uma estrutura hierárquica que possibilita uma boa avaliação dos documentos, pois estes estão sumarizados e com alto grau de informatividade.

Para um melhor entendimento do modelo Cassiopeia, segue descrição dos passos de seu funcionamento, dividido em 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 decrescente de frequência (da maior para a menor);

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

2.3 Métricas

Conforme a literatura, os métodos para avaliação do processo de agrupamento podem ser distribuídos em três principais categorias: métricas externas ou supervisionadas, métricas internas ou não supervisionadas e as métricas relativas (GUELPELI, 2012), (BONATO,2008). As últimas possuem como objetivo encontrar o melhor conjunto de grupos formados por um algoritmo de agrupamento, onde em cada simulação efetuada, deve-se inserir parâmetros de entrada diferentes. Devido suas finalidades, métricas relativas não serão utilizadas neste trabalho.

Nas métricas externas ou supervisionadas, os grupos gerados pelo processo de agrupamento são avaliados em comparação a uma estrutura de classes pré-definida. Segundo Bonato (2008), o difícil para este tipo de validação é criar uma estrutura externa, sendo que esta deve ser criada por um especialista humano. As métricas externas utilizadas neste trabalho são mencionadas na seção 2.3.1. Segundo Fan *et al*

(2006), para esta finalidade, as medidas mais usadas são: *Recall*, *Precision* e *F-Measure*.

Nas métricas internas ou não supervisionadas, não existe nenhuma influência externa para a avaliação dos grupos formados, tais grupos são avaliados pelas informações contidas nos próprios grupos. Segundo Fan et al (2006), para esta ocasião, as medidas mais usadas são: Coesão, Acoplamento e *Coeficiente Silhouette*. As métricas internas utilizadas neste trabalho são abordadas na seção 2.3.2.

2.3.1 Métricas Externas

Recall(R): **Equação (2):**

$$\frac{\mathbf{tlcd_i}}{\mathbf{tgcd_i}} * 100 \tag{2}$$

O *Recall* mede a proporção de objetos alocados corretamente em relação ao número total de objetos da classe do agrupamento (GUELPELI *et al*, 2011), (ALMEIDA, 2007), (LOPES, 2004).

Onde *tlcd* é o total local da categoria dominante do *cluster* i e *tgcd* é o total global da categoria dominante do *cluster* i no processo.

Precision(P): **Equação (3):**

$$\frac{\mathbf{tlcd_i}}{\mathbf{te_i}} * 100 \tag{3}$$

O *Precision* mede a proporção de objetos alocados corretamente em relação ao número total de objetos do agrupamento (GUELPELI, 2012), (ALMEIDA, 2007), (LOPES, 2004).

Onde *tlcd* é o total local da categoria dominante do *cluster* i e *te* é o total de elementos no *cluster* i.

F-Measure(F): **Equação (4):**

$$2 * \frac{Precision(P) * Recall(R)}{Precision(P) + Recall(R)} \tag{4}$$

Como *Recall* e *Precision* tendem a ser inversamente proporcionais, a medida F-Measure representa uma média harmônica entre estas duas medidas (GUELPELI, 2012), (ALMEIDA, 2007), (BONATO, 2008).

2.3.2 Métricas Internas

Coesão(C): **Equação (5):**

$$\frac{\sum_{i>j} sim\,(P_i, P_j)}{n(n-1)/2} \tag{5}$$

A coesão mede o grau de similaridade entre os elementos de um mesmo grupo (GUELPELI, 2012), (BONATO, 2008).

Onde *Sim* (P_i, P_j) é o calculo da similaridade entre 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 (6):**

$$\frac{\sum_{i>j} sim(C_i, C_j)}{n_a(n_a-1)/2} \tag{6}$$

O acoplamento mede a similaridade média de todos os pares de documentos, sendo um elemento pertencente a um grupo e o outro pertencente a um outro grupo (GUELPELI, 2012), (BONATO, 2008).

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 pertencente a P, C_i é o centroide do agrupamento P e C_j é o centroide do agrupamento P_i e n_a é o número de agrupamentos presentes em P.

Coeficientte Silhouette(S): **Equação (7):**

$$\frac{\mathbf{b(i)} - \mathbf{a(i)}}{\max(\mathbf{a(i)}, \mathbf{b(i)})} \tag{7}$$

O *Coeficientte Silhouette* mostra o quanto um objeto é similar aos outros objetos de seu grupo e dissimilar aos objetos de outros grupos (GUELPELI, 2012), (ALMEIDA, 2007), (BONATO, 2008).

Onde a(i) é a distância média entre o i-ésimo elemento do grupo e os outros elementos 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 *Coeficientte Silhouette* de um grupo é média aritmética dos coeficientes calculados para cada elemento pertencente ao grupo.

2.4 Testes Estatísticos

A análise estatística dos resultados obtidos numa pesquisa é um processo indispensável para atestar a veracidade proporcionando aceitabilidade no meio científico. É necessário avaliar os dados em estudo para escolher quais testes deverão ser utilizados para a validação dos dados. Para isso, Normando (2010), analisa-se o tipo dos dados; a distribuição dos dados e os tipos de amostras analisadas.

Neste trabalho, a identificação dos testes estatísticos foi feita com base no Anexo D, o qual apresenta um diagrama proposto por Callegari-Jacques (2007). Segundo Oliveira (2014), os testes são também usados na conferência internacional TAC (*Text Analysis Conference*). Analisando as amostras obtidas, verificou-se que elas são independentes e possuem *k* variáveis, além de apresentar uma distribuição anormal. Assim os testes a serem utilizados são os não paramétricos, onde, de acordo com o diagrama de Callegari-Jacques (2007), os testes mais indicados foram de ANOVA, de Friedman, e o coeficiente de concordância de Kendall.

2.4.1 ANOVA de Friedman

O ANOVA de Friedman é um teste não paramétrico utilizado para comparar os resultados de três ou mais amostras relacionadas, numa distribuição bivariada. O ANOVA ordena os resultados para cada um dos casos e depois calcula a média das ordens para cada uma das amostras (CALLEGARI-JACQUES, 2007). O teste ANOVA utiliza a ordem ocupada pelos dados da amostra e não dos dados diretamente.

2.4.2 Coeficiente de Concordância de Kendall

O teste de concordância de Kendall 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 das classificações estarem na mesma ordem e a de estarem em ordem diferentes. Quanto mais próximo de zero, menor é a concordância, e quanto mais próximo de um, maior é a concordância. O teste de concordância de Kendall tem a finalidade de normalizar o teste estatístico ANOVA de Friedman (CALLEGARI-JACQUES, 2007).

Capítulo 3 - METODOLOGIA

Neste capítulo será apresentada a metodologia adotada para realização deste trabalho. Será explicado sobre a composição do corpora inicial, formada pelos textos originais e os novos textos formados sem as *stopwords,* bem como o processo de retirada das *stopwords* com a utilização do algoritmo criado para tal finalidade.

Será abordado sobre a utilização do algoritmo Cassiopeia no processo de agrupamento dos textos e a avaliação e comparação dos resultados obtidos pelo algoritmo com textos com e sem *stopwords*. Por fim, será comentado sobre a utilização dos testes estatísticos para a validação dos resultados obtidos.

3.1 Corpora

A corpora utilizado neste trabalho foi fornecida pelo Grupo de pesquisa Mineração de Texto, Processamento de Linguagem Natural e Aprendizado de Máquina (MTPLNAM), que pode ser acessado em http://www.mtplnam.com. Tal corpora é composto por textos sumarizados nas compressões 50%,70%,80% e 90%, sendo que estes valores de compressões foram propostos por Guelpeli (2012), nos idiomas: Inglês, nos domínios Jornalístico e Médico; português nos domínios Jornalístico, Jurídico e Médico, sendo 100 textos em cada compressão de cada domínio, totalizando 6000 textos. A taxa de compressão especificada, é o percentual que foi retirado do texto de maneira que quanto maior a compressão utilizada menor este fica.

3.1.1 Corpora em Inglês

A corpora em Inglês é constituída por 2400 documentos textos, sendo estes divididos em categorias como apresenta a Figura 05. Primeiramente a corpora é dividida em duas partes onde cada parte se refere a um domínio diferente, sendo 1200 textos pertencentes ao domínio jornalístico e os outros 1200 pertencentes ao domínio médico.

Dentro de cada domínio, os textos são divididos em três novas categorias, pois tais textos são resultado do processo de sumarização de três sumarizadores diferentes: o sumarizador *Copernic,* o sumarizador *Intellexer Summarizer Pro* e o sumarizador *SweSum*.

Assim, cada categoria referente a um determinado sumarizador possui 400 textos sumarizados, ou seja, textos que passaram por um processo de compressão reduzindo o seu tamanho de forma que a informatividade do texto continue igual a informação do texto original. Cada texto passou por quatro taxas de compressões diferentes: compressão de 50%, compressão de 70%, compressão de 80% e compressão de 90%. Assim, conforme esta distribuição, em cada categoria referente a uma taxa de compressão existe um total de 100 textos, sendo estes de assuntos diversos.

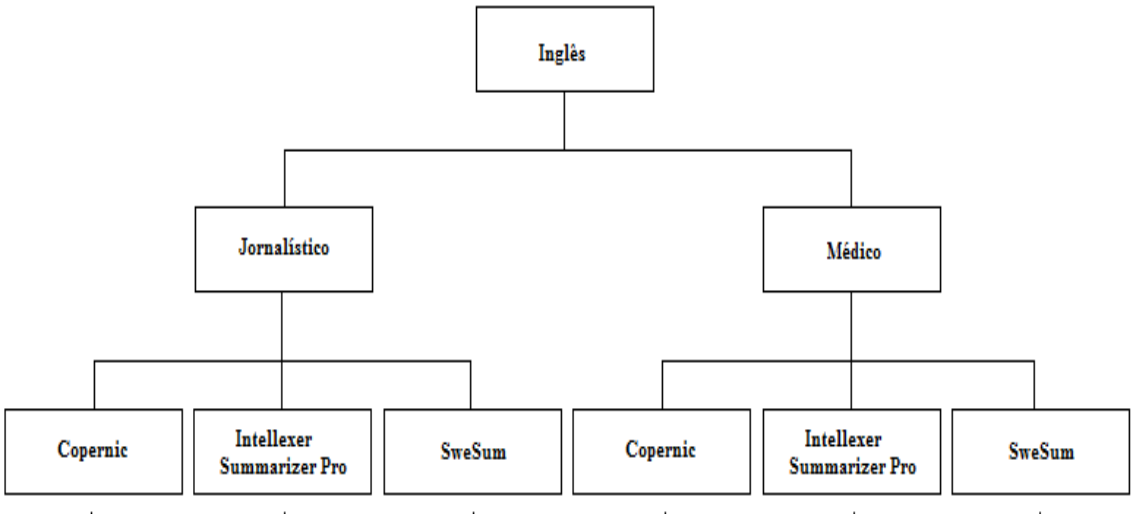


Figura 05: Diagrama do corpora em Inglês utilizado neste trabalho.

Todos os 2400 textos do corpora em inglês, passaram por um algoritmo para a retirada de todas as *stopwords*, dando origem a 2400 novos textos. O algoritmo mencionado é apresentado no item 3.2 deste capítulo. Dessa forma é gerado um novo corpora com os textos sem *stopwords*, sendo que tal corpora possui a mesma estrutura apresentada na Figura 05.

3.1.2 Corpora em Português

A corpora em Português é constituída por 3600 documentos textos, sendo estes divididos em categorias como apresenta a figura 06. Primeiramente a corpora é dividida em três partes onde cada parte se refere a um domínio diferente, sendo 1200 textos pertencentes ao domínio jornalístico, 1200 textos pertencentes ao domínio jurídico e 1200 textos pertencentes ao domínio médico.

Dentro de cada domínio, os textos são divididos em três novas categorias, pois tais textos são resultado do processo de sumarização de três sumarizadores diferentes: o sumarizador *Gist Average Keyword*, o sumarizador *Gist Intrasenteca* e o sumarizador *Supor2*.

Assim, cada categoria referente a um determinado sumarizador possui 400 textos sumarizados. Cada texto, assim como no corpora em inglês, passou por quatro taxas de compressões diferentes: compressão de 50%, compressão de 70%, compressão de 80% e compressão de 90%. Assim, conforme esta distribuição, em cada categoria referente a uma taxa de compressão existe um total de 100 textos, sendo estes de assuntos diversos.

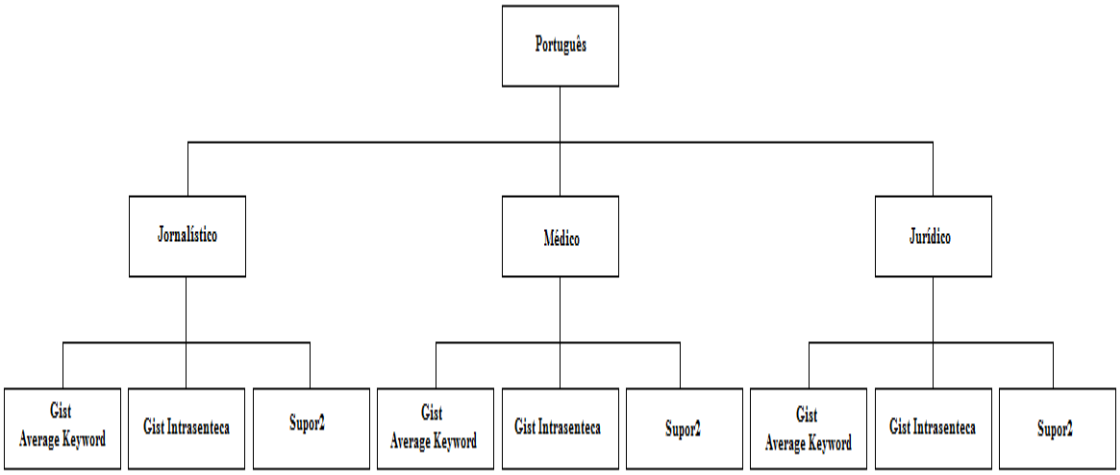


Figura 06: Diagrama do corpora em Português utilizado neste trabalho.

Todos os 3600 textos do corpora em português, também passaram pelo algoritmo para a retirada de todas as *stopwords* do texto, dando origem a 3600 novos textos. Dessa forma é gerado um novo corpora com os textos sem *stopwords*, sendo que tal corpora possui a mesma estrutura apresentada na figura 06.

3.2 Programa para a Retirada de *Stopwords*

Para o desenvolvimento do trabalho, um algoritmo foi desenvolvido utilizando a linguagem de programação C++ juntamente à IDE *Visual Studio*, para a remoção das *stopwords* da corpora com a finalidade de resolver o grande problema da alta dimensionalidade, conforme visto em (REZENDE *et al*, 2011), (WIVES, 2004), (SALAZAR, 2012), (DELGADO *et al*, 2012), reduzindo o número de palavras do texto mantendo as palavras mais significativas para sua representação.

Para eliminar estas palavras, foram utilizadas *stoplists* propostas por Stanley Loh (2008), listas estas encontradas em Anexo A, lista de *stopwords* em Português; e em Anexo B, lista de *stopwords* em Inglês.

Para um melhor entendimento do algoritmo desenvolvido, segue abaixo descrição dos passos de seu funcionamento:

1. Receber nome do arquivo que contém as *stopwords*;
2. Receber nome do arquivo que contém a lista com nome de todos os textos;
3. Abrir arquivo que contém as *stopwords*;
4. Armazenar cada *stopword* em uma posição do vetor (**S**);
5. Abrir arquivo que contém a lista com nome dos textos;
6. Enquanto existir texto, faça;

7. Abrir arquivo com o texto;
8. Armazenar cada palavra do texto em uma posição do vetor (**T**);
9. Enquanto existir palavra no vetor **T**, faça;
10. Enquanto existir *stopword* no vetor **S**, faça;
11. Se palavra for igual *stopword*;
12. Retirar palavra do vetor **T**;
13. Fim se;
14. Fim do enquanto;
15. Fim do enquanto;
16. Abrir novo arquivo texto em branco;
17. Passar todas as palavras do vetor **T** para o novo arquivo;
18. Salvar novo arquivo;
19. Fim do enquanto;

Assim a outra metade da base de dados utilizada neste trabalho é formada por mais 6000 textos sem *stopwords*, ou seja, cada texto da corpora fornecida pelo grupo MTPLNAM passou por um processo onde foi retirado todas as *stopwords* do mesmo dando origem a um novo texto.

3.3 Uso do Cassiopeia

Para obter os resultados deste trabalho, foi usado para o processo de agrupamento dos textos, o algoritmo desenvolvido por Guelpeli (2012). Tal algoritmo, denominado Cassiopeia, seleciona em seu processamento os termos para representar os documentos de forma diferente aos demais agrupadores de textos, conforme tópico 2.2.2 deste trabalho.

Com o objetivo de identificar a influência ou não das *stopwords* nos resultados do Cassiopeia, todos textos do corpora cedida para este trabalho, juntamente com todos os novos textos gerados com a retirada das *stopwords*, foram todos submetidos ao processo de *clusterização* do modelo Cassiopeia. Cada processo de *clusterização* foi submetido a uma repetição de 30 simulações com o objetivo de obter uma média dos resultados.

A cada simulação, o modelo Cassiopeia fornece um arquivo com extensão .html contendo todos os *clusters* formados no processo de *clusterização* e quais textos ficaram

em cada *cluster*. Um outro arquivo, sendo este em E*xcel*, é fornecido pelo algoritmo contendo os valores das métricas internas ou externas de cada simulação, conforme o processo ao qual os textos foram submetidos.

3.4 Textos Com *Stopwords Versus* Textos Sem *Stopwords*

Os resultados obtidos usando o Modelo Cassiopeia para criar os agrupamentos com os textos de toda corpora, foram todos exportados para gráficos para ser realizada análise e comparação destes resultados, ou seja, além de exportar os resultados para um gráfico, os resultados dos textos com *stopwords* foram colocados em um mesmo gráfico que os resultados dos textos sem *stopwords*, conforme apresentação dos resultados no Capítulo 4. Assim, os resultados obtidos com os textos com *stopwords* foram todos comparados com os resultados dos textos sem *stopwords*, respeitando as compressões e os domínios.

3.5 Uso dos testes estatísticos para validação dos resultados

Para avaliação e confirmação dos resultados, estes foram submetidos a avaliações estatísticas através dos métodos escolhidos: ANOVA de Friedman e coeficiente de concordância de Kendall. Sendo assim, foi utilizado o *software statplus* para obter os resultados estatísticos, visto que, tal software nos fornece os métodos estatísticos sugeridos para utilização neste trabalho assim como mencionado no capítulo 2.

O *software statplus* fornece uma tabela contendo valores como: número de simulações realizadas, grau de liberdade, coeficiente de concordância de Kendall, ordem médio, soma de ordens e média. Os valores obtidos nos testes estatísticos foram tabelados e estão apresentados no Apêndice C deste trabalho.

Capítulo 4 - RESULTADOS

4.1 Experimentos

No capítulo 4, serão apresentados todos os experimentos realizados para a obtenção dos resultados deste trabalho, bem como a apresentação dos resultados dos testes estatísticos ANOVA Friedman e Coeficiente de Concordância de Kendall, testes estes usados para a validação do trabalho em questão. Em seguida será apresentado sobre a hipótese deste trabalho e o capítulo é finalizado com uma discussão dos resultados obtidos nos estudos realizados.

Os experimentos foram organizados em duas partes; sendo a primeira apresentada no item 4.1.1, refere-se aos resultados obtidos com as métricas externas: *recall*, *precision* e a média harmônica entre estas duas, o *F-measure*. Em seguida, no item 4.1.2, são apresentados os resultados obtidos com as métricas internas: coesão, acoplamento e a média harmônica entre estas duas, *Coeficientte Silhouette*. Em ambas as partes, tais resultados são agrupados respeitando o idioma; inglês e português, e a taxa de compressão dos textos utilizados nas simulações com o modelo Cassiopeia, sendo que, a taxa de compressão variou entre os valores: 50%, 70%, 70% e 90%.

4.1.1 Métrica Externa: *Recall*, *Precision* e F-*Measure*

Para organizar a apresentação dos resultados, serão apresentados nesta seção apenas os resultados obtidos com o *F-measure*, a média harmônica das medidas *recall* e *precision*. Os melhores resultados para esta medida serão os que mais se aproximarem de um. Os resultados obtidos com a métrica *recall* e a métrica *precision*, serão colocados no apêndice A, juntamente com os comentários referentes a cada resultado.

4.1.1.1 Compressão de 50% no Idioma Inglês

No domínio jornalístico, como mostra Figura 07, os maiores resultados obtidos foram com os textos sem *stopwords* na maioria das simulações juntamente com algumas poucas simulações que textos com *stopwords* do sumarizador *Copernic* apareceram, sendo que todos os resultados ficaram entre 0,20 e 0,28. No domínio médico, conforme Figura 08, os resultados com os textos com e sem *stopwords* foram bem próximos, sendo que todos os resultados ficaram entre 0,18 e 0,22.

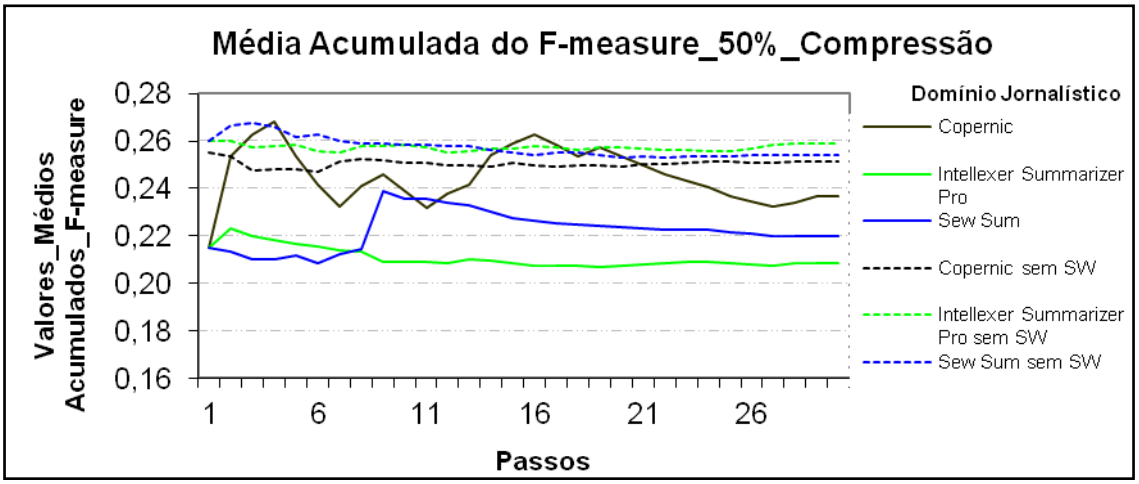


Figura 07: Resultados obtidos pelo modelo Cassiopeia, usando a medida *F-measure* com 50% de compressão no idioma Inglês no domínio jornalístico.

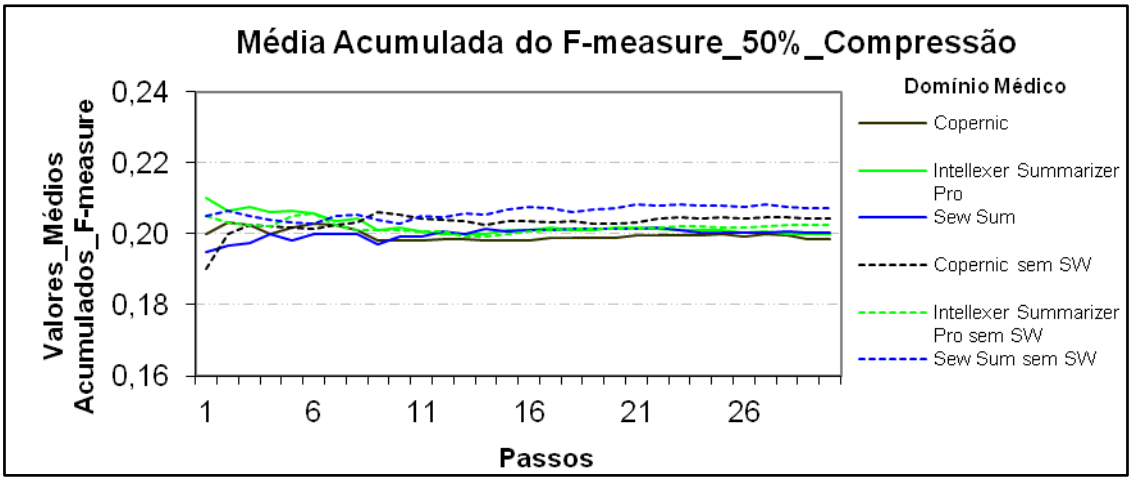


Figura 08: Resultados obtidos pelo modelo Cassiopeia, usando a medida *F-measure* com 50% de compressão no idioma Inglês no domínio médico.

4.1.1.2 Compressão de 70% no Idioma Inglês

No domínio jornalístico, como apresentado na Figura 09, os maiores resultados foram obtidos com os textos sem *stopwords* dos sumarizadores *Intellexer Summarizer Pro* e *Copernic* na maioria das simulações, sendo que todos os resultados ficaram entre 0,20 e 0,30. No domínio médico, como mostra Figura 10, os maiores resultados foram obtidos com os textos sem *stopwords*, com resultados bem próximos dos textos com *stopwords* do sumarizador *Sew Sum*, sendo que todos os resultados ficaram próximos de 0,20.

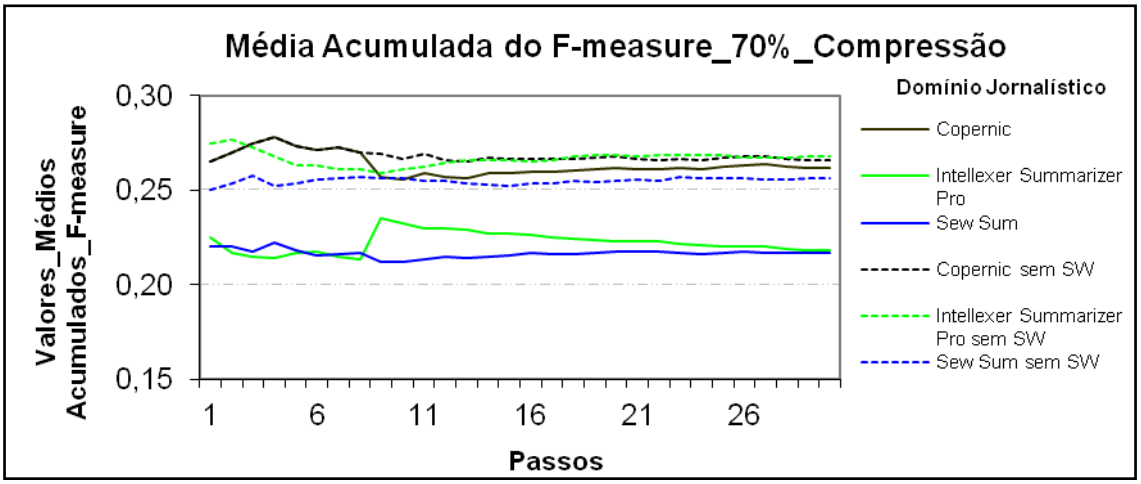


Figura 09: Resultados obtidos pelo modelo Cassiopeia, usando a medida *F-measure* com 70% de compressão no idioma Inglês no domínio jornalístico.

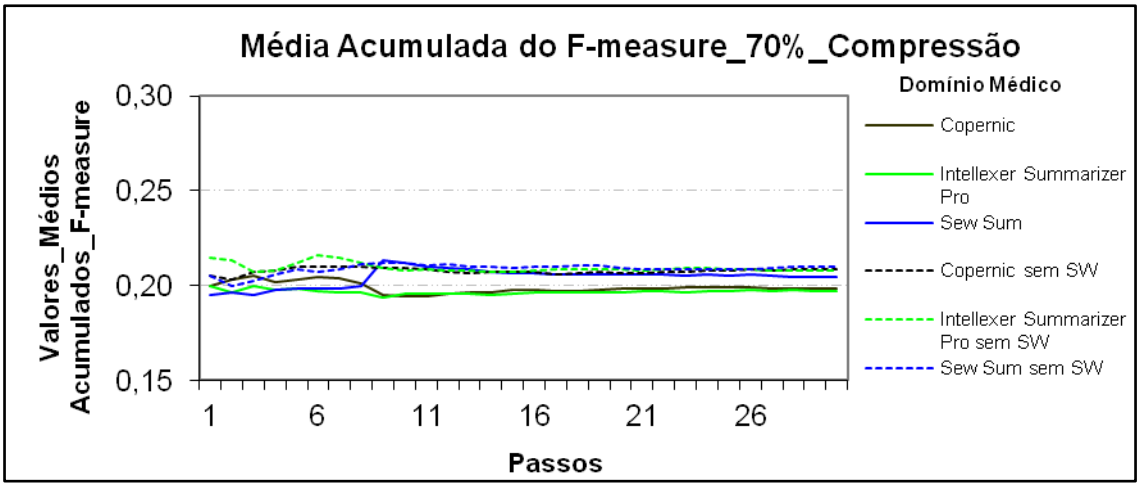


Figura 10: Resultados obtidos pelo modelo Cassiopeia, usando a medida *F-measure* com 70% de compressão no idioma Inglês no domínio médico.

4.1.1.3 Compressão de 80% no Idioma Inglês

Tanto no domínio jornalístico, quanto no médico, assim como demonstram respectivamente as Figuras 11 e 12, os maiores resultados obtidos foram com textos sem *stopwords*, sendo que no domínio jornalístico todos os resultados ficaram entre 0,20 e 0,30 e no domínio médico resultados próximos de 0,21.

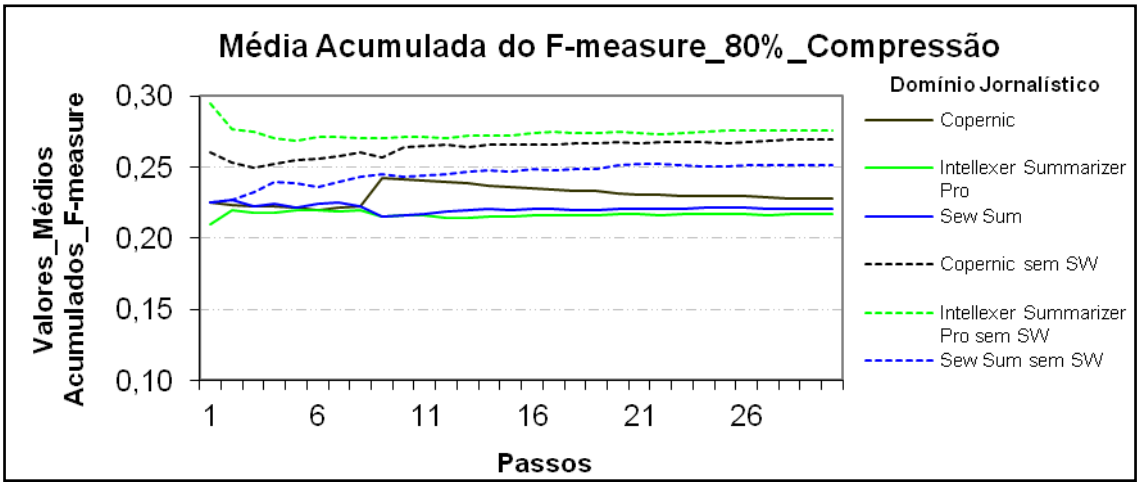


Figura 11: Resultados obtidos pelo modelo Cassiopeia, usando a medida *F-measure* com 80% de compressão no idioma Inglês no domínio jornalístico.

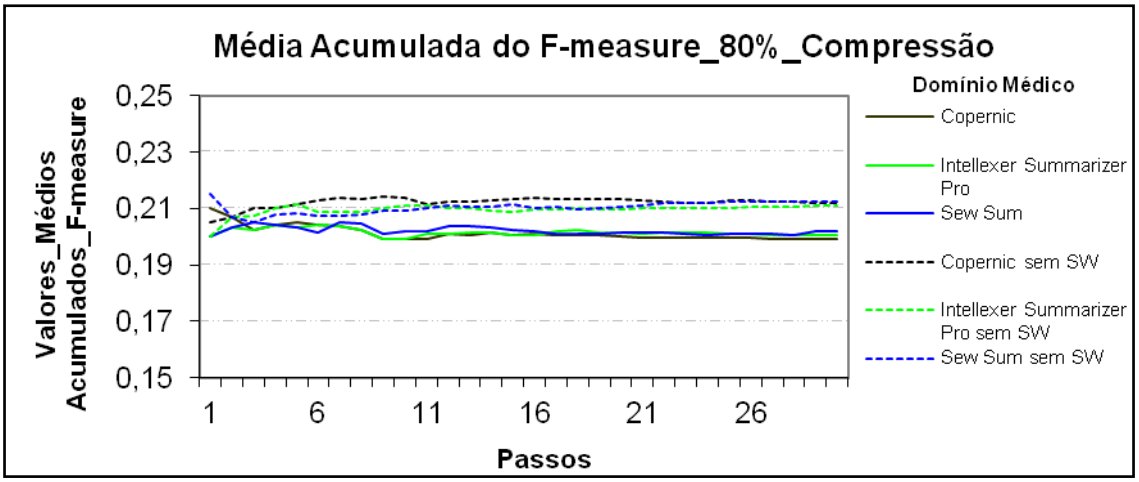


Figura 12: Resultados obtidos pelo modelo Cassiopeia, usando a medida *F-measure* com 80% de compressão no idioma Inglês no domínio médico.

4.1.1.4 Compressão de 90% no Idioma Inglês

No domínio jornalístico, como demonstra Figura 13, os maiores resultados obtidos foram com os textos sem *stopwords*. No domínio médico, como apresentado na Figura 14, os maiores resultados obtidos foram com os textos sem *stopwods* dos sumarizadores *Copernic* e *Sew Sum.* Em ambos os domínios todos os resultados ficaram entre 0,20 e 0,30.

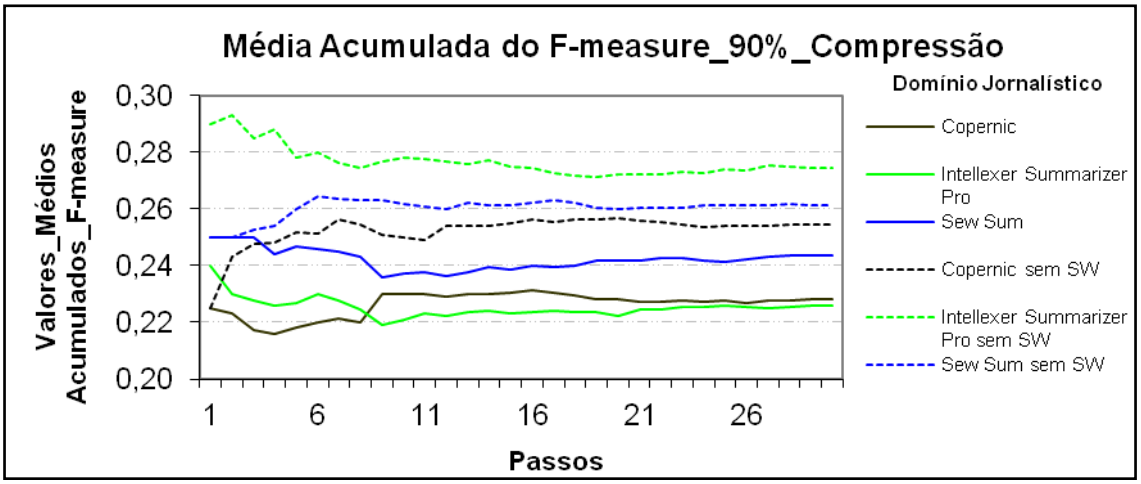


Figura 13: Resultados obtidos pelo modelo Cassiopeia, usando a medida *F-measure* com 90% de compressão no idioma Inglês no domínio jornalístico.

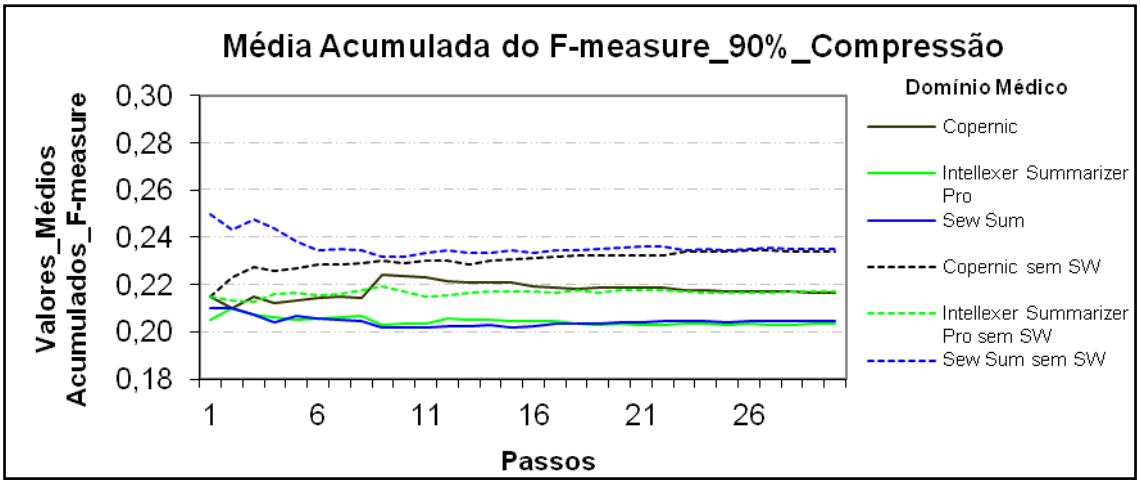


Figura 14: Resultados obtidos pelo modelo Cassiopeia, usando a medida *F-measure* com 90% de compressão no idioma Inglês no domínio médico.

4.1.1.5 Compressão de 50% no Idioma Português

No domínio jornalístico, como mostra Figura 15, os maiores resultados obtidos foram com os textos sem *stopwords*. No domínio jurídico, apresentado na Figura 16 e no médico apresentado na Figura 17, os resultados obtidos com os textos com e sem *stopwords* foram bem próximos. Em ambos os domínios todos os resultados ficaram entre 0,12 e 0,20.

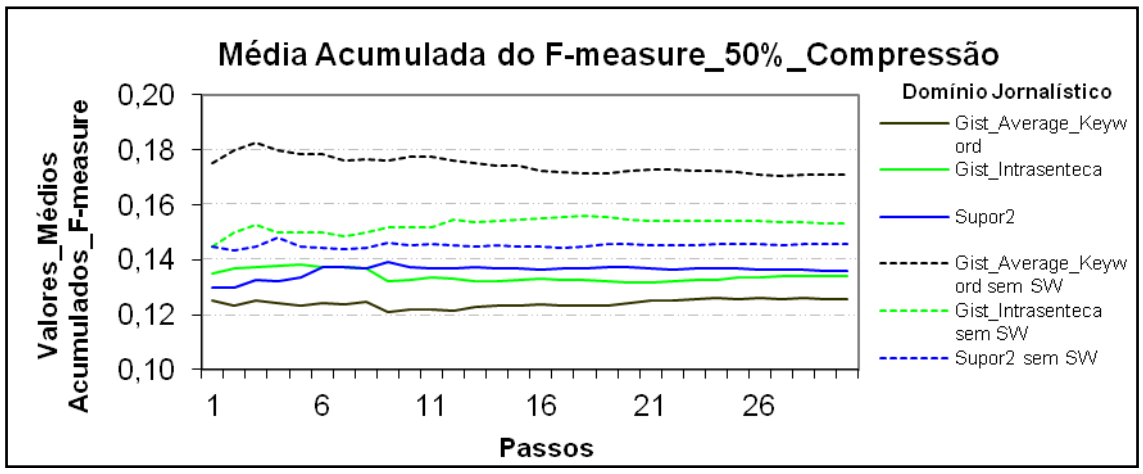


Figura 15: Resultados obtidos pelo modelo Cassiopeia, usando a medida *F-measure* com 50% de compressão no idioma Português no domínio jornalístico.

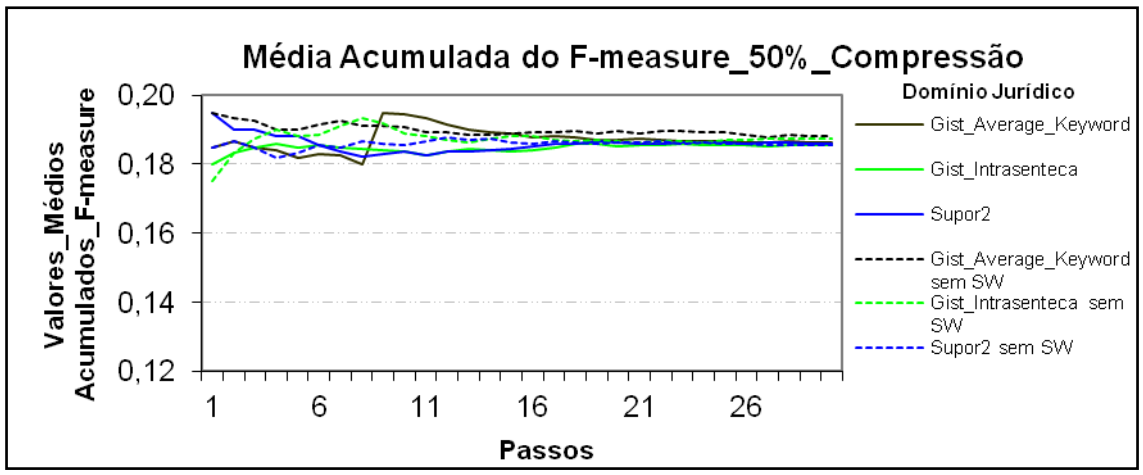


Figura 16: Resultados obtidos pelo modelo Cassiopeia, usando a medida *F-measure* com 50% de compressão no idioma Português no domínio jurídico.

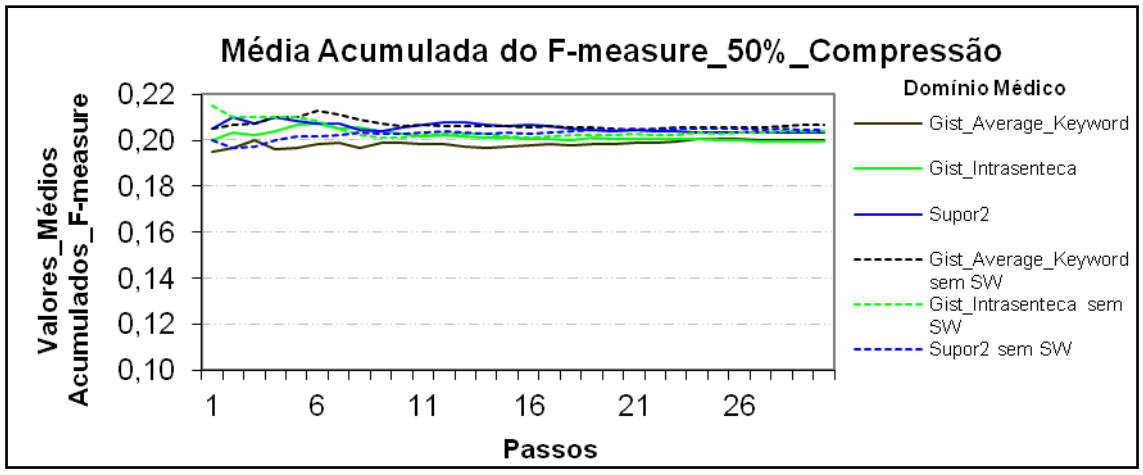


Figura 17: Resultados obtidos pelo modelo Cassiopeia, usando a medida *F-measure* com 50% de compressão no idioma Português no domínio médico.

4.1.1.6 Compressão de 70% no Idioma Português

No domínio jornalístico, como mostra Figura 18, os maiores resultados obtidos foram com os textos sem *stopwords* dos sumarizadores *Gist Intrasenteca e Supor2*, sendo que todos os resultados ficaram entre 0,08 e 0,20. No domínio jurídico, como demonstra Figura 19, os maiores resultados obtidos foram com os textos sem *stopwords*. No domínio médico, como apresentado na Figura 20, os maiores resultados obtidos foram com os textos sem *stopwords* dos sumaridores *Gist Average Keyword* e *Supor2* na maioria dos resultados. Tanto no domínio jurídico como no médico os resultados ficaram entre 0,17 e 0,23.

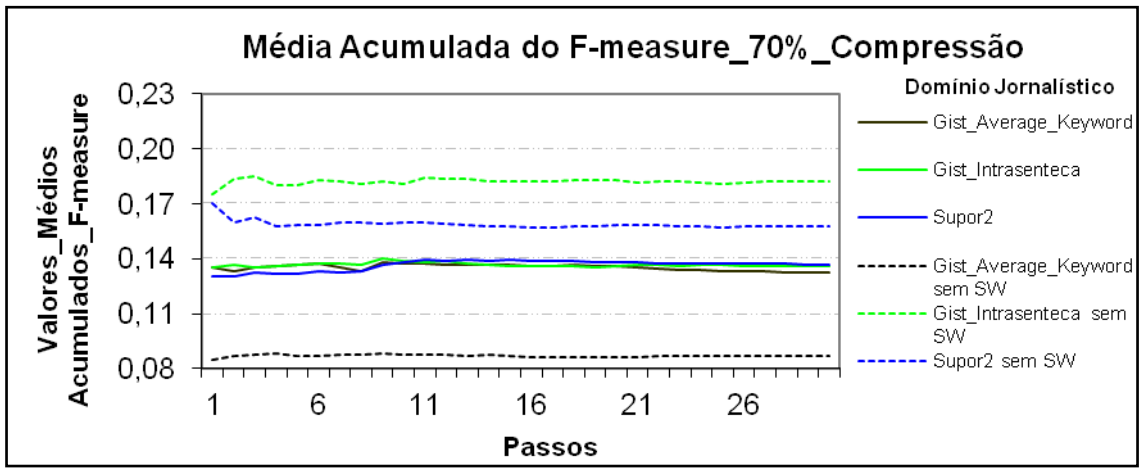


Figura 18: Resultados obtidos pelo modelo Cassiopeia, usando a medida *F-measure* com 70% de compressão no idioma Português no domínio jornalístico.

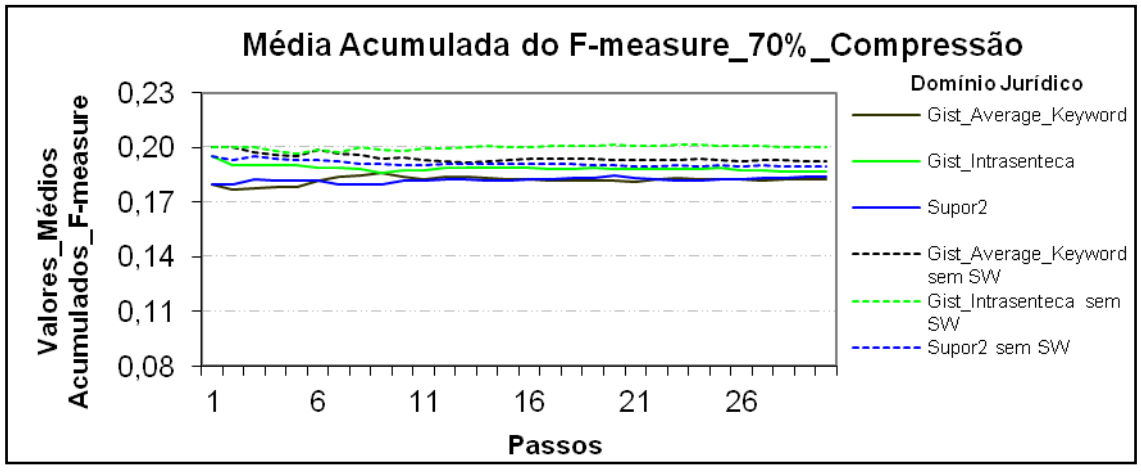


Figura 19: Resultados obtidos pelo modelo Cassiopeia, usando a medida *F-measure* com 70% de compressão no idioma Português no domínio jurídico.

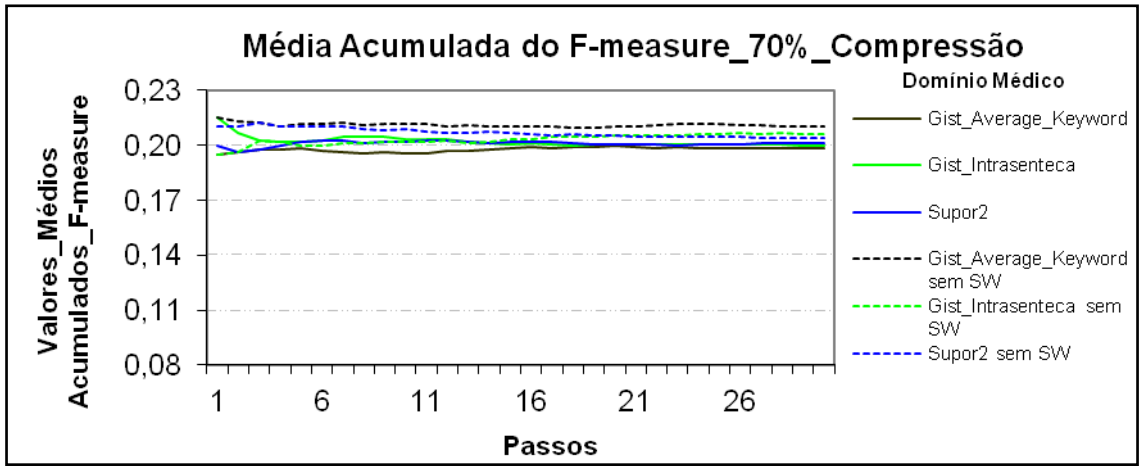


Figura 20: Resultados obtidos pelo modelo Cassiopeia, usando a medida *F-measure* com 70% de compressão no idioma Português no domínio médico.

4.1.1.7 Compressão de 80% no Idioma Português

No domínio jornalístico, como apresentado na Figura 21, os maiores resultados obtidos foram com os textos sem *stopwords* dos sumarizadores *Gist Intrasenteca e Supor2*, sendo que todos os resultados ficaram entre 0,10 e 0,20. No domínio jurídico, como demonstra Figura 22, os maiores resultados obtidos foram com os textos sem *stopwords na maioria das simulações*. No domínio médico, como apresentado na Figura 23, os maiores resultados obtidos foram com os textos sem *stopwords* dos sumaridores *Gist Average Keyword* e *Gist Intrasenteca* na maioria dos resultados. Tanto no domínio jurídico como no médico os resultados ficaram entre 0,17 e 0,22.

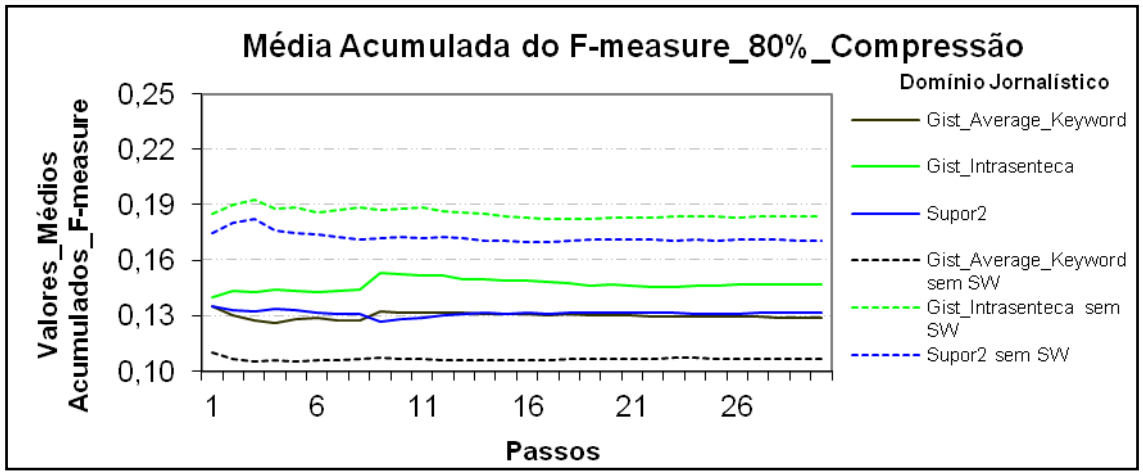


Figura 21: Resultados obtidos pelo modelo Cassiopeia, usando a medida *F-measure* com 80% de compressão no idioma Português no domínio jornalístico.

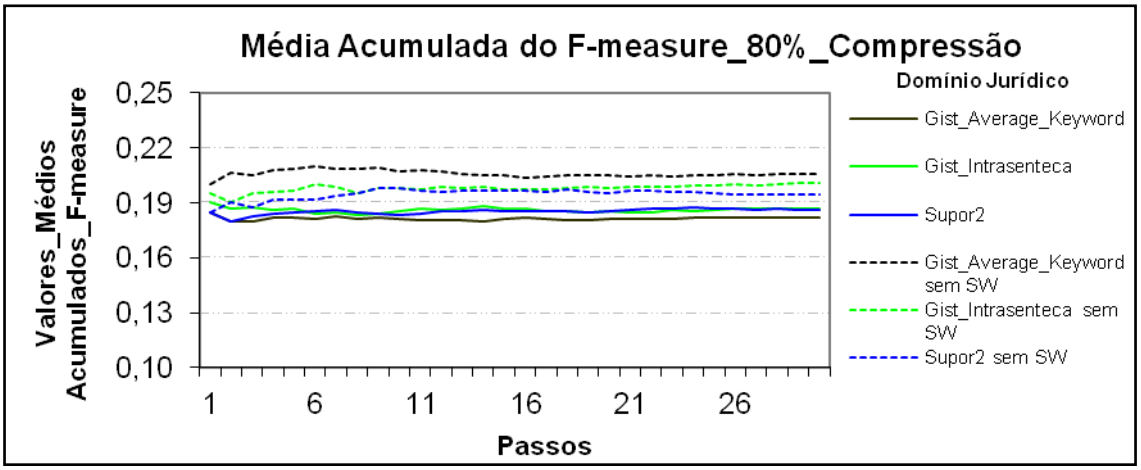


Figura 22: Resultados obtidos pelo modelo Cassiopeia, usando a medida *F-measure* com 80% de compressão no idioma Português no domínio jurídico.

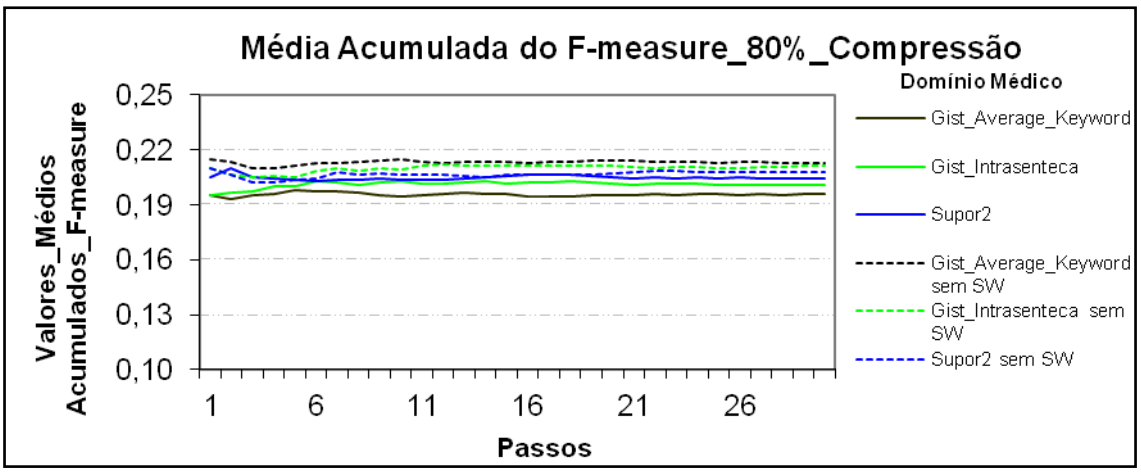


Figura 23: Resultados obtidos pelo modelo Cassiopeia, usando a medida *F-measure* com 80% de compressão no idioma Português no domínio médico.

4.1.1.8 Compressão de 90% no Idioma Português

No domínio jornalístico, como apresentado na Figura 24, os maiores resultados obtidos foram com os textos sem *stopwords* do sumarizador *Gist Average Keyword*, sendo que todos os resultados ficaram entre 0,12 e 0,20. No domínio jurídico, como mostra Figura 25 e no médico, como mostra Figura 26, os maiores resultados obtidos foram com os textos sem *stopwords*. Tanto no domínio jurídico como no médico os resultados ficaram próximos de 0,20.

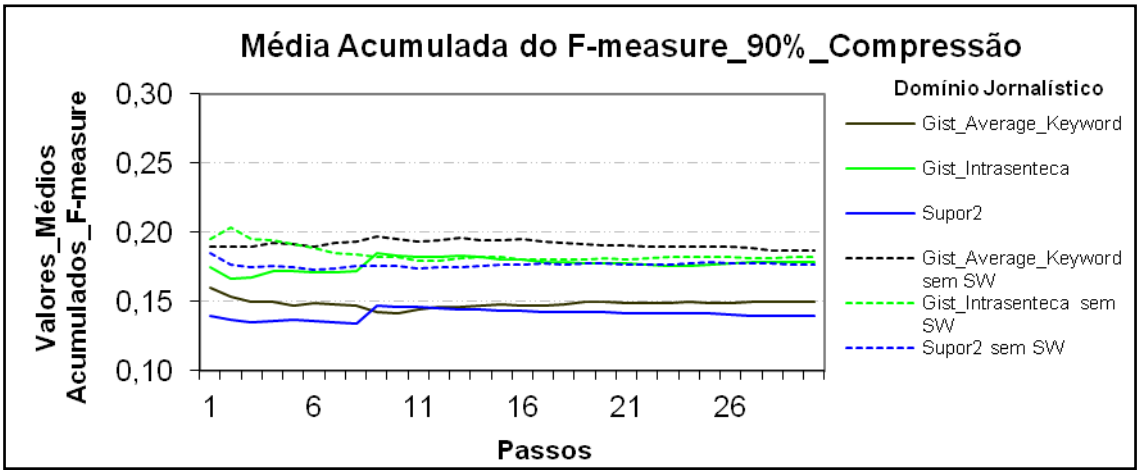


Figura 24: Resultados obtidos pelo modelo Cassiopeia, usando a medida *F-measure* com 90% de compressão no idioma Português no domínio jornalístico.

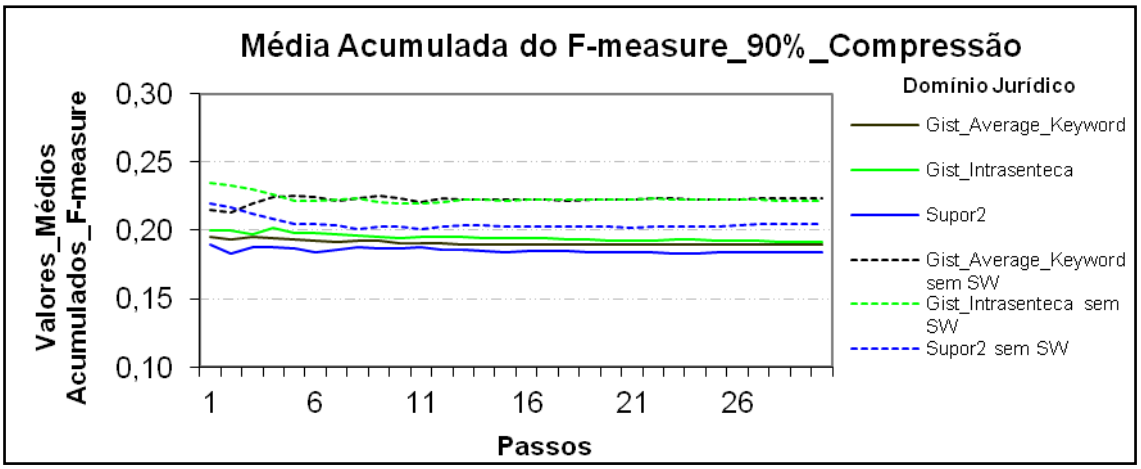


Figura 25: Resultados obtidos pelo modelo Cassiopeia, usando a medida *F-measure* com 90% de compressão no idioma Português no domínio jurídico.

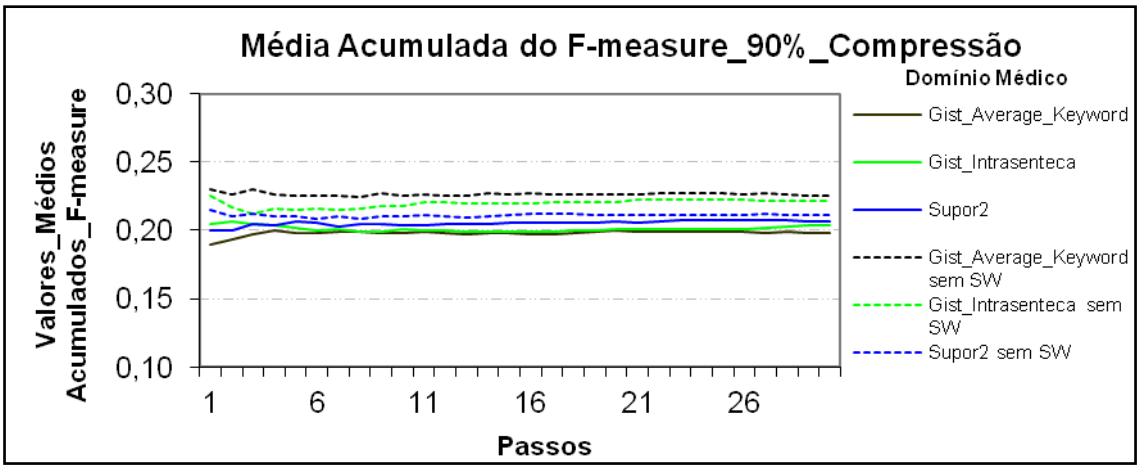


Figura 26: Resultados obtidos pelo modelo Cassiopeia, usando a medida *F-measure* com 90% de compressão no idioma Português no domínio médico.

4.1.2 Métricas Internas: Coesão, Acoplamento e *Coeficiente Silhouette*

4.1.2.1 Compressão de 50% no Idioma Inglês

Tanto no domínio jornalístico quanto no médico, assim como apresentado respectivamente nas Figuras 27 e 28, os resultados obtidos pelos textos com *stopwords* foram os maiores resultados, sendo que no domínio médico os resultados ficaram entre 0,96 e 0,99.

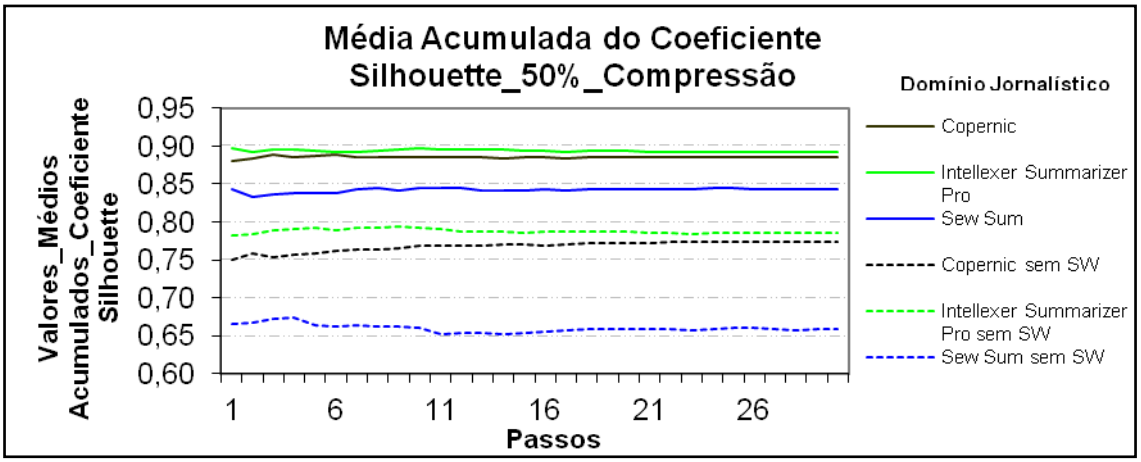


Figura 27: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Coeficiente Silhouette* com 50% de compressão no idioma Inglês no domínio jornalístico.

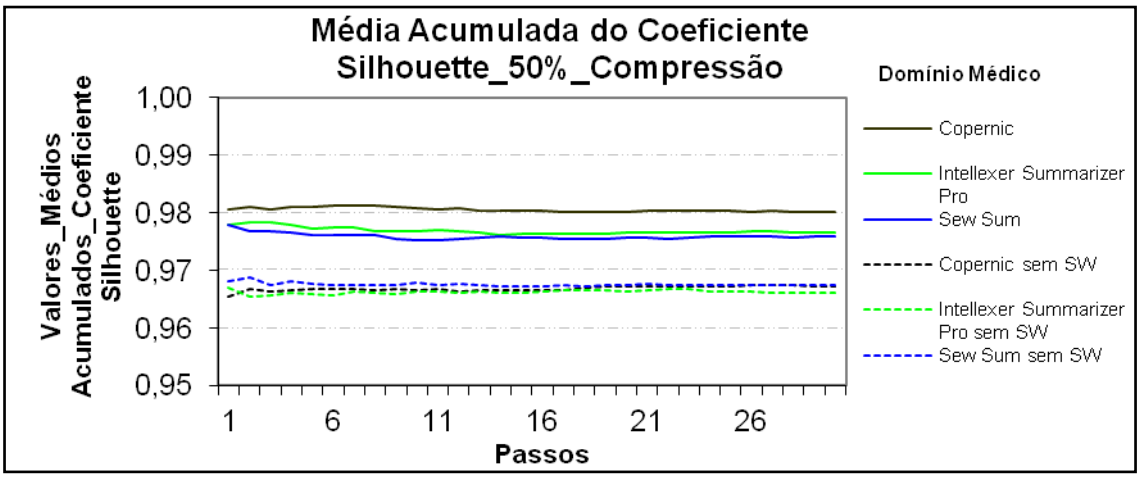


Figura 28: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Coeficiente Silhouette* com 50% de compressão no idioma Inglês no domínio médico.

4.1.2.2 Compressão de 70% no Idioma Inglês

Tanto no domínio jornalístico quanto no médico, assim como mostra respectivamente as Figuras 29 e 30, os resultados obtidos pelos textos com *stopwords* foram os maiores resultados, sendo que no domínio médico os resultados ficaram entre 0,94 e 0,98.

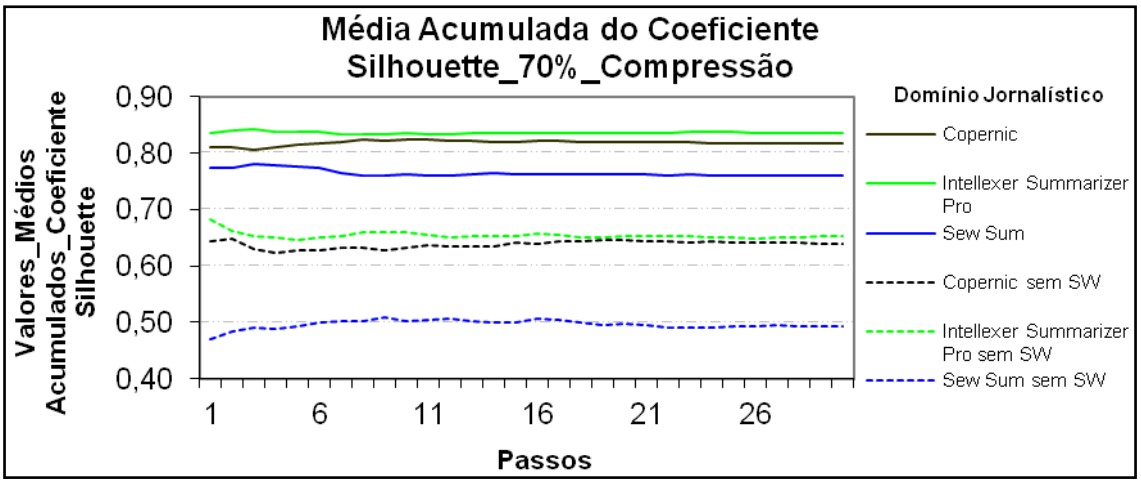


Figura 29: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Coeficiente Silhouette* com 70% de compressão no idioma Inglês no domínio jornalístico.

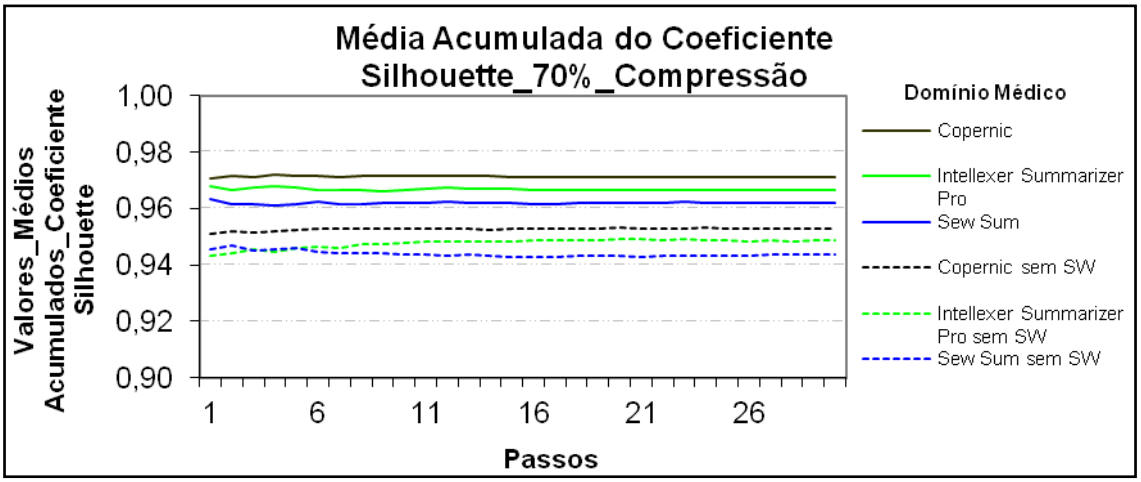


Figura 30: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Coeficiente Silhouette* com 70% de compressão no idioma Inglês no domínio médico.

4.1.2.3 Compressão de 80% no Idioma Inglês

Tanto no domínio jornalístico quanto no médico, assim como mostra respectivamente as Figuras 31 e 32, os resultados obtidos pelos textos com *stopwords* foram os maiores resultados, sendo que no domínio médico os resultados ficaram entre 0,91 e 0,97.

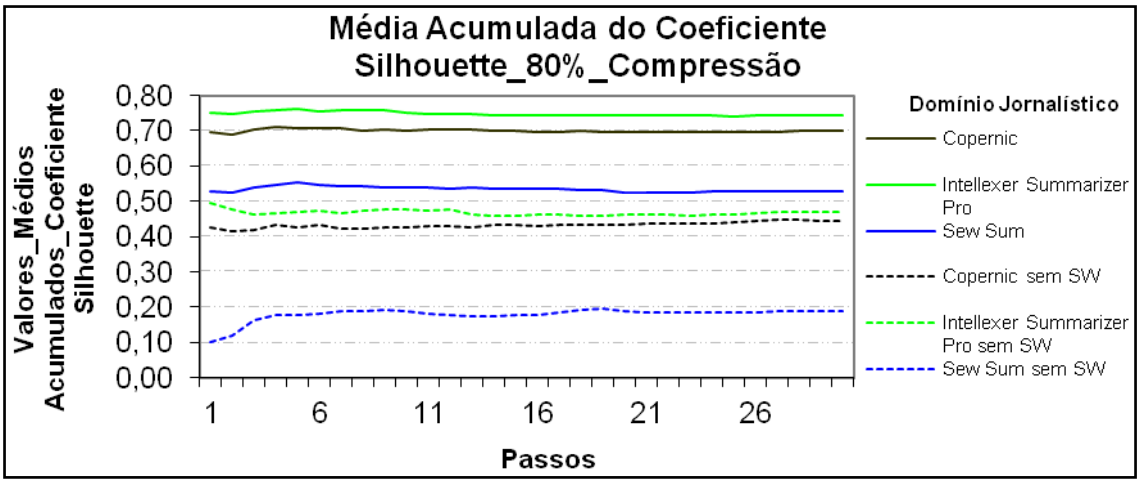


Figura 31: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Coeficiente Silhouette* com 80% de compressão no idioma Inglês no domínio jornalístico.

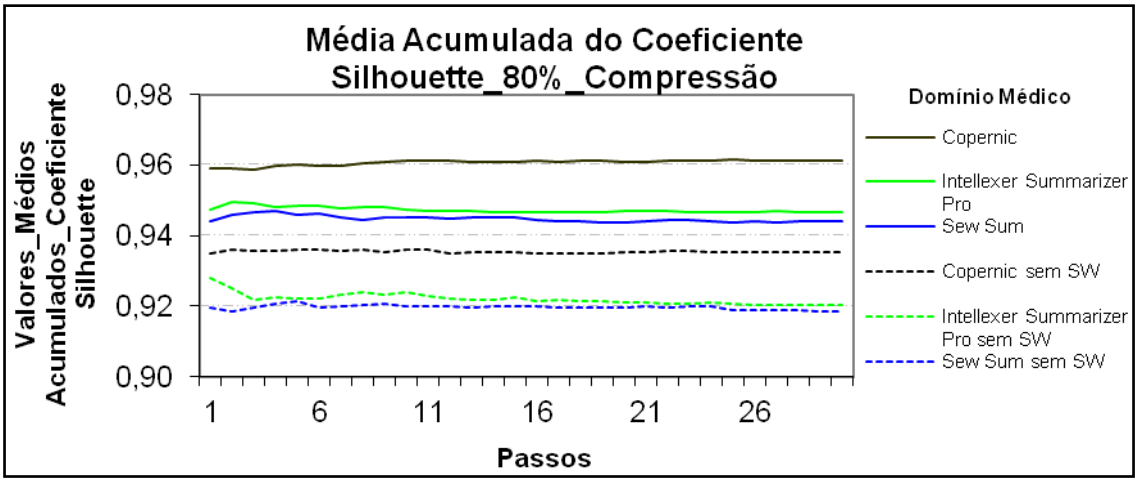


Figura 32: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Coeficiente Silhouette* com 80% de compressão no idioma Inglês no domínio médico.

4.1.2.4 Compressão de 90% no Idioma Inglês

No domínio jornalístico, como apresentado na Figura 33, os maiores resultados foram obtidos com os textos sem *stopwords* do sumarizador *Intellexer Summarizer Pro*. No domínio médico, conforme Figura 34, os maiores resultados foram obtidos com os

textos com *stopwords*, sendo que todos os resultados neste domínio ficaram entre 0,80 e 1,00.

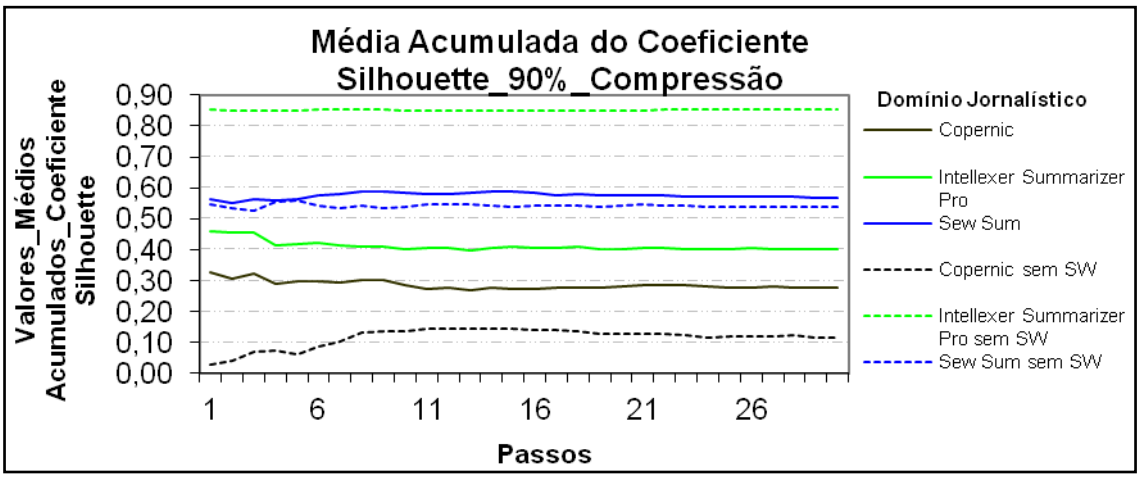


Figura 33: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Coeficiente Silhouette* com 90% de compressão no idioma Inglês no domínio jornalístico.

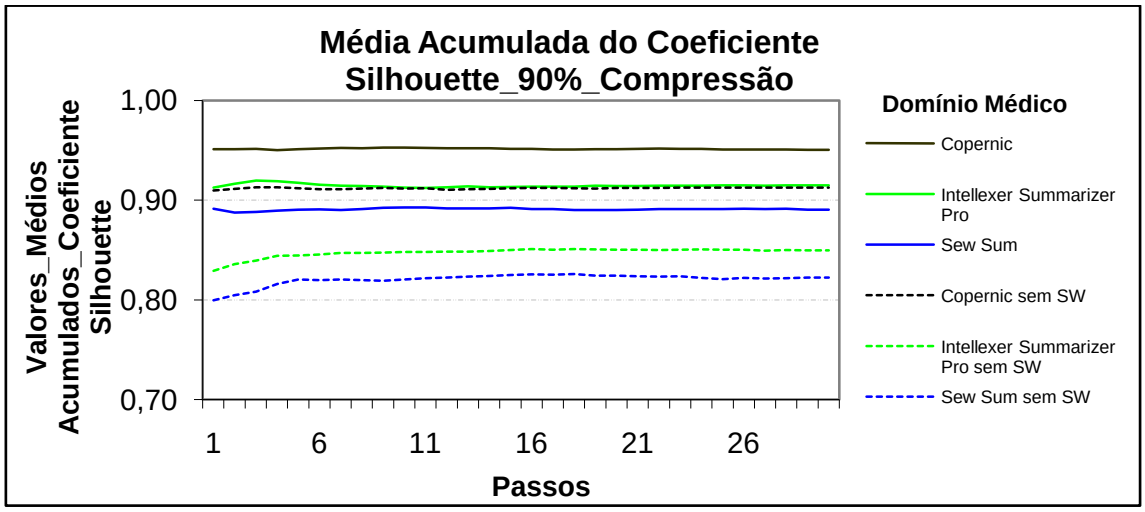


Figura 34: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Coeficiente Silhouette* com 90% de compressão no idioma Inglês no domínio médico.

4.1.2.5 Compressão de 50% no Idioma Português

No domínio jornalístico e no médico, como demonstram respectivamente as Figuras 35 e 36, os maiores resultados obtidos foram com os textos com *stopwords*. No domínio jurídico, como apresentado na Figura 37, os maiores resultados obtidos foram com os textos com *stopwords* do sumarizadores *Gist Average Keyword* e *Supor2*. Tanto no domínio jurídico como no médico, os resultados ficaram todos entre 0,95 e 0,99.

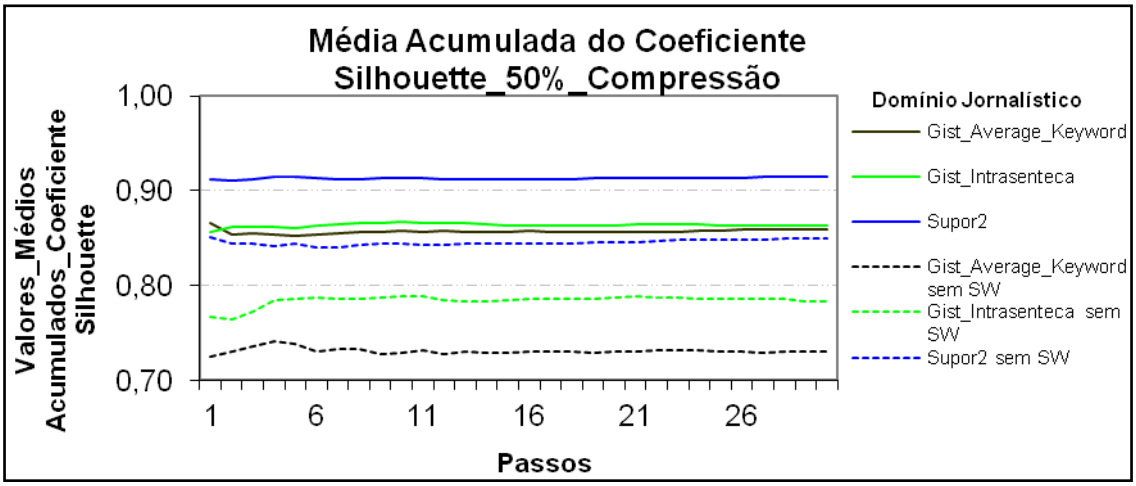


Figura 35: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Coeficiente Silhouette* com 50% de compressão no idioma Português no domínio jornalístico.

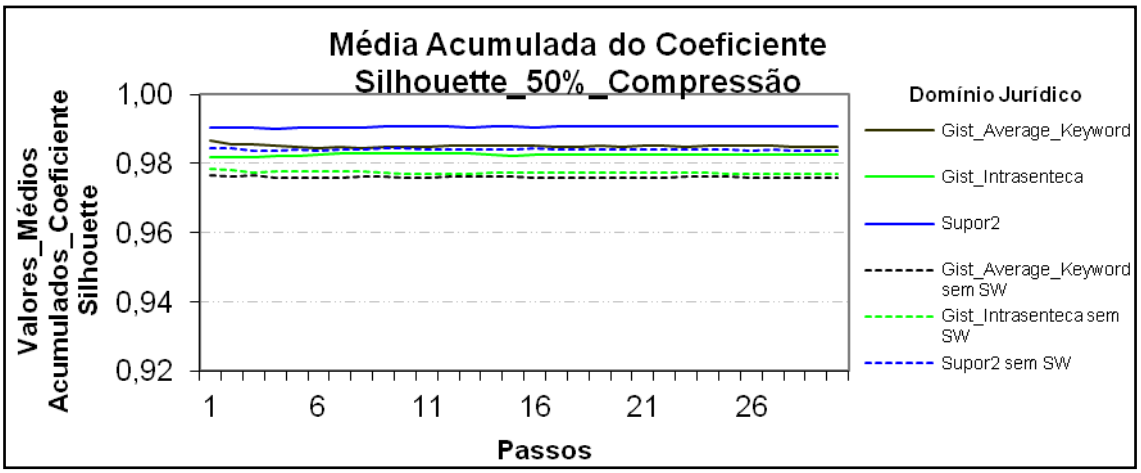


Figura 36: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Coeficiente Silhouette* com 50% de compressão no idioma Português no domínio jurídico.

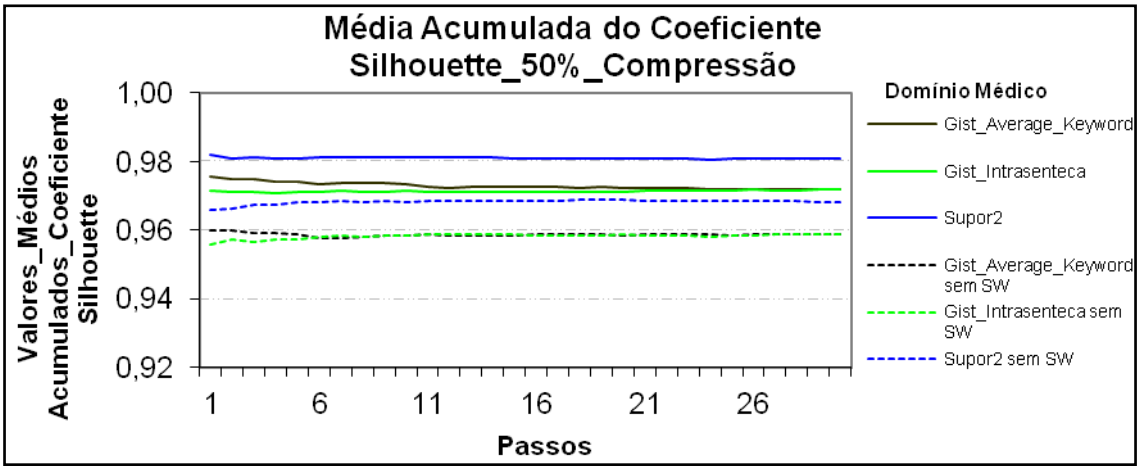


Figura 37: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Coeficiente Silhouette* com 50% de compressão no idioma Português no domínio médico.

4.1.2.6 Compressão de 70% no Idioma Português

No domínio jornalístico, conforme Figura 38, os maiores resultados obtidos foram com os textos com *stopwords* dos sumarizadores *Gist Average Keyword* e *Supor2*. No domínio jurídico e no médico, como apresentado respectivamente nas Figuras 39 e 40, os maiores resultados obtidos foram com os textos com *stopwords* do sumarizador *Supor2*. Tanto no domínio jurídico como no médico, os resultados ficaram todos entre 0,93 e 0,99.

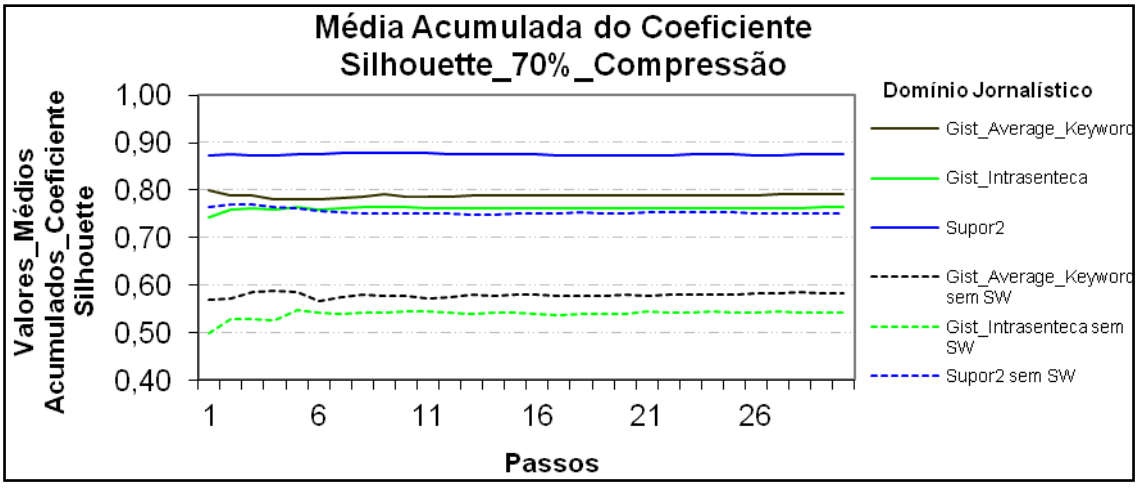


Figura 38: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Coeficiente Silhouette* com 70% de compressão no idioma Português no domínio jornalístico.

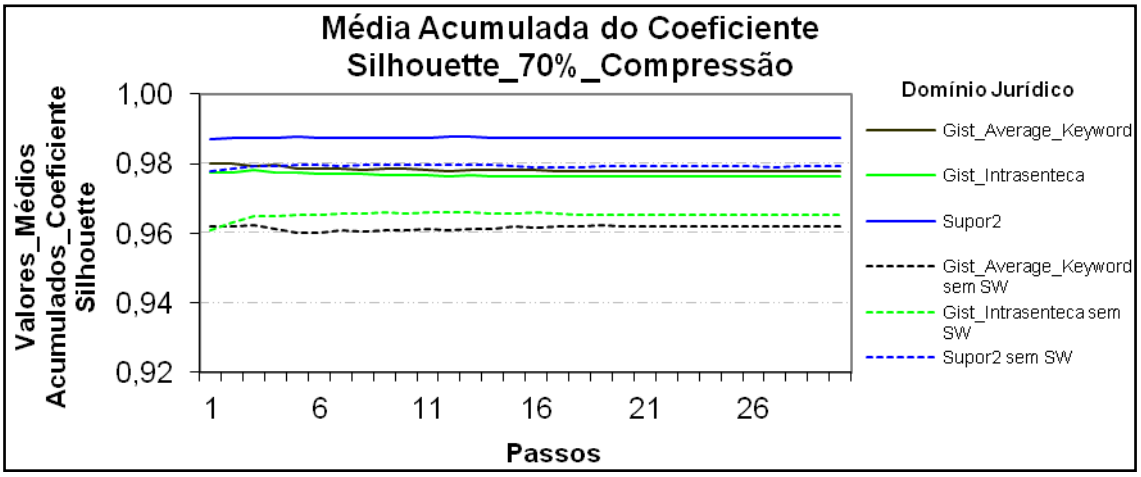


Figura 39: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Coeficiente Silhouette* com 70% de compressão no idioma Português no domínio jurídico.

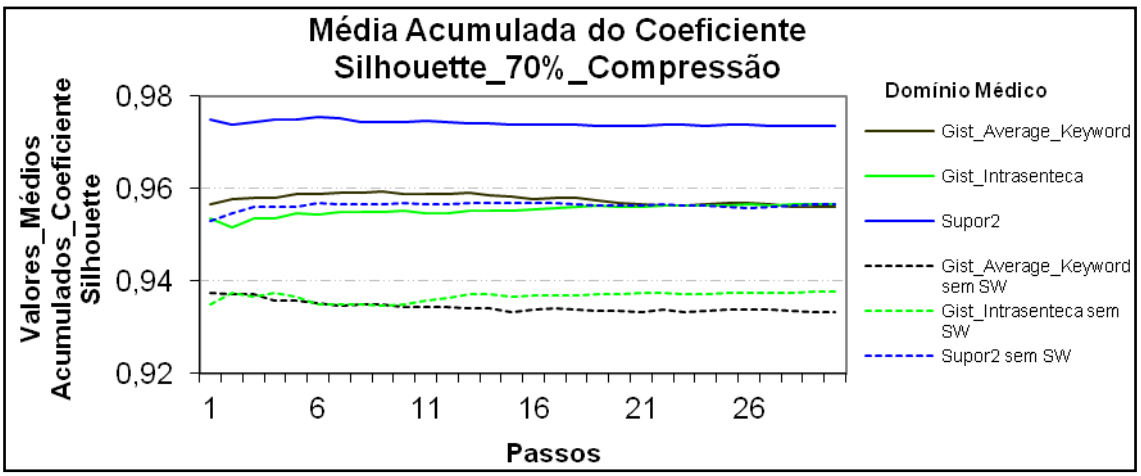


Figura 40: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Coeficiente Silhouette* com 70% compressão no idioma Português no domínio médico.

4.1.2.7 Compressão de 80% no Idioma Português

No domínio jornalístico, como mostra a Figura 41, os maiores resultados obtidos foram com os textos com *stopwords* dos sumarizadores *Gist Average Keyword* e *Supor2*. No domínio jurídico e no médico, como apresentado respectivamente nas Figuras 42 e 43, os maiores resultados obtidos foram com os textos com *stopwords* do sumarizador *Supor2*. Tanto no domínio jurídico como no médico, os resultados ficaram todos entre 0,89 e 0,99.

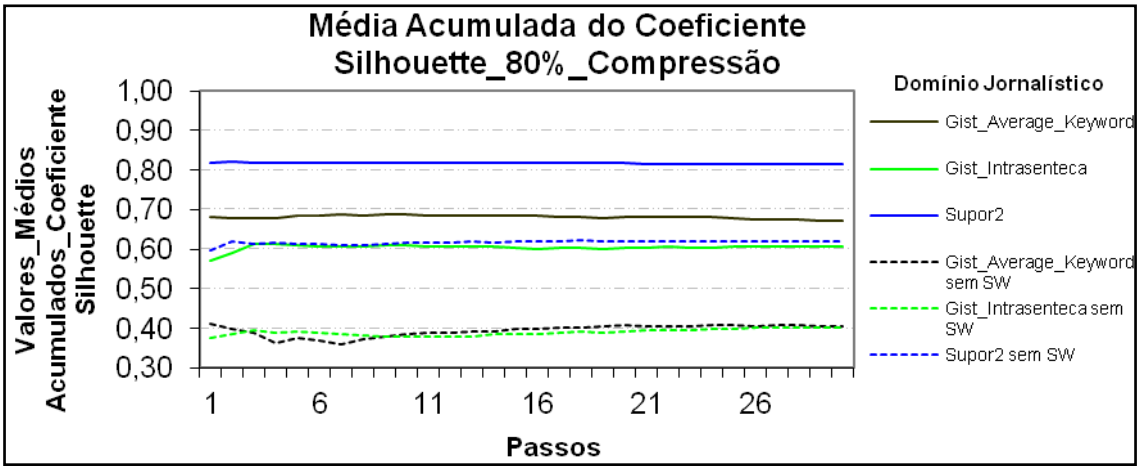


Figura 41: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Coeficiente Silhouette* com 80% de compressão no idioma Português no domínio jornalístico.

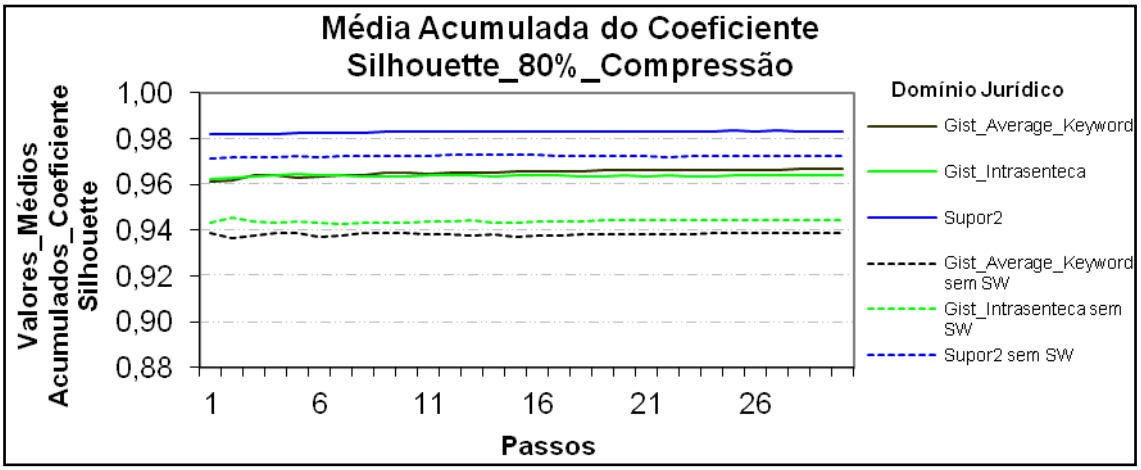


Figura 42: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Coeficiente Silhouette* com 80% de compressão no idioma Português no domínio jurídico.

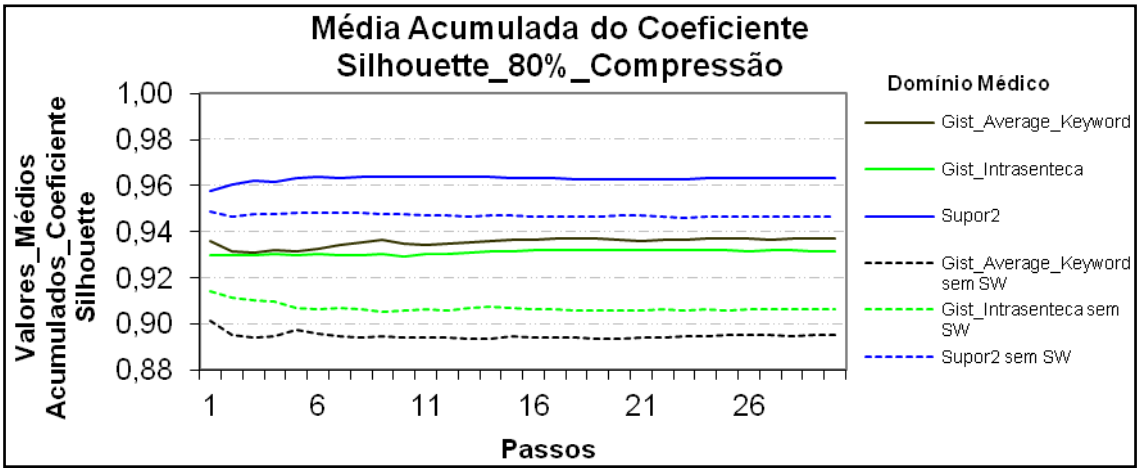


Figura 43: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Coeficiente Silhouette* com 80% de compressão no idioma Português no domínio médico.

4.1.2.8 Compressão de 90% no Idioma Português

No domínio jornalístico e no jurídico, como mostra respectivamente as Figuras 44 e 45, os maiores resultados obtidos foram com os textos com *stopwords* do sumarizador *Supor2*. No domínio médico, conforme Figura 46, os maiores resultados obtidos foram com os textos com *stopwords*.

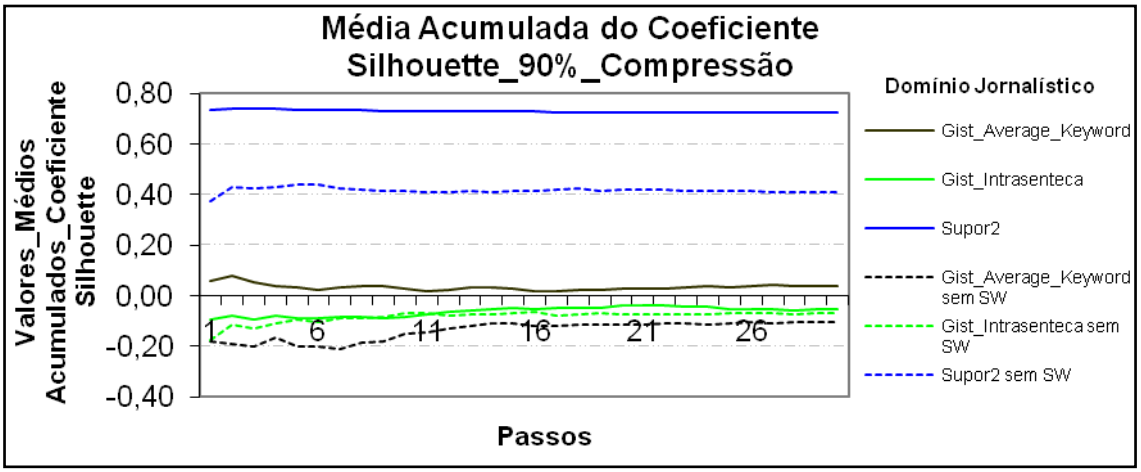


Figura 44: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Coeficiente Silhouette* com 90% de compressão no idioma Português no domínio jornalístico.

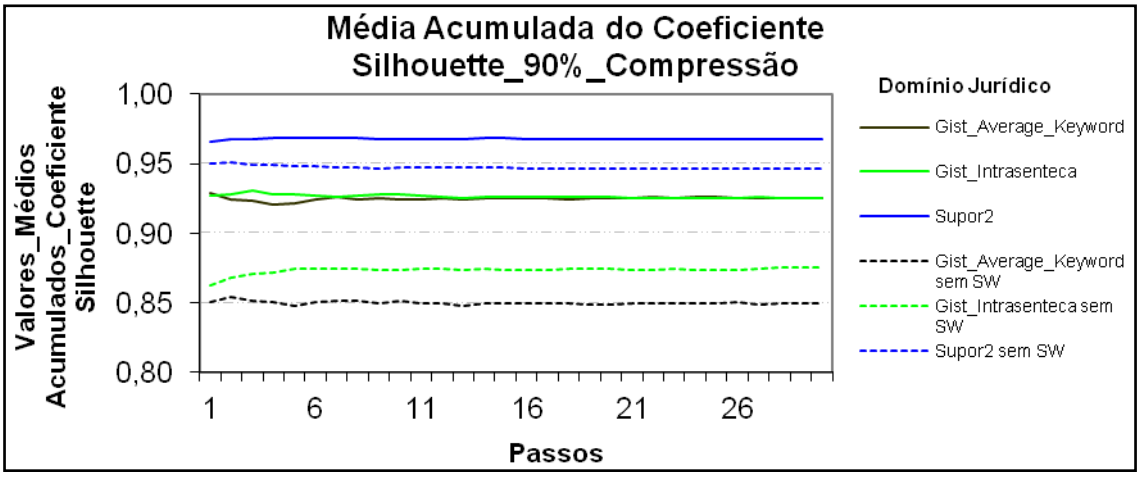


Figura 45: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Coeficiente Silhouette* com 90% de compressão no idioma Português no domínio jurídico.

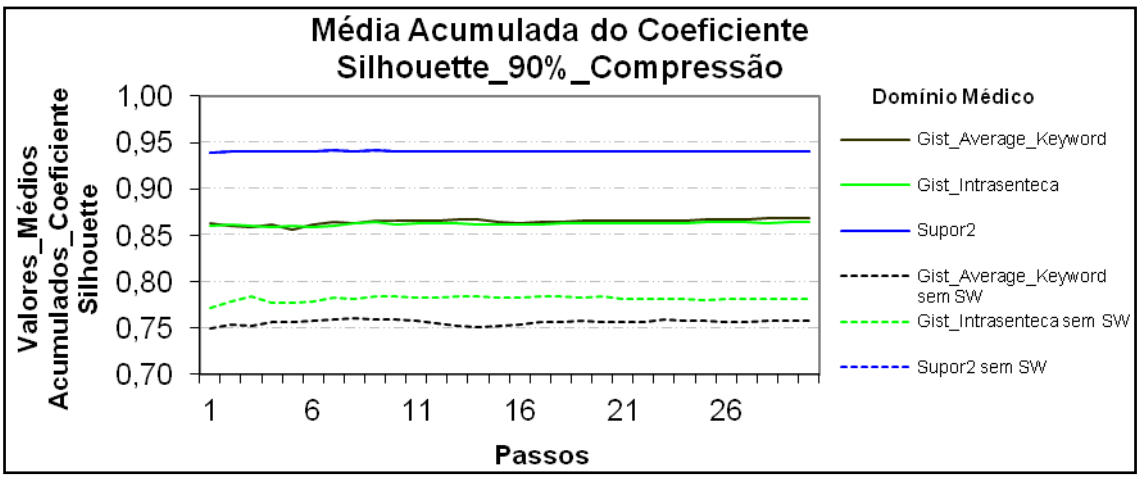


Figura 46: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Coeficiente Silhouette* com 90% de compressão no idioma Português no domínio médico.

4.2 Hipótese

A hipótese nula deste trabalho consiste que a retirada das *stopwords* de todos os textos do corpus influenciará nos resultados do modelo Cassiopeia. Formulando em um teste de hipóteses tem-se que a hipótese nula deste trabalho consiste que os resultados obtidos pelo modelo Cassiopeia são influenciados pelas *stopwords* dos textos. A equação 8 traz a representação desta hipótese nula:

Equação 8:

$$H_0: K_{\text{agrupamento com } stopwords} \neq K_{\text{agrupamento sem } stopwords} \quad (8)$$

Onde:

$H_{0:}$ hipótese nula;

$K_{\text{agrupamento com } stopwords}$: Distribuição das K amostras referente aos textos com *stopwords.*

$K_{\text{agrupamento sem } stopwords}$: Distribuição das K amostras referente aos textos sem *stopwords.*

Se a hipótese nula for considerada falsa, outra alternativa deve ser verdadeira. Este trabalho propõe a alternativa H_1, na qual os resultados de agrupamentos de textos, avaliados por métricas para estes fins destinadas, não são influenciados pela permanência das *stopwords* nos textos.

A hipótese alternativa está formalmente representada pela **Equação 9:**

$$H_1: K_{\text{agrupamento com } stopwords} = K_{\text{agrupamento sem } stopwords} \quad (9)$$

4.3 Análises dos Testes Estatísticos

Para análise dos testes estatísticos, foram geradas as tabelas apresentadas no Apêndice C. Nelas estão contidos os valores gerados para o teste ANOVA de Friedman, e para o de concordância, de Kendall, obtidos com o software citado e explicado no mesmo apêndice. São dez tabelas, cinco referentes às métricas externas, e cinco referentes às métricas internas.

As tabelas são compostas de informações, como: compressões (50%, 70%, 80% e 90%); *N* (o número de amostras); *GL* (grau de liberdade); o valor de ordem médio, soma de ordens e média, são usados para ANOVA de Friedman calculado conforme se explica na subseção 2.4.1; Coeficiente de Concordância de Kendall, calculado conforme se explica na subseção 2.4.2, e por fim o *desvio padrão*.

Submetendo os resultados obtidos com as medidas externas aos testes estatísticos, dos 20 valores obtidos do coeficiente de concordância de Kendall, apenas 3 foram um pouco mais baixos ficando entre 0,607 e 0,693, não sendo estes resultados ruins, onde os demais valores superaram 0,826. Já para as métricas internas, todos os valores de coeficiente de concordância de Kendall ficaram acima de 0,929, chegando a alcançar 1 em alguns casos.

Com estes valores de coeficiente de concordância de Kendall, observa-se a rejeição da hipótese nula (H_0) e a aceitação da hipótese alternativa (H_1).

4.4 Discussão dos resultados

Serão apresentados os valores apenas das medidas harmônicas, sendo *F-Measure* na métrica externa e *Coeficientte Silhouette* na métrica interna. Foram gerados resultados para os idiomas em inglês e em português, onde os gráficos apresentam diferentes níveis de compressão, sendo 50%, 70%, 80% e 90%, nos domínios: jornalístico, jurídico

e médico. Para identificar o sumarizador, foi utilizada uma cor específica para cada um e para diferenciar os valores obtidos com os textos com e sem *stopwords*, os contornos das colunas foram contínuos e pontilhados, respectivamente.

No domínio jornalístico, mostrado na Figura 47, observam-se valores de *F-Measure* maiores do que os valores obtidos no domínio médico, os quais praticamente não saíram da faixa de valores de 0,20 e 0,21.

Pode-se observar que, independente do domínio, os textos sem *stopwords* possibilitaram ao modelo Cassiopeia gerar resultados com valores de *F-Measure* maiores que os textos com *stopwords*. A Figura 47, que apresenta 83,33% dos resultados dos textos sem *stopwords*, foram maiores, sendo com pouca variação, todos os resultados ficaram apenas entre 0,20 e 0,29.

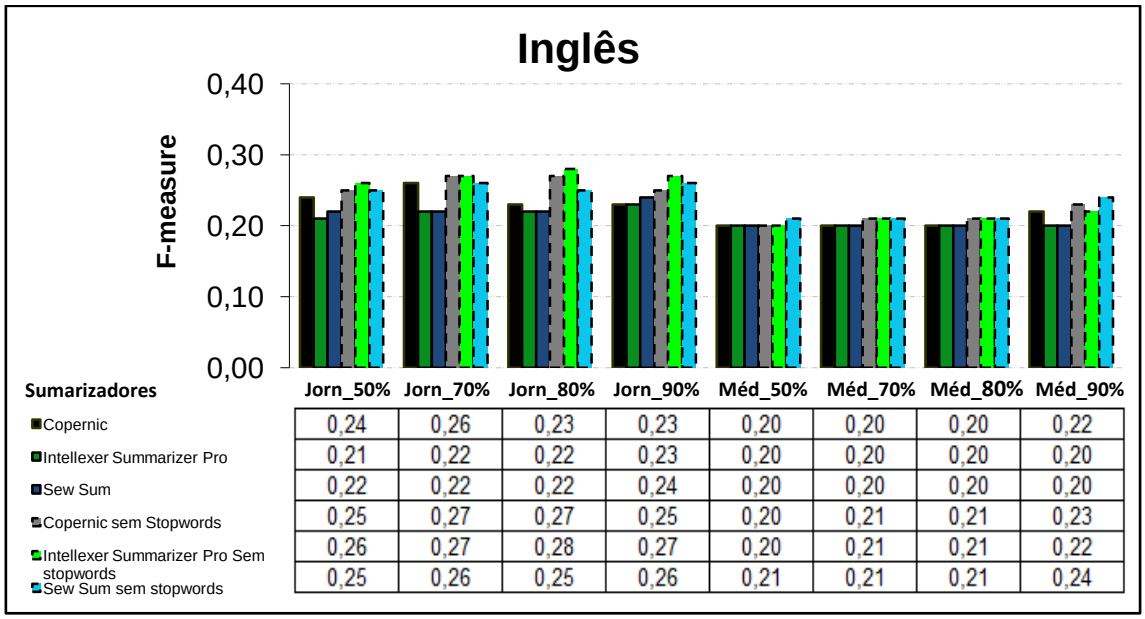
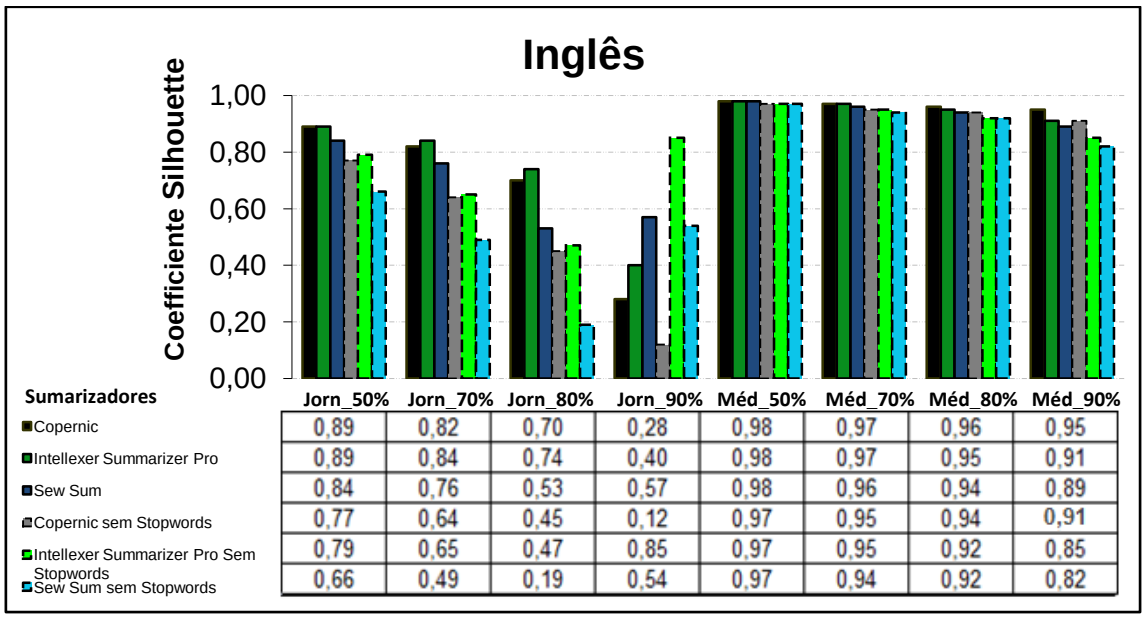


Figura 47: Mostra as médias acumuladas em 30 interações da medida *F-Measure*, no idioma Inglês, nos domínios jornalístico e médico, usando compressões de 50%, 70%, 80% e 90%. Os resultados foram através de textos com e sem *stopwords*.

Observa-se para o domínio jornalístico, na figura 48, que os agrupamentos formados pelo modelo Cassiopeia melhoram os valores da medida *Coefficiente Silhouette*, conforme a taxa de compressão diminui.

Em aproximadamente 95,8% dos resultados, observando os domínios jornalístico e médico, os textos com *stopwords* alcançaram valores de *Coefficiente Silhouette* maiores que os textos sem *stopwords*, sendo que os valores *de Coefficiente Silhouette* dos textos sem *stopwords* superaram o dos textos com *stopwords*, na compressão de 90% no domínio jornalístico.

No domínio médico praticamente todos os resultados ficaram entre 0,80 e 1,00, tanto dos textos com e sem *stopwords*. Por outro lado, no domínio jornalístico houve variações maiores, resultados entre 0,12 e 0,89.



Sumarizadores	Jorn_50%	Jorn_70%	Jorn_80%	Jorn_90%	Méd_50%	Méd_70%	Méd_80%	Méd_90%
Copernic	0,89	0,82	0,70	0,28	0,98	0,97	0,96	0,95
Intellexer Summarizer Pro	0,89	0,84	0,74	0,40	0,98	0,97	0,95	0,91
Sew Sum	0,84	0,76	0,53	0,57	0,98	0,96	0,94	0,89
Copernic sem Stopwords	0,77	0,64	0,45	0,12	0,97	0,95	0,94	0,91
Intellexer Summarizer Pro Sem Stopwords	0,79	0,65	0,47	0,85	0,97	0,95	0,92	0,85
Sew Sum sem Stopwords	0,66	0,49	0,19	0,54	0,97	0,94	0,92	0,82

Figura 48: Mostra as médias acumuladas em 30 interações da medida *Coefficiente Silhouette*, no idioma Inglês, nos domínios jornalístico e médico, usando compressões de 50%, 70%, 80% e 90%. Os resultados foram através de textos com e sem *stopwords*.

Observa-se que nos domínios jornalístico, médico e jurídico, na figura 49, os valores de *F-Measure* referentes aos textos com *stopwords* não ultrapassaram os textos sem *stopwords.* A exceção foi para os textos sumarizados pelo *Copernic* na compressão de 70% e 80%.

No Geral, os valores de agrupamentos gerados pelo modelo Cassiopeia na medida *F-Measure* dos textos sem *stopwords*, foram maiores que os textos com *stopwords* em 69,45% dos casos, sendo que, em 25% dos casos, os valores de *F-Measure* dos textos com e sem stopwords foram equivalentes. Tanto no domínio médico como no jurídico os resultados foram bastante próximos, ficando sempre entre 0,18 e 0,23, no domínio jornalistico houve variações entre 0,13 e 0,18.

A figura 50 mostra que os textos sem *stopwords* não conseguiram aumentar os valores de *Coefficiente Silhouette* nos agrupamentos gerados pelo modelo Cassiopeia. O máximo que os textos sem *stopwords* alcançaram foi, em 8,33% dos casos, manter valores equivalentes aos textos com *stopwords*, não conseguindo em nenhum caso obter valores de *Coefficiente Silhouette* mais significativos.

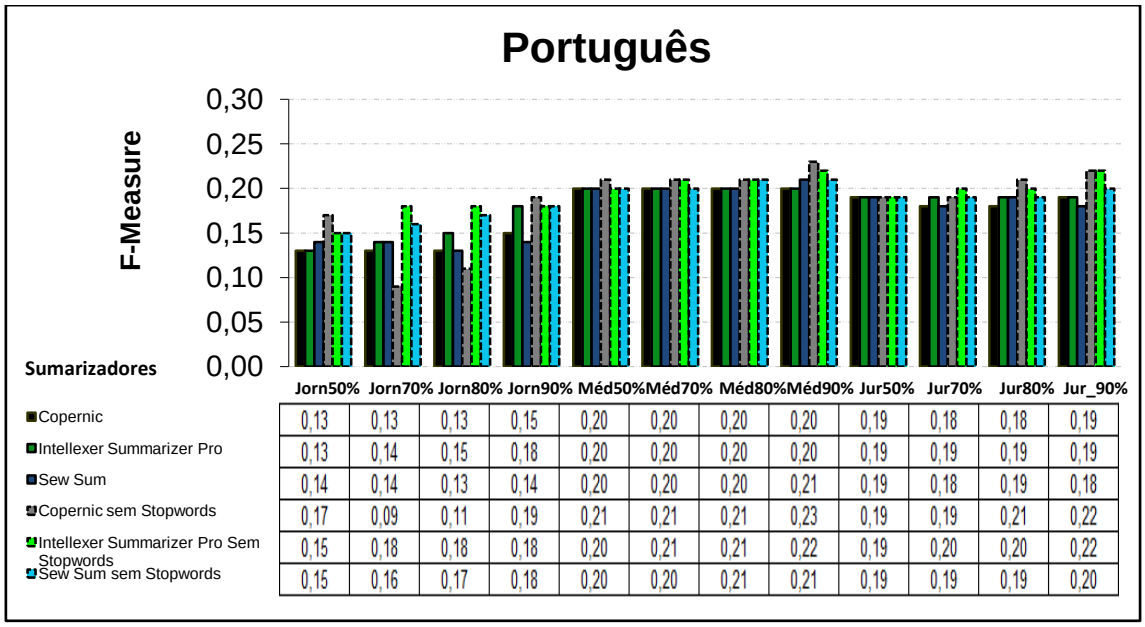
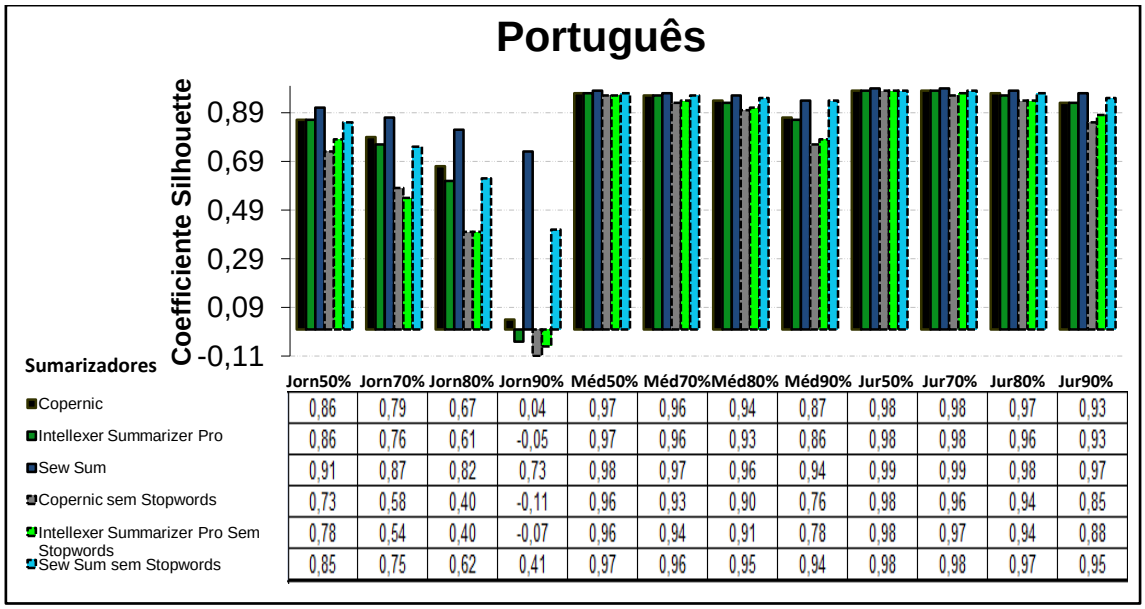


Figura 49: Mostra as médias acumuladas em 30 interações da medida *F-Measure*, no idioma português, nos domínios jornalístico, médico e jurídico, usando compressões de 50%, 70%, 80% e 90%. Os resultados foram com os textos com e sem *stopwords*.

Observa-se, no domínio jornalístico, que conforme a taxa de compressão dos textos sumarizados aumenta, os valores de *Coefficiente Silhouette* gerados pelo modelo Cassiopeia tendem a diminuir. Na compressão de 90%, nota-se que os valores de *Coefficiente Silhouette* diminuiram em uma proporção bem mais significativa do que nas compressões anteriores.

Observando os domínios médico e jurídico, nota-se pouca variação dos resultados, sendo que tais resultados ficaram na maioria das vezes entre 0,85 e 0,98.



Sumarizadores	Jorn50%	Jorn70%	Jorn80%	Jorn90%	Méd50%	Méd70%	Méd80%	Méd90%	Jur50%	Jur70%	Jur80%	Jur90%
Copernic	0,86	0,79	0,67	0,04	0,97	0,96	0,94	0,87	0,98	0,98	0,97	0,93
Intellexer Summarizer Pro	0,86	0,76	0,61	-0,05	0,97	0,96	0,93	0,86	0,98	0,98	0,96	0,93
Sew Sum	0,91	0,87	0,82	0,73	0,98	0,97	0,96	0,94	0,99	0,99	0,98	0,97
Copernic sem Stopwords	0,73	0,58	0,40	-0,11	0,96	0,93	0,90	0,76	0,98	0,96	0,94	0,85
Intellexer Summarizer Pro Sem Stopwords	0,78	0,54	0,40	-0,07	0,96	0,94	0,91	0,78	0,98	0,97	0,94	0,88
Sew Sum sem Stopwords	0,85	0,75	0,62	0,41	0,97	0,96	0,95	0,94	0,98	0,98	0,97	0,95

Figura 50: Mostra as médias acumuladas em 30 interações da medida *Coefficiente Silhouette*, no idioma Português, nos domínios jornalístico, médico e jurídico, usando compressões de 50%, 70%, 80% e 90%. Os resultados foram através de textos com e sem *stopwords*.

Analisando todos os experimentos, pode-se observar que, na medida *F-Measure*, em quase todos os resultados obtidos indiferente de idioma e de domínio, os textos sem stopwords apresentaram valores maiores. Já os resultados de *Coefficiente Silhouette*, também indiferente de idioma e domínio, apresentaram valores maiores com os textos com stopwords. Resalta-se que quando os valores foram maiores, independente dos textos serem com ou sem stopwords, a diferença foi considerada pequena.

Uma consideração importante é feita quanto ao domínio jornalístico. Este domínio, independente do idioma português ou inglês e independente da métrica de avaliação, *F-Measure* ou *Coefficiente Silhouette*, foi o que obteve mais variações em seus resultados.

Capítulo 5 – CONCLUSÕES

O principal objetivo deste trabalho foi apresentar uma avaliação do desempenho do modelo Cassiopeia, quando utilizado como *clusterizador*. Foram avaliados os resultados dos agrupamentos formados com textos sem *stopwords* comparando-os com os sem *stopwords*.

Avaliando a medida *Coefficiente Silhouette* no domínio jornalístico no idioma inglês e no idioma português apresentados na Figura 48 e 50, observa-se a queda dos valores da medida harmônica da métrica interna conforme a taxa de compressão dos textos aumenta. Este fato pode ser justificado pelo pequeno tamanho dos textos jornalísticos originais, assim quanto maior a compressão aplicada ao texto maior será a perda de informatividade.

Avaliando os resultados das métricas internas, observados nas figuras 48 e 50 o uso da medida *Coefficiente Silhouette*, que é a medida harmônica da coesão e do acoplamento, nota-se bons resultados para o modelo Cassiopeia, afinal na maioria dos resultados os textos com e sem *stopwords* alcançaram valores da medida *Coefficiente Silhouette* mais elevados, sendo pequenas as variações entre estes. Esses resultados foram bem significativos para a pesquisa, pois utilizando o Cassiopeia como *clusterizador*, o presente trabalho tem por finalidade mostrar a grande relação existente entre os textos de um mesmo cluster e a distância existente entre textos de clusters diferentes.

Avaliando os resultados das métricas externas, observados nas Figuras 47 e 49 o uso da medida *F-Measure*, que é a medida harmônica do *recall* e da *precision*, nota-se resultados mais baixos, onde estes não ultrapassaram 0,28. Portanto a variação dos resultados obtidos com os textos com e sem *stopwords* foi muito pequena para a métrica avaliada. Tais métricas não seriam adequadas para avaliar um *clusterizador*, para utilizar as métricas externas para avaliar o modelo Cassiopeia, o modelo deveria ser utilizado como ferramenta de busca, porém, acredita-se que quando utilizado para recuperação de informação, o modelo terá resultados satisfatórios, pois se as métricas externas não apresentaram variações significativas no agrupamento, não apresentarão também na Recuperação de Informação .

Assim pode-se concluir que as *stopwords* presentes nos centroides que representam os textos no modelo Cassiopeia não atrapalham os resultados do algoritmo quando o mesmo é utilizado para formação de agrupamento de textos.

5.1 Limitações

Neste trabalho, foi utilizada apenas um corpora em inglês e outro em português, sendo cada domínio possuindo 100 textos distintos. Acredita-se que esse fator seja limitante para o trabalho, uma vez que os textos poderiam ser em um número maior e com uma maior variação entre idiomas.

5.2 Trabalhos futuros

Todas as simulações realizadas para este trabalho, foram utilizando o Modelo Cassiopéia. Algorítmo este, conforme capítulo 2, agrupa os textos diferentemente a outros algoritmos da literatura. Como este algoritmo apresenta resultados satisfatórios, sugere-se para trabalhos futuros, utilizar outros algoritmos para comparação de resultados de agrupamento de textos com e sem *stopwords*.

REFERÊNCIAS BIBLIOGRÁFICAS

ALMEIDA, L. G. P. **Análise de algoritmos de agrupamento para base de dados textuais.** 2007, 161f. Tese (Mestrado) – Laboratório Nacional de Computação Científica, 2007.

ARORA R., BANGALORE P. **Text Mining: Classification & Clustering of articles related to sports.** Proceedings of the 43rd annual Southeast regional conference, vol. 1, 2005.

BITTENCOURT, A. C.; GALHO, T. S.; MORAES, S. M. W. **Um estudo comparativo de uma Rede Backpropagation e de Similaridade Difusa para a Identificação de Spam.** In: XXV CONGRESSO DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO. São Leopoldo, RS, 22 a 29 de julho, 2005.

BONATO, J. **Automatizando o Processo de Estimativa de Revocação e Precisão de Funções de Similaridade.** 2008. Dissertação (Mestrado) – Programa de Pós Graduação em Computação, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2008.

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

CORRÊA, G. N.; MARCACINI, R. M.; REZENDE, S. O. **Uso da mineração de textos na análise exploratória de artigos científicos.** São Carlos, 2012.

DELGADO, C. C. N.; DIAS, H. D. **Utilização de sumários humanos no modelo Cassiopeia.** 2012. 71f. Monografia – Curso Bacharel em Engenharia da Computação, Centro universitário de Barra Mansa, Barra Mansa, 2012.

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

GALHO, T, S. **Categorização automática de documentos de textos utilizando lógica difusa.** 2003, 79f. Monografia – Curso Bacharel em Ciência da Computação, Universidade Luterana do Brasil, câmpus Gravataí, Gravataí, 2003.

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, China, 2009. p. 24 - 27.

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**, USA, v. 3, p. 110-121, setembro 2011.

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: proceedinds of the 6th international conferense for INTERNET TECHNOLOGY and SECURED TRANSACTIONS. Abu Dhabi,UAE 2011. p 388 – 395.

GUELPELI, M. V. C. **Cassiopeia: Um modelo de agrupamento de textos baseado em sumarização.** 2012, 220f**.** Tese (Doutorado) – Programa de Pós Graduação em computação, Universidade Federal Fluminense, Niterói, 2012.

LOH, S. Text mining por Stanley Loh. Disponível em: http://miningtext.blogspot.com.br/2008/11/listas-de-stopwords-stoplist-portugues.html. Acesso em: 04 de Março de 2013.

LOPES, M. C. S. **Mineração de dados textuais utilizando técnicas de clustering para o idioma português**. 2004, 191f. Tese (Doutorado) – Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2004.

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

METZ, J.; MONARD, M. C. **Clustering Hierárquico: uma metodologia para auxiliar na interpretação de clusters**. In: XXV CONGRESSO DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO. São Leopoldo, RS, 22 a 29 de julho, 2005.

NASSIF, L. F. C. **Técnicas de agrupamento de textos aplicadas à computação forense**. 2010. 71f. Dissertação (Mestrado) – Faculdade de Tecnologia. Departamento de Engenharia Elétrica, Universidade de Brasília, Brasília, 2011.

NORMANDO, D.; TJÄDERHANE, L.; QUINTÃO C. C. A. A escolha do teste estatístico – um tutorial em forma de apresentação em PowerPoint. Dental Press J. Orthod. vol. 15, n. 1, pp. 101-106. 2010.

OLIVEIRA, E.; CIARELLI, P. M.; SANTOS, M. H.; COSTA B. O. Um modelo algébrico para representação, indexação e classificação automática de documentos digitais. **Revista Brasileira de Biblioteconomia e Documentação**, São Paulo, v. 3, n. 1, p. 73-98, jan./jun. 2007.

PRIOR, A. K. F. **Exame de partículas aplicado ao agrupamento de textos**. 2010. 61f. Dissertação (Mestrado) – Programa de pós Graduação em Engenharia Elétrica, Universidade Presbiteriana Mackenzie, São Paulo, 2010.

REZENDE, S. O.; MARCACINI, R. M.; MOURA, M. F. O uso da Mineração de Textos para Extração e Organização Não Supervisionada de Conhecimento. **Revista de Sistemas de Informação da FSMA**, Macaé, n.7, p. 7-21, 2011.

REZENDE, S. O.; MARCACINI, R. M.; CORRÊA G. N. Uso da mineração de textos na análise exploratória de artigos científicos, 2012.

SÁ, H. R. **Seleção de características para classificação de textos**. 2008. 57f. Monografia – Graduação em Ciência da Computação, Centro de Informática da Universidade Federal de Pernambuco, Recife, 2008.

SALAZAR, F. A. S. R. **Um estudo sobre o papel de medidas de similaridade em visualização de coleções de documentos**. 2012, 125f. Tese (Mestrado) – Instituto de Ciências Matemáticas e computação, Universidade de São Paulo, São Carlos 2012.

SANTOS, M. N.; COSTA, B. O.; OLIVEIRA, E. **Utilizando Comparações Ponderadas em Classificação Automática de Documentos**. In: SIMPÓSIO INTERNACIONAL DE BIBLIOTECAS DIGITAIS, Campinas, São Paulo, 2005.

ZOU, F., WANG, F. L., DENG, X., HAN, S. **Automatic Identification of Chinese Stop Words.** Conference on Intelligent Text Processing and Computational Linguistics, Kathmandu, 2006. P. 151-162.

WIVES, L. K.; LOH, S. **Recuperação de Informações Usando a Expansão Semântica e a Lógica Difusa**. In: CONGRESO INTERNACIONAL DE INGENIERÍA INFORMÁTICA and IV INTERNATIONAL CONGRESS OF INFORMATION ENGINEERING, Buenos Aires, 1998. p. 201-211.

WIVES, L. K. **Um estudo sobre Agrupamento de Documentos Textuais em Processamento de Informações não Estruturadas Usando Técnicas de "Clustering".** 1999. Dissertação (Mestrado) – Programa de Pós Graduação em Computação, Universidade Federal do Rio Grande do Sul, Porto Alegre, 1999.

WIVES, L. K. **Utilizando Conceitos como Descritores de Textos para o Processo de Identificação de Conglomerados (Clustering) de Documentos.** 2004, 136f. Tese (Doutorado) - Universidade Federal do Rio Grande do Sul, Porto Alegre, 2004.

APÊNDICES

Apêndice A - Métricas externas

O Apêndice A mostra a continuidade dos resultados descritos na subseção 4.1.1, onde foi apresentada a medida *F Measure*. 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 4.1.1 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*.

Média acumulada do recall

Compressão de 50% no idioma inglês

Conforme figura 7a, os maiores resultados foram obtidos no domínio jornalístico com os textos sem *stopwords*, acompanhados em alguns pequenos intervalos pelos resultados dos textos com *stopwords* sumarizados pelo *Copernic*. No domínio médico, como apresentado na figura 8a, os resultados foram bem próximos, variando apenas entre 0,12 e 0,13, mas na maioria dos resultados os textos sem *stopwords* sumarizados pelo *Copernic* e pelo *Sew Sum* obtiveram os resultados maiores.

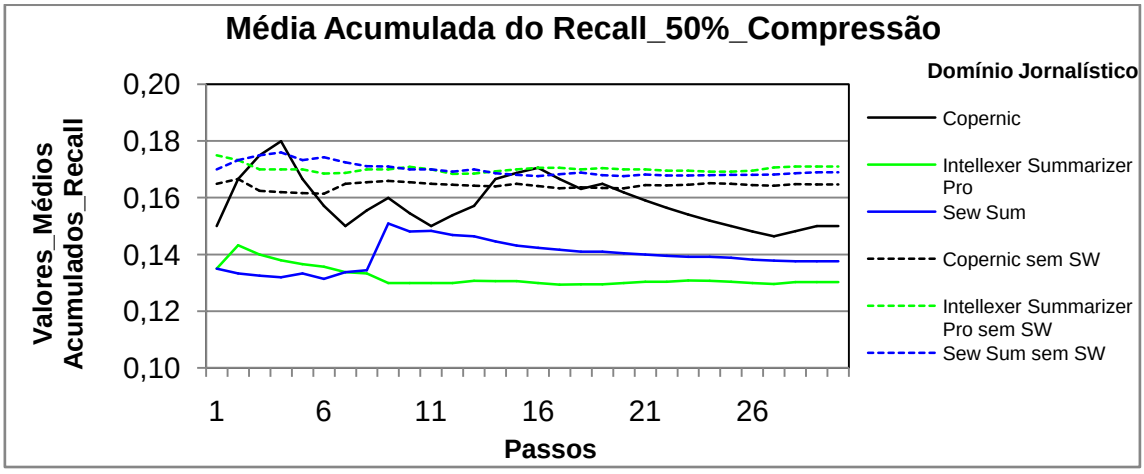


Figura 7a: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Recall* com 50% de compressão no idioma Inglês no domínio jornalístico.

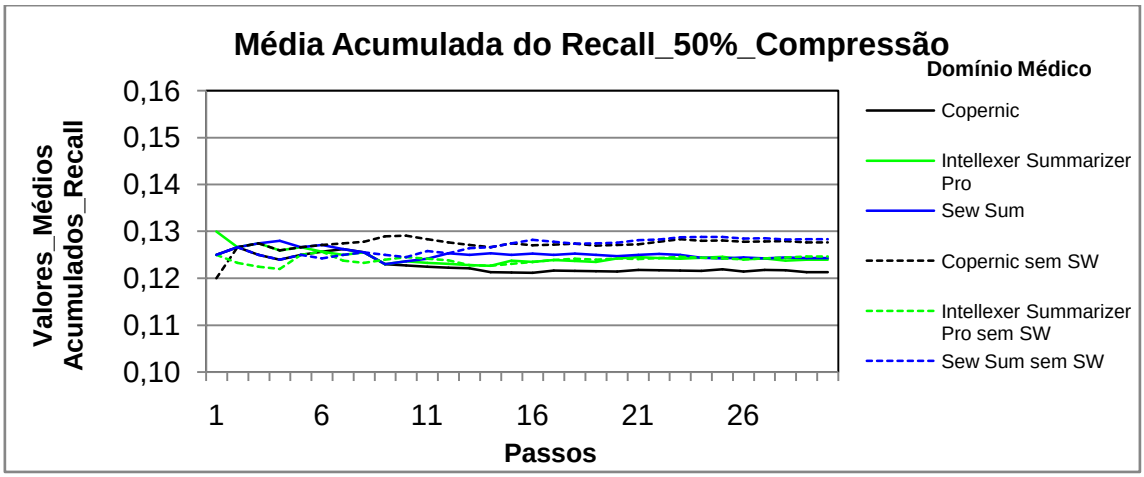


Figura 8a: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Recall* com 50% compressão no idioma Inglês no domínio médico.

Compressão de 70% no idioma inglês

Os maiores resultados em ambos os domínios foram obtidos pelos textos sem *stopwords*. No domínio jornalístico, conforme figura 9a, tais resultados tiveram a presença próxima dos resultados dos textos com *stopwords* sumarizados pelo *Copernic*. Já no domínio médico, assim como mostra figura 10a, os maiores resultados tiveram em alguns intervalos a presença dos resultados dos textos com *stopwords* sumarizados pelo *Sew Sum*.

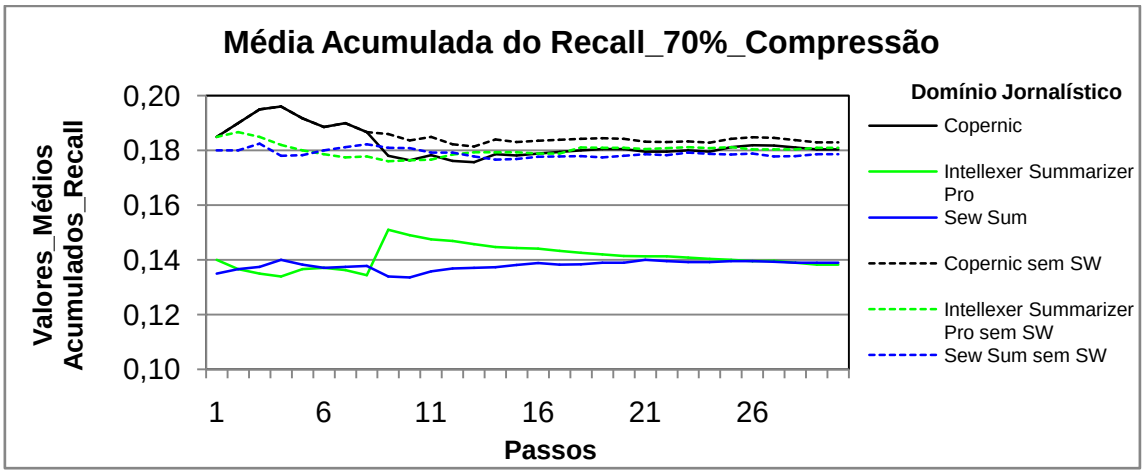


Figura 9a: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Recall* com 70% de compressão no idioma Inglês no domínio jornalístico.

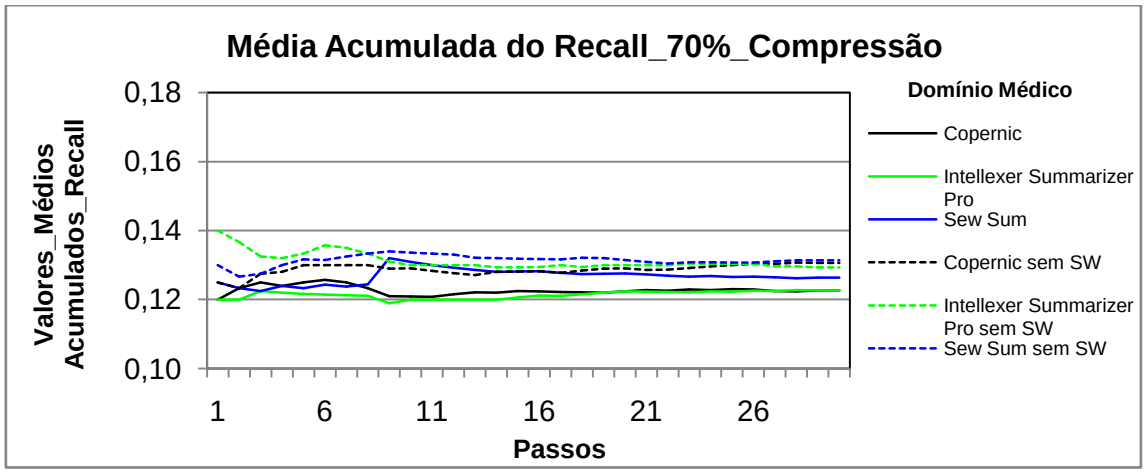


Figura 10a: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Recall* com 70% de compressão no idioma Inglês no domínio médico.

Compressão de 80% no idioma inglês

Tanto no domínio jornalístico quanto no médico, assim como demonstram respectivamente as figuras 11a e 12a, os maiores resultados obtidos foram com os textos sem *stopwords*.

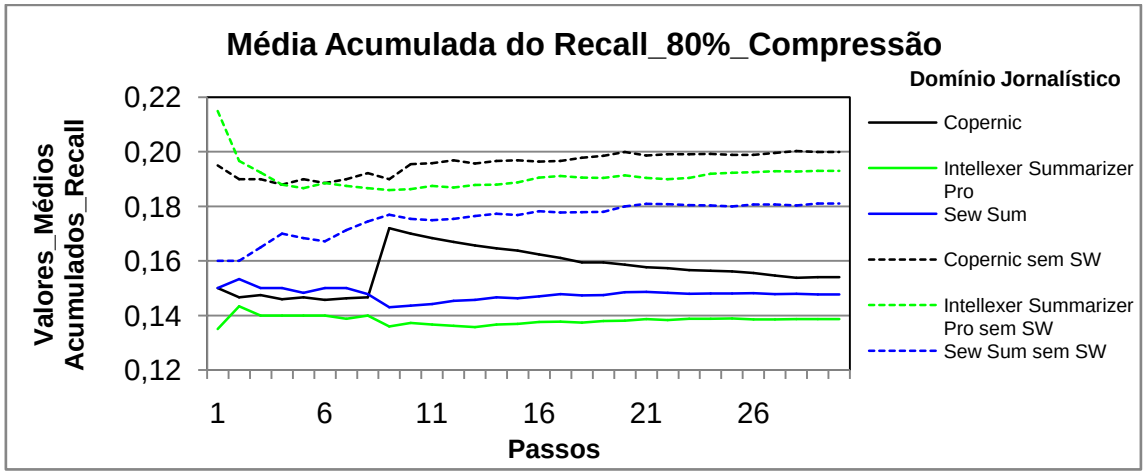


Figura 11a: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Recall* com 80% de compressão no idioma Inglês no domínio jornalístico.

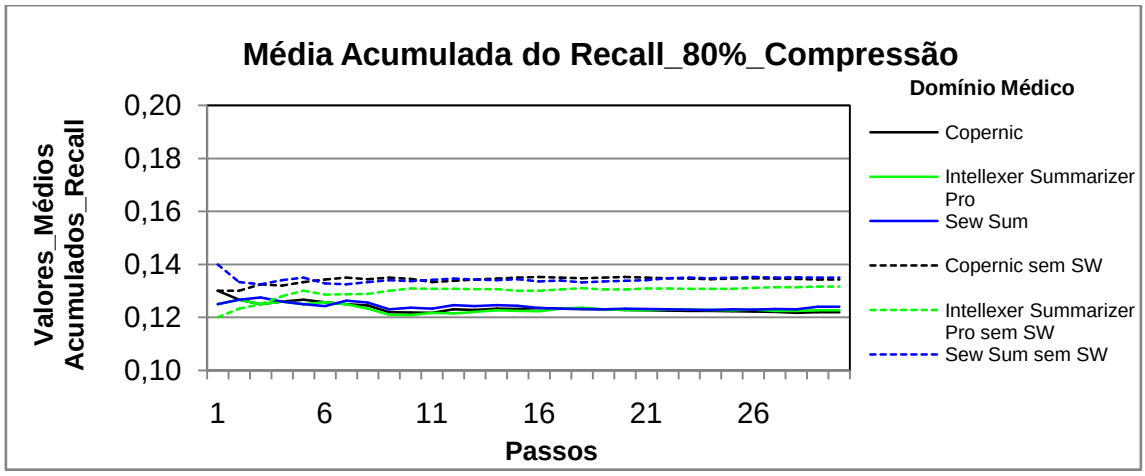


Figura 12a: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Recall* com 80% de compressão no idioma Inglês no domínio médico.

Compressão de 90% no idioma inglês

Tanto no domínio jornalístico quanto no médico, assim como mostram as figuras 13a e 14a, os maiores resultados obtidos foram com os textos sem *stopwords* na maioria das simulações. Sendo que apenas em um pequeno intervalo o resultado obtido com os textos com *stopwords* alcançou o resultado dos textos sem *stopwords*.

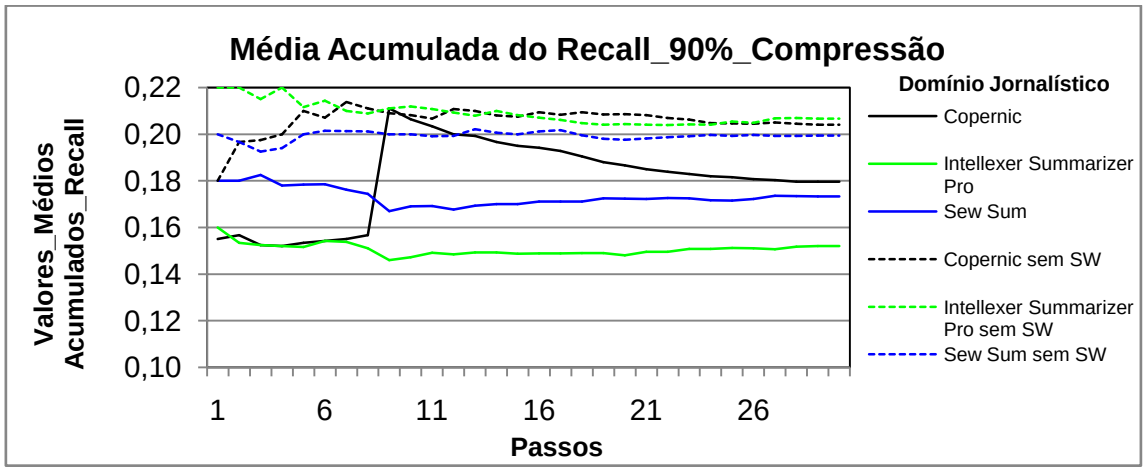


Figura 13a: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Recall* com 90% de compressão no idioma Inglês no domínio jornalístico.

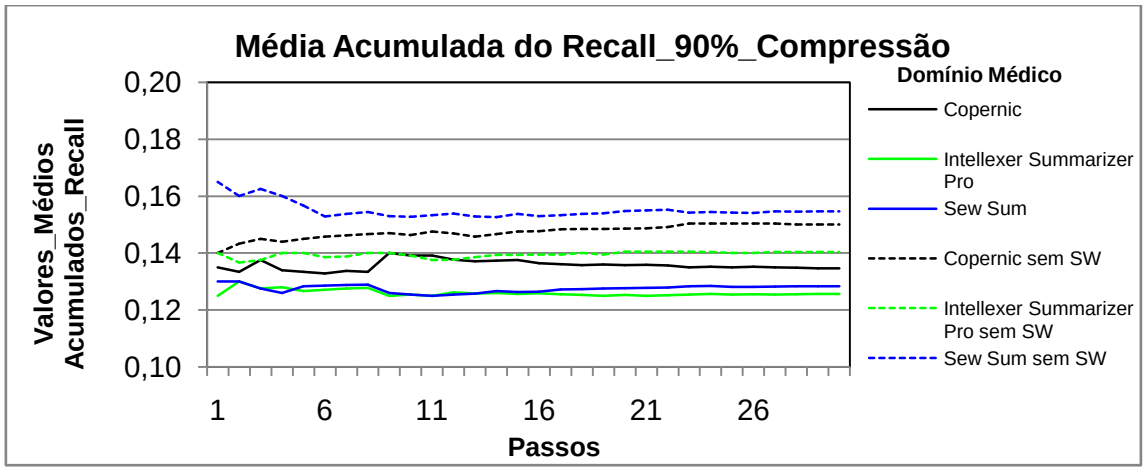


Figura 14a: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Recall* com 90% de compressão no idioma Inglês no domínio médico.

Média acumulada do precision

Compressão de 50% no idioma inglês

Tanto no domínio jornalístico quanto no médico, como apresentado respectivamente nas figuras 7b e 8b, os resultados obtidos com os textos com e sem *stopwords* foram bem próximos, porém os textos sem *stopwords* do domínio jornalístico mantiveram-se com resultados maiores a todo tempo, exceto nos resultados dos textos sumarizados pelo *Sew Sum*.

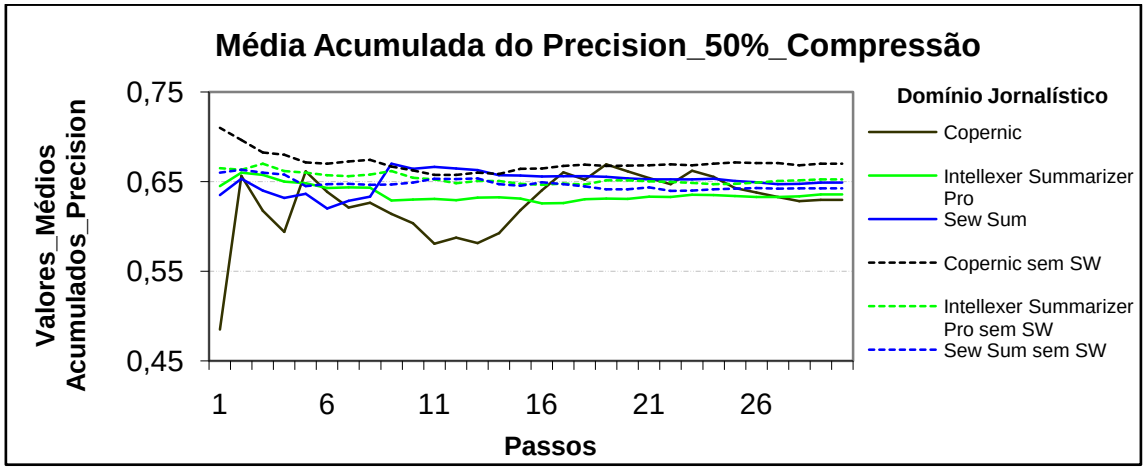


Figura 7b: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Precision* com 50% de compressão no idioma Inglês no domínio jornalístico.

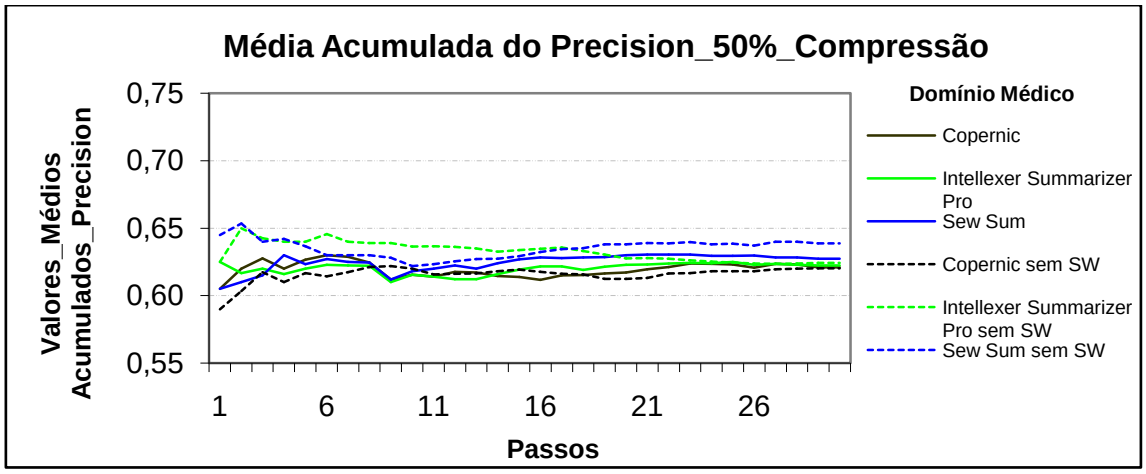


Figura 8b: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Precision* com 50% de compressão no idioma Inglês no domínio médico.

Compressão de 70% no idioma inglês

Tanto no domínio jornalístico quanto no médico, assim como mostram respectivamente as figuras 9b e 10b, os resultados obtidos com os textos com e sem *stopwords* foram bem próximos, sendo que no domínio médico os maiores resultados foram com os textos com *stopwords* do sumarizador *Sew Sum.*

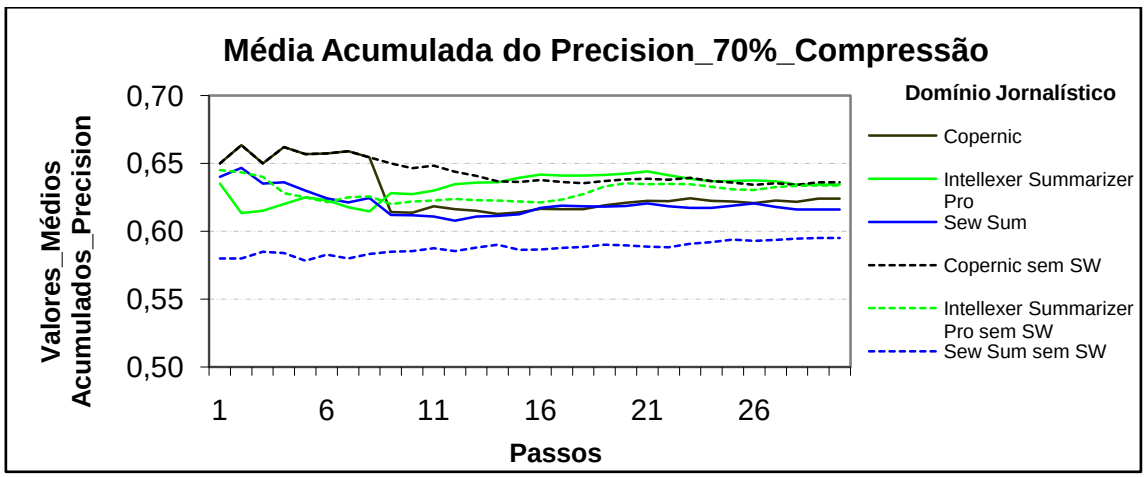


Figura 9b: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Precision* com 70% de compressão no idioma Inglês no domínio jornalístico.

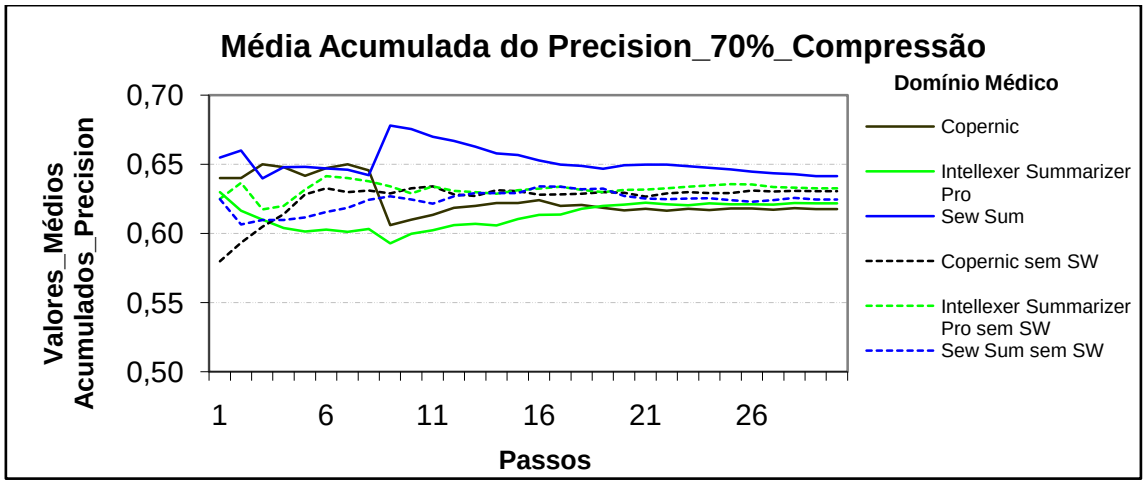


Figura 10b: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Precision* com 70% de compressão no idioma Inglês no domínio médico.

Compressão de 80% no idioma inglês

Tanto no domínio jornalístico quanto no médico, conforme apresentado respectivamente nas figuras 11b e 12b, os resultados obtidos com os textos com e sem *stopwords* foram bem próximos.

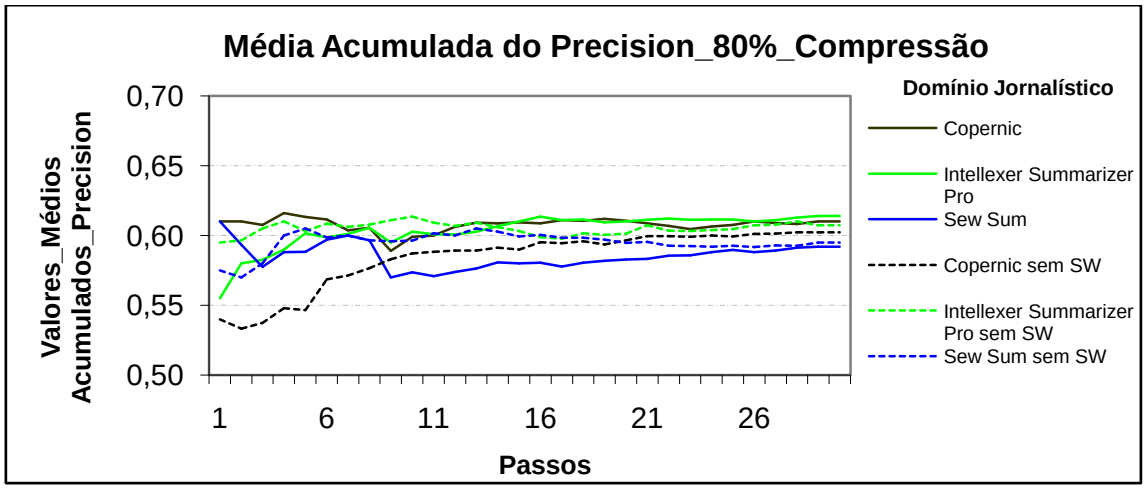


Figura 11b: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Precision* com 80% de compressão no idioma Inglês no domínio jornalístico.

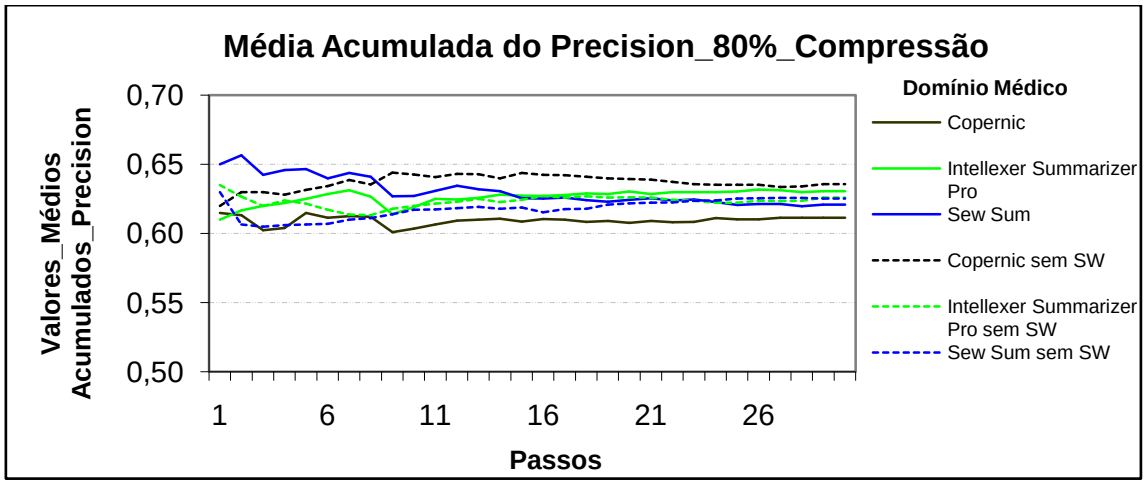


Figura 12b: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Precision* com 80% de compressão no idioma Inglês no domínio médico.

Compressão de 90% no idioma inglês

No domínio jornalístico, conforme figura 13b, os maiores resultados obtidos foram com os textos sem *stopwords* do sumarizador *Sew Sum*. No domínio médico, como mostra figura 14b, os resultados dos textos com *stopwords* foram mais significantes, seguidos dos resultados dos textos sem *stopwords* sumarizados pelo *Copernic*.

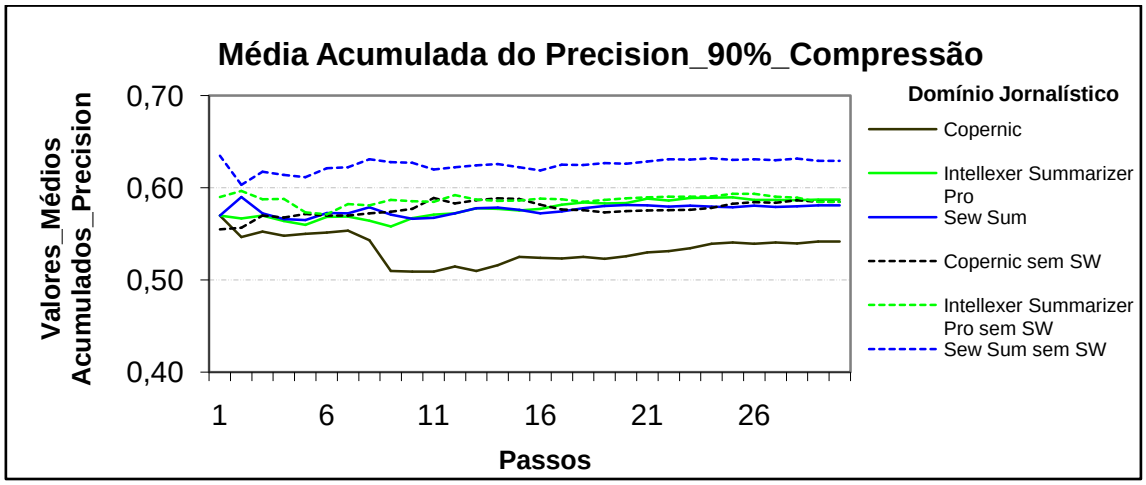


Figura 13b: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Precision* com 90% de compressão no idioma Inglês no domínio jornalístico.

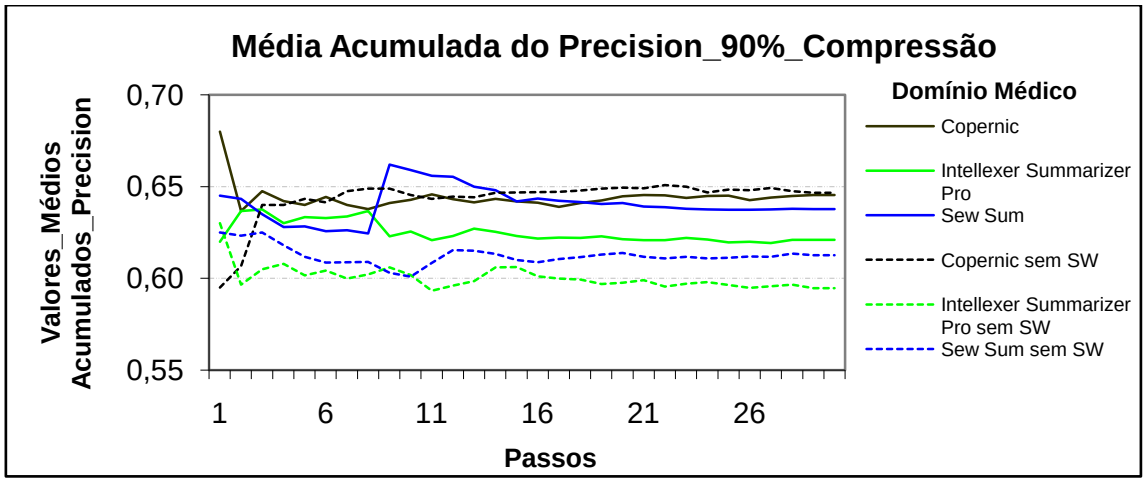


Figura 14b: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Precision* com 90% de compressão no idioma Inglês no domínio médico.

Média acumulada do recall

Compressão de 50% no idioma português

No domínio jornalístico, como apresentado na figura 15a, os maiores resultados obtidos foram com os textos sem *stopwords*. No domínio jurídico e no médico, assim como mostram respectivamente as figuras 16a e 17a, os resultados obtidos com os textos com e sem *stopswords* foram bem próximos, sendo que os textos sem *stopwords* do sumarizador *Gist Average Keyword* obtiveram maiores resultados na maioria das simulações.

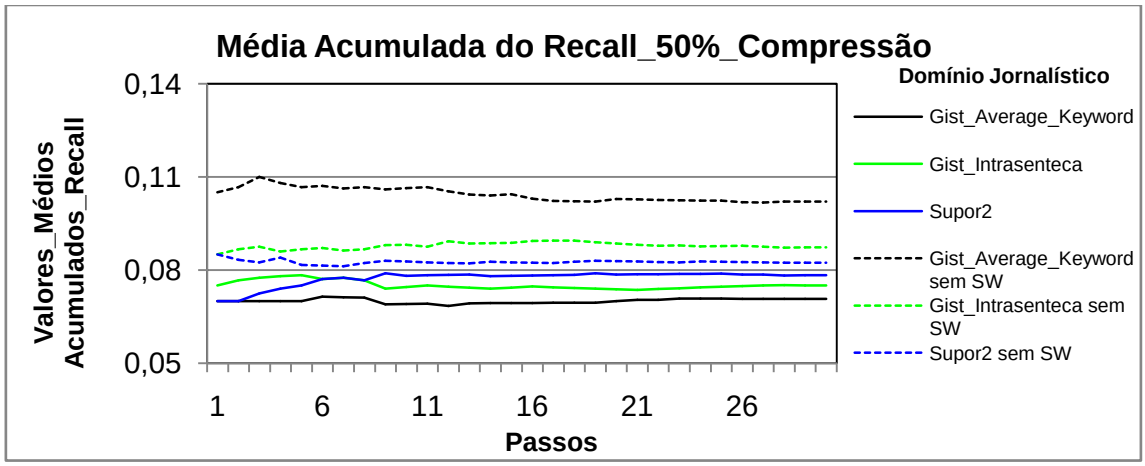


Figura 15a: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Recall* com 50% de compressão no idioma Português no domínio jornalístico.

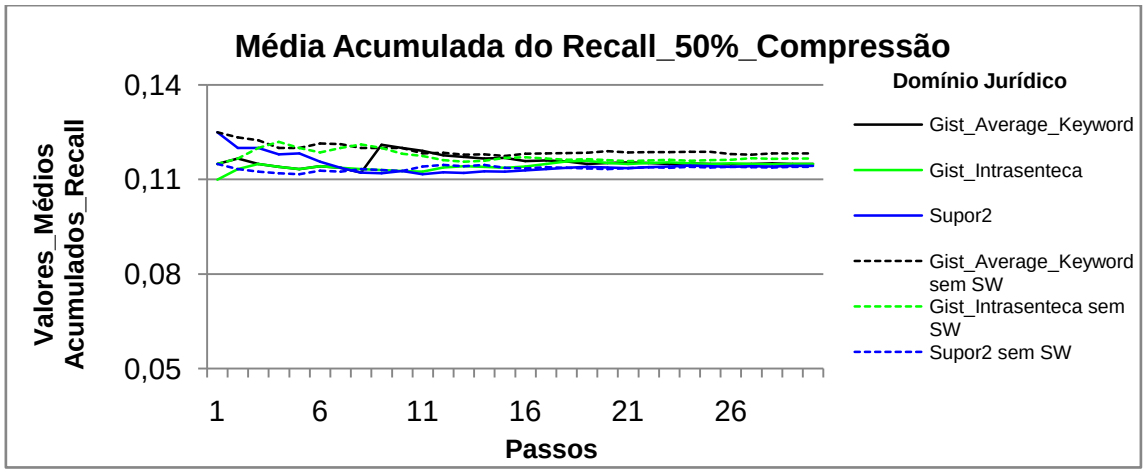


Figura 16a: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Recall* com 50% de compressão no idioma Português no domínio jurídico.

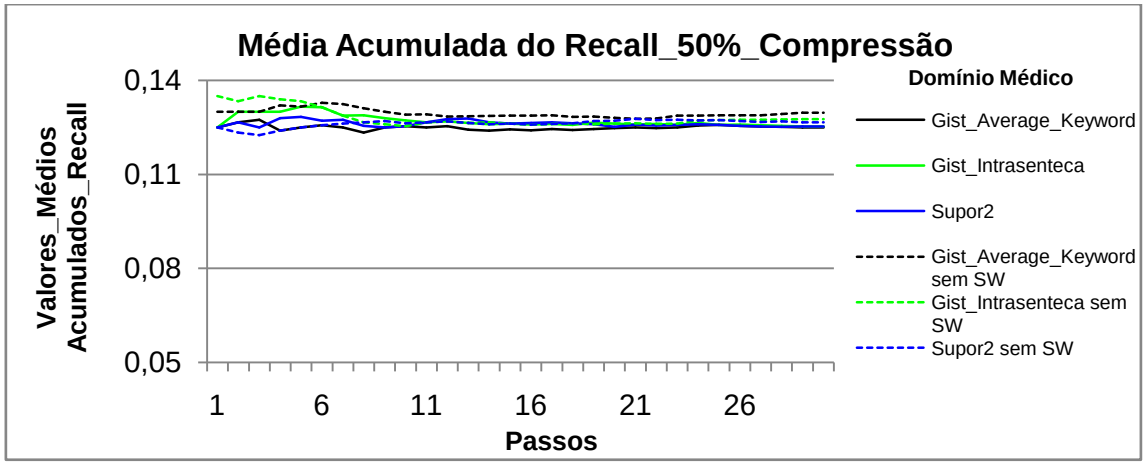


Figura 17a: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Recall* com 50% de compressão no idioma Português no domínio médico.

Compressão de 70% no idioma português

No domínio jornalístico, conforme figura 18a, os maiores resultados obtidos foram com os textos sem stopwords dos sumarizadores Gist Intrasenteca e Supor2. No domínio jurídico e no médico, como apresentado respectivamente nas figuras 19a e 20a, os maiores resultados foram obtidos com os textos sem stopwords na maioria das simulações.

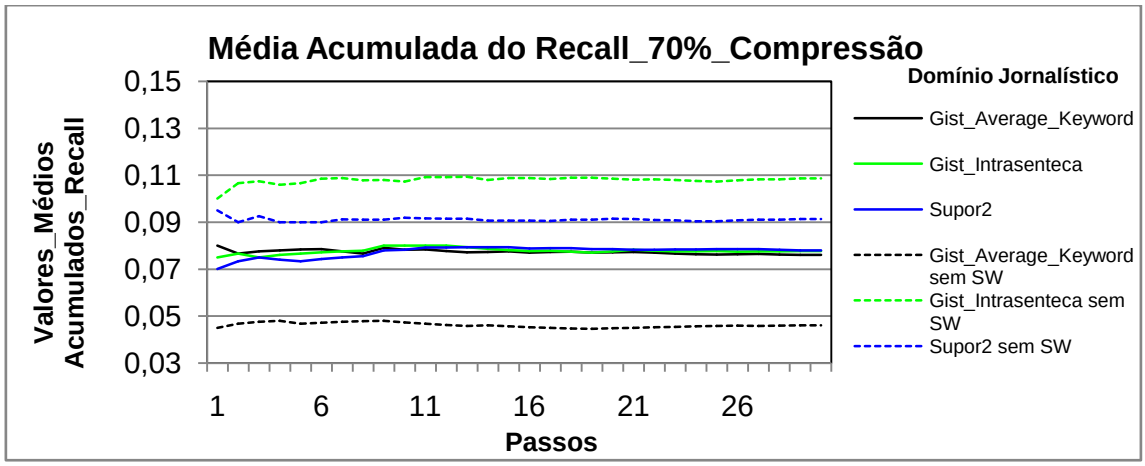


Figura 18a: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Recall* com 70% de compressão no idioma Português no domínio jornalístico.

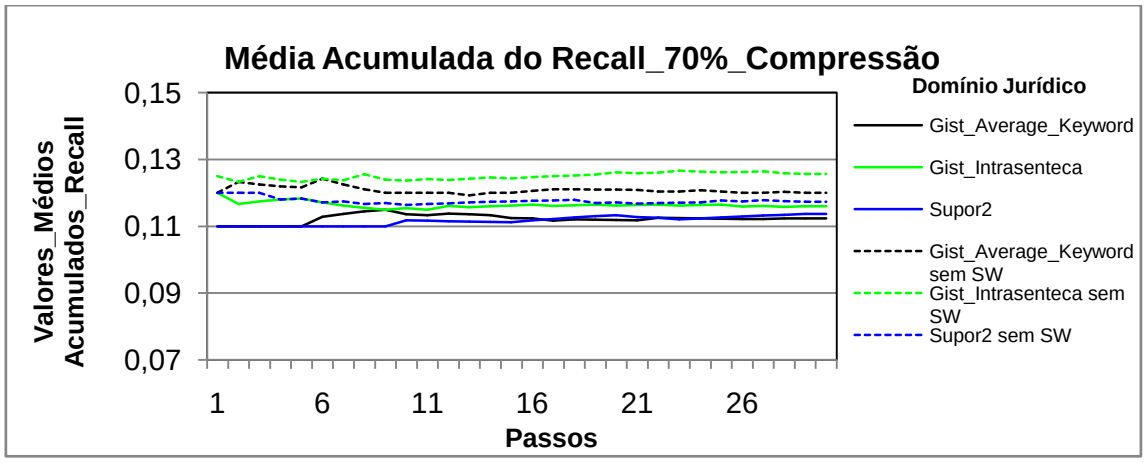


Figura 19a: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Recall* com 70% de compressão no idioma Português no domínio jurídico.

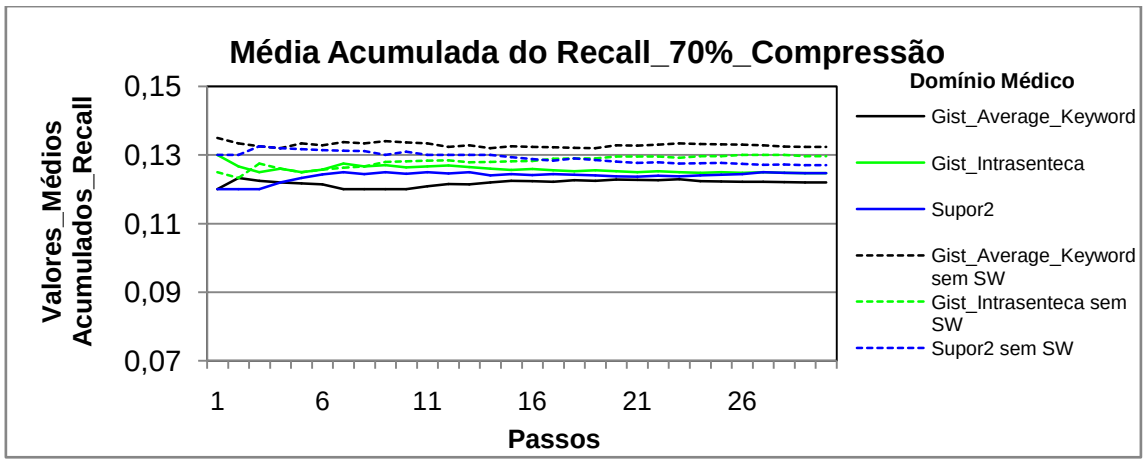


Figura 20a: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Recall* com 70% de compressão no idioma Português no domínio médico.

Compressão de 80% no idioma português

No domínio jornalístico, como apresentado na figura 21a, os maiores resultados obtidos foram com os textos sem *stopwords* dos sumarizadores *Gist Intrasenteca* e Supor2. No domínio jurídico e no médico, assim como mostram respectivamente as figuras 22a e 23a, os maiores resultados foram obtidos com os textos sem *stopwords* na maioria das simulações.

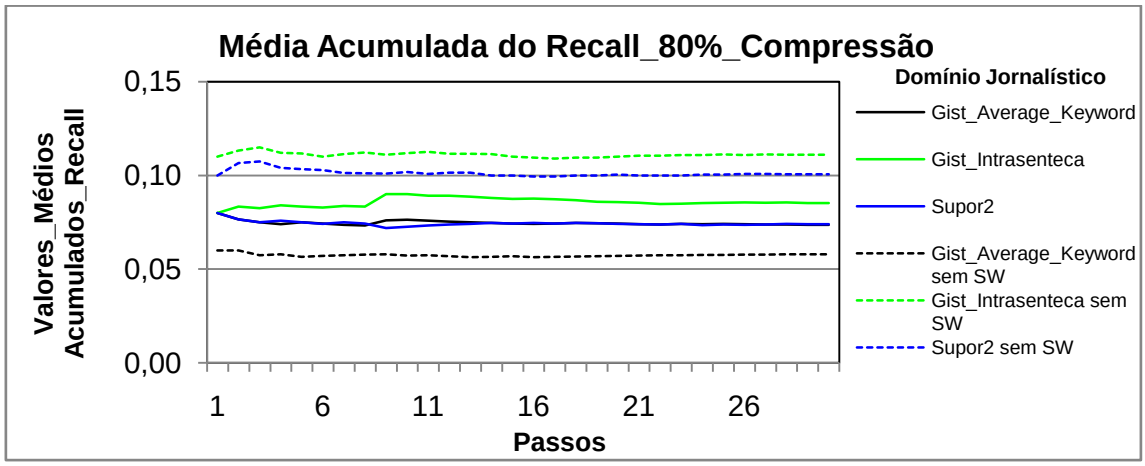


Figura 21a: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Recall* com 80% de compressão no idioma Português no domínio jornalístico.

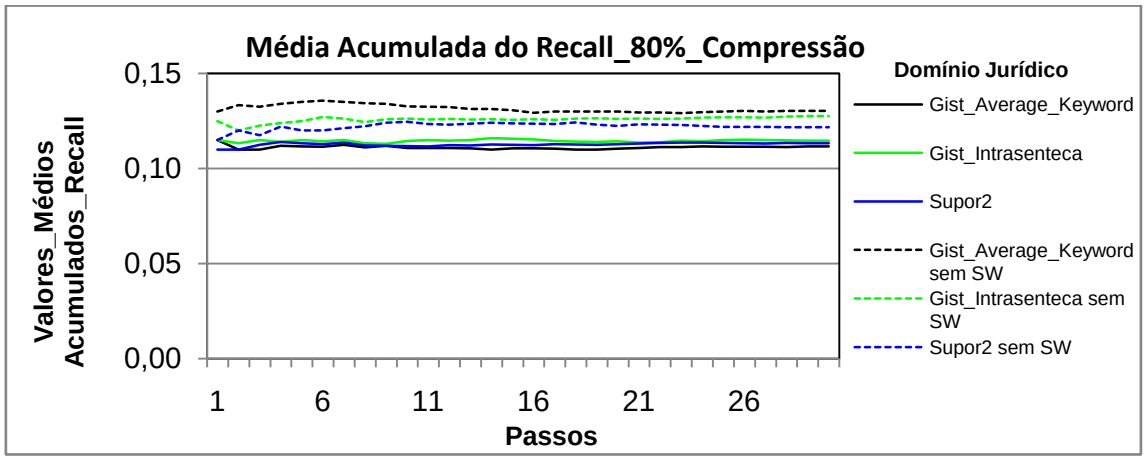


Figura 22a: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Recall* com 80% de compressão no idioma Português no domínio jurídico.

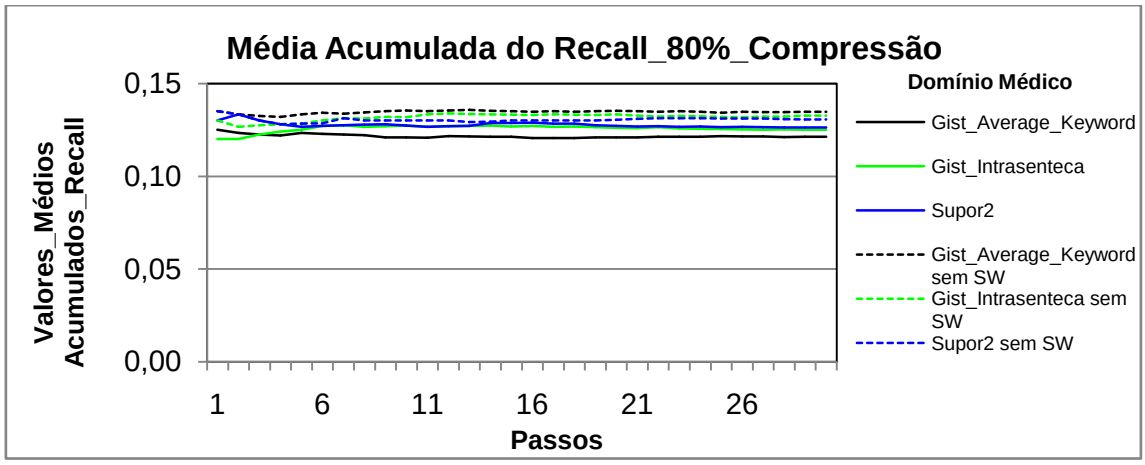


Figura 23a: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Recall* com 80% de compressão no idioma Português no domínio médico.

Compressão de 90% no idioma português

No domínio jornalístico, conforme figura 24a, os maiores resultados obtidos foram com os textos sem *stopwords* dos sumarizadores *Gist Intrasenteca* e *Gist Average Keyword*. No domínio jurídico e no médico, como apresentado nas figuras 25a e 26a, os maiores resultados foram obtidos com os textos sem *stopwords*.

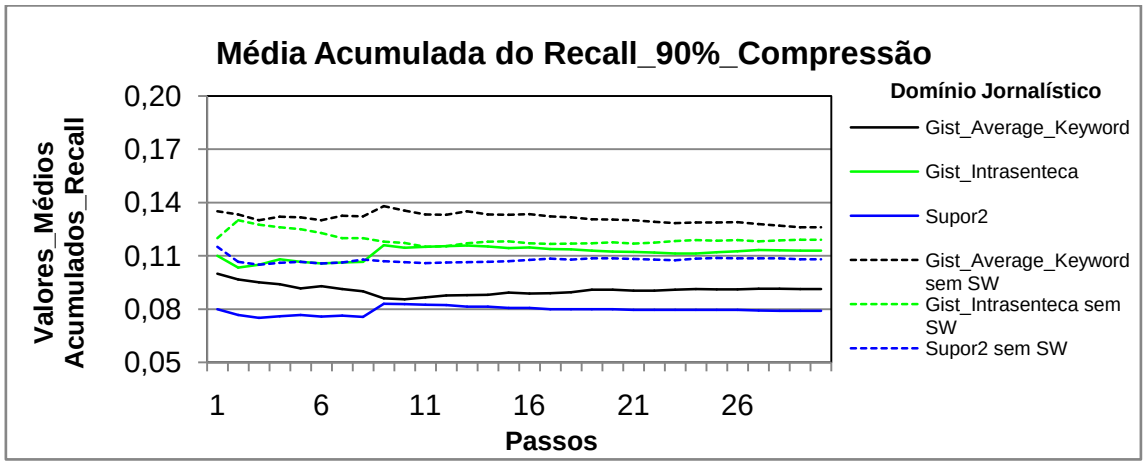


Figura 24a: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Recall* com 90% de compressão no idioma Português no domínio jornalístico.

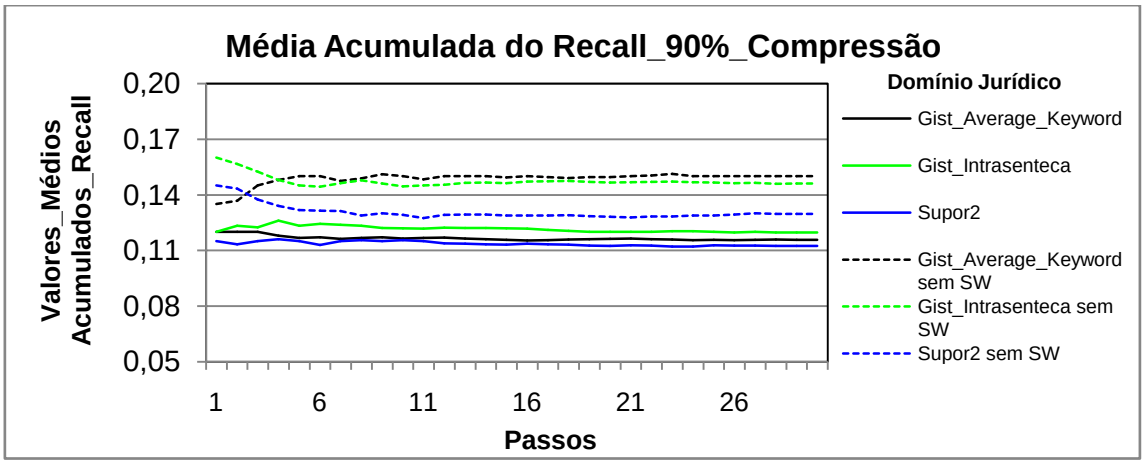


Figura 25a: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Recall* com 90% de compressão no idioma Português no domínio jurídico.

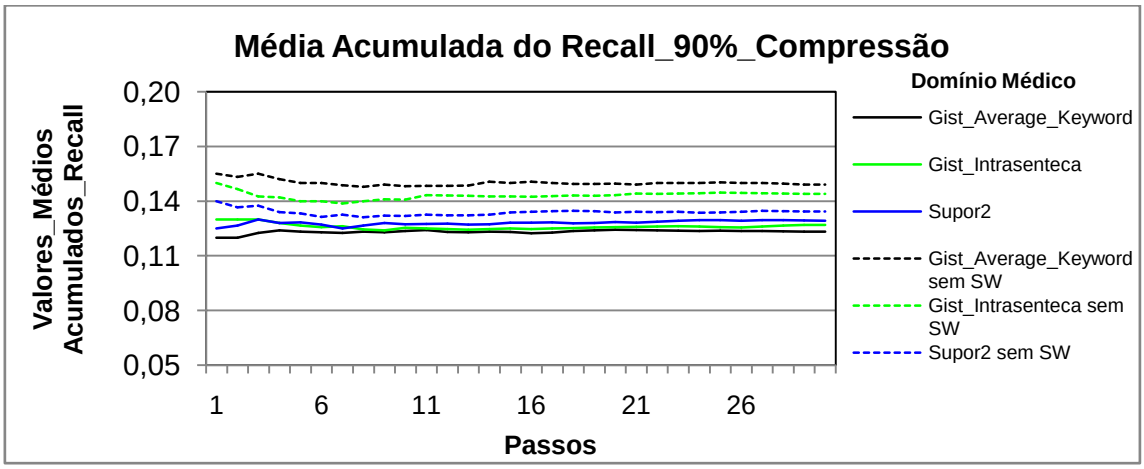


Figura 26a: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Recall* com 90% de compressão no idioma Português no domínio médico.

Média acumulada precision

Compressão de 50% no idioma português

No domínio jornalístico, conforme figura 15b, os maiores resultados obtidos foram com os textos sem *stopwords* dos sumarizadores *Gist Intrasenteca* e *Supor2* na maioria das simulações. No domínio jurídico, como mostra figura 16b, os maiores resultados obtidos foram com os textos com *stopwords* dos sumarizadores *Gist Average Keyword* e *Supor2* na maioria das simulações. No domínio médico, como demonstra figura 17b, os resultados obtidos com os textos com e sem *stopwords* foram bem próximos.

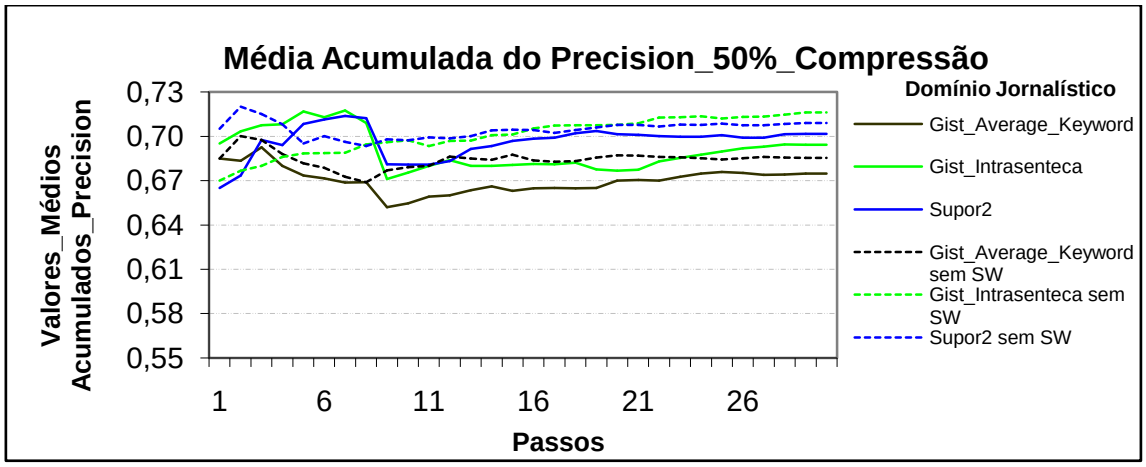


Figura 15b: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Precision* com 50% de compressão no idioma Português no domínio jornalístico.

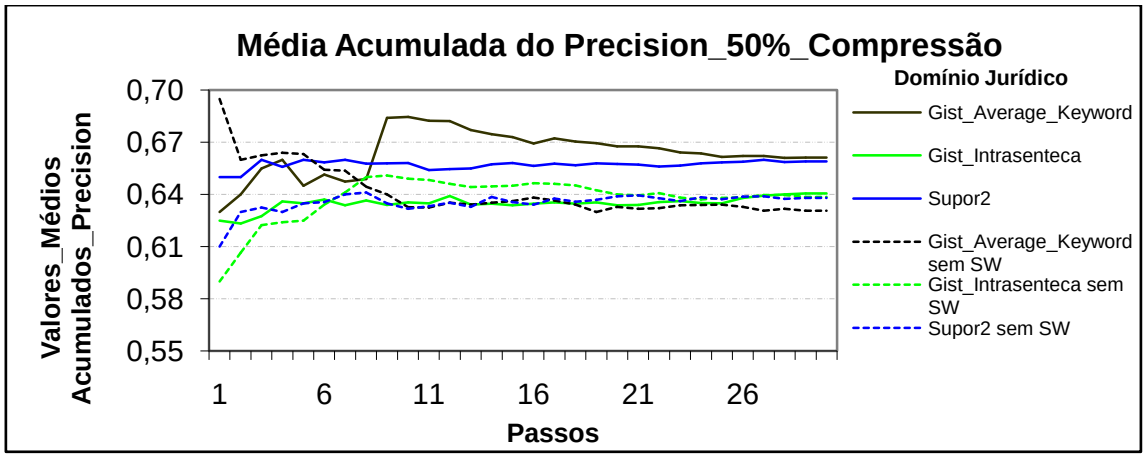


Figura 16b: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Precision* com 50% de compressão no idioma Português no domínio jurídico.

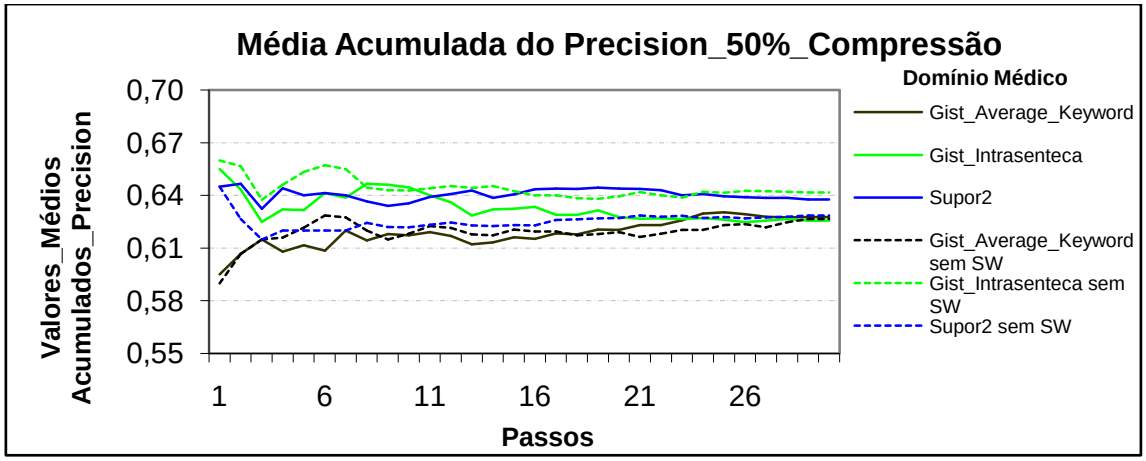


Figura 17b: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Precision* com 50% de compressão no idioma Português no domínio médico.

Compressão de 70% no idioma português

No domínio jornalístico, assim como mostra figura 18b, os maiores resultados obtidos foram com os textos sem *stopwords* do sumarizador *Gist Average Keyword*. No domínio jurídico, conforme figura 19b, os resultados obtidos com os textos com e sem *stopwords* foram bem próximos. No domínio médico, como apresentado figura 20b, os maiores resultados obtidos foram com os textos com *stopwords* dos sumarizadores *Gist Average Keyword* e *Gist Intrasenteca* na maioria dos resultados.

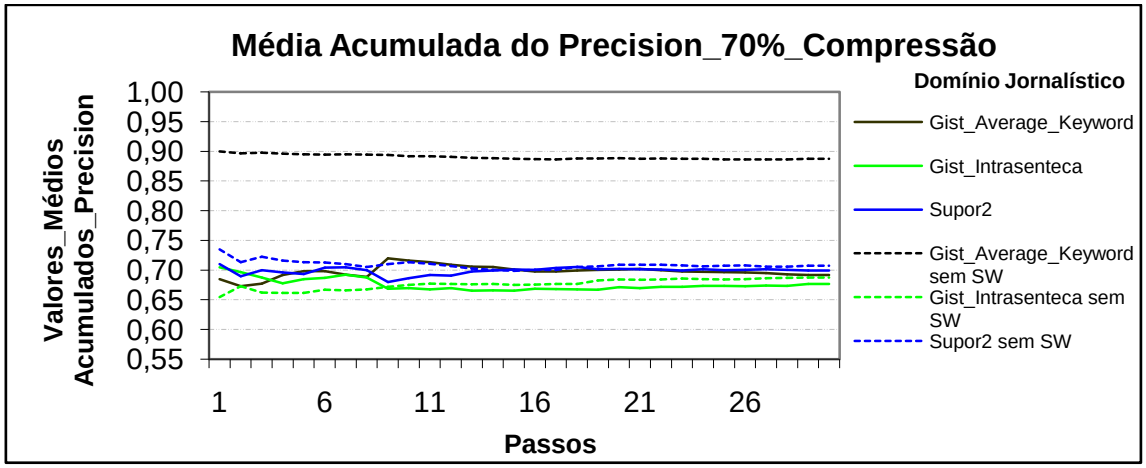


Figura 18b: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Precision* com 70% de compressão no idioma Português no domínio jornalístico.

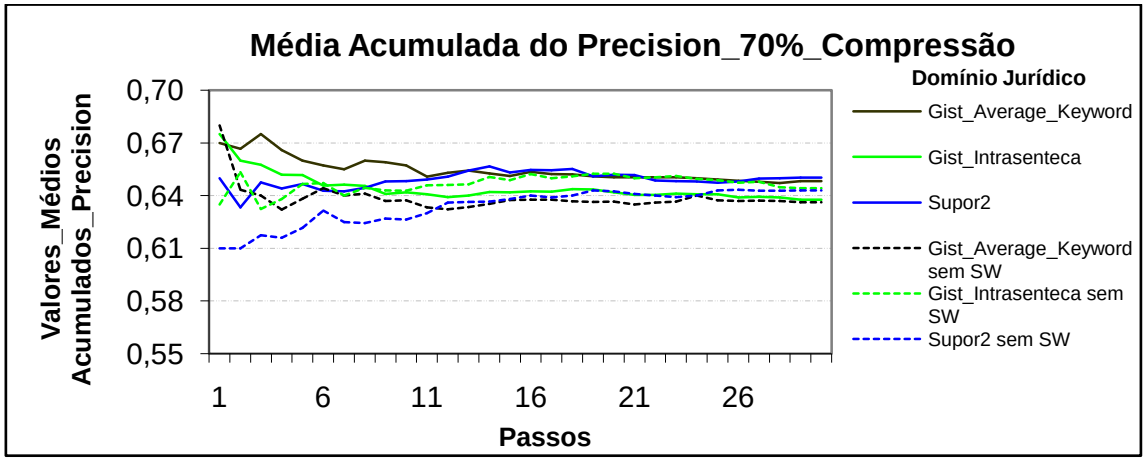


Figura 19b: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Precision* com 70% de compressão no idioma Português no domínio jurídico.

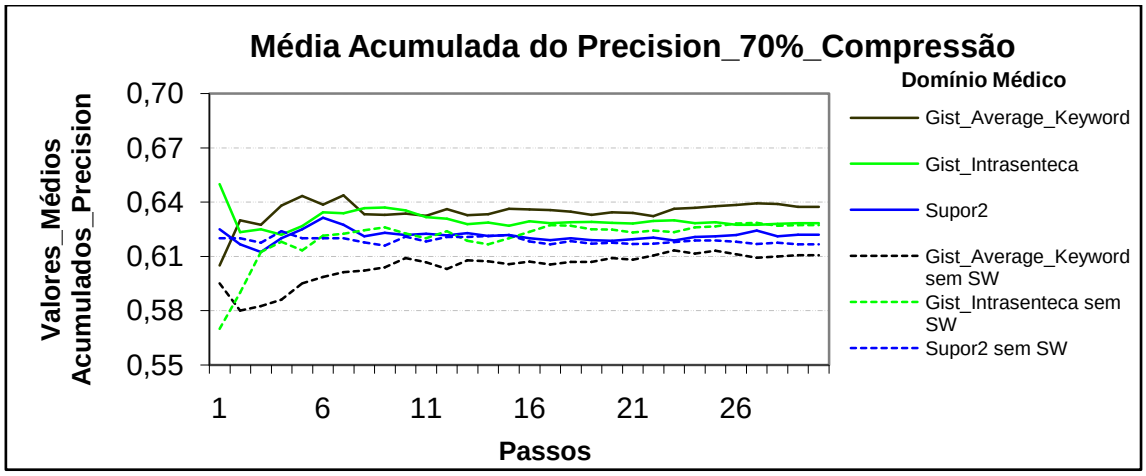


Figura 20b: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Precision* com 70% de compressão no idioma Português no domínio médico.

Compressão de 80% no idioma português

No domínio jornalístico, como apresentado na figura 21b, os maiores resultados obtidos foram com os textos sem *stopwords* do sumarizador *Supor2*. No domínio jurídico, como demonstra figura 22b, os resultados obtidos com os textos com e sem *stopwords* foram bem próximos. No domínio médico, conforme figura 23b, os maiores resultados obtidos foram com os textos com *stopwords* do sumarizador *Supor2*.

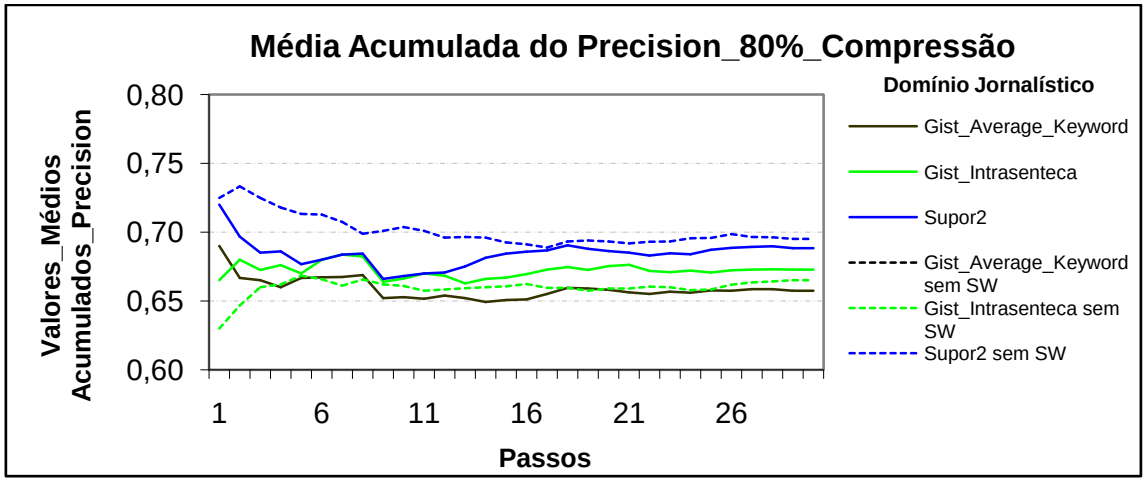


Figura 21b: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Precision* com 80% de compressão no idioma Português no domínio jornalístico.

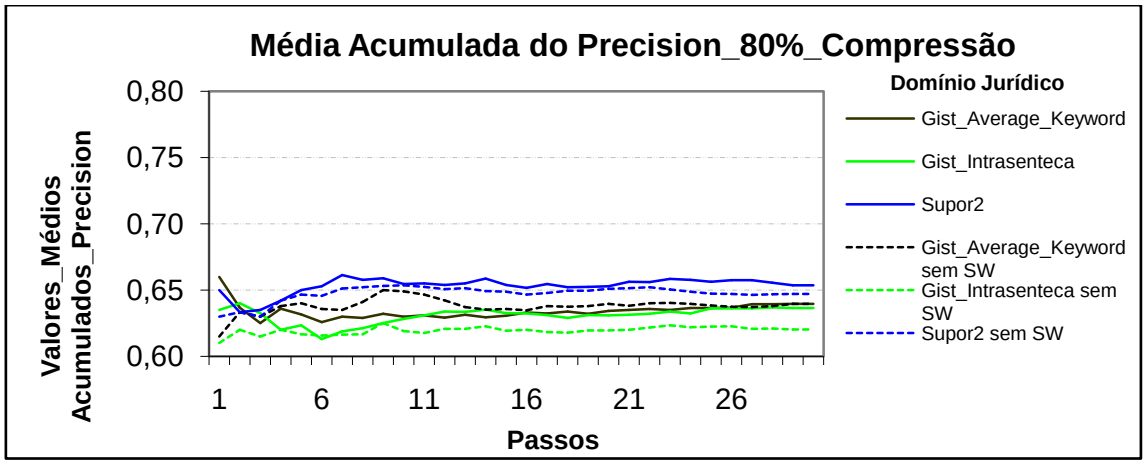


Figura 22b: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Precision* com 80% de compressão no idioma Português no domínio jurídico.

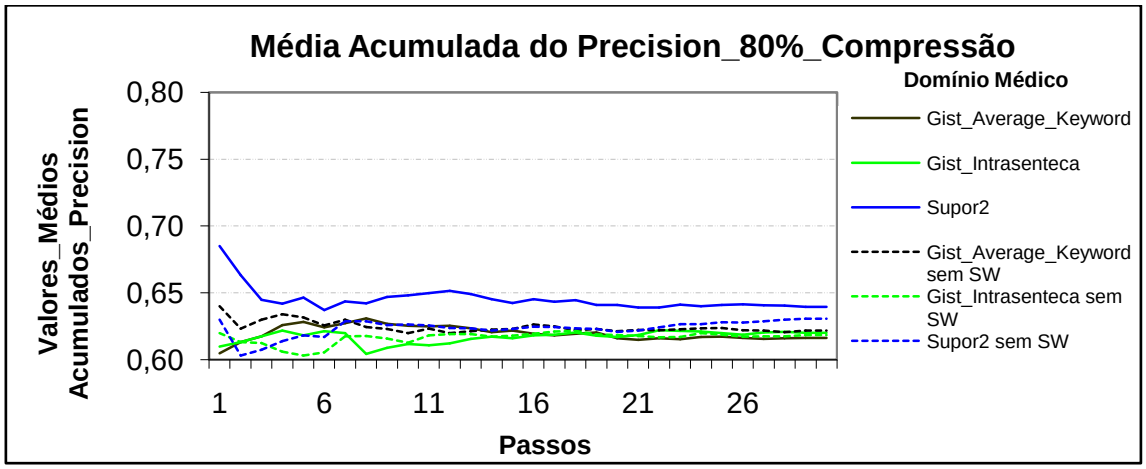


Figura 23b: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Precision* com 80% de compressão no idioma Português no domínio médico.

Compressão de 90% no idioma português

Nos três domínios: jornalístico, jurídico e médico, como apresentado respectivamente nas figuras 24b, 25b e 26b, os maiores resultados obtidos foram com os textos com *stopwords* do sumarizador Supor2.

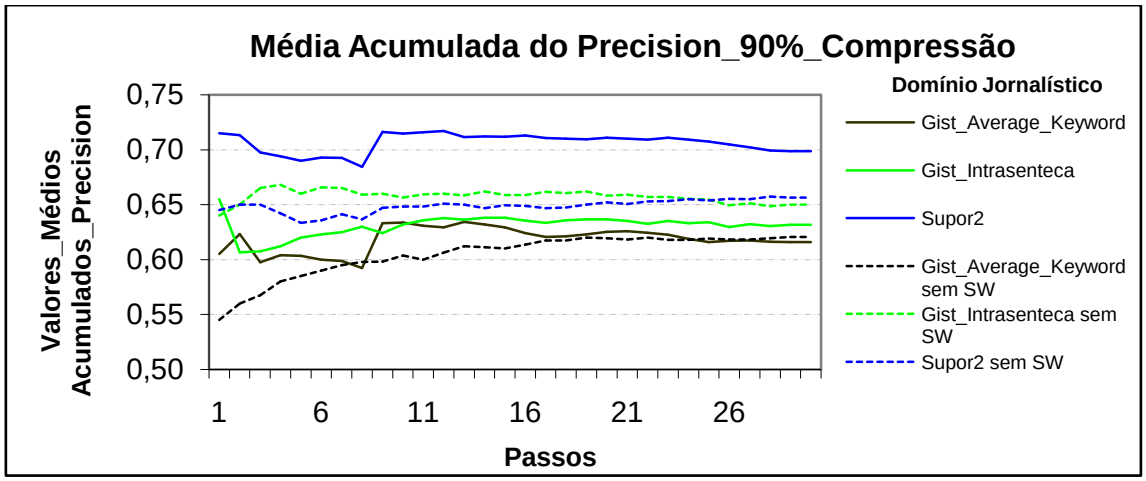


Figura 24b: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Precision* com 90% de compressão no idioma Português no domínio jornalístico.

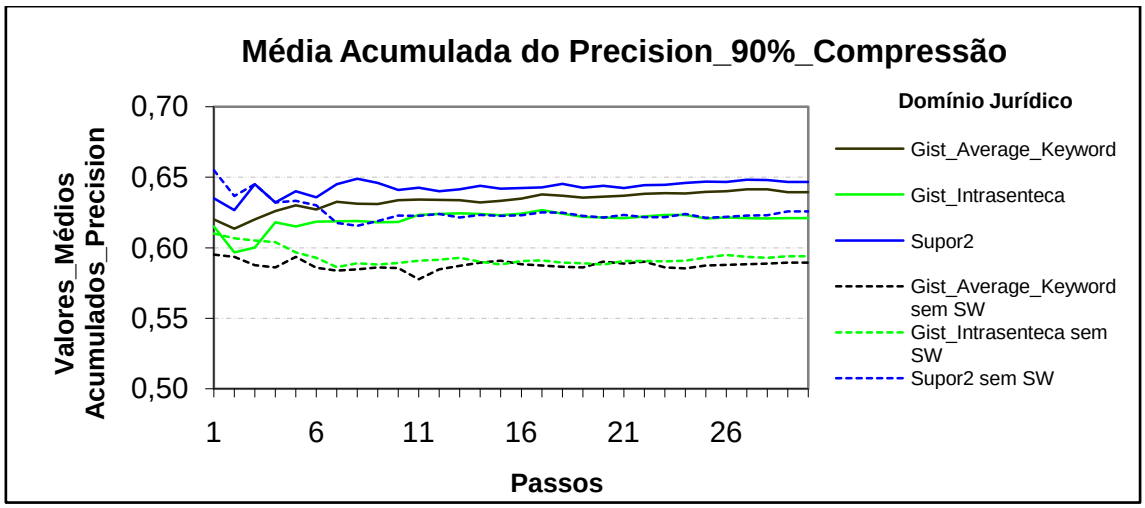


Figura 25b: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Precision* com 90% de compressão no idioma Português no domínio jurídico.

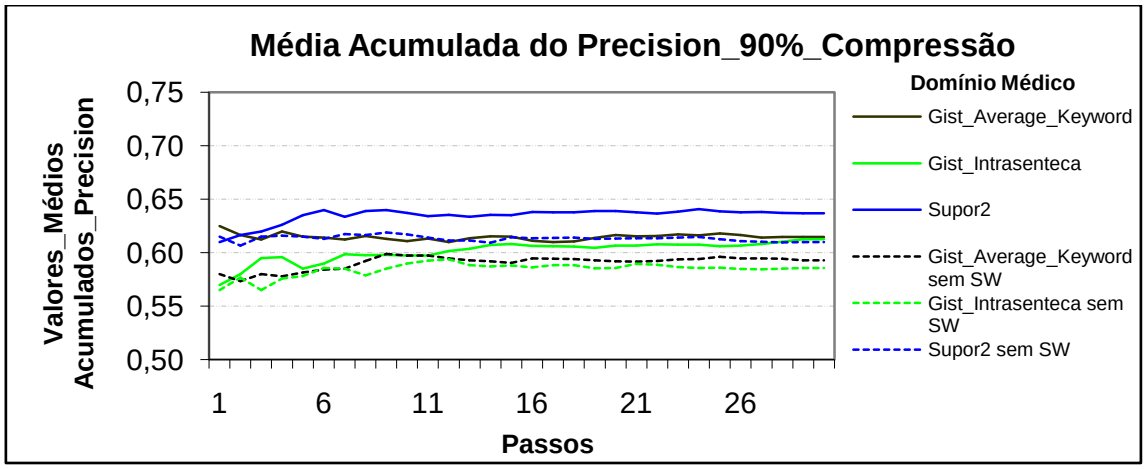


Figura 26b: Resultados obtidos pelo modelo Cassiopeia, usando a medida *Precision* com 90% de compressão no idioma Português no domínio médico.

Apêndice B - Métricas internas

O Apêndice B mostra a continuidade dos resultados descritos na subseção 4.1.2, onde foi apresentada a medida *Coeficiente Silhouette*. As medidas Coesão e Acoplamento, 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 4.1.2. 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 *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..

Média acumulada da coesão

Compressão de 50% no idioma inglês

No domínio jornalístico, conforme figura 27a, os maiores resultados foram obtidos com os textos sem *stopwords* do sumarizador Intellexer Summarizer Pro, sendo que os demais resultados sem *stopwords* também se destacaram na maioria das simulações. No domínio médico, assim como mostra figura 28a, os resultados dos textos com e sem *stopwords* foram bem próximos, com os menores resultados obtidos com os textos com *stopwords* do sumarizador *Sew Sum*.

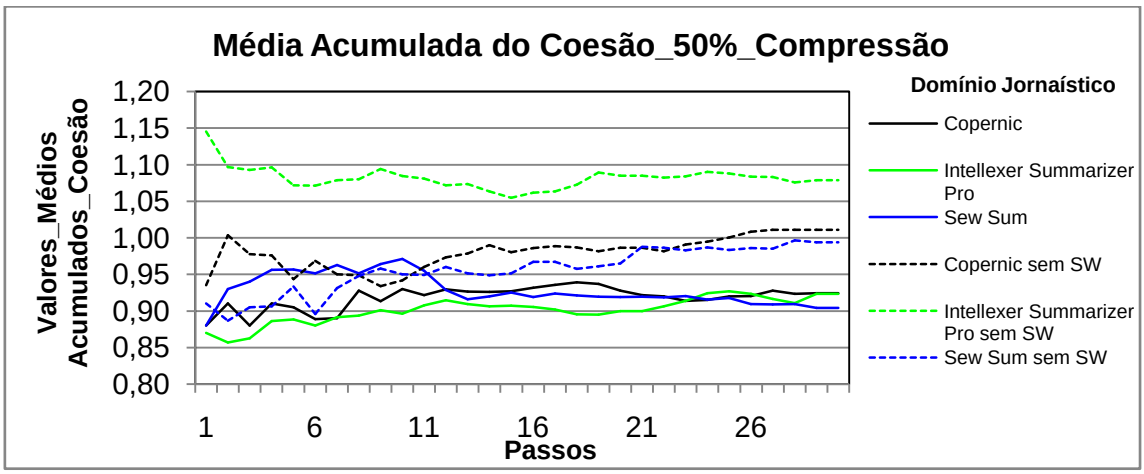


Figura 27a: Resultados obtidos pelo modelo Cassiopeia, usando a medida Coesão com 50% de compressão no idioma Inglês no domínio jornalístico.

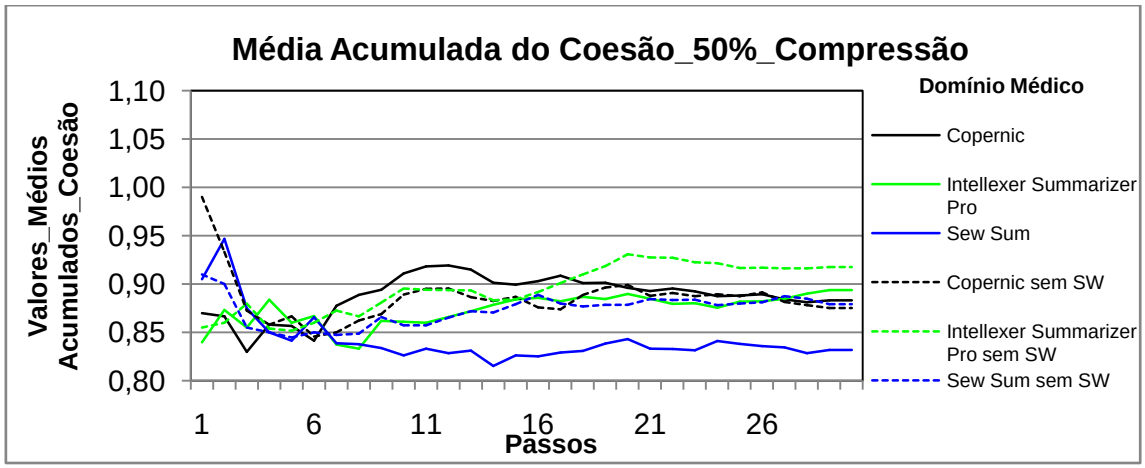


Figura 28a: Resultados obtidos pelo modelo Cassiopeia, usando a medida Coesão com 50% de compressão no idioma Inglês no domínio médico.

Compressão de 70% no idioma inglês

No domínio jornalístico, conforme figura 29a, os resultados obtidos com os textos sem *stopwords* foram os maiores. No domínio médico, como mostra figura 30a, os resultados obtidos foram bem próximos.

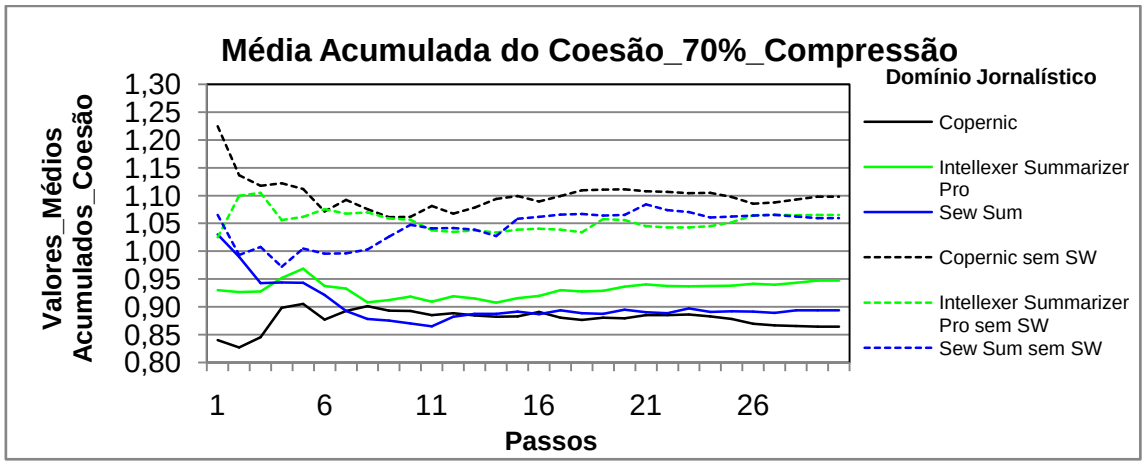


Figura 29a: Resultados obtidos pelo modelo Cassiopeia, usando a medida Coesão com 70% de compressão no idioma Inglês no domínio jornalístico.

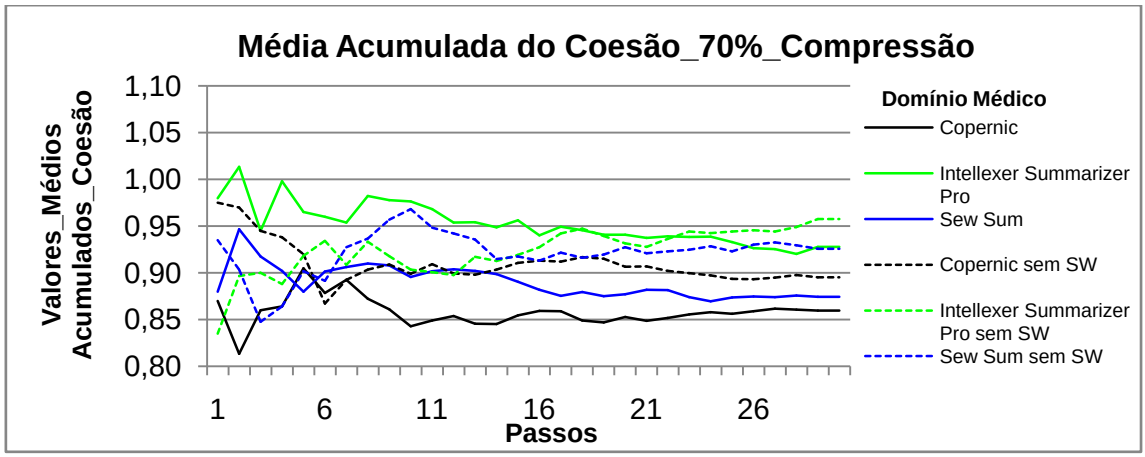


Figura 30a: Resultados obtidos pelo modelo Cassiopeia, usando a medida Coesão com 70% de compressão no idioma Inglês no domínio médico.

Compressão de 80% no idioma inglês

Tanto no domínio jornalístico quanto no médico, assim como apresentado nas figuras 31a e 32a, os resultados obtidos pelos textos sem *stopwords* com os sumarizadores *Intellexer Summarizer Pro* e *Sew Sum* foram os maiores do que os resultados obtidos com os textos sem *stopwords, sendo que os resultados dos textos sem stopwords do sumarizador Copernic também teve destaques.*

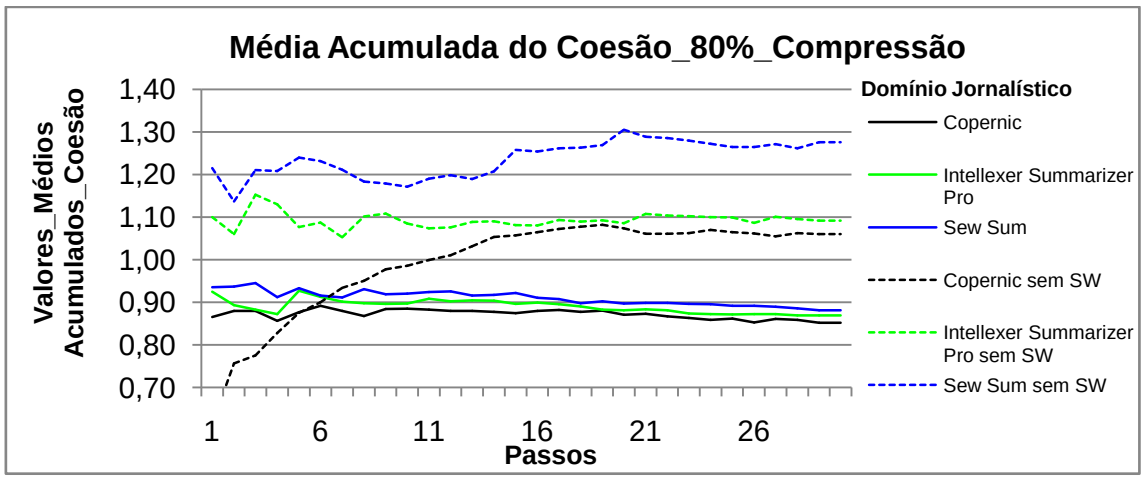


Figura 31a: Resultados obtidos pelo modelo Cassiopeia, usando a medida Coesão com 80% de compressão no idioma Inglês no domínio jornalístico.

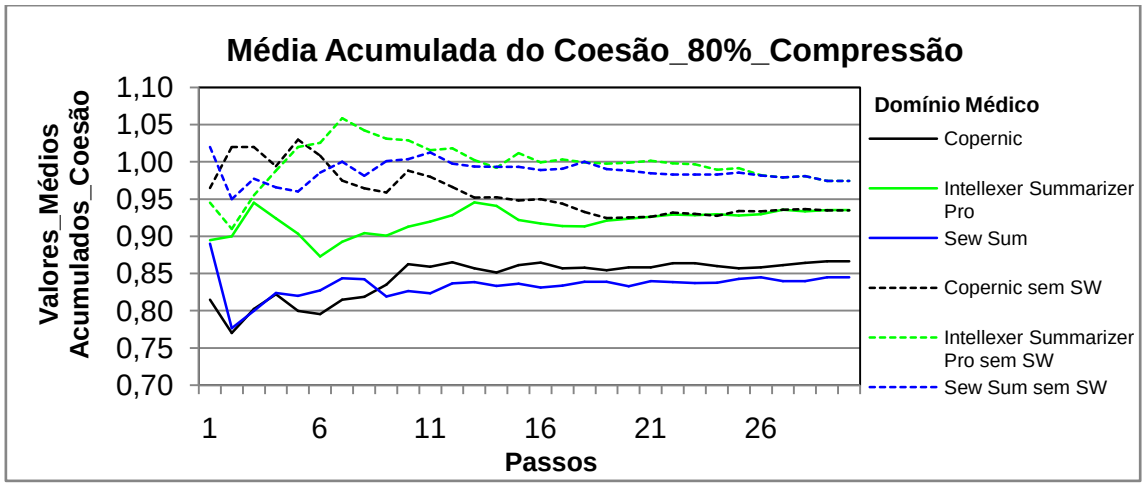


Figura 32a: Resultados obtidos pelo modelo Cassiopeia, usando a medida Coesão com 80% de compressão no idioma Inglês no domínio médico.

Compressão de 90% no idioma inglês

No domínio jornalístico, como demonstra a figura 33a, os maiores resultados foram obtidos com os textos sem *stopwords* dos sumarizadores *Copernic* e *Sew Sum*. No domínio médico, conforme figura 34a, os maiores resultados foram obtidos com os textos sem *stopwords* do sumarizador *Copernic*.

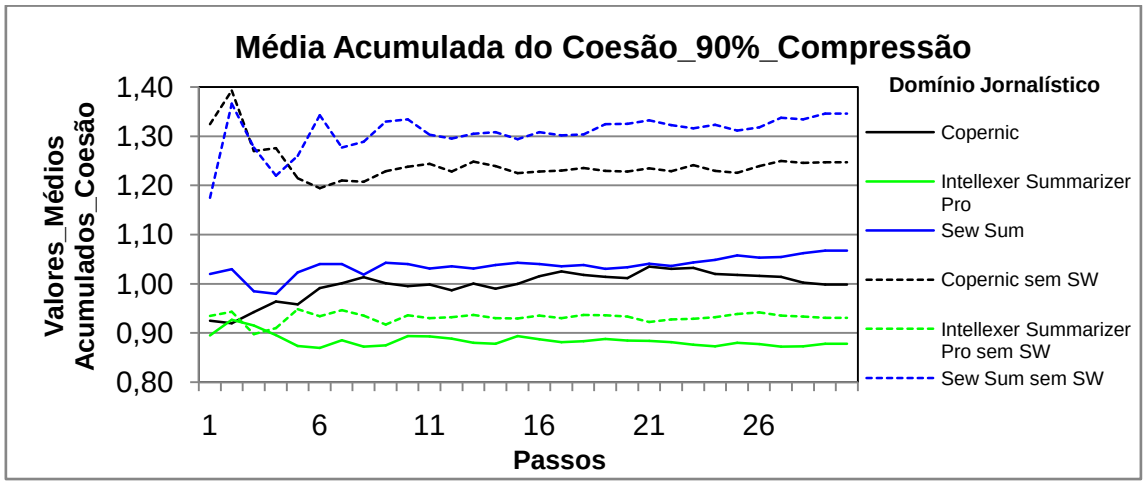


Figura 33a: Resultados obtidos pelo modelo Cassiopeia, usando a medida Coesão com 90% de compressão no idioma Inglês no domínio jornalístico.

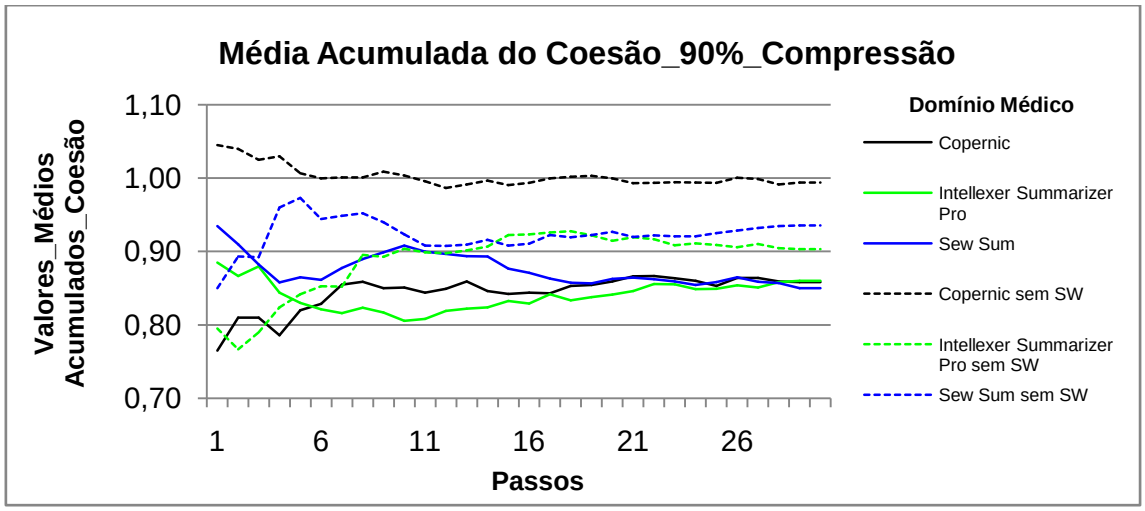


Figura 34a: Resultados obtidos pelo modelo Cassiopeia, usando a medida Coesão com 90% de compressão no idioma Inglês no domínio médico.

Média acumulada do acoplamento

Compressão de 50% no idioma inglês

No domínio jornalístico, conforme figura 27b, os resultados obtidos com os textos com e sem *stopwords* foram próximos, exceto os resultados do sumarizador *Copernic*, esses foram menores. No domínio médico, como mostra figura 28b, os resultados obtidos com os textos com *stopwords* foram os maiores.

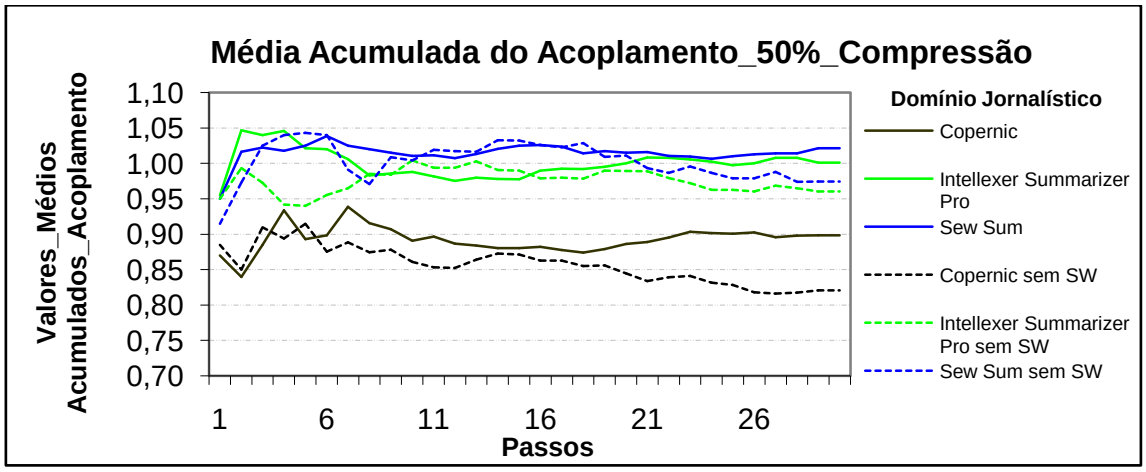


Figura 27b: Resultados obtidos pelo modelo Cassiopeia, usando a medida Acoplamento com 50% de compressão no idioma Inglês no domínio jornalístico.

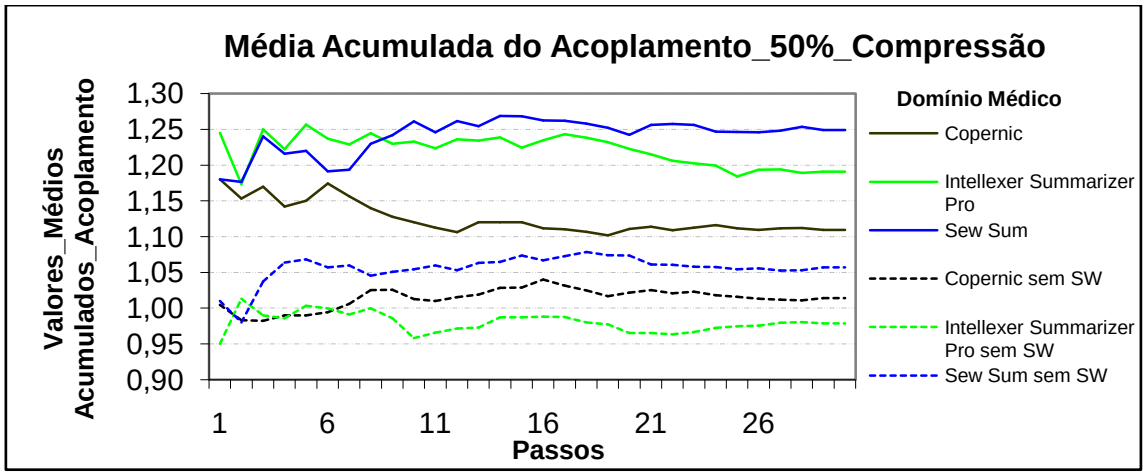


Figura 28b: Resultados obtidos pelo modelo Cassiopeia, usando a medida Acoplamento com 50% de compressão no idioma Inglês no domínio médico.

Compressão de 70% no idioma inglês

No domínio jornalístico, como demonstra figura 29b, os resultados obtidos com os textos com e sem *stopwords* foram próximos, exceto os resultados dos textos sem *stopwords* do sumarizador *Copernic*, esses foram menores. No domínio médico, assim como apresentado na figura 30b, na maioria das simulações, os maiores resultados dos textos com *stopwords.*

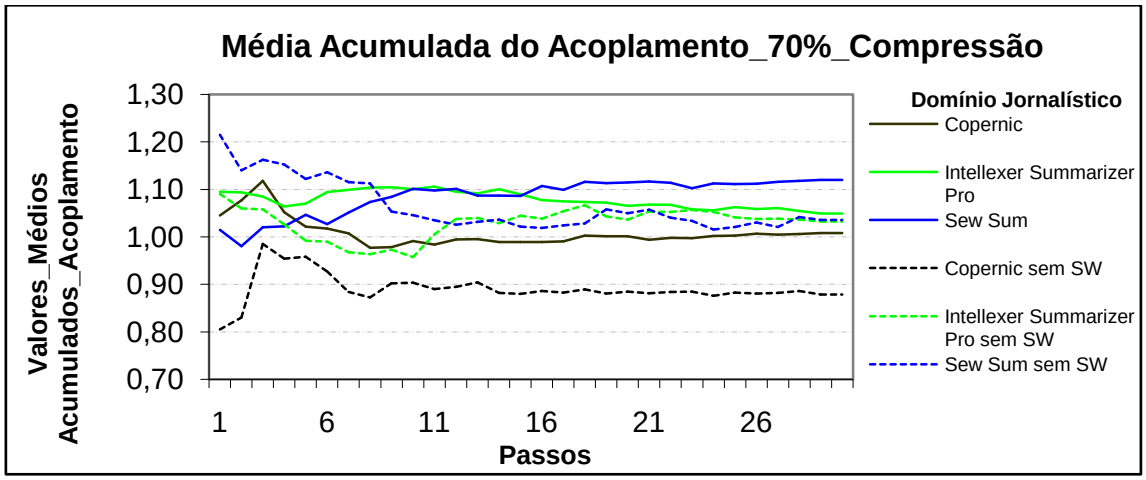


Figura 29b: Resultados obtidos pelo modelo Cassiopeia, usando a medida Acoplamento com 70% de compressão no idioma Inglês no domínio jornalístico.

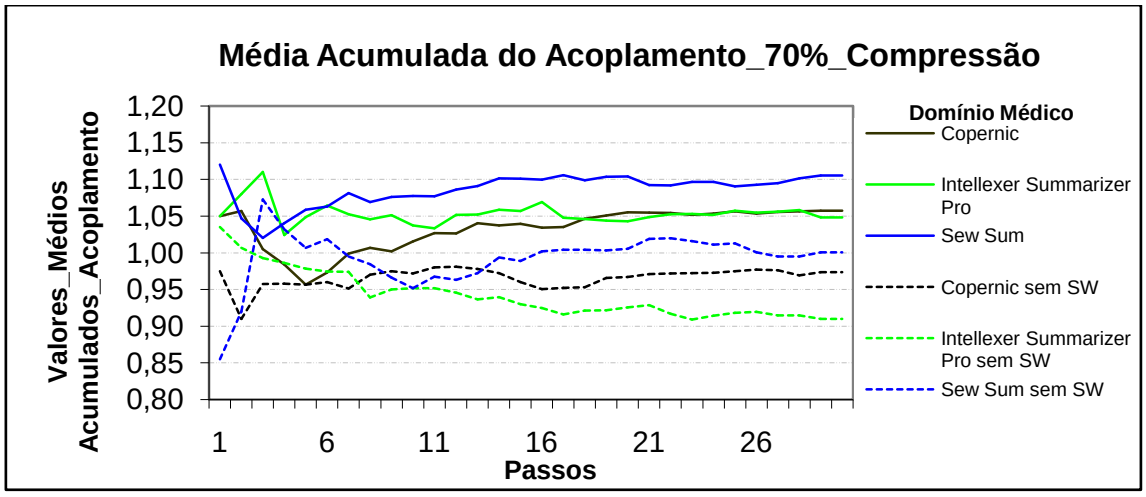


Figura 30b: Resultados obtidos pelo modelo Cassiopeia, usando a medida Acoplamento com 70% de compressão no idioma Inglês no domínio médico.

Compressão de 80% no idioma inglês

No domínio jornalístico, como apresentado na figura 31b, os resultados obtidos com os textos com *stopwords* foram os maiores na maioria das simulações, sendo que os textos sem *stopwords* do sumarizadores *Intellexer Summarizer Pro* e *Sew Sum* obtiveram resultados significativos em poucas simulações. No domínio médico, conforme figura 32b, os maiores resultados foram obtidos com os textos com *stopwords* dos sumarizadores *Copernic* e *Sew Sum*.

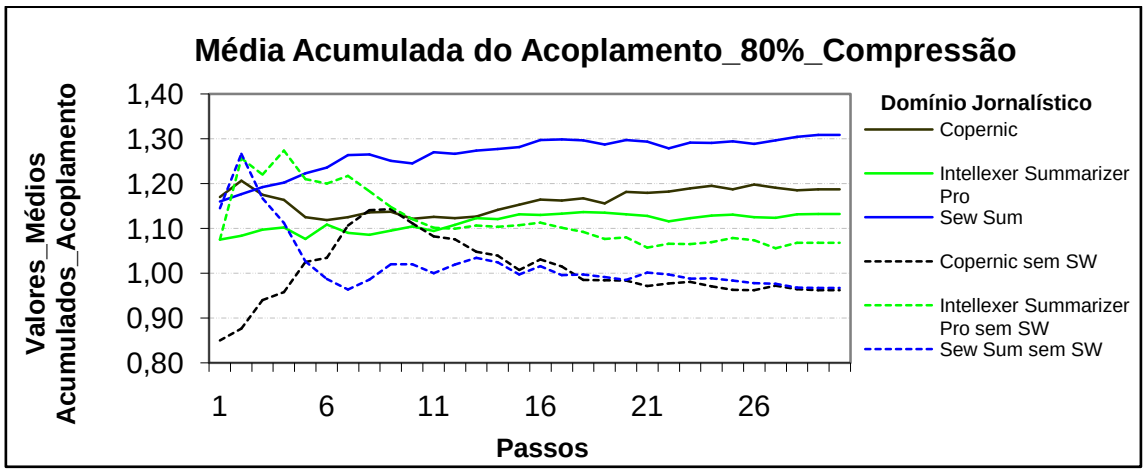


Figura 31b: Resultados obtidos pelo modelo Cassiopeia, usando a medida Acoplamento com 80% de compressão no idioma Inglês no domínio jornalístico.

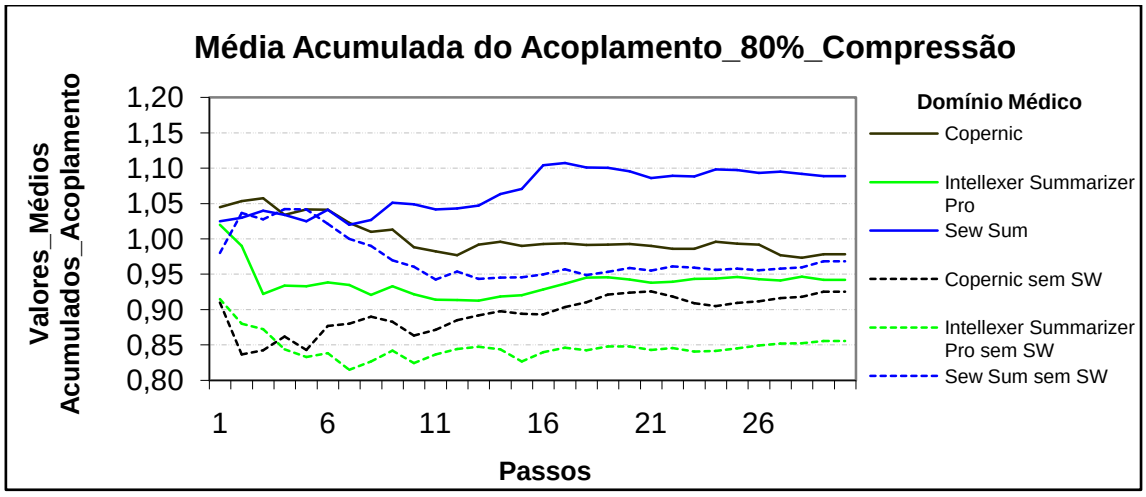


Figura 32b: Resultados obtidos pelo modelo Cassiopeia, usando a medida Acoplamento com 80% de compressão no idioma Inglês no domínio médico.

Compressão de 90% no idioma inglês

No domínio jornalístico, como apresentado na figura 33b, os maiores resultados foram obtidos com os textos com *stopwords*. No domínio médico, conforme figura 34b, os resultados obtidos com os textos com e sem *stopwords* foram bem próximos.

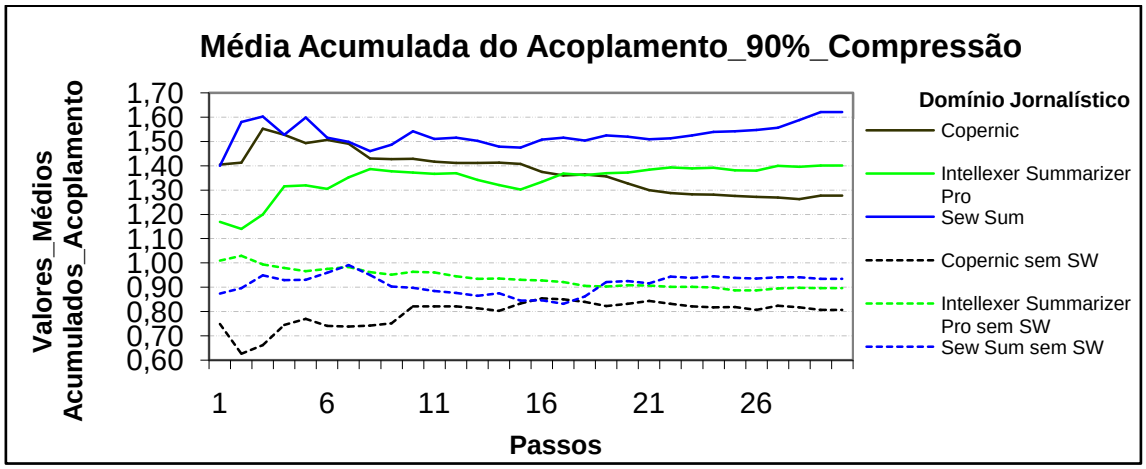


Figura 33b: Resultados obtidos pelo modelo Cassiopeia, usando a medida Acoplamento com 90% de compressão no idioma Inglês no domínio jornalístico.

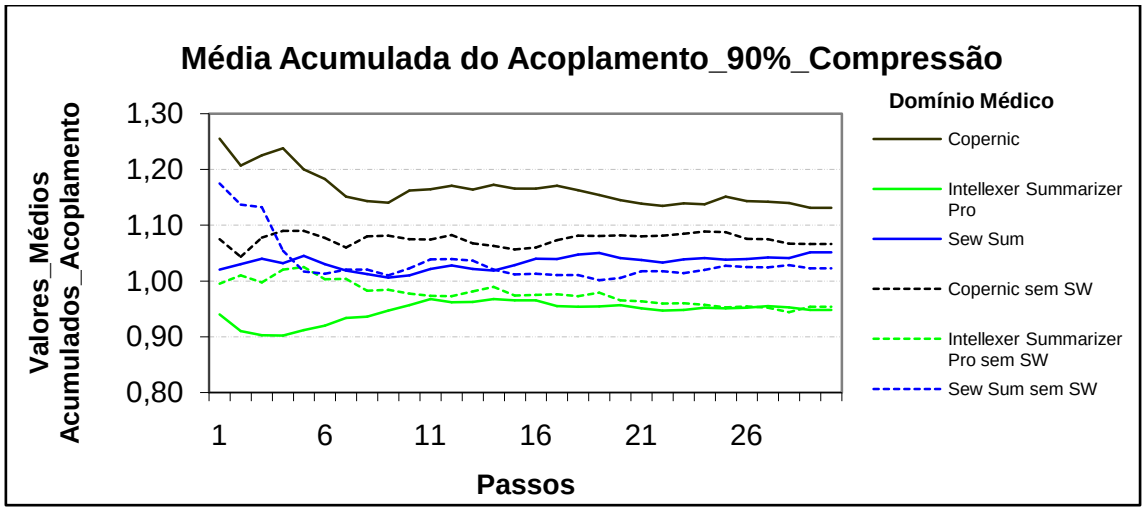


Figura 34b: Resultados obtidos pelo modelo Cassiopeia, usando a medida Acoplamento com 90% de compressão no idioma Inglês no domínio médico.

Média acumulada coesão

Compressão de 50% no idioma português

No domínio jornalístico, conforme figura 35a, os maiores resultados obtidos foram com os textos sem *stopwords* do sumarizador *Gist Intrasenteca*, os demais resultados foram bem próximos. No domínio jurídico e no médico, como mostram respectivamente as figuras 36a e 37a, os resultados dos textos com e sem *stopwords* foram bem próximos.

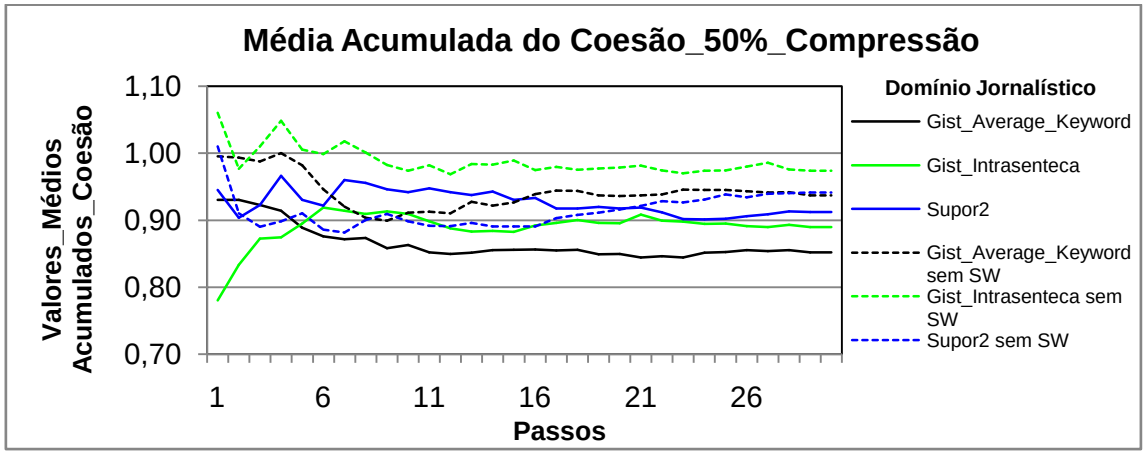


Figura 35a: Resultados obtidos pelo modelo Cassiopeia, usando a medida Coesão com 50% de compressão no idioma Português no domínio jornalístico.

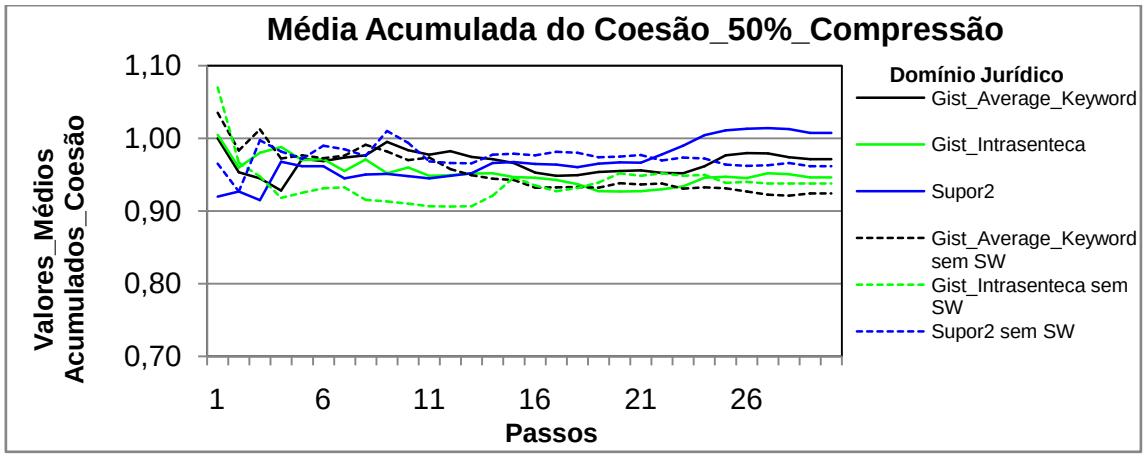


Figura 36a: Resultados obtidos pelo modelo Cassiopeia, usando a medida Coesão com 50% de compressão no idioma Português no domínio jurídico.

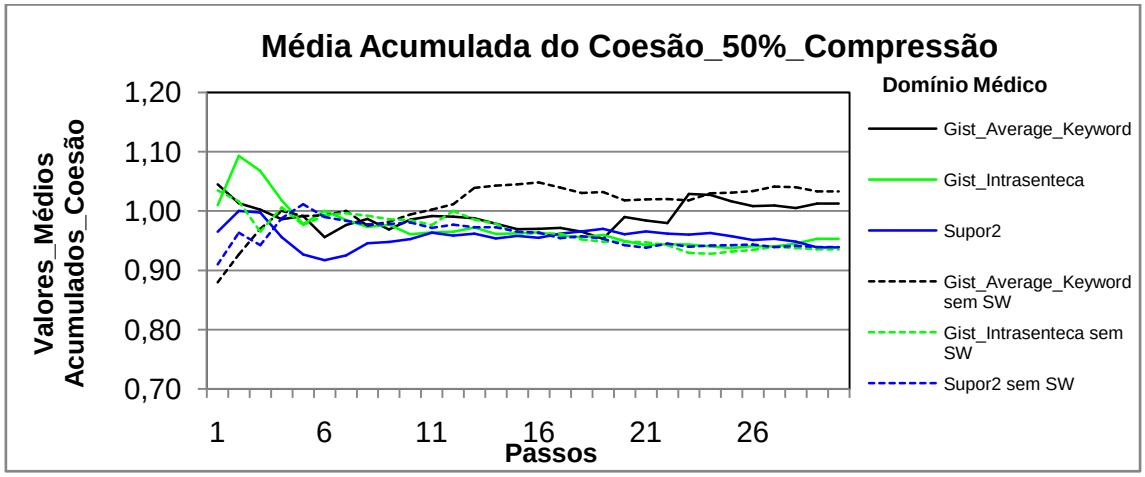


Figura 37a: Resultados obtidos pelo modelo Cassiopeia, usando a medida Coesão com 50% de compressão no idioma Português no domínio médico.

Compressão de 70% no idioma português

No domínio jornalístico, como mostra figura 38a, os maiores resultados foram obtidos na maioria das simulações com os textos sem *stopwords* dos sumarizadores *Gist Average Keyword* e *Gist Intrasenteca*. No domínio jurídico, conforme figura 39a, os resultados obtidos com os textos com e sem *stopwords* foram bem próximos. No domínio médico, assim como apresentado na figura 40a, os maiores resultados obtidos foram com os textos sem *stopwords* do sumarizador *Gist Average Keyword.*

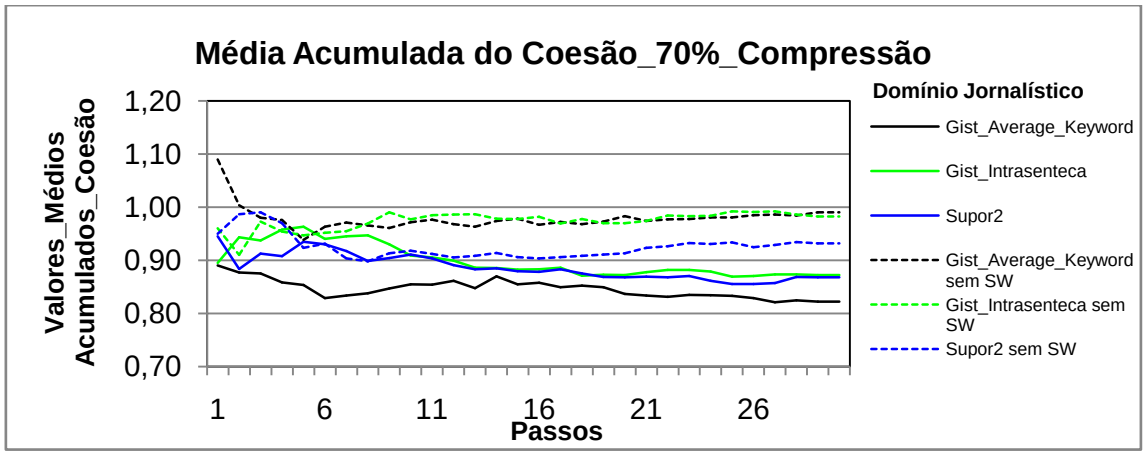


Figura 38a: Resultados obtidos pelo modelo Cassiopeia, usando a medida Coesão com 70% de compressão no idioma Português no domínio jornalístico.

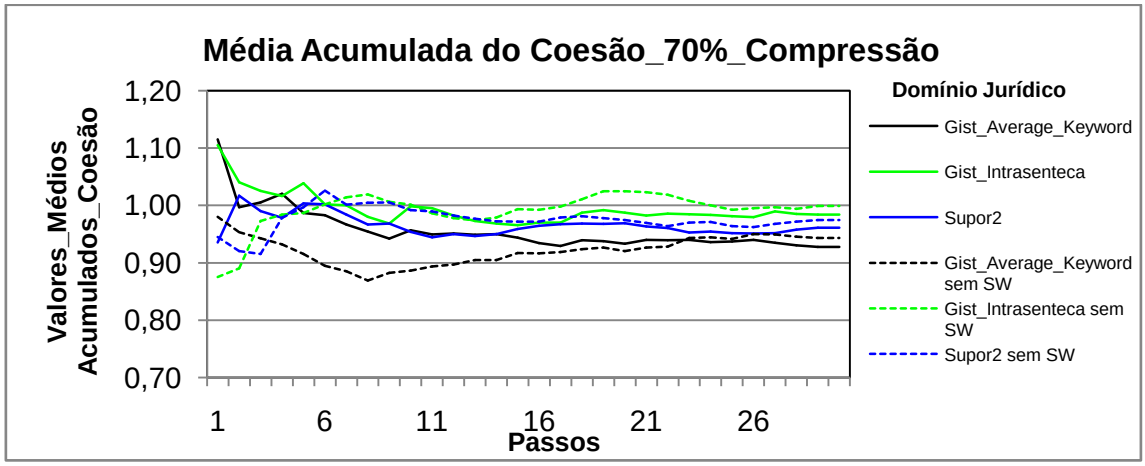


Figura 39a: Resultados obtidos pelo modelo Cassiopeia, usando a medida Coesão com 70% de compressão no idioma Português no domínio jurídico.

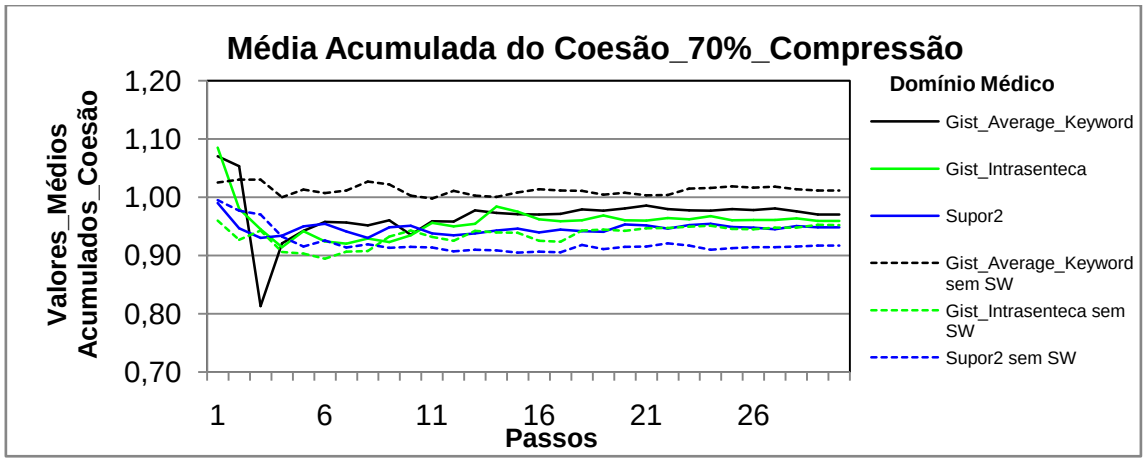


Figura 40a: Resultados obtidos pelo modelo Cassiopeia, usando a medida Coesão com 70% de compressão no idioma Português no domínio médico.

Compressão de 80% no idioma português

No domínio jornalístico, como demonstra a figura 41a, os maiores resultados obtidos foram com os textos sem stopwords. No domínio jurídico e no médico, como apresentado nas figuras 42a e 43a, os resultados obtidos foram bem próximos.

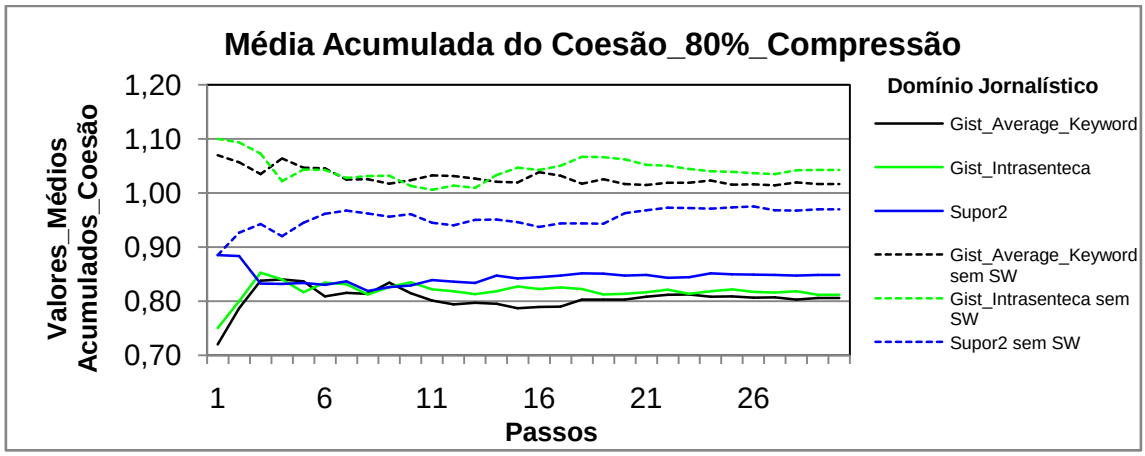


Figura 41a: Resultados obtidos pelo modelo Cassiopeia, usando a medida Coesão com 80% de compressão no idioma Português no domínio jornalístico.

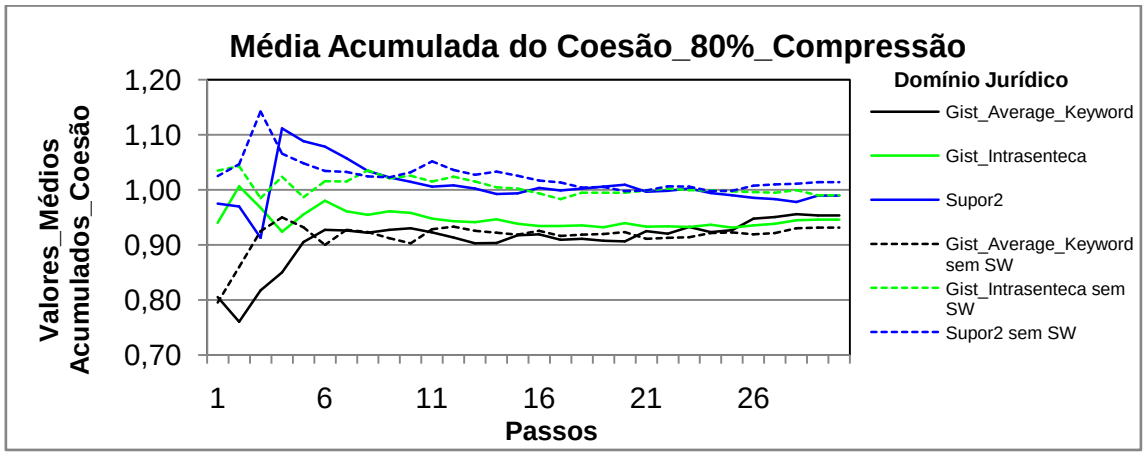


Figura 42a: Resultados obtidos pelo modelo Cassiopeia, usando a medida Coesão com 80% de compressão no idioma Português no domínio jurídico.

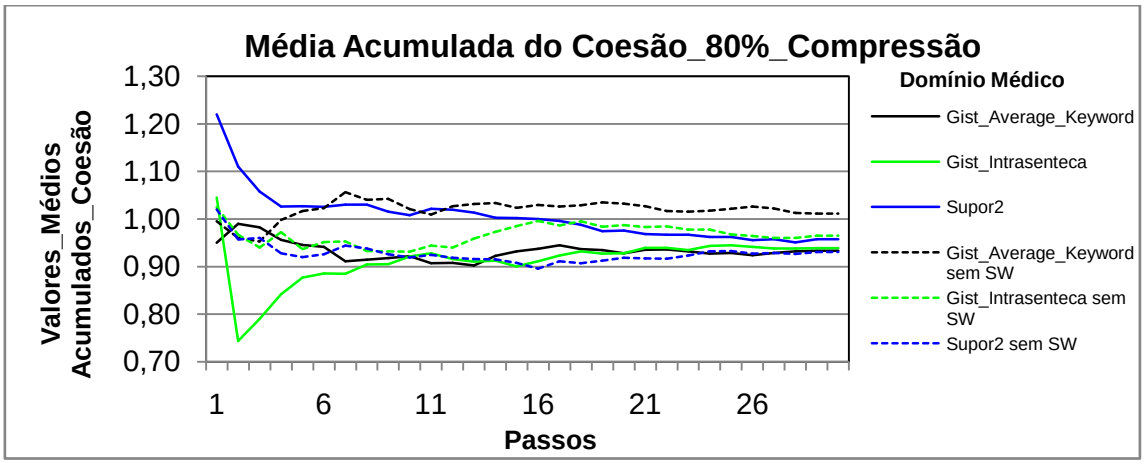


Figura 43a: Resultados obtidos pelo modelo Cassiopeia, usando a medida Coesão com 80% de compressão no idioma Português no domínio médico.

Compressão de 90% no idioma português

No domínio jornalístico, conforme figura 44a, os maiores resultados obtidos foram com os textos sem *stopwords*. No domínio jurídico e no médico, como demonstram as figuras 45a e 46a, os maiores resultados obtidos foram com os textos sem *stopwords* dos sumarizadores *Gist Average Keyword* e *Gist Intrasenteca.*

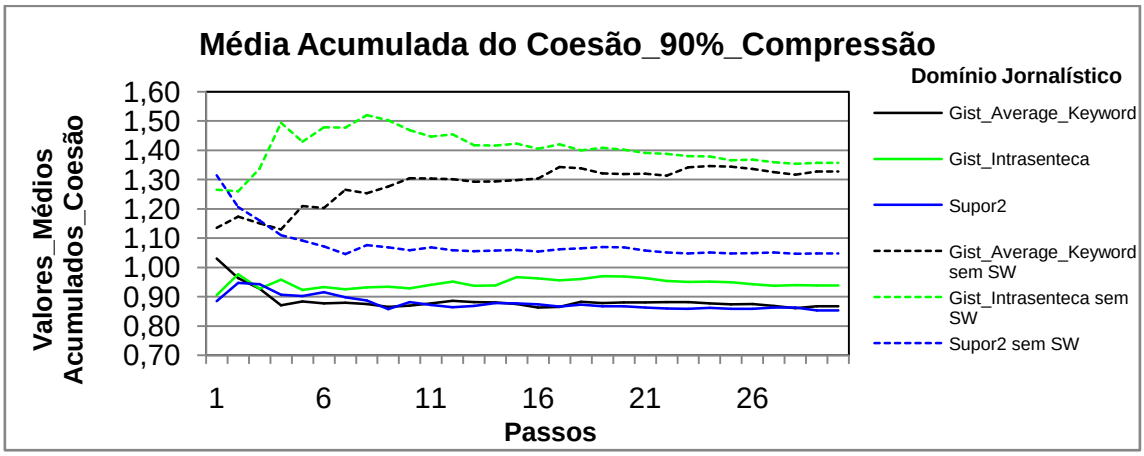


Figura 44a: Resultados obtidos pelo modelo Cassiopeia, usando a medida Coesão com 90% de compressão no idioma Português no domínio jornalístico.

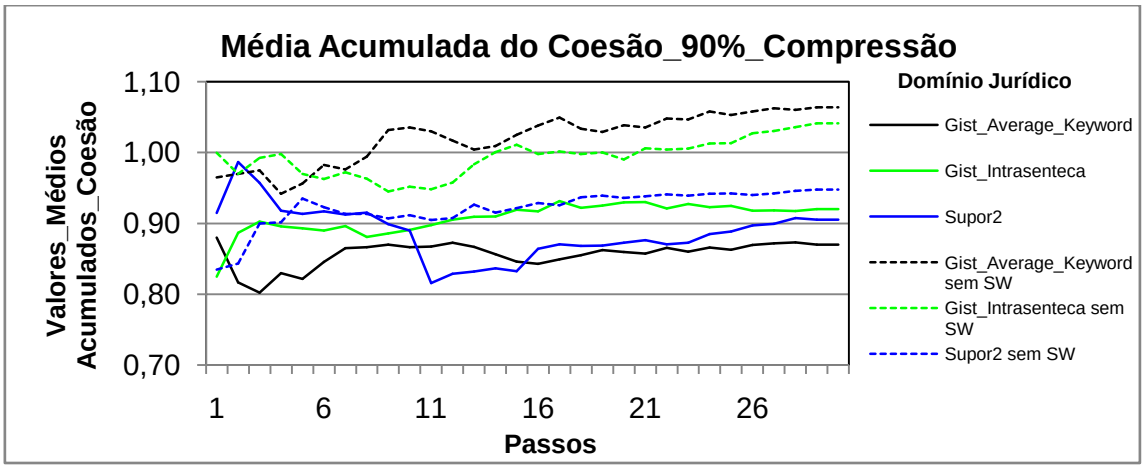


Figura 45a: Resultados obtidos pelo modelo Cassiopeia, usando a medida Coesão com 90% de compressão no idioma Português no domínio jurídico.

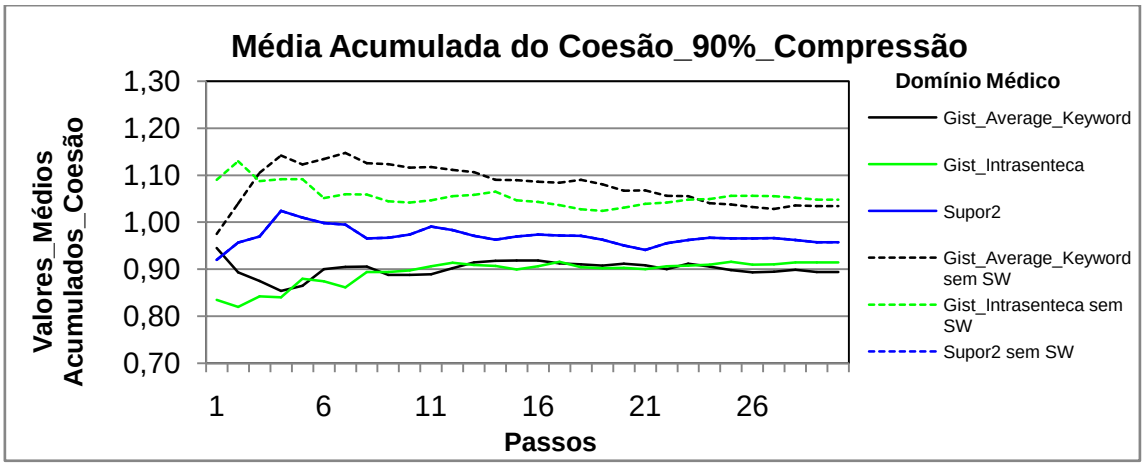


Figura 46a: Resultados obtidos pelo modelo Cassiopeia, usando a medida Coesão com 90% de compressão no idioma Português no domínio médico.

Média acumulada do acoplamento

Compressão de 50% no idioma português

No domínio jornalístico, conforme figura 35b, os maiores resultados obtidos foram com os textos com *stopwords* do sumarizador *Gist Average Keyword*. No domínio jurídico, como mostra a figura 36b, os maiores resultados obtidos foram com os textos com *stopwords* dos sumarizadores *Gist Average Keyword* e *Supor2*. No domínio médico, como apresentado na figura 37b, os maiores resultados obtidos foram com os textos com *stopwords* do sumarizador *Supor2*.

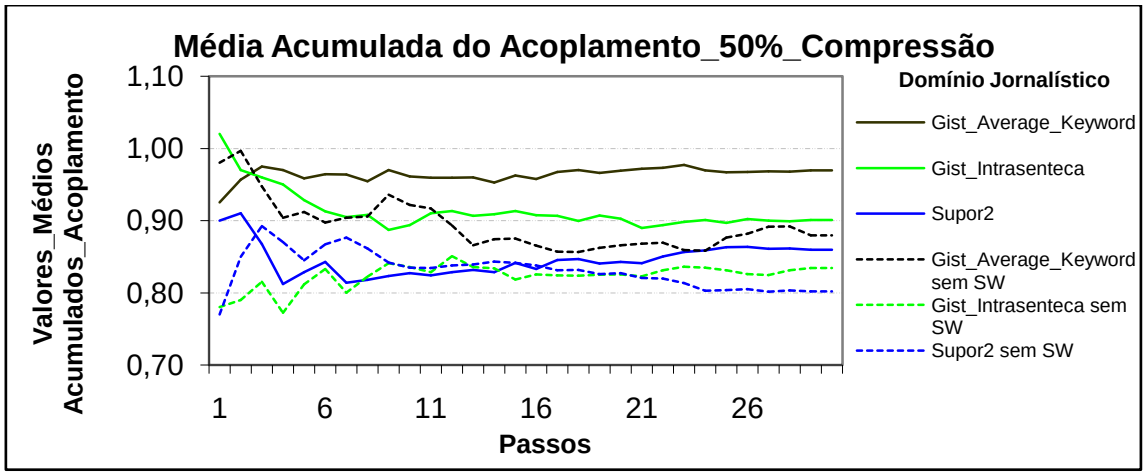


Figura 35b: Resultados obtidos pelo modelo Cassiopeia, usando a medida Acoplamento com 50% de compressão no idioma Português no domínio jornalístico.

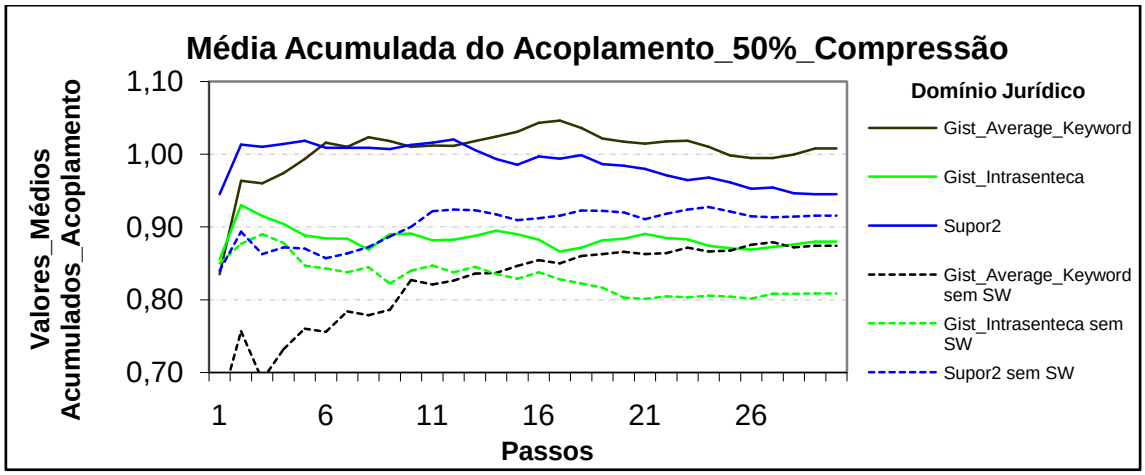


Figura 36b: Resultados obtidos pelo modelo Cassiopeia, usando a medida Acoplamento com 50% de compressão no idioma Português no domínio jurídico.

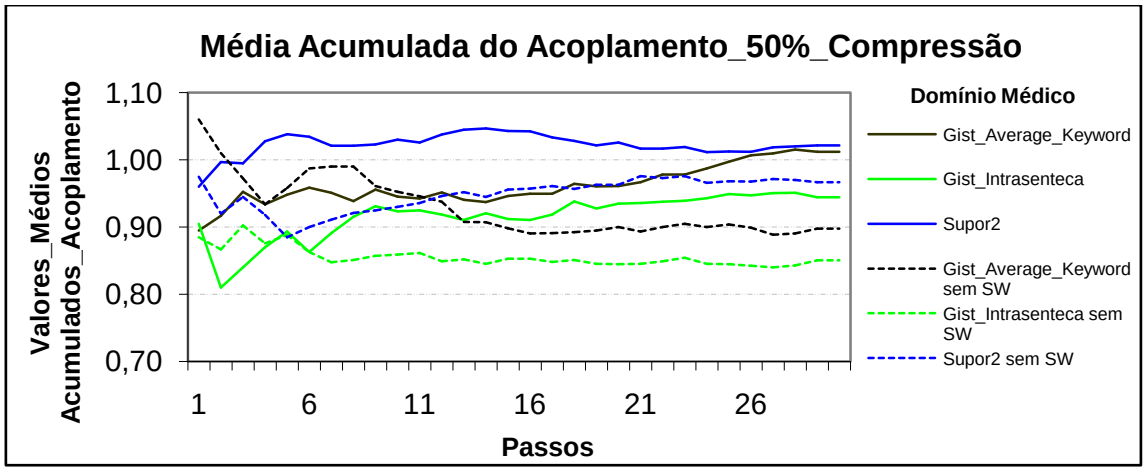


Figura 37b: Resultados obtidos pelo modelo Cassiopeia, usando a medida Acoplamento com 50% de compressão no idioma Português no domínio médico.

Compressão de 70% no idioma português

No domínio jornalístico e no jurídico, como demonstram respectivamente as figuras 38b e 39b, os maiores resultados obtidos foram com os textos com *stopwords* do sumarizador *Gist Average Keyword.* No domínio médico, conforme figura 40b, os resultados obtidos com os textos com e sem *stopwords* foram bem próximos.

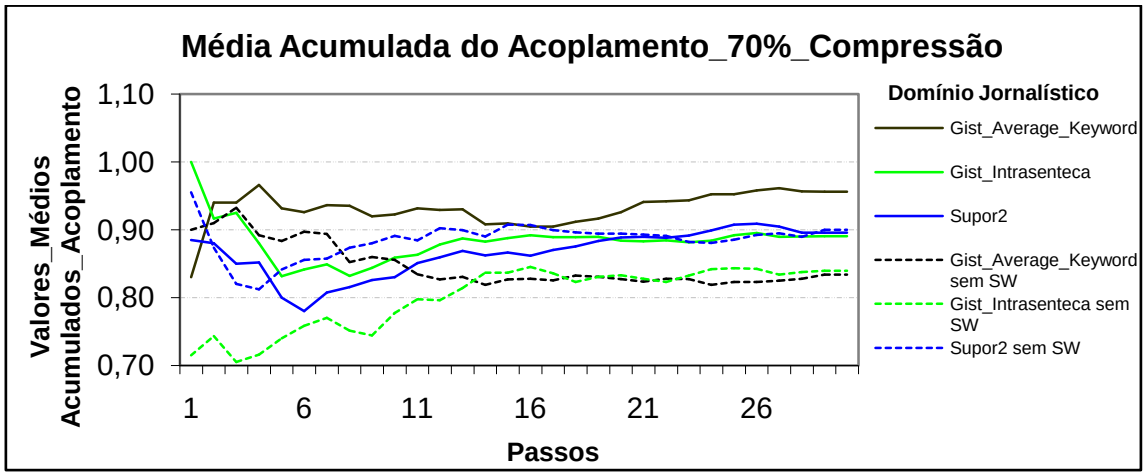


Figura 38b: Resultados obtidos pelo modelo Cassiopeia, usando a medida Acoplamento com 70% de compressão no idioma Português no domínio jornalístico.

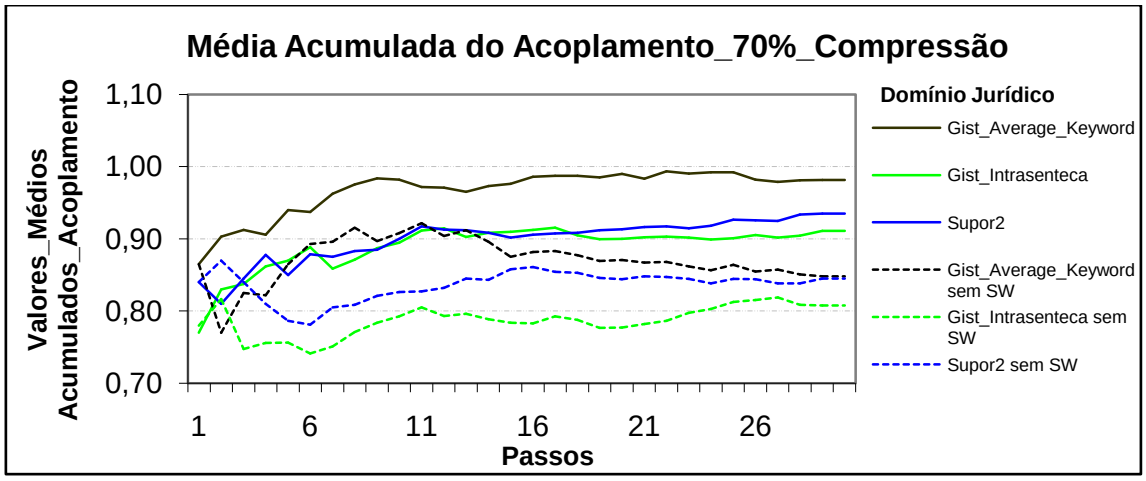


Figura 39b: Resultados obtidos pelo modelo Cassiopeia, usando a medida Acoplamento com 70% de compressão no idioma Português no domínio jurídico.

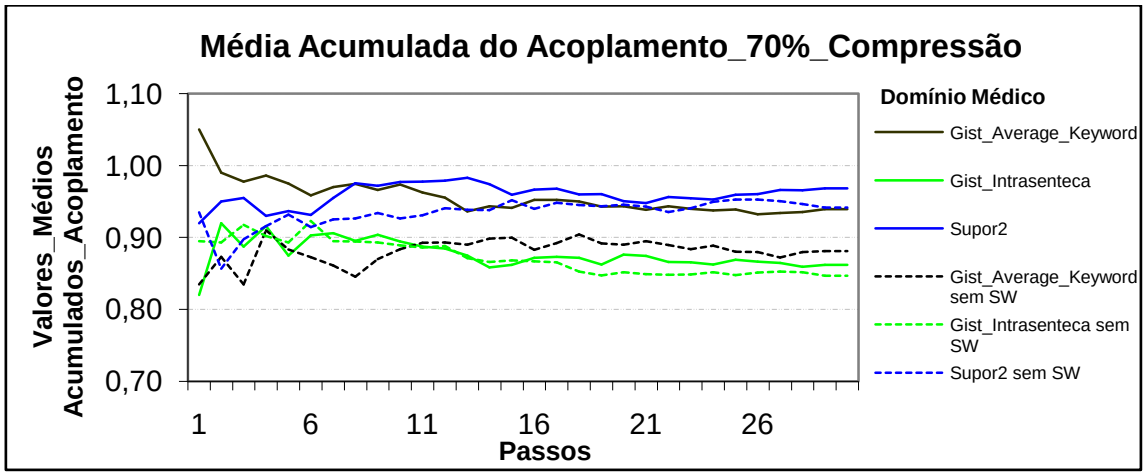


Figura 40b: Resultados obtidos pelo modelo Cassiopeia, usando a medida Acoplamento com 70% de compressão no idioma Português no domínio médico.

Compressão de 80% no idioma português

No domínio jornalístico, conforme figura 41b, os maiores resultados obtidos foram com os textos com *stopwords* dos sumarizadores *Gist Average Keyword* e *Gist Intrasenteca*. No domínio jurídico, assim como apresentado na figura 42b, os maiores resultados obtidos foram com os textos com *stopwords* dos sumarizadores *Gist Average Keyword* e *Supor2*. No domínio médico, como mostra a figura 43b, os maiores resultados obtidos foram com os textos com *stopwords* do sumarizador *Gist Average Keyword.*

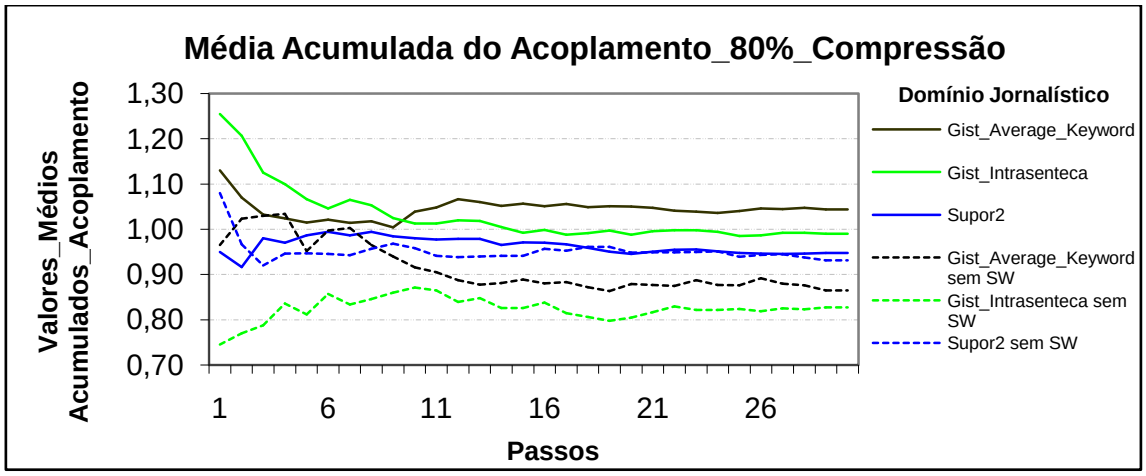


Figura 41b: Resultados obtidos pelo modelo Cassiopeia, usando a medida Acoplamento com 80% de compressão no idioma Português no domínio jornalístico.

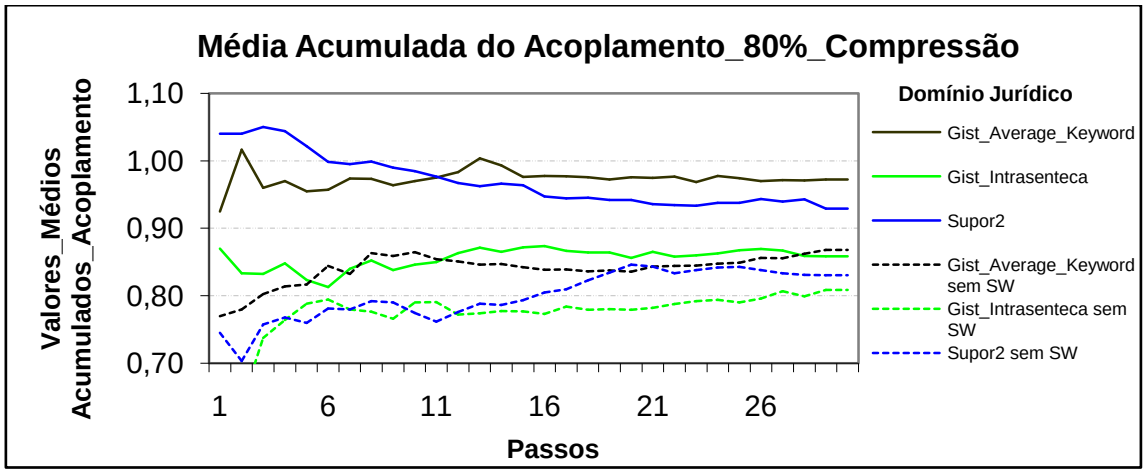


Figura 42b: Resultados obtidos pelo modelo Cassiopeia, usando a medida Acoplamento com 80% de compressão no idioma Português no domínio jurídico.

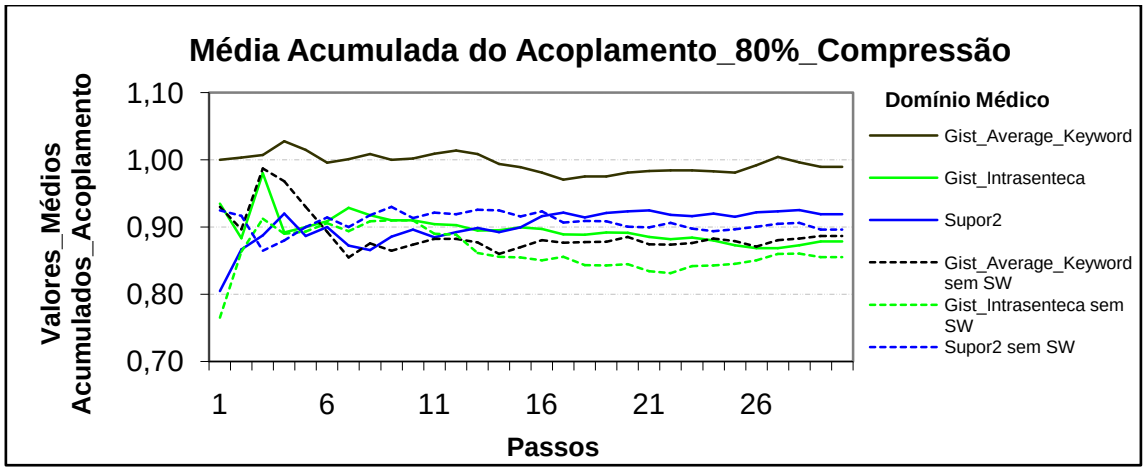


Figura 43b: Resultados obtidos pelo modelo Cassiopeia, usando a medida Acoplamento com 80% de compressão no idioma Português no domínio médico.

Compressão de 90% no idioma português

No domínio jornalístico, como apresentado na figura 44b, os maiores resultados obtidos foram com os textos com *stopwords* dos sumarizadores *Gist Average Keyword* e *Gist Intrasenteca*. No domínio jurídico, conforme figura 45b, os maiores resultados obtidos foram com os textos com *stopwords* do sumarizador *Gist Average Keyword*. No domínio médico, assim como mostra a figura 46b, os maiores resultados obtidos foram com os textos com *stopwords* do sumarizador *Gist Average Keyword*.

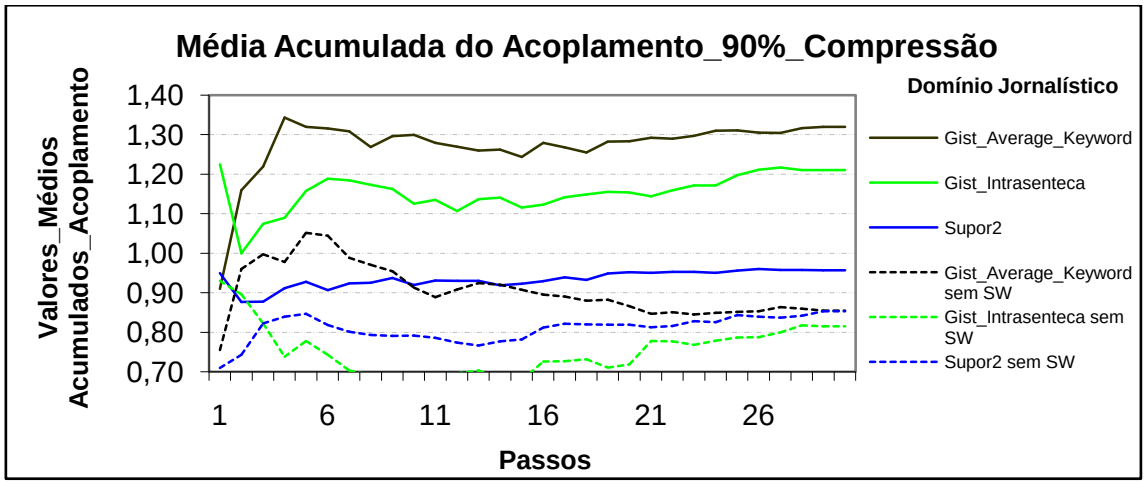


Figura 44b: Resultados obtidos pelo modelo Cassiopeia, usando a medida Acoplamento com 90% de compressão no idioma Português no domínio jornalístico.

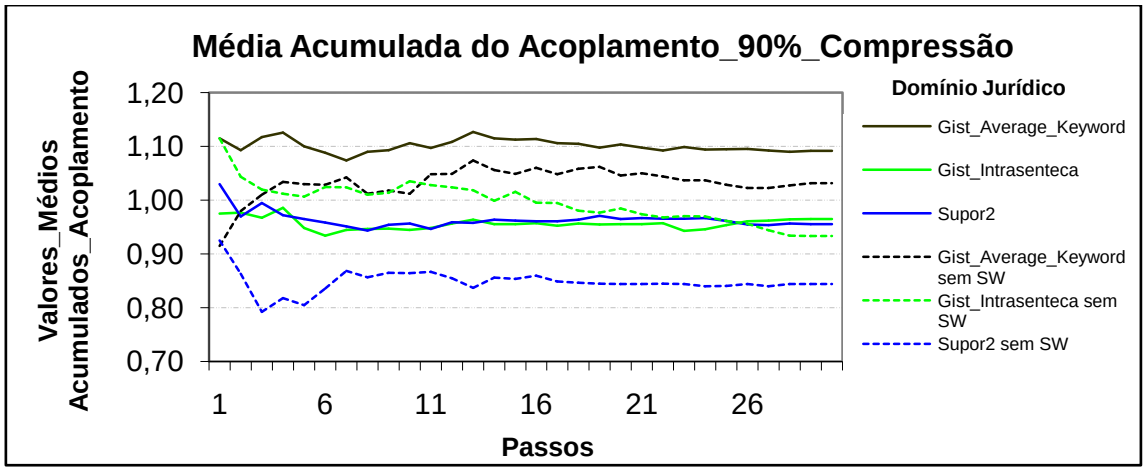


Figura 45b: Resultados obtidos pelo modelo Cassiopeia, usando a medida Acoplamento com 90% de compressão no idioma Português no domínio jurídico.

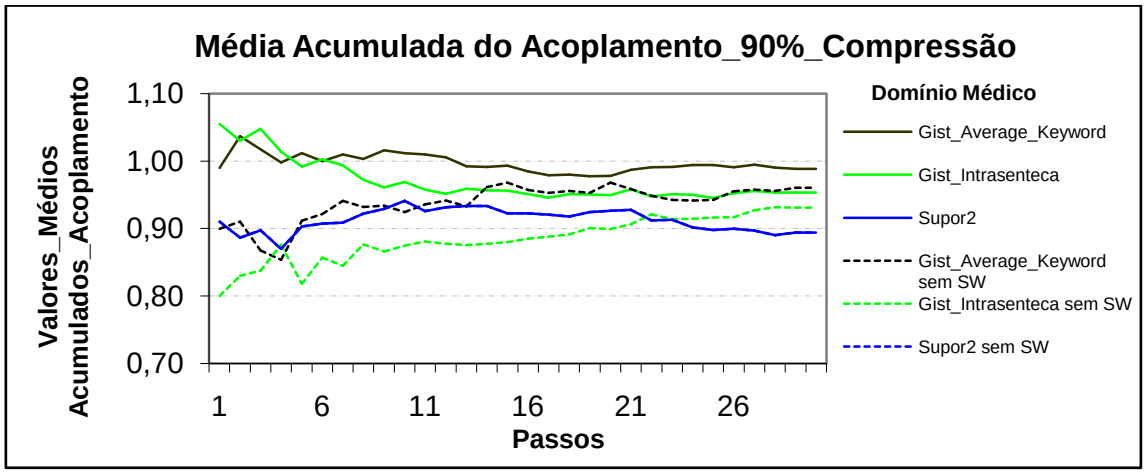


Figura 46b: Resultados obtidos pelo modelo Cassiopeia, usando a medida Acoplamento com 90% de compressão no idioma Português no domínio médico.

Apêndice C - Software 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 foi usado para realizar os testes estatísticos dos experimentos e comprovação da hipótese o seguinte software, *StatPlus*® (http://www.analystsoft.com/en/products/statplus/) uma versão *Trial*. Esse software foi escolhido porque tem os testes estatísticos ANOVA de Friedman e o coeficiente de concordância de Kendall adotado neste trabalho.

Tabela 01: Teste estatístico das amostras das medidas internas do domínio jornalístico do idioma inglês.

Teste Estatístico ANOVA de Friedman e Coeficiente de Concordância de Kendall Comparando amostras múltiplas relacionadas Inglês - Jornalístico 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=30									**Coeficiente de concordância de Kendall**				**Ordem médio**			
GL=5									1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
Métodos Seleção atributos	**Ordem médio**				**Soma de ordens**				**Média**				**Desvio padrão**			
Copernic	5,000	5,000	5,000	2,000	150,000	150,000	150,000	60,000	0,885	0,818	0,700	0,287	0,001	0,004	0,004	0,014
Intellexer Summarizer Pro	6,000	6,000	6,000	3,000	180,000	180,000	180,000	90,000	0,894	0,836	0,748	0,411	0,001	0,002	0,006	0,016
Sew Sum	4,000	4,000	4,000	5,000	120,000	120,000	120,000	150,000	0,842	0,764	0,534	0,575	0,003	0,006	0,008	0,009
Copernic sem STW	2,000	2,000	2,000	1,000	60,000	60,000	60,000	30,000	0,768	0,638	0,432	0,116	0,007	0,006	0,008	0,031
Intellexer Summarizer Pro sem STW	3,000	3,000	3,000	6,000	90,000	90,000	90,000	180,000	0,788	0,654	0,468	0,852	0,003	0,006	0,008	0,001
Sew Sum sem STW	1,000	1,000	1,000	4,000	30,000	30,000	30,000	120,000	0,660	0,496	0,179	0,542	0,005	0,008	0,020	0,006

Tabela 02: Teste estatístico das amostras das medidas internas do domínio médico do idioma inglês.

Teste Estatístico ANOVA de Friedman e Coeficiente de Concordância de Kendall Comparando amostras múltiplas relacionadas Inglês - Médico 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=30									**Coeficiente de concordância de Kendall**				**Ordem médio**			
GL=5									0,994	0,990	1,000	1,000	0,994	0,989	1,000	1,000
Métodos Seleção atributos	**Ordem médio**				**Soma de ordens**				**Média**				**Desvio padrão**			
Copernic	6,000	6,000	6,000	6,000	180,000	180,000	180,000	180,000	0,981	0,971	0,961	0,951	0,000	0,000	0,001	0,001
Intellexer Summarizer Pro	4,983	5,000	5,000	5,000	149,500	150,000	150,000	150,000	0,977	0,967	0,947	0,915	0,001	0,000	0,001	0,002
Sew Sum	4,017	4,000	4,000	4,000	120,500	120,000	120,000	120,000	0,976	0,962	0,945	0,891	0,001	0,000	0,001	0,001
Copernic sem STW	1,983	3,000	3,000	1,000	59,500	90,000	90,000	30,000	0,967	0,953	0,935	0,116	0,000	0,000	0,000	0,031
Intellexer Summarizer Pro sem STW	1,033	1,900	2,000	3,000	31,000	57,000	60,000	90,000	0,966	0,948	0,922	0,847	0,000	0,002	0,002	0,005
Sew Sum sem STW	2,983	1,100	1,000	2,000	89,500	33,000	30,000	60,000	0,968	0,944	0,920	0,820	0,000	0,001	0,001	0,006

Tabela 03: Teste estatístico das amostras das medidas internas do domínio jornalístico do idioma português.

Teste Estatístico ANOVA de Friedman e Coeficiente de Concordância de Kendall Comparando amostras múltiplas relacionadas Português - Jornalístico 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=30									**Coeficiente de concordância de Kendall**				**Ordem médio**			
GL=5									0,996	0,987	0,980	0,990	0,996	0,986	0,979	0,989
Métodos Seleção atributos	**Ordem médio**				**Soma de ordens**				**Média**				**Desvio padrão**			
Gist_Average_Keyword	4,033	5,000	5,000	4,000	121,000	150,000	150,000	120,000	0,857	0,788	0,681	0,034	0,002	0,004	0,004	0,012
Gist_Intrasenteca	4,967	3,867	3,000	2,900	149,000	116,000	90,000	87,000	0,864	0,762	0,605	-0,063	0,002	0,004	0,008	0,019
Supor2	6,000	6,000	6,000	6,000	180,000	180,000	180,000	180,000	0,913	0,875	0,818	0,730	0,001	0,002	0,002	0,005
Gist_Average_Keyword sem STW	1,000	2,000	1,767	1,000	30,000	60,000	53,000	30,000	0,731	0,579	0,394	-0,138	0,003	0,005	0,015	0,037
Gist_Intrasenteca sem STW	2,000	1,000	1,233	2,100	60,000	30,000	37,000	63,000	0,784	0,539	0,390	-0,082	0,006	0,009	0,008	0,024
Supor2 sem STW	3,000	3,133	4,000	5,000	90,000	94,000	120,000	150,000	0,845	0,754	0,618	0,416	0,003	0,006	0,005	0,011

Tabela 04: Teste estatístico das amostras das medidas internas do domínio jurídico do idioma português.

Teste Estatístico ANOVA de Friedman e Coeficiente de Concordância de Kendall Comparando amostras múltiplas relacionadas Português - Juridico 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=30									**Coeficiente de concordância de Kendall**				**Ordem médio**			
GL=5									1	0,986	0,981	0,984	1	0,985	0,981	0,984
Métodos Seleção atributos	**Ordem médio**				**Soma de ordens**				**Média**				**Desvio padrão**			
Copernic_S_St	5,000	4,117	3,800	3,167	150,000	123,500	114,000	95,000	0,985	0,978	0,965	0,925	0,000	0,001	0,001	0,001
Intellexer Summarizer Pro_S_St	3,000	3,000	3,200	3,833	90,000	90,000	96,000	115,000	0,983	0,977	0,964	0,926	0,000	0,000	0,000	0,001
Sew Sum_S_St	6,000	6,000	6,000	6,000	180,000	180,000	180,000	180,000	0,991	0,987	0,983	0,968	0,000	0,000	0,000	0,000
Copernic_C_St	1,000	1,033	1,000	1,000	30,000	31,000	30,000	30,000	0,976	0,962	0,938	0,850	0,000	0,001	0,001	0,001
Intellexer Summarizer Pro_C_St	2,000	1,967	2,000	2,000	60,000	59,000	60,000	60,000	0,977	0,965	0,944	0,873	0,000	0,001	0,001	0,002
Sew Sum_C_St	4,000	4,883	5,000	5,000	120,000	146,500	150,000	150,000	0,984	0,979	0,972	0,947	0,000	0,000	0,000	0,001

Tabela 05: Teste estatístico das amostras das medidas internas do domínio médico do idioma português.

Teste Estatístico ANOVA de Friedman e Coeficiente de Concordância de Kendall Comparando amostras múltiplas relacionadas Português - Médico 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=30									**Coeficiente de concordância de Kendall**				**Ordem médio**			
GL=5									0,975	0,929	1	0,989	0,974	0,927	1	0,989
Métodos Seleção atributos	**Ordem médio**				**Soma de ordens**				**Média**				**Desvio padrão**			
Copernic_S_St	5,000	4,767	4,000	3,900	150,000	143,000	120,000	117,000	0,973	0,958	0,935	0,865	0,001	0,001	0,002	0,003
Intellexer Summarizer Pro_S_St	4,000	3,367	3,000	3,100	120,000	101,000	90,000	93,000	0,971	0,955	0,931	0,862	0,000	0,001	0,001	0,002
Sew Sum_S_St	6,000	6,000	6,000	5,500	180,000	180,000	180,000	165,000	0,981	0,974	0,963	0,940	0,000	0,001	0,001	0,000
Copernic_C_St	1,583	1,167	1,000	1,000	47,500	35,000	30,000	30,000	0,959	0,934	0,895	0,756	0,000	0,001	0,001	0,003
Intellexer Summarizer Pro_C_St	1,417	1,833	2,000	2,000	42,500	55,000	60,000	60,000	0,958	0,937	0,907	0,782	0,001	0,001	0,002	0,003
Sew Sum_C_St	3,000	3,867	5,000	5,500	90,000	116,000	150,000	165,000	0,968	0,956	0,947	0,940	0,001	0,001	0,001	0,000

Tabela 06: Teste estatístico das amostras das medidas externas do domínio jornalístico do idioma inglês.

Teste Estatístico ANOVA de Friedman e Coeficiente de Concordância de Kendall Comparando amostras múltiplas relacionadas Inglês - Jornalístico 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=30									**Coeficiente de concordância de Kendall**				**Ordem médio**			
GL=5									0,837	0,908	0,975	0,949	0,831	0,905	0,974	0,947
Métodos Seleção atributos	**Ordem médio**				**Soma de ordens**				**Média**				**Desvio padrão**			
Copernic	3,667	4,300	2,800	1,750	110,000	129,000	84,000	52,500	0,245	0,263	0,231	0,226	0,012	0,006	0,007	0,004
Intellexer Summarizer Pro	1,233	1,800	1,050	1,300	37,000	54,000	31,500	39,000	0,211	0,222	0,217	0,225	0,004	0,006	0,002	0,004
Sew Sum	1,833	1,200	2,200	3,133	55,000	36,000	66,000	94,000	0,222	0,216	0,221	0,242	0,008	0,002	0,002	0,004
Copernic sem STW	3,650	5,400	5,000	3,850	109,500	162,000	150,000	115,500	0,250	0,268	0,263	0,252	0,002	0,003	0,006	0,006
Intellexer Summarizer Pro sem STW	5,417	5,267	6,000	6,000	162,500	158,000	180,000	180,000	0,257	0,267	0,274	0,277	0,001	0,004	0,005	0,005
Sew Sum sem STW	5,200	3,033	3,950	4,967	156,000	91,000	118,500	149,000	0,257	0,255	0,245	0,260	0,004	0,002	0,007	0,004

Tabela 07: Teste estatístico das amostras das medidas externas do domínio médico do idioma inglês.

Teste Estatístico ANOVA de Friedman e Coeficiente de Concordância de Kendall Comparando amostras múltiplas relacionadas Inglês - Médico 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=30									**Coeficiente de concordância de Kendall**				**Ordem médio**			
GL=5									0,637	0,743	0,832	0,943	0,624	0,734	0,826	0,941
Métodos Seleção atributos	**Ordem médio**				**Soma de ordens**				**Média**				**Desvio padrão**			
Copernic	1,517	2,250	1,567	3,667	45,500	67,500	47,000	110,000	0,200	0,199	0,201	0,218	0,002	0,003	0,003	0,003
Intellexer Summarizer Pro	3,667	1,217	2,067	1,483	110,000	36,500	62,000	44,500	0,202	0,197	0,201	0,204	0,003	0,001	0,001	0,002
Sew Sum	2,333	3,050	2,533	1,550	70,000	91,500	76,000	46,500	0,200	0,204	0,202	0,204	0,002	0,005	0,001	0,002
Copernic sem STW	4,433	4,283	5,733	4,967	133,000	128,500	172,000	149,000	0,203	0,208	0,212	0,230	0,003	0,002	0,002	0,004
Intellexer Summarizer Pro sem STW	3,400	4,967	4,317	3,333	102,000	149,000	129,500	100,000	0,202	0,209	0,209	0,217	0,001	0,002	0,002	0,001
Sew Sum sem STW	5,650	5,233	4,783	6,000	169,500	157,000	143,500	180,000	0,206	0,209	0,210	0,236	0,002	0,003	0,002	0,004

Tabela 08: Teste estatístico das amostras das medidas externas do domínio jornalístico do idioma português.

Teste Estatístico ANOVA de Friedman e Coeficiente de Concordância de Kendall Comparando amostras múltiplas relacionadas Português - Jornalístico 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=30									**Coeficiente de concordância de Kendall**				**Ordem médio**			
GL=5									0,982	0,916	0,984	0,920	0,982	0,913	0,983	0,917
Métodos Seleção atributos	**Ordem médio**				**Soma de ordens**				**Média**				**Desvio padrão**			
Gist_Average_Keyword	1,000	2,517	2,183	1,900	30,000	75,500	65,500	57,000	0,124	0,135	0,130	0,148	0,001	0,002	0,002	0,003
Gist_Intrasenteca	2,217	3,117	4,000	3,717	66,500	93,500	120,000	111,500	0,134	0,137	0,147	0,177	0,002	0,001	0,003	0,005
Supor2	2,783	3,367	2,817	1,100	83,500	101,000	84,500	33,000	0,136	0,136	0,131	0,141	0,002	0,003	0,002	0,004
Gist_Average_Keyword sem STW	6,000	1,000	1,000	5,850	180,000	30,000	30,000	175,500	0,174	0,087	0,107	0,191	0,003	0,001	0,001	0,003
Gist_Intrasenteca sem STW	4,983	6,000	6,000	4,967	149,500	180,000	180,000	149,000	0,153	0,182	0,185	0,184	0,002	0,002	0,003	0,006
Supor2 sem STW	4,017	5,000	5,000	3,467	120,500	150,000	150,000	104,000	0,145	0,159	0,172	0,177	0,001	0,002	0,003	0,002

Tabela 09: Teste estatístico das amostras das medidas externas do domínio jurídico do idioma português.

Teste Estatístico ANOVA de Friedman e Coeficiente de Concordância de Kendall Comparando amostras múltiplas relacionadas Português - Juridico 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=30									**Coeficiente de concordância de Kendall**				**Ordem médio**			
GL=5									0,607	0,971	0,958	0,968	0,594	0,970	0,957	0,967
Métodos Seleção atributos	**Ordem médio**				**Soma de ordens**				**Média**				**Desvio padrão**			
Gist_Average_Keyword	4,100	1,467	1,050	2,000	123,000	44,000	31,500	60,000	0,187	0,182	0,181	0,191	0,003	0,002	0,006	0,002
Gist_Intrasenteca	1,667	3,000	2,650	3,000	50,000	90,000	79,500	90,000	0,185	0,188	0,186	0,195	0,001	0,002	0,006	0,003
Supor2	2,667	1,550	2,383	1,000	80,000	46,500	71,500	30,000	0,186	0,182	0,185	0,185	0,003	0,001	0,006	0,002
Gist_Average_Keyword sem STW	5,683	5,050	6,000	5,533	170,500	151,500	180,000	166,000	0,190	0,194	0,206	0,222	0,002	0,002	0,006	0,002
Gist_Intrasenteca sem STW	4,167	5,950	4,933	5,400	125,000	178,500	148,000	162,000	0,187	0,200	0,198	0,223	0,003	0,001	0,003	0,003
Supor2 sem STW	2,717	3,983	3,983	4,067	81,500	119,500	119,500	122,000	0,186	0,191	0,195	0,205	0,001	0,002	0,006	0,004

Tabela 10: Teste estatístico das amostras das medidas externas do domínio médico do idioma português.

Teste Estatístico ANOVA de Friedman e Coeficiente de Concordância de Kendall Comparando amostras múltiplas relacionadas Português - Médico 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=30									**Coeficiente de concordância de Kendall**				**Ordem médio**			
GL=5									0,693	0,826	0,974	0,990	0,682	0,820	0,973	0,989
Métodos Seleção atributos	**Ordem médio**				**Soma de ordens**				**Média**				**Desvio padrão**			
Gist_Average_Keyword	1,250	1,067	1,017	1,017	37,500	32,000	30,500	30,500	0,198	0,198	0,196	0,198	0,001	0,001	0,001	0,002
Gist_Intrasenteca	2,283	3,017	2,000	2,083	68,500	90,500	60,000	62,500	0,202	0,202	0,201	0,201	0,002	0,003	0,002	0,002
Supor2	4,517	2,683	3,217	2,900	135,500	80,500	96,500	87,000	0,206	0,201	0,205	0,205	0,002	0,001	0,001	0,002
Gist_Average_Keyword sem STW	5,550	5,950	6,000	6,000	166,500	178,500	180,000	180,000	0,207	0,211	0,213	0,226	0,002	0,001	0,001	0,001
Gist_Intrasenteca sem STW	3,583	3,650	4,917	4,983	107,500	109,500	147,500	149,500	0,204	0,203	0,210	0,220	0,003	0,003	0,002	0,003
Supor2 sem STW	3,817	4,633	3,850	4,017	114,500	139,000	115,500	120,500	0,203	0,207	0,207	0,211	0,002	0,002	0,002	0,001

ANEXOS

Anexo A

Tabela 04: Lista de palavras e sua frequência no texto.

Palavra e sua respectiva freqüência							
Palavra	**Frequência**	**Palavra**	**Frequência**	**Palavra**	**Frequência**	**Palavra**	**Frequência**
de	146	entre	7	serviços	4	situação	3
a	109	idéias	7	social	4	sua	3
e	104	pelo	7	só	4	sujeito	3
o	70	pelos	7	tinha	4	tempo	3
que	69	centrais	6	trabalho	4	termos	3
da	51	central	6	tudo	4	tratado	3
do	51	desta	6	vem	4	tratam	3
em	41	este	6	ver	4	três	3
um	31	estudo	6	vida	4	unidade	3
os	30	família	6	acima	3	vai	3
dos	29	filho	6	análise	3	valorização	3
no	29	foram	6	apesar	3	embora	3
é	27	outros	6	aspecto	3	sus	3
com	25	paciente	6	aspectos	3	uti	3
para	24	qualidade	6	assistencial	3	abordando	2
na	23	seu	6	atender	3	acordo	2
uma	23	tal	6	bastante	3	adolescente	2
atendimento	22	ações	5	carinho	3	agente	2
dsc	22	cada	5	coisas	3	agrupamento	2
as	20	comunicação	5	conforme	3	alguns	2
criança	20	condições	5	discursos	3	assistência	2
ser	20	experiência	5	ele	3	associados	2
como	19	falta	5	enfermeiras	3	atendem	2
das	19	neste	5	enfrentamento	3	atendido	2
não	19	nos	5	enquanto	3	atividades	2
por	18	participação	5	entendimento	3	atual	2
saúde	18	quando	5	entrevistados	3	até	2
é	18	sendo	5	era	3	ação	2
profissionais	16	sistema	5	estão	3	baixa	2
ao	15	sobre	5	eu	3	boa	2
humanizado	14	sujeitos	5	foi	3	bruto	2
acompanhantes	13	todos	5	fundamental	3	característica	2
bem	12	às	5	humanizada	3	caridade	2
hospitalar	11	acerca	4	há	3	casos	2
humanização	11	além	4	identificação	3	coletivo	2
mais	11	apoio	4	importância	3	competência	2
tem	11	avaliação	4	internação	3	comportamento	2
atenção	10	bom	4	maioria	3	compreensão	2
cuidado	10	contexto	4	mal	3	comum	2
idéia	10	deste	4	manifestações	3	condição	2
ou	10	diagnóstico	4	minha	3	considerar	2
relação	10	dias	4	muito	3	contextos	2
são	10	estudos	4	mães	3	conteúdo	2
à	10	expressões-chave	4	percepção	3	controle	2
assim	9	forma	4	permanência	3	conversar	2
equipe	9	humano	4	permitiram	3	criação	2
hospital	9	já	4	perspectiva	3	dados	2
pela	9	medida	4	pesquisa	3	dando	2
processo	9	modo	4	pessoas	3	dar	2
se	9	necessidades	4	problemas	3	definidas	2
atitudes	8	nem	4	profissional	3	deixar	2
hospitalização	8	parte	4	receber	3	depoimento	2
mãe	8	pessoa	4	regras	3	desde	2
pais	8	prestado	4	relatos	3	desenvolvidas	2
acompanhante	7	relacionadas	4	sem	3	desenvolvimento	2
ambiente	7	respeito	4	seus	3	destes	2
crianças	7	sentido	4	singular	3	deve	2

direito	2	médico	2	acentuada	1	buscar	1
discurso	2	médicos	2	achados	1	campo	1
disso	2	método	2	achei	1	capacitados	1
diversos	2	natureza	2	acompanhamento	1	capacitação	1
durante	2	nesse	2	acompanhá-lo	1	capazes	1
dó	2	olhar	2	acontecendo	1	caracterizam	1
educação	2	onde	2	acontecimento	1	carinhoso	1
empatia	2	origem	2	adaptações	1	caráter	1
ensino	2	parceria	2	adequada	1	certa	1
entender	2	participantes	2	adequados	1	chamando	1
envolvendo	2	particular	2	adotado	1	cinco	1
específico	2	pediátrica	2	afago	1	cirúrgicos	1
esta	2	pequenas	2	afetuoso	1	citados	1
estavam	2	percebidas	2	agentes	1	citação	1
estes	2	pergunta	2	agirem	1	civilizar	1
estratégias	2	pode	2	agradável	1	ciência	1
extração	2	porque	2	ainda	1	cliente	1
fala	2	possuíam	2	ajuda	1	clientela	1
familiares	2	presença	2	alcance	1	coisa	1
fatores	2	primeira	2	alguma	1	coletiva	1
fazem	2	programa	2	alta	1	coletivos	1
ficar	2	projeto	2	amamentação	1	comer	1
filha	2	público	2	amigável	1	cometidas	1
função	2	públicos	2	amor	1	comida	1
gente	2	qualquer	2	analisados	1	comparado	1
gerando	2	quatro	2	angéstia	1	compararem	1
gesto	2	realidade	2	ano	1	completado	1
gostam	2	realizado	2	anos	1	composição	1
grupos	2	realmente	2	ansiedade	1	compreender	1
homogêneas	2	recursos	2	aparecem	1	compreensiva	1
hospital-escola	2	referência	2	aparência	1	comunidade	1
hospitalares	2	relacionados	2	apontados	1	conceito	1
hospitalizadas	2	representam	2	apresentam	1	conceitos	1
humana	2	restrito	2	apresentaram	1	conduta	1
humanizadas	2	segurança	2	aprovado	1	conferem	1
humanos	2	sentimentos	2	após	1	confirmar	1
identificar	2	seres	2	assinatura	1	conflito	1
igual	2	serviço	2	assistir	1	conformidade	1
implica	2	significa	2	associadas	1	conhecer	1
importante	2	suas	2	associam	1	conhecimento	1
infantil	2	suporte	2	associar	1	conhecimentos	1
informação	2	também	2	atencioso	1	conjunto	1
informações	2	tange	2	atenda	1	conotação	1
instituição	2	tecnologias	2	atendeu	1	conseguia	1
internadas	2	todas	2	atenuar	1	conseguiram	1
inverso	2	toque	2	ativa	1	conseqüente	1
isso	2	tratar	2	atividade	1	considerado	1
lado	2	verbal	2	atribuída	1	considerarmos	1
lençol	2	vez	2	avaliam	1	consigo	1
lhe	2	vezes	2	avançadas	1	consta	1
mas	2	várias	2	avanço	1	constituir	1
materna	2	nacional	2	avariada	1	constituído	1
medicamentos	2	abarcar	1	avó	1	construção	1
melhor	2	abordagens	1	benefícios	1	conta	1
menos	2	abrir	1	binêmio	1	contato	1
mesmo	2	acabam	1	bons	1	continuidade	1
momentos	2	aceite	1	burocracia	1	contornos	1

contribuem	1	disparidade	1	esteja	1	global	1
contribuição	1	disponibilidade	1	estilo	1	gostar	1
controlando	1	disponibilizadas	1	estressante	1	governamental	1
contudo	1	disse	1	está	1	grandes	1
convulsiva	1	distanciamento	1	estúpidos	1	gravados	1
cooperação	1	diziam	1	etária	1	grosso	1
correspondam	1	doença	1	evolução	1	haviam	1
correspondentes	1	dois	1	examinar	1	histórico	1
crise	1	dominação-subordinação	1	excluir	1	hoje	1
criticar	1	dor	1	exemplificarem	1	holástica	1
critério	1	dupla	1	exercem	1	hora	1
cronicidade	1	dúvidas	1	exigem	1	horário	1
crítica	1	educado	1	existem	1	hospitais	1
críticas	1	educados	1	existência	1	hospitalizados	1
crítico	1	educativo	1	expectativas	1	humanizadora	1
cuidador	1	efetividade	1	experienciavam	1	humanizar	1
cuidados	1	eficaz	1	experiências	1	ia	1
cuidar	1	elaborado	1	explicar	1	identificada	1
cuidar-assistir	1	elas	1	explicava	1	identificadas	1
cuidá-la	1	elementos	1	exposição	1	identificado	1
cujo	1	eles	1	expressões	1	implementadas	1
cujos	1	emitido	1	extensivos	1	importa	1
culminou	1	emocionais	1	facilmente	1	imposto	1
dada	1	emocional	1	faixa	1	imprescindível	1
decisões	1	encontra-se	1	falas	1	incluindo	1
define	1	encontrar	1	falhas	1	inclusão	1
definem	1	encontro	1	familiar	1	incompleto	1
definir	1	encontros	1	familiaridade	1	individualidades	1
definição	1	enfermaria	1	fará	1	individualizada	1
defrontamos	1	enfermeira	1	fato	1	individualizado	1
deixa	1	enquadrava-se	1	fator	1	indícios	1
deixam	1	entende	1	faz	1	inexistente	1
delas	1	entendida	1	fazendo	1	inexistência	1
delicada	1	entonação	1	feito	1	infinita	1
demonstrando	1	então	1	ficou	1	informativo	1
depoimentos	1	envolver	1	figuras	1	iniciativa	1
depois	1	envolvidas	1	filhos	1	iniciativas	1
derivar	1	envolvimento	1	firma	1	integral	1
desafiadora	1	equilibrada	1	fonte	1	integraram	1
desconsidera	1	errôneo	1	fontes	1	intenção	1
descritivo	1	esclarecimentos	1	formas	1	internada	1
desejável	1	escolaridade	1	formação	1	internamento	1
desencadear	1	escuta	1	fornece	1	interpessoais	1
desenvolver	1	especiais	1	fornecidos	1	interpretadas	1
desfavoráveis	1	especial	1	fortalecimento	1	intrínseco	1
despeito	1	específica	1	fosse	1	intuito	1
destacados	1	espera	1	fotos	1	invariavelmente	1
desvantagem	1	esperadas	1	frente	1	investigada	1
determinar	1	esperar	1	fui	1	investigar	1
deve-se	1	espirituais	1	fundamentais	1	investigarmos	1
devendo	1	esquecemos	1	fundamentalmente	1	investigação	1
dever	1	essenciais	1	físicas	1	investimentos	1
deverão	1	estabelecendo	1	físico	1	junto	1
diagnósticos	1	estabelecida	1	físicos	1	lacunas	1
difícil	1	estabelecimentos	1	gastrintestinais	1	lançado	1
direcionado	1	estas	1	geralmente	1	legais	1
discurso-síntese	1	estava	1	gestos	1	leitos	1

leitura	1	observação	1	preciso	1	relações	1
levam	1	observou-se	1	preconiza	1	relembrar	1
levantar	1	obteve	1	preconizado	1	renais	1
lidar	1	obtidos	1	predominante	1	representa	1
limpeza	1	ocasionar	1	prematuridade	1	representada	1
literais	1	ocorreu	1	preocupação	1	representante	1
longa	1	oferecer	1	presente	1	representantes	1
lugar	1	olham	1	prestativo	1	respeitosa	1
maiores	1	olhavam	1	prevalente	1	respiratórios	1
manhã	1	organizada	1	prevê	1	responsável	1
manifestado	1	otimização	1	primeiro	1	ressaltar	1
manutenção	1	outra	1	primordial	1	retratada	1
me	1	outro	1	principal	1	revelaram	1
mecanicismo	1	pacientes	1	principalmente	1	rotina	1
mecanismo	1	paciência	1	prioridade	1	rotinas	1
mediadores	1	pai	1	priorização	1	roupeiro	1
mediante	1	palavra	1	priorizações	1	ruim	1
meia	1	panorama	1	privado	1	ruins	1
meio	1	papel	1	pro	1	satisfatória	1
meios	1	parceira	1	procedimentos	1	satisfação	1
melhora	1	parece	1	procuramos	1	segmentos	1
melhoria	1	parecem	1	propiciem	1	seguida	1
melhorias	1	parecer	1	proporcionam	1	seguintes	1
membros	1	participar	1	proporcionar	1	seguiu	1
memoráveis	1	partir	1	prova	1	segundo	1
mensagens	1	passivamente	1	prática	1	sempre	1
merecem	1	passo	1	própria	1	senso	1
mesma	1	passos	1	psicológicos	1	sentimento	1
metas	1	paterna	1	psíquico	1	septicemia	1
metodológicas	1	pauta	1	publicação	1	seqüência	1
modelo	1	pedia	1	quais	1	setores	1
mostrava	1	pediatria	1	qualificação	1	sido	1
motivos	1	pequeno	1	qualitativa	1	significativa	1
mudança	1	percebemos	1	quase	1	sim	1
mudanças	1	perceber	1	questionador	1	sine	1
mudar	1	percebidos	1	questionados	1	situações	1
muitas	1	percepções	1	questão	1	sob	1
má	1	periódica	1	questões	1	sobremaneira	1
máquina	1	permanecer	1	rapidez	1	solicitados	1
mão	1	permanente	1	realizadas	1	soluções	1
médio	1	permeiam	1	realizados	1	somente	1
nada	1	permitiu	1	reações	1	suave	1
nas	1	plano	1	receita	1	substanciais	1
necessitam	1	podem	1	recentes	1	subsídios	1
necessárias	1	poder	1	reconhecerem	1	supervalorização	1
necessário	1	pois	1	reconhecíveis	1	surpresa	1
necessários	1	ponto	1	rede	1	tangente	1
negativamente	1	população	1	redigido	1	tanto	1
negociação	1	portanto	1	referido	1	tava	1
nesta	1	positivos	1	reflete	1	tecnológicos	1
noite	1	possuem	1	reformulações	1	temática	1
nom	1	possui	1	regulamentando	1	tendem	1
nome	1	possuía	1	rela	1	tendo	1
nãohumanizadas	1	possível	1	relacionado	1	tenham	1
níveis	1	posturas	1	relacional	1	tenta	1
número	1	pouco	1	relacionamento	1	terapêutica	1
objetiva	1	poucos	1	relatadas	1	terapêutico	1

teria	1	universo	1	voz	1	umas	1
terminologia	1	uns	1	vários	1	expressão-chave	1
teórica	1	urgência	1	vão	1	hu	1
toda	1	usando	1	vá	1	ic	1
todo	1	usuário	1	vínculo	1	intensiva	1
tolerável	1	usuários	1	xixi	1	isto	1
tornando	1	utilizando	1	ética	1	lei	1
tornar	1	vaga	1	énico	1	livre	1
torno	1	valorizado	1	área	1	lá	1
total	1	valorizam	1	véspera	1	maringá	1
trabalhar	1	variados	1	último	1	meu	1
traduz-se	1	variedade	1	únicas	1	ministério	1
traduzidas	1	velha	1	único	1	nestes	1
traduzido	1	verbo	1	ancoragem	1	noroeste	1
transcritos	1	visitas	1	artigo	1	qualidades	1
transformar	1	vislumbra	1	brasil	1	resolução	1
transmitir	1	vista	1	comitê	1	tais	1
trata	1	vistas	1	conselho	1	ter	1
tratamento	1	vivencias	1	consentimento	1	terapia	1
tratar-se	1	vocacional	1	devido	1	termo	1
troca	1	volta	1	esclarecido	1	trata-se	1
trás	1	voltado	1	estadual	1	uem	1
técnica	1	vontade	1	estatuto	1	universidade	1
têm	1						

Anexo B

Tabela 05: Lista de stopwords em português.

STOPWORDS						
a	de	elas	isso	nossas	pouca	tampouco
à	dela	ele	isto	nosso	poucas	te
agora	delas	eles	já	nossos	pouco	tem
ainda	dele	em	la	num	poucos	tendo
alguém	deles	enquanto	la	numa	primeiro	tenha
algum	depois	entre	lá	nunca	primeiros	ter
alguma	dessa	era	lhe	o	própria	teu
algumas	dessas	essa	lhes	os	próprias	teus
alguns	desse	essas	lo	ou	próprio	ti
ampla	desses	esse	mas	outra	próprios	tido
amplas	desta	esses	me	outras	quais	tinha
amplo	destas	esta	mesma	outro	qual	tinham
amplos	deste	está	mesmas	outros	quando	toda
ante	deste	estamos	mesmo	para	quanto	todas
antes	destes	estão	mesmos	pela	quantos	todavia
ao	deve	estas	meu	pelas	que	todo
aos	devem	estava	meus	pelo	quem	todos
após	devendo	estavam	minha	pelos	são	tu
aquela	dever	estávamos	minhas	pequena	se	tua
aquelas	deverá	este	muita	pequenas	seja	tuas
aquele	deverão	estes	muitas	pequeno	sejam	tudo
aqueles	deveria	estou	muito	pequenos	sem	última
aquilo	deveriam	eu	muitos	per	sempre	últimas
as	devia	fazendo	na	perante	sendo	último
até	deviam	fazer	não	pode	será	últimos
através	disse	feita	nas	pôde	serão	um
cada	disso	feitas	nem	podendo	seu	uma
coisa	disto	feito	nenhum	poder	seus	umas
coisas	dito	feitos	nessa	poderia	si	uns
com	diz	foi	nessas	poderiam	sido	vendo
como	dizem	for	nesta	podia	só	ver
contra	do	foram	nestas	podiam	sob	vez
contudo	dos	fosse	ninguém	pois	sobre	vindo
da	e	fossem	no	por	sua	vir
daquele	é	grande	nos	porém	suas	vos
daqueles	e'	grandes	nós	porque	talvez	vós
das	ela	há	nossa	posso	também	

Anexo C

Tabela 06: Lista de stopwords em inglês.

STOPWORDS						
a	back	few	later	on	t	us
about	back	for	least	once	than	very
above	be	from	less	one	that	very
according	because	further	less	only	the	was
across	been	get	let	or	their	were
actually	before	going	little	other	them	what
after	behind	got	many	our	then	when
again	being	great	may	out	there	where
against	below	has	maybe	over	therefore	whether
all	besides	have	me	perhaps	these	which
almost	better	he	might	put	they	while
along	between	her	more	rather	thing	whole
already	beyond	here	most	really	this	whose
also	both	high	much	set	those	will
although	but	his	must	several	though	with
always	by	how	neither	she	three	within
among	can	however	never	should	through	without
an	certain	i	new	since	till	would
and	could	if	no	snot	to	yet
another	do	in	non	snt	together	you
any	does	instead	nor	so	too	your
anything	during	into	not	some	toward	
are	each	is	nothing	something	towards	
aren	else	it	of	sometimes	two	
as	enough	its	off	soon	under	
at	even	itself	often	still	up	
away	ever	just	old	such	upon	

Anexo D

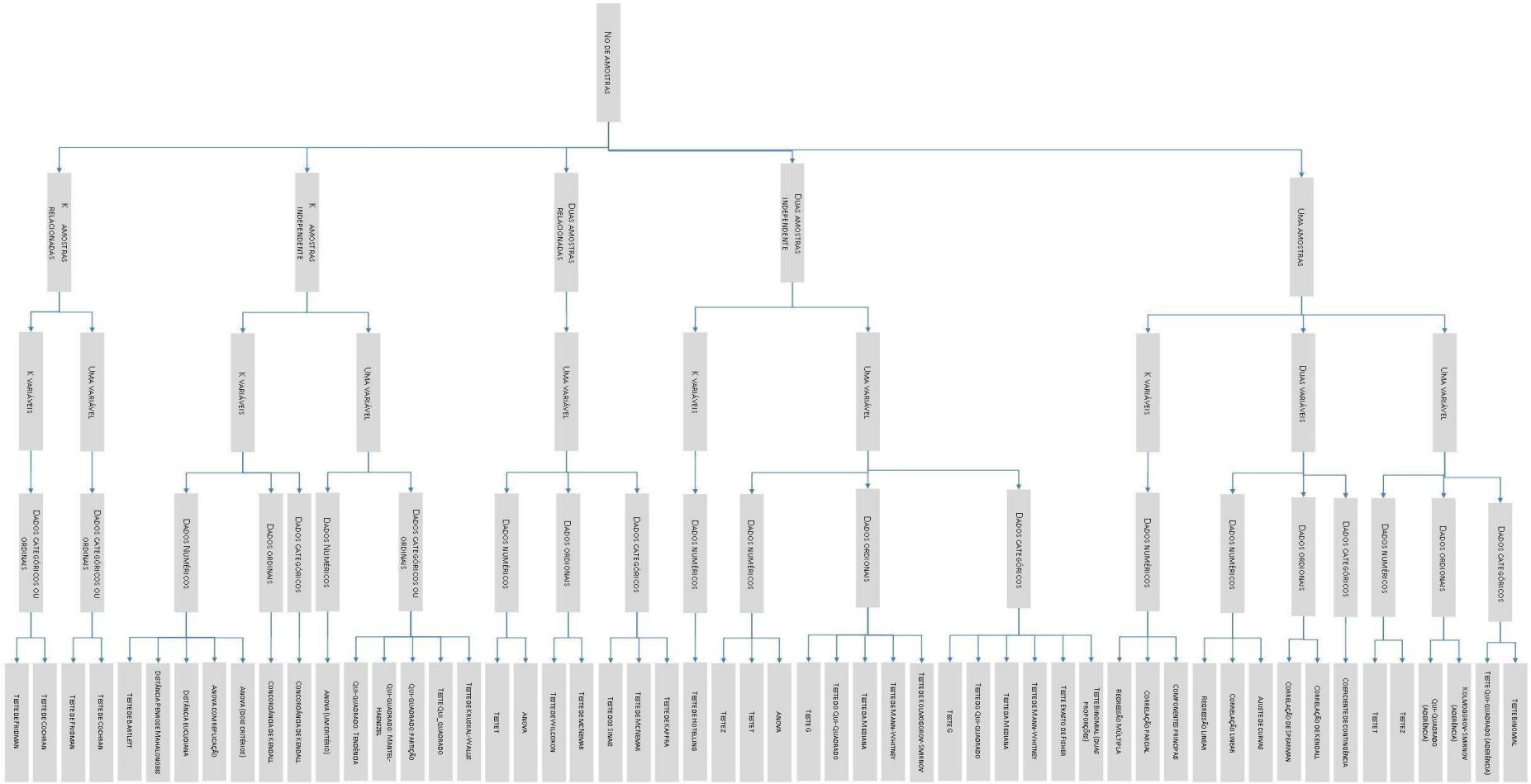


Figura 51: Diagrama para escolha da técnica teste estatístico a partir do número de amostra (CALLEGARI E JACQUES, 2007).