UNIVERSIDADE FEDERAL DOS VALES DO JEQUITINHONHA E MUCURI

Bacharelado em Sistemas de Informação

Marcos Vinícius Santos Cruz

ESTUDO EXPERIMENTAL SOBRE O EFEITO DA REMOÇÃO DE STOPWORDS NA QUALIDADE DA CLUSTERIZAÇÃO COM O MODELO CASSIOPEIA

Diamantina-MG

2025

Marcos Vinícius Santos Cruz

ESTUDO EXPERIMENTAL SOBRE O EFEITO DA REMOÇÃO DE STOPWORDS NA QUALIDADE DA CLUSTERIZAÇÃO COM O MODELO CASSIOPEIA

Trabalho de Conclusão de Curso apresentado à Faculdade de Ciências Exatas da Universidade Federal dos Vales do Jequitinhonha e Mucuri, como parte dos requisitos para a obtenção do título de Bacharel em Sistemas de Informação.

Orientador: Prof. Dr. Marcus Vinícius Carvalho Guelpeli.

Diamantina-MG

2025

AGRADECIMENTOS

Primeiramente, queria agradecer à minha família, em particular à minha mãe Maria José e às minhas irmãs Marina e Marcela, pelo amor, suporte e encorajamento constantes que me acompanharam durante toda essa trajetória. Agradeço também aos meus amigos queridos, tanto aos de infância, Natan, Victor e David, que sempre estiveram presentes nos momentos mais importantes da minha vida, quanto aos que conquistei na faculdade, João Gabriel, Rayssa e André, cuja amizade e companheirismo tornaram esta etapa mais leve e significativa. Agradeço imensamente aos técnicos e docentes do curso e um agradecimento especial ao meu orientador, professor Dr. Marcus Guelpeli, pela dedicação, paciência e pelas orientações valiosas que foram fundamentais para a realização deste trabalho. E, acima de tudo, agradeço a Deus, por me conceder força, sabedoria e oportunidades que tornaram possível a concretização deste sonho.

RESUMO

Este estudo examina de forma sistemática o impacto da remoção de *stopwords* na qualidade dos agrupamentos textuais gerados pelo modelo Cassiopeia. Utilizou-se um corpus composto por artigos científicos da área de Química e um protocolo experimental pareado. Nesse protocolo, para cada combinação de sumarizador (IMP-Py, PragmaSUM e um sumarizador randômico) e nível de compressão (50%, 70% e 90%), foram analisadas as versões com e sem a filtragem de *stopwords*. A análise das representações textuais e dos agrupamentos foi conduzida por meio de métrica interna, como o coeficiente de *Silhouette*, coesão e acoplamento. Os resultados indicam que, em compressões moderadas (50% e 70%), as variações médias no *Silhouette* entre as duas variantes são mínimas e não afetam de forma significativa a qualidade do agrupamento. As alterações mais relevantes foram observadas apenas na compressão extrema (90%), particularmente ao empregar o sumarizador randômico. Nesse cenário, a eliminação de *stopwords* intensificou a diferenciação entre os *clusters*, porém de forma dependente da qualidade do resumo gerado. Ademais, notou-se que as variações nos valores de coesão e acoplamento nem sempre levam a melhorias no *Silhouette*, sugerindo que essa métrica reflete outros aspectos da estrutura dos agrupamentos. De modo geral, os resultados sugerem que o Cassiopeia é bastante resistente à remoção de *stopwords* nas condições analisadas. Assim, a aplicação dessa filtragem deve ser vista como uma opção experimental ou operacional, fundamentada em evidências empíricas e critérios de eficiência computacional.

Palavras-chave: *stopwords*; sumarização automática; Cassiopeia; clusterização; coeficiente *Silhouette*; pré-processamento textual.

ABSTRACT

This study systematically examines the impact of stopword removal on the quality of textual clusters generated by the Cassiopeia model. A corpus composed of scientific articles in the field of Chemistry and a paired experimental protocol were used. In this protocol, for each combination of summarizer (IMP-Py, PragmaSUM, and a random summarizer) and compression level (50%, 70%, and 90%), versions with and without stopword filtering were analyzed. The analysis of textual representations and clusters was conducted using internal metrics such as the Silhouette coefficient, cohesion, and coupling. The results indicate that, at moderate compressions (50% and 70%), the average variations in Silhouette between the two variants are minimal and do not significantly affect the quality of the cluster. The most relevant changes were observed only at extreme compression (90%), particularly when using the random summarizer. In this scenario, the elimination of stopwords intensified the differentiation between clusters, but in a way that depended on the quality of the generated summary. Furthermore, it was noted that variations in cohesion and coupling values do not always lead to improvements in Silhouette, suggesting that this metric reflects other aspects of the cluster structure. Overall, the results indicate that Cassiopeia is quite resistant to stopword removal under the analyzed conditions. Thus, the application of this filtering should be seen as an experimental or operational option, based on empirical evidence and computational efficiency criteria.

Keywords: stopwords; automatic summarization; Cassiopeia; clustering; Silhouette coefficient; text preprocessing.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO

A aplicação conjunta de técnicas de sumarização automática e clusterização tem se revelado uma estratégia eficiente para organizar coleções de textos científicos e diminuir a excessiva quantidade de informações. Em domínios específicos, como o conjunto de artigos em Química analisado neste estudo, a aplicação de sumarizadores e métodos de pré-processamento (como a eliminação de *stopwords*) altera a distribuição das frequências lexicais e, consequentemente, a representação que embasa os algoritmos de agrupamento (Salton e McGill, 1983; Manning, Raghavan e Schütze, 2008). Assim, escolhas que aparentam ser simples no pré-processamento podem influenciar consideravelmente a qualidade das partições resultantes, exigindo, portanto, uma análise empírica minuciosa.

A motivação para este estudo decorre da necessidade prática de orientar as escolhas metodológicas em *pipelines* que combinam sumarização e clusterização: a remoção de lexicais e os limiares de frequência visam reduzir o ruído e o custo computacional. No entanto, seu efeito nas métricas de qualidade depende em grande parte da interação entre o sumarizador, o domínio e o nível de compressão aplicado aos textos. Assim, é necessário investigar de maneira sistemática quando a filtragem lexical contribui para aprimorar o agrupamento e quando ela atua apenas como um mecanismo de redução do espaço amostral.

A questão principal em investigação é a seguinte: a eliminação de *stopwords*, realizada após a sumarização, altera de maneira prática e consistente a qualidade dos agrupamentos gerados pelo modelo Cassiopeia (Guelpeli, 2012) quando o *corpus* é submetido a diversos níveis de compressão e tratado por diferentes sumarizadores? Para investigar essa questão, foram formuladas hipóteses estatísticas explícitas, que seriam examinadas de forma pareada para cada combinação de sumarizador e compressão. Na abordagem clássica, considerou-se a Hipótese Nula (H_0), não há diferença estatisticamente significativa nas medianas das métricas Silhouette nas condições 'S' e 'C' e a Hipótese Alternativa (H_1), há uma diferença estatisticamente significativa entre as medianas da métrica nas condições 'S' e 'C':

$$H_{0,\,dif}: \mu_S - \mu_C = 0$$

e a hipótese alternativa associada à diferença:

$$H_{1,\,dif}: \mu_S - \mu_C \neq 0$$

Com base nessas suposições, a contribuição do estudo consiste em (i) propor e aplicar um protocolo experimental pareado em 120 artigos científicos; (ii) examinar três níveis de compressão e três tipos de sumarizadores; (iii) avaliar o impacto da filtragem usando métrica interna; e (iv) oferecer diretrizes práticas para o uso de *stoplists* em processos de sumarização e clusterização. Os experimentos foram conduzidos no ambiente Cassiopeia v.17 (Guelpeli, 2012), com análises comparativas efetuadas pelo sumarizador e pelo grau de compressão.

O objetivo final é oferecer evidências empíricas que ajudem a tomar decisões metodológicas: quando a remoção de *stopwords* deve ser usada como medida de eficiência, quando deve ser tratada como variável experimental e quando deve ser evitada para preservar a informação discriminativa.

2. ESTADO DA ARTE

2.1. Clusterização

A clusterização de dados é a técnica de agrupar documentos em conjuntos (*clusters*) de forma que os itens de um mesmo conjunto sejam mais parecidos entre si do que com os itens de outros conjuntos, conforme definido por Manning et al. (2008). Os autores afirmam que, no âmbito da Recuperação de Informação e Mineração de Textos, essa técnica tem como objetivo organizar coleções não estruturadas, simplificar a navegação, detectar tópicos emergentes e diminuir a sobrecarga de informações para usuários e sistemas automatizados.

Em condições de funcionamento, conforme os autores, a clusterização textual normalmente abrange as seguintes etapas: (i) pré-processamento (normalização, tokenização, remoção de ruído e, opcionalmente, remoção de *stopwords*); (ii) processamento (construção de vetores de características por TF–IDF, modelos probabilísticos ou *embeddings* distribuídos); (iii) métrica de similaridade (cosseno, distância Euclidiana, medidas baseadas em sobreposição de termos ou em espaço denso); (iv) algoritmo de agrupamento (particional, hierárquico, baseado em densidade ou métodos híbridos); e (v) pós-processamento (métricas internas e externas).

A clusterização textual enfrenta desafios específicos tanto do ponto de vista teórico quanto prático, como a alta dimensionalidade e esparsidade da matriz termo-documento, ambiguidade lexical, diversidade de gêneros e estilos textuais, além da necessidade de escalabilidade em coleções de grande porte. Esses obstáculos levaram ao desenvolvimento de estratégias complementares, como redução dimensional, modelagem de tópicos, representações densas aprendidas e abordagens híbridas que incluem sumarização e seleção de atributos antes do agrupamento, conforme discutido por Blei, Ng e Jordan (2003) e Reimers e Gurevych (2019).

2.2. Cassiopeia

O modelo Cassiopeia, proposto por Guelpeli (2012), propõe uma estratégia poli-estrutural para o agrupamento automático de textos. Ele incorpora a sumarização automática no pré-processamento e emprega um método de agrupamento hierárquico. A meta é reduzir a alta dimensionalidade, diminuir a esparsidade dos dados e aumentar a robustez do modelo em vários domínios e idiomas. A principal motivação é reduzir a dependência de domínio e idioma presente em muitos agrupadores textuais convencionais, aprimorando as

métricas de recuperação de informação (*precision*/*recall*/*F-measure*) e a métrica interna de qualidade (coesão, acoplamento e coeficiente *Silhouette*).

No pré-processamento, utiliza-se métodos tradicionais de limpeza (como *case-folding*, remoção de figuras/tabelas/marcações) e, de forma importante, inclui a sumarização automática dos textos-fonte para diminuir a quantidade de sentenças e atributos. No processamento, cada documento resumido é representado por um vetor limitado (truncado) de 50 posições, organizado de acordo com a frequência relativa dos termos.

Com base nesses vetores, o modelo executa a identificação e a seleção de atributos, calculando a média de frequência e aplicando um corte centrado na distribuição, um "corte médio" fundamentado e aprimorado do corte de *Luhn* (Luhn, 1958). Empregando centróides - vetores de 50 palavras que sintetizam cada grupo - para representar os agrupamentos e estabelecer regras de fusão que mantêm as 25 palavras mais comuns de cada centróide ao unir os agrupamentos, gerando centróides consolidados com 50 termos.

Na fase de agrupamento, o modelo utiliza o método hierárquico aglomerativo, juntamente com uma versão adaptada do algoritmo *Cliques*, a fim de assegurar que os textos em um mesmo grupo possuam um nível mínimo de similaridade. A abordagem em duas etapas - inicialmente agrupando textos com base nos vetores e, posteriormente, reagrupando utilizando centróides - permite a criação de uma estrutura hierárquica explorável (grupos e subgrupos), preservando, para cada agrupamento, tanto os textos originais quanto seus respectivos resumos, assim descrito por Guelpeli (2012).

No pós-processamento, o modelo fornece uma hierarquia de tópicos que inclui os textos agrupados e seus respectivos resumos, com um alto nível de informatividade. Essa arquitetura simplifica a busca por documentos específicos e a generalização para documentos com temas relacionados, diminuindo a sobrecarga de informações durante a recuperação. Guelpeli (2012) sugere um novo critério de corte para a seleção de atributos: um corte médio na distribuição de frequências. Quando esse critério é combinado com a sumarização e a representação por centróides, o modelo se torna menos dependente do domínio e do idioma.

Por fim, o Cassiopeia (Guelpeli, 2012) é descrito como um modelo não supervisionado que, de acordo com os testes realizados, obteve melhorias consideráveis em precisão, recuperação e medidas internas de qualidade. Isso valida a hipótese de que a inclusão da sumarização no pré-processamento e o corte médio no processamento aumentam a qualidade do agrupamento de documentos em domínios heterogêneos.

2.2.1. Métrica de avaliação

Neste estudo, o coeficiente de *Silhouette* foi empregado como a principal métrica de avaliação. Considerado uma métrica interna utilizada na análise de agrupamentos, que avalia a coesão de cada objeto i dentro de seu cluster e a distinção em relação aos demais *clusters*. De acordo com Guelpeli (2012), o coeficiente de *Silhouette* para i é calculado pela Equação (1), em que $a(i)$ denota a distância média entre i e os outros elementos de seu *cluster*, e $b(i)$ indica a menor distância média entre i e os elementos de qualquer outro *cluster*.

Enquanto isso, em termos interpretativos, a coesão diz respeito à união interna do cluster: quanto menor for $a(i)$, mais coeso será o grupo ao qual i pertence. Por outro lado, acoplamento diz respeito ao isolamento do *cluster* em relação aos outros: quanto maior for $b(i)$, maior será a distância entre i e o *cluster* mais próximo, o que indica uma melhor separação entre os grupos. Dessa forma, valores de *Silhouette* próximos a +1 indicam alta coesão e boa separação, enquanto valores próximos a 0 sinalizam pontos na fronteira entre *clusters*. É importante observar que $a(i)$ e $b(i)$ dependem da métrica de distância utilizada.

Equação (1):

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)} \quad (1)$$

Valores próximos de zero indicam fronteiras ambíguas, enquanto valores negativos sugerem possíveis atribuições errôneas. Na prática experimental, o coeficiente *Silhouette* é utilizado como uma medida complementar à métrica externa (como *Recall*, *Precision* e *F-Measure*), oferecendo uma visão não supervisionada da qualidade estrutural dos agrupamentos. O *Silhouette* é calculado para cada instância dos agrupamentos, e a média das observações é frequentemente apresentada para reduzir a variabilidade decorrente de inicializações aleatórias ou múltiplas iterações.

A interpretação dos resultados precisa ser contextualizada: os ganhos em *Silhouette* indicam não apenas maior compacidade dos grupos, mas também melhor isolamento entre eles. Isso faz com que essa métrica seja adequada para comparar diferentes variantes de pré-processamento (como textos-fonte e textos sumarizados), técnicas de seleção de atributos e ajustes do agrupador. Por outro lado, a forma geométrica dos *clusters* e o domínio dos documentos afetam o valor absoluto do *Silhouette*.

Nos experimentos conduzidos, o coeficiente *Silhouette* foi empregado como a principal medida interna: as pontuações foram computados para cada execução e exibidos

como médias finais acumuladas ao longo das repetições (50 iterações por configuração). Isso permitiu a comparação tanto dos métodos de seleção de atributos quanto dos diferentes sumarizadores e taxas de compressão.

Por fim, recomenda-se utilizar o *Silhouette* em conjunto com análises qualitativas e métricas externas sempre que possível, uma vez que essa combinação oferece um diagnóstico mais robusto da coerência interna dos clusters e da fidelidade à estrutura esperada dos dados.

2.3. Sumarização e Nível de Compressão

Nesta seção, apresenta-se uma definição geral de sumarização e nível de compressão textual, bem como suas características fundamentais, a fim de facilitar a compreensão ao longo do texto, de acordo com o enquadramento proposto por Rino e Pardo (2002). A sumarização consiste em criar um texto condensado a partir de um texto-fonte, preservando as ideias centrais e as unidades informativas essenciais, além de garantir a coerência necessária para o objetivo comunicativo pretendido. Na prática, isso significa identificar unidades significativas no texto e reestruturá-las em um novo enredo que mantenha o significado original.

O nível de compressão refere-se à quantidade de redução aplicada ao texto original para criar o sumário. Formalmente, é o quociente entre o tamanho do sumário e o tamanho do texto original. Além disso, atua como um parâmetro que influencia as decisões de seleção e geração, bem como a avaliação da informatividade do resultado. Na pesquisa, sua definição e controle possibilitam a comparação de métodos e o ajuste do equilíbrio entre retenção de conteúdo, coerência e objetivo comunicativo.

A próxima seção fornece uma descrição concisa dos sumarizadores utilizados neste estudo: o IMP-Py, um sumarizador implementado em Python pelo autor; o PragmaSUM proposto por Rocha e Guelpeli (2017); e um sumarizador randômico. Serão abordadas suas abordagens, componentes e parâmetros experimentais.

2.3.1. Sumarizador implementado em Python (IMP-Py)

O IMP-Py é um sumarizador extrativo desenvolvido de autoria própria e implementado em Python 3.12.3, fundamentado no algoritmo *TextRank*. Ele pode ser utilizado por meio da biblioteca *Sumy*[1], possuindo uma interface de linha de comando para processamento em lote. Estruturalmente, possui três componentes fundamentais: (i) *summarize_text*(*text*, *compression*), que aceita um texto e um parâmetro de compressão,

[1] *SUMY developers*. *Sumy* (v. 0.11.0). Disponível em: https://pypi.org/project/sumy/

retornando um resumo; (ii) *process_files*(*input_dir*, *output_dir*, *compression*), que percorre arquivos .txt (UTF-8), aplica a sumarização e registra os resultados; e (iii) o bloco principal, responsável pelo parsing de argumentos com argparse e pela validação do parâmetro de compressão (intervalo permitido).

A função *summarize_text* gera um *PlaintextParser* com tokenização em português, utiliza o sumarizador *TextRank* e determina a quantidade de sentenças do resumo de acordo com o parâmetro de compressão, observando limites (mínimo de uma sentença e máximo do total disponível); se o resumo calculado não diminuir o texto, o texto original é retornado. Para manter a coerência, as sentenças escolhidas são reorganizadas de acordo com sua posição original no texto antes de serem incluídas no resumo final.

Ao processar arquivos, o código garante que o diretório de saída exista, lê cada arquivo .txt em UTF-8 e salva o resumo com nomes de arquivo que mostram o arquivo original e o nível de compressão, o que facilita o rastreamento. O tratamento de entradas vazias e a validação básica de parâmetros são abordados; no entanto, operações de E/S não dispõem de um tratamento de exceções robusto, o processamento é sequencial (sem possibilidade de paralelização) e não há suporte para formatos além de .txt, aspectos que restringem a escalabilidade e a resiliência.

2.3.2. Sumarizador PragmaSUM

O PragmaSUM, desenvolvido por Rocha e Guelpeli (2017), é um sistema de sumarização automática do tipo extractivo que escolhe sentenças representativas do texto original sem modificá-las. Ele foi projetado para funcionar de forma independente, independentemente do idioma ou do domínio do documento.

Suas principais funcionalidades são: (i) utilização do modelo de clusterização Cassiopeia (Guelpeli, 2012) em conjunto com critérios estatísticos para avaliação de termos e sentenças; (ii) implementação em Java com truncamento de vetores (*cut-off* de até 50 posições) para diminuir a dimensionalidade; (iii) opção de personalizar o resumo com até cinco palavras-chave ponderadas pelo usuário; e (iv) controle da razão de compressão para determinar a quantidade de sentenças extraídas.

Rocha e Guelpeli (2017) analisaram o PragmaSUM em um conjunto de 500 artigos científicos em português, abrangendo 10 domínios, com o uso das métricas *ROUGE*. Os autores notaram um aumento na precisão quando o método utiliza palavras-chave, especialmente em níveis mais altos de compressão. Isso faz com que o PragmaSUM seja

apropriado para resumir artigos, notícias e para sistemas que necessitam de resumos personalizados.

2.3.3. Sumarizador Randômico

Uma técnica básica de resumo automático é o sumarizador randômico, que elabora um resumo ao selecionar aleatoriamente algumas sentenças do texto original. Suponha que temos um texto composto por diversas sentenças e desejamos escolher um número reduzido delas para elaborar um resumo. O método seleciona essas sentenças de forma aleatória e sem repetições. Ele pode fazer isso de maneira completamente uniforme, oferecendo a mesma probabilidade para todas as sentenças, ou pode aplicar critérios que atribuam mais relevância a sentenças vistas como mais importantes, como aquelas que aparecem no início do texto ou que são mais extensas. De acordo com Gambhir e Gupta (2017), após selecionar as sentenças, elas costumam ser reordenadas na sequência em que aparecem no texto original para preservar a coerência local.

Esse tipo de sumarizador, por ser fundamentado em escolhas aleatórias, não analisa o significado do texto nem busca identificar as partes mais importantes. Portanto, é usado principalmente como uma referência básica para comparação com métodos de sumarização mais sofisticados, como discutido por Gambhir e Gupta (2017) e Nenkova e McKeown (2012). Sob amostragem uniforme, a expectativa da sobreposição - que segue distribuição hipergeométrica - com um resumo de referência O é dada pela Equação (2), o que fornece uma base analítica para interpretar medidas de similaridade como *ROUGE* e para construir testes estatísticos que verificam se um método supera o acaso, conforme Ross (2014).

Equação (2):

$$E[|R \cap O|] = k \cdot \left(\frac{|O|}{n}\right) \quad (2)$$

Entretanto, o sumarizador randômico não leva em conta o conteúdo, a coerência nem evita repetições, o que resulta em resumos que costumam ser pouco práticos, confusos ou com pouca informação. Além disso, em textos em que a informação está concentrada no início, a simples seleção das primeiras sentenças pode dar uma falsa impressão de qualidade. Por esse motivo, Mani (2001) e Nenkova (2012) aconselham cautela ao usar esses métodos como *baseline* em avaliações.

2.4. *Stopwords*

A remoção de *stopwords* é uma das etapas de pré-processamento mais comuns em *pipelines* de processamento de linguagem natural (PLN). De forma prática, elas se manifestam como *tokens* de alta frequência e baixo valor informativo em atividades que demandam a ocorrência lexical, como artigos, preposições, pronomes e conjunções. A remoção de *stopwords* tem como objetivo principal reduzir o ruído e a dimensionalidade, melhorando assim a capacidade discriminatória das representações textuais e a eficiência computacional dos modelos, conforme apontado por Salton e McGill (1983) e Manning *et al.* (2008).

Historicamente, a utilização de listas fixas (*stoplists*, conjunto que contém as *stopwords*) tem sido defendida devido ao comportamento indesejado de termos funcionais em esquemas de ponderação simples. No entanto, a prática evoluiu para abordagens mais adaptativas: além de listas manuais e genéricas disponibilizadas em bibliotecas, são desenvolvidos métodos baseados em estatísticas do próprio *corpus*, filtragem por classe gramatical e seleção condicional por meio de métricas como TF-IDF médio ou entropia do *token*, conforme discutido por Manning *et al.* (2008).

Embora a literatura clássica associe a remoção de *stopwords* à melhoria de desempenho em tarefas que usam representações esparsas, conforme Salton e McGill (1983) e Manning, Raghavan e Schütze (2008), pesquisas mais recentes indicam que, em modelos baseados em representações distribuídas, esse impacto tende a ser menor, como demonstrado por Mikolov *et al.* (2013) e Devlin *et al.* (2019). De fato, os pesos e as dependências contextuais desses modelos são capazes de reduzir o impacto dos *tokens* funcionais. Isso implica que a eliminação de *stopwords* é uma escolha condicionada ao contexto experimental, e não uma exigência universal.

No presente estudo, os resultados empíricos corroboram essa tendência: a eliminação de *stopwords* não resultou em melhorias significativas nas métricas de desempenho, indicando que, em determinados contextos, essa etapa pode ser dispensável. Esse resultado enfatiza a necessidade de avaliar empiricamente cada etapa do pré-processamento conforme o tipo de representação empregado, evitando generalizações fundamentadas apenas em tradições metodológicas.

Ao traduzir para o português, é fundamental considerar fenômenos morfológicos e sintáticos, como a incorporação de formas contraídas (do/da/no/na), o tratamento de clíticos e as variações regionais, além de garantir que a tokenização preserve unidades significativas

antes da filtragem. Ferramentas e listas genéricas podem servir como ponto de partida, porém devem ser adaptadas com base na análise de frequência e validação qualitativa no corpus específico.

Por fim, os avanços mais recentes indicam tendências promissoras: (a) integração de sinais linguísticos com métricas estatísticas para filtragem contextualizada; (b) uso de seleção de características supervisionada para identificar *tokens* pouco informativos em tarefas específicas; e (c) estratégias de *down-weighting* (em vez de remoção completa) que preservam a estrutura sintática enquanto reduzem a influência estatística de *tokens* funcionais.

3. METODOLOGIA

Neste estudo, a remoção de *stopwords* foi tratada como uma variável experimental controlada de maneira explícita no fluxo do modelo Cassiopeia (Guelpeli, 2012) modificado (Figura 1). A filtragem ocorre após a sumarização e normalização. Assim, conforme estabelecido por Guelpeli (2012), a operação impacta diretamente a distribuição de frequências, que está sujeita ao corte de *Luhn* (Luhn, 1958). Para cada configuração de compressão, foram executadas duas versões do *pipeline* conforme apresentado pela Figura 1: (A) com remoção de *stopwords* (seta pontilhada) e (B) sem remoção de *stopwords* (seta bidirecional), para que o impacto da filtragem pudesse ser isolado.

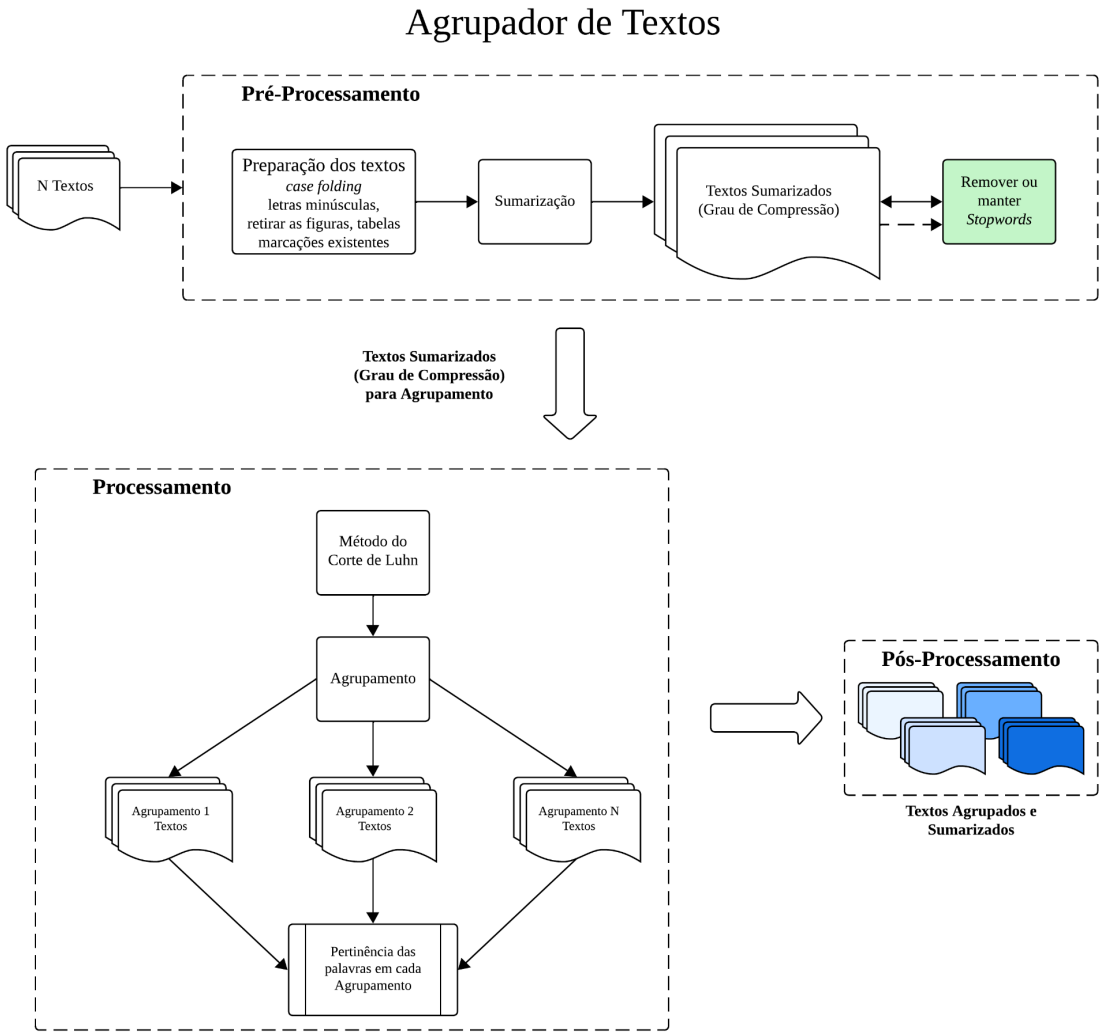

Figura 1 - Modelo Cassiopeia com pré-processamento adaptado para *stopwords*.
Fonte: Guelpeli, 2012

As medições abrangem métricas internas, como coesão, acoplamento e o coeficiente *Silhouette*; e métricas externas, como o *Precision*, *Recall* e *F-Measure*, onde ambas métricas

são medidas harmônicas; e indicadores operacionais, como o tamanho do vocabulário resultante. Os resultados são apresentados para cada combinação de compressão e estratégia de *stopwords*, permitindo uma comparação direta entre as versões com remoção e sem remoção.

O teste de Friedman proposto em 1937, (estatística x^2) foi empregado para medir o grau de concordância entre as classificações dadas às seis condições experimentais (50%C, 50%S, 70%C, 70%S, 90%C, 90%S). O coeficiente de concordância de Kendall (W) (Kendall, M.G. and Babington Smith, B., 1939) foi calculado a partir de x^2, conforme a seguinte relação:

$$W = \frac{x^2}{n(k-1)}, \quad (3)$$

em que n é o número de blocos (documentos) e k é o número de condições. Os cálculos foram realizados utilizando o Google Planilhas, valores de W próximos a 0 indicam baixa concordância nas ordens, ao passo que valores próximos a 1 sinalizam alta concordância. Um valor de α de 0,05 foi utilizado nas decisões de significância.

3.1. *Corpus*

Em relação à base textual, empregou-se um *corpus* composto por 120 artigos científicos na área de Química, coletados previamente e integralmente pré-processados por Amariz e Guelpeli (2024), os quais disponibilizaram a coleção para esta pesquisa. Os autores executaram um pré-processamento que incluiu a conversão para o formato de texto simples (.txt) com codificação UTF-8, tokenização, eliminação de caracteres não imprimíveis, normalização de espaços e outras etapas de normalização necessárias para garantir a uniformidade no tratamento inicial dos dados. Nesta pesquisa, empregou-se a versão do *corpus* disponibilizada por eles, efetuando apenas a verificação de integridade e a normalização dos metadados quando necessário.

3.2. Ferramentas e ambiente experimental

As execuções foram realizadas no ambiente do programa Cassiopeia v.17, utilizado para orquestrar *pipelines* de pré-processamento e de clusterização (Figura 1).

3.3. Configurações testadas

Para começar, quatro configurações de processamento textual foram estabelecidas diretamente no modelo Cassiopeia (Guelpeli, 2012) para possibilitar comparações diretas entre as estratégias de pré-processamento:

1. Clusterização pura: implementação do algoritmo de clusterização de forma direta no texto completo;
2. Sumarização seguida de Clusterização: elaboração prévia de resumos pelo próprio modelo, seguida da fase de agrupamento;
3. Remoção de *stopwords* seguida de clusterização: uso de *stopwords* prévio ao agrupamento; e
4. A combinação da condensação e da filtragem léxica é a sumarização, seguida da remoção de *stopwords* e da clusterização.

A organização e a sequência dessas configurações, bem como a posição da operação de remoção de *stopwords* no fluxo, estão esquematizadas na Figura 1. Para os experimentos descritos neste estudo, o *corpus* foi previamente resumido por três sumarizadores distintos em três níveis de compressão, que serão explicados na próxima seção. Em cada versão resumida do *corpus*, apenas duas configurações de processamento foram executadas: (1) clusterização pura; e (3) remoção de *stopwords* seguida de clusterização. As comparações entre políticas (com/sem remoção de *stopwords*) são realizadas de forma pareada para cada combinação, assegurando que eventuais diferenças sejam avaliadas considerando o sumarizador e o nível de compressão.

3.4. Sumarização externa

Na etapa complementar, os métodos anteriores foram reduzidos pela metade, mantendo apenas a clusterização pura e a remoção ou não das *stopwords* seguida de clusterização. Dessa forma, três sumarizadores externos ao Cassiopeia foram incorporados ao experimento: (i) um sumarizador desenvolvido pelo autor em Python, (ii) a ferramenta PragmaSUM, implementada por Rocha e Guelpeli (2017) e (iii) um sumarizador randômico, utilizado como *baseline* probabilístico. Para investigar como o processo de clusterização reage à quantidade de conteúdo preservada pelo resumo, cada sumarizador foi ajustado para funcionar em três níveis de compressão (50%, 70% e 90%).

A seleção dos sumarizadores para este estudo foi baseada nos seguintes critérios: (i) adequação ao Português e suporte à tokenização/segmentação em língua portuguesa; (ii)

representatividade de paradigmas - incluir um método baseado em grafos (*TextRank*), um método estatístico/híbrido (PragmaSUM) e um *baseline* probabilístico (randômico) permite comparar mecanicamente diferentes fontes; (iii) parametrização por compressão, essencial para avaliar desempenho em múltiplos níveis de compressão; e (iv) reprodutibilidade e operacionalização, visando implementações com execução em lote, registro de parâmetros e formatos padronizados (UTF-8 .txt).

Em resumo, a combinação escolhida garante a cobertura conceitual, a viabilidade experimental e o rigor avaliativo, cumprindo diretamente os objetivos de comparar a qualidade informativa, a robustez por domínio e a sensibilidade à compressão na coleção analisada.

4. RESULTADOS

Para cada combinação de sumarizador e nível de compressão testados, as variantes do *pipeline* com e sem remoção de *stopwords* foram comparadas de forma pareada e reprodutível. A seguir, os resultados mais relevantes são apresentados, agrupados por blocos temáticos.

4.1. Qualidade geral

Em relação ao coeficiente *Silhouette* médio acumulado, as diferenças entre as versões **com** e **sem** remoção de *stopwords* foram, de modo geral, insignificantes. Ao levar em conta todas as execuções, a variação média absoluta observada no *Silhouette* foi de apenas alguns centésimos. Em diversas combinações de método e compressão, as variações foram inferiores a aproximadamente de 0,01, valor próximo da resolução experimental, o que indica que não houve impacto prático significativo.

Vendo a análise do coeficiente *Silhouette* foi realizada tanto individualmente para cada nível de compressão (50%, 70% e 90%) quanto de forma resumida, considerando as médias das 50 execuções. Como ilustrado na Figura 2.a, para uma compressão de 50%, as médias de *Silhouette* nas variantes com e sem remoção de *stopwords* são bastante semelhantes. Para uma compressão de 70%, observa-se uma maior variabilidade entre os sumarizadores, com variações de apenas algumas centésimas entre os métodos com e sem *stopwords* (Figura 2.b). Para uma compressão de 90%, as variações se tornam mais pronunciadas em casos particulares (sumarizador randômico), mas ainda são afetadas pelo sumarizador, como esquematizado na Figura 2.c e na Tabela 3 posteriormente para todas as compressões.

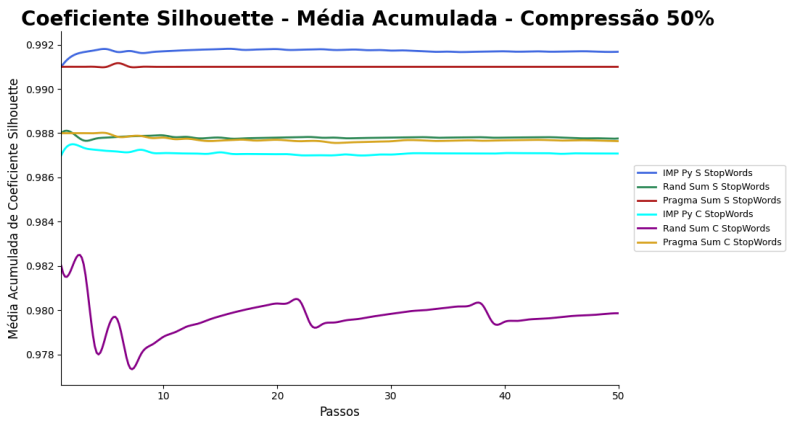

Figura 2.a - Resultados obtidos pelo modelo Cassiopeia, média acumulada Coeficiente *Silhouette* com 50% de compressão**.**

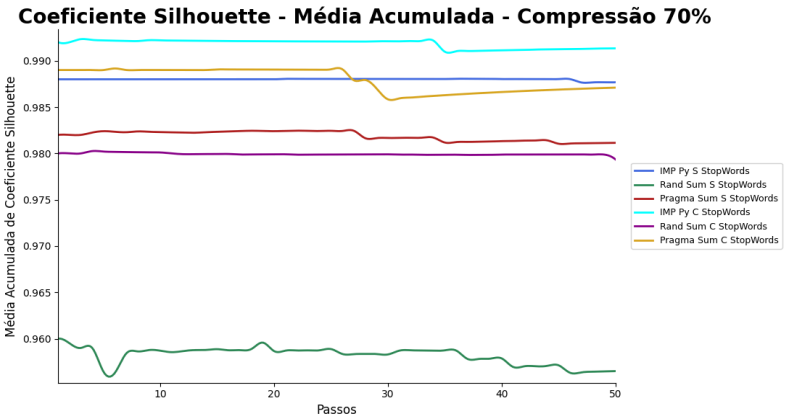

Figura 2.b - Resultados obtidos pelo modelo Cassiopeia, média acumulada Coeficiente *Silhouette* com 70% de compressão**.**

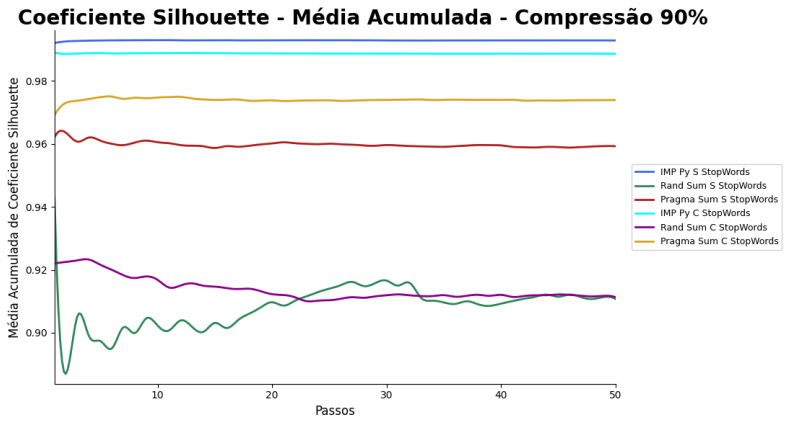

Figura 2.c - Resultados obtidos pelo modelo Cassiopeia, média acumulada Coeficiente *Silhouette* com 90% de compressão**.**

Em termos interpretativos, embora existam variações pontuais, a evidência geral sugere que a média acumulada do valor de *Silhouette* não apresenta mudanças significativas com a remoção de *stopwords* no modelo, nas condições analisadas. Em resumo, o *Silhouette* sugere que a qualidade geral dos agrupamentos permanece estável na prática, independentemente da utilização da *stopwords*, exceto em casos excepcionais que dependem da qualidade do sumarizador em condições de compressão extrema.

A Tabela 1 apresenta os desvios-padrão do coeficiente *Silhouette* por método e nível de compressão, distinguindo as variantes sem (S) e com (C) remoção de *stopwords*, além do delta (Δ) (diferença entre S e C). Observa-se que, para compressões moderadas (50% e 70%), as variações entre as duas condições são, em regra, menores, indicando que o modelo exibe um comportamento semelhante, a despeito da filtragem lexical. Por exemplo, no método

IMP-Py a 50%, os desvios-padrão das duas condições permanecem próximos (0,00048 para S e 0,00053 para C), resultando em um delta (Δ) de pequena magnitude (-0,00005).

Tabela 1 - Desvio Padrão para cada sumarizador, compressão e condição (S e C)

	Compressão (%)	**Silhouette DP (S)**	**Silhouette DP (C)**	**Δ (S-C)**
IMP-Py		0,00048	0,00053	-0,00005
RandSum	50	0,0004	0,0069	-0,0065
PragmaSUM		0,0002	0,0006	-0,0004
IMP-Py		0,0060	0,0024	0,0036
RandSum	70	0,0037	0,0102	-0,0065
PragmaSUM		0,0078	0,0048	0,0031
IMP-Py		0,0004	0,0005	-0,0001
RandSum	90	0,0388	0,0118	0,0270
PragmaSUM		0,0031	0,0060	-0,0029

Fonte: elaboração própria
Legenda: (S = sem stopwords; C = com stopwords)

Em contrapartida, na compressão extrema (90%), alguns métodos mostram variações mais notáveis, particularmente o sumarizador randômico, que apresenta um delta (Δ) mais alto entre as condições S e C. Esse comportamento sugere que a remoção de *stopwords* pode intensificar as flutuações na estabilidade dos agrupamentos quando o texto é consideravelmente reduzido, particularmente em resumos com qualidade informativa inferior. De maneira geral, os valores de delta (Δ) não apresentam uma tendência consistente entre os métodos e graus de compressão. Isso corrobora a noção de que, nas condições experimentais estudadas, o modelo Cassiopeia é, de fato, bastante indiferente à eliminação de *stopwords*.

4.2. Coesão e acoplamento

As métricas auxiliares de coesão (compacidade interna) e acoplamento (separação relativa) exibiram variações entre as variantes, com as alterações na coesão sendo mais notórias em certos contextos. Entretanto, essas alterações não levaram de maneira consistente a melhorias no *Silhouette* nem a ganhos práticos na qualidade das partições. Em outras

palavras, mudanças locais em coesão/acoplamento foram feitas sem que houvesse um benefício consistente e aplicável a todos os casos na avaliação final.

Para compreender como a filtragem lexical altera a geometria dos *clusters*, foram examinadas as métricas internas de coesão e acoplamento. Na remoção de *stopwords* em 50%, notaram-se alterações relevantes na coesão em certos métodos (Figura 4.a), enquanto o acoplamento exibiu variações menos significativas (Figura 3.a). Podemos ver um comportamento diverso em 70% das situações: a filtragem melhorou a coesão em algumas combinações, mas a diminuiu em outras, sem seguir um padrão claro, como representado nas Figuras 3.a e 4.b. Em 90% dos casos, as Figuras 3.c e 4.c mostram que a compressão extrema intensifica essas flutuações, em alguns casos, há um aumento significativo da coesão com a remoção, enquanto em outros há uma diminuição da coesão e alterações no acoplamento.

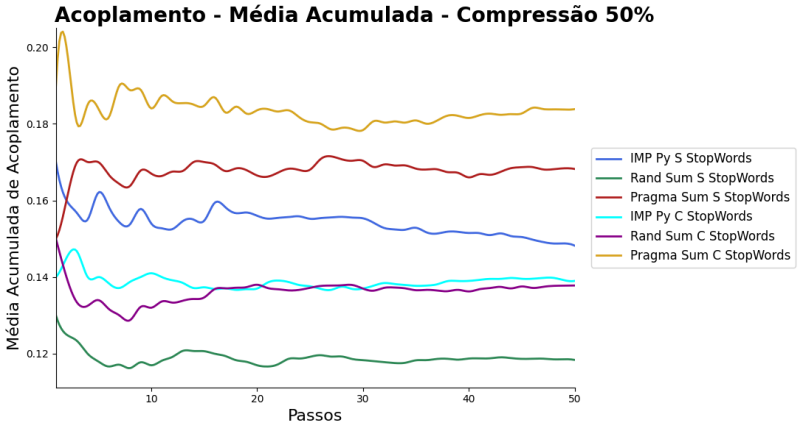

Figura 3.a - Resultados obtidos pelo modelo Cassiopeia, média acumulada de acoplamento com 50% de compressão.

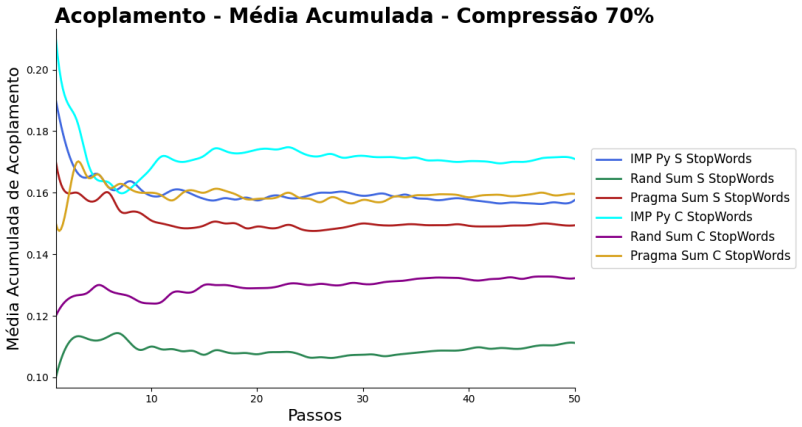

Figura 3.b - Resultados obtidos pelo modelo Cassiopeia, média acumulada de acoplamento com 70% de compressão.

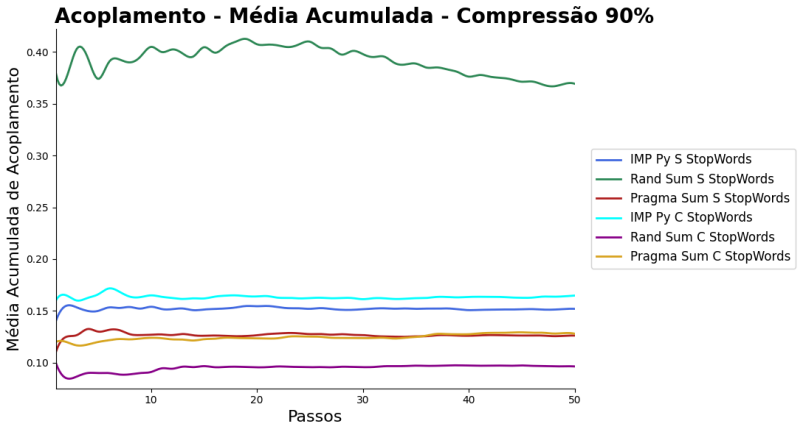

Figura 3.c - Resultados obtidos pelo modelo Cassiopeia, média acumulada de acoplamento com 90% de compressão.

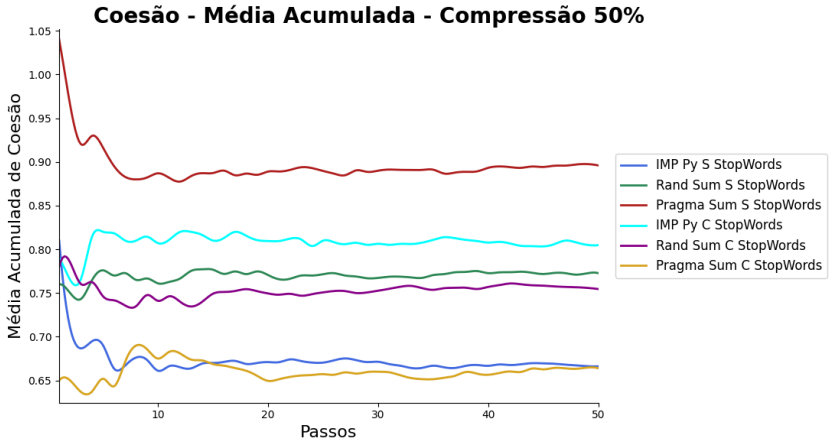

Figura 4.a - Resultados obtidos pelo modelo Cassiopeia, média acumulada de coesão com 50% de compressão.

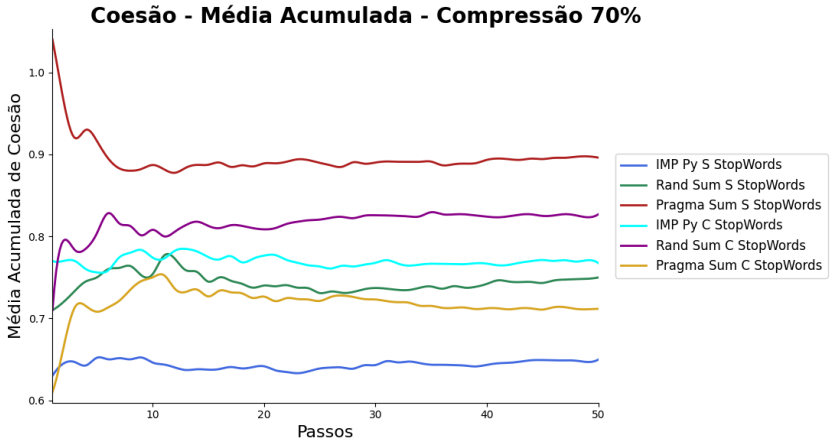

Figura 4.b - Resultados obtidos pelo modelo Cassiopeia, média acumulada de coesão com 70% de compressão.

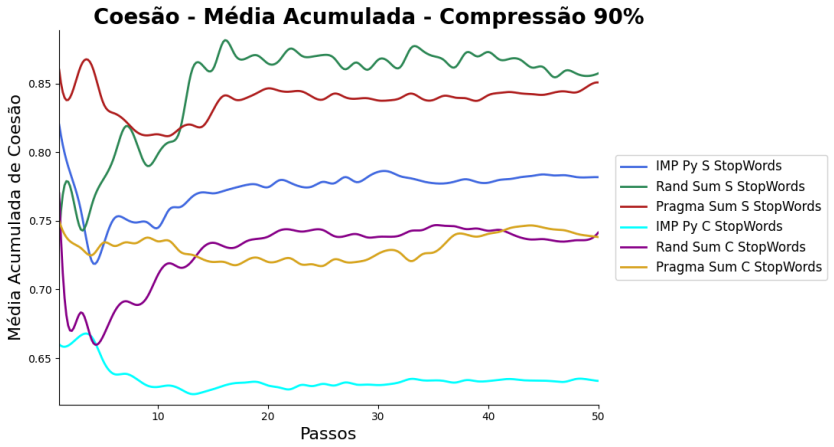

Figura 4.c - Resultados obtidos pelo modelo Cassiopeia, média acumulada de coesão com 90% de compressão.

Embora haja essas variações internas, o ponto principal é que mudanças na coesão e no acoplamento raramente levam a uma mudança prática significativa na qualidade dos agrupamentos, conforme avaliado pelo *Silhouette*. Em suma, a remoção de *stopwords* altera a geometria local dos *clusters*, porém o efeito na separabilidade global - que é o que realmente influencia a performance do *clustering* no Cassiopeia - foi, em geral, irrelevante nas configurações analisadas.

4.3. Comportamento por método e compressão

O efeito da filtragem variou conforme o sumarizador e o grau de compressão: alguns pares exibiram leve melhora com a remoção, outros leve piora, e poucos mostraram diferenças mais relevantes. Não se identificou um padrão consistente que indicasse uma vantagem sistemática de uma política (manter ou remover) quando implementada de forma geral em todos os métodos e níveis de compressão.

4.4. Interpretação geral e conclusão dos resultados

Os resultados indicam que a remoção de *stopwords* não proporciona um benefício consistente. Embora essa operação provoque variações mensuráveis em métricas como coesão, acoplamento e, ocasionalmente, *Silhouette*, não se observa uma melhoria consistente e prática na qualidade dos agrupamentos gerados pelo modelo Cassiopeia (Guelpeli, 2012) nas condições testadas. Essas condições englobam um conjunto de 120 artigos em Química, três sumarizadores e diferentes níveis de compressão.

Pode-se concluir isso graças a análise global por Friedman, que retornou $x^2 = 0,2857$ e $p = 0,9979$. A partir dessa estatística, o coeficiente de concordância de

Kendall apresentou um valor de $W = 0,0011$, conforme demonstrado na Tabela 2. Esses resultados indicam que não existe uma diferença estatisticamente significativa entre as condições e que as ordens apresentam uma baixa concordância. Isso sugere que não há um padrão consistente de desempenho entre as seis condições.

Tabela 2 - Estatística de Friedman e coeficiente de concordância de Kendall (W)

n (simulações)	k (condições)	X^2	df	p-valor	Kendall's W
50	6	0,2857	5	0,9979	0,0011

Fonte: elaboração própria

De forma geral, o modelo utilizado demonstrou resistência à política de *stopwords*, apresentando variações mínimas nas métricas internas e externas na maioria das configurações. Ainda que essas variações tenham sido estatisticamente significativas, seu efeito prático foi restrito. Além disso, observou-se que o efeito da filtragem depende do tipo de sumarizador utilizado e do nível de compressão aplicado. Isso destaca a importância de verificar empiricamente a necessidade de remover *stopwords*, levando em consideração o contexto experimental específico.

Ao levar em consideração a redução dos custos operacionais decorrente da remoção de *stopwords*, a decisão de adotar ou não essa filtragem pode ser guiada por critérios práticos, como tempo de processamento ou restrições de memória, especialmente quando não há necessidade de preservar unidades funcionais para tarefas futuras.

Conclui-se então que os experimentos realizados com o modelo usado indicam que a qualidade dos agrupamentos é praticamente inalterada pela presença ou ausência de *stopwords*, considerando o corpus e as configurações testadas. Embora a filtragem provoque mudanças heterogêneas em aspectos internos, essas alterações não geram ganhos práticos relevantes. Portanto, a exclusão de *stopwords* pode ser considerada uma escolha operacional e experimental, cuja eficácia deve ser validada empiricamente em cada novo domínio.

5. DISCUSSÃO

A análise dos resultados confirma um ponto essencial para a interpretação do experimento: a limitação do espaço amostral - entendida aqui como a redução do vocabulário disponível para pontuação, vetorização e seleção de termos - um elemento fundamental que pode se manifestar tanto pela exclusão explícita de *stopwords* quanto pela definição temática do *corpus*. Apesar de esses dois mecanismos operarem em níveis diferentes, eles convergem no mesmo efeito prático: alteram a distribuição de frequências e, como consequência, alteram quais *tokens* podem ser utilizados em resumos e vetores documentais. Neste estudo, verificamos que, embora houvesse mudanças léxicas, o modelo proposto por Guelpeli em 2012 mostrou, na prática, ser pouco influenciado pela presença ou ausência de *stopwords* nas condições avaliadas. A seguir, a discussão esclarece os motivos dessa solidez e as consequências metodológicas.

O coeficiente de concordância de Kendall calculado globalmente foi muito baixo ($W = 0,0011$), o que mostra que as seis condições para cada nível de compressão não concordam muito entre si ao longo das simulações. Na prática, isso significa que não há um padrão consistente que favoreça uma política (remoção ou não de *stopwords*) ou um nível de compressão; as variações observadas são locais e irregulares. Essa evidência estatística corrobora a conclusão principal deste estudo: o modelo Cassiopeia demonstra robustez em relação à remoção de *stopwords* nas condições analisadas.

5.1. Evidência empírica

Os gráficos que mostram as médias do coeficiente Silhouette por sumarizador em cada nível de compressão oferecem uma evidência clara da sensibilidade do sistema à eliminação de *stopwords*. A Tabela 2 apresenta as médias por método para cada compressão. Essa tabela apresenta que as variantes com e sem *stopwords* exibem diferenças médias mínimas para a compressão de 50%. Na próxima compressão sugere uma maior diversidade entre os sumarizadores para 70%, no entanto, as variações médias continuam sendo de apenas algumas centésimas. Por fim, para 90%, as variações mais notáveis acontecem em casos particulares (especialmente para o sumarizador randômico). Contudo, essas variações também estão relacionadas ao tipo de sumarizador utilizado e não demonstram um efeito consistente.

Tabela 3 - Média do coeficiente Silhouette por método e compressão

	50% (S)	**50% (C)**	**70% (S)**	**70% (S)**	**90% (S)**	**90% (C)**
IMP-Py	**0.9917**	0.9871	0.9880	0.9910	0.9930	0.9890
RandSum	**0.9878**	0.9799	0.9560	0.9790	0.9110	0.9110
PragmaSUM	**0.9910**	0.9876	0.9810	0.9870	0.9590	0.9740

Fonte: elaboração própria
Legenda: (S = sem stopwords; C = com stopwords)

Essas médias por métodos confirmam a principal conclusão prática: o modelo, nas configurações e no corpus avaliados, demonstra robustez - ou seja, na maioria das condições testadas, é insensível à remoção de *stopwords*. As diferenças mais notáveis aparecem em situações de compressão extrema ou em sumarizadores que geram resumos de baixa qualidade, momentos em que a filtragem lexical pode tanto eliminar ruídos quanto, inversamente, remover termos de suporte.

5.2. Corte de frequência e sensibilidade ao domínio

O método de *Luhn* (Luhn, 1958) emprega dois cortes diferentes na distribuição de frequência: um corte inferior e outro superior. Esses cortes definem uma faixa de termos considerados "relevantes" para a sumarização. Essa metodologia é inerentemente sensível ao domínio e ao idioma: limiares fixos ou heurísticos podem eliminar termos significativos de uma área em um *corpus* técnico (como nomes de técnicas ou termos usuais em Química) ou, inversamente, incluir termos que são pouco relevantes para a discriminação no domínio. Na prática, isso implica que a banda selecionada pelo método pode variar consideravelmente de acordo com o *corpus*, e a utilização de uma *stopwords* externa pode ser desnecessária ou, inversamente, remover *tokens* que o corte deveria preservar.

Em contrapartida, o critério de corte médio empregado no modelo usado costuma ser mais estável e menos afetado por extremos na distribuição de frequência. Ao evitar decisões fundamentadas em limites extremos, o corte médio preserva um núcleo lexical mais consistente entre as variantes. Essa variação metodológica explica por que, em nossos experimentos, a inclusão de uma *stopwords* adicional teve um efeito prático limitado: o mecanismo de corte do sumarizador/extrator já realiza parte do trabalho de eliminação de termos de alta frequência ou, quando isso não ocorre, a redução do corte médio do Cassiopeia (Guelpeli, 2012) torna a abordagem menos sensível a alterações marginais no vocabulário.

5.3. Filtragem de *stopwords*: melhoria da qualidade ou apenas redução do volume?

Os dados indicam que, em muitos casos, o uso de uma *stopwords* atua principalmente como um mecanismo de redução de volume (reduzindo o vocabulário e o custo computacional), em vez de ser uma estratégia que, por si só, melhora de maneira consistente a qualidade do agrupamento. Duas observações sustentam essa inferência:

1. Há uma grande sobreposição entre o corte superior de *Luhn* (Luhn, 1958) e *stopwords*, pois o corte superior já elimina termos de altíssima frequência, e a *stopwords* não fornece informações novas, tornando-se, assim, dispensável. Isso explica a semelhança das médias por método nos percentuais de 50% e 70% (Tabela 3 e Figura 2.a-2.b).
2. Amplificação em compressões extremas: quando a lista de *tokens* úteis é reduzida em 90%, a remoção de *stopwords* pode tanto melhorar a separação quanto prejudicar a partição. A Tabela 3 e Figura 2.c (90%) demonstram essa dinâmica, destacando uma maior variação nas médias por método sob essa condição.

Portanto, a filtragem pode ser tratada como uma variável experimental em vez de um pressuposto metodológico; sua utilização deve ser fundamentada em evidências (como médias por método, coesão/acoplamento).

5.4. Limitações e direções futuras

A limitação do espaço amostral opera por meio de duas frentes interativas: (i) exclusão léxica explícita (*stopwords*) e (ii) parametrização do extrator/sumarizador (cortes de frequência). O método de *Luhn* emprega dois cortes (superior e inferior) para estabelecer uma banda de frequência considerada "útil". Essa banda é extremamente sensível à língua e ao domínio temático: palavras que são usuais em um campo técnico podem ser pertinentes, mas em um contexto diferente podem ser vistas como desnecessárias. Em contrapartida, o critério de corte médio do Cassiopeia (Guelpeli, 2012) tende a preservar um núcleo lexical mais estável e menos sujeito a extremos, reduzindo a sensibilidade do *pipeline* a alterações sutis no conjunto de *tokens*.

Os gráficos de média acumudala de coesão e acoplamento comprovam essa análise, mostram variações locais na compacidade interna sem que isso resulte, obrigatoriamente, em uma mudança média no *Silhouette*; as Figuras 2.b e 2.c, focando mais especificamente nas compressões de 70% e 90%, exibem um comportamento semelhante, porém com uma amplitude maior à medida que a compressão se intensifica.

5.5. Considerações finais

A análise de concordância (Kendall 's W) confirmou que as ordens das condições não apresentam concordância consistente ao longo do corpus ($W = 0,0011$). Isso corrobora a noção de que, no ambiente testado, a remoção de *stopwords* não apresenta um impacto prático sistemático na qualidade do agrupamento. Para a criação de experimentos em sumarização e agrupamento textual, é essencial distinguir entre filtragem focada na qualidade e filtragem focada no controle de volume. Neste estudo, a evidência empírica sugere que o uso ou não de *stopwords* não altera significativamente a qualidade dos agrupamentos gerados pelo Cassiopeia (Guelpeli, 2012). Isso se deve principalmente à forma como o espaço léxico é organizado pelo critério de corte do modelo e pelos processos de sumarização. Assim, a decisão de eliminar *stopwords* deve ser tratada como uma variável experimental e operacional, em vez de uma suposição metodológica. A prática científica recomendada continua sendo testar, documentar e justificar essa decisão para cada domínio e configuração.

REFERÊNCIAS

AMARIZ, Diego Sampaio; GUELPELI, Marcus Vinícius Carvalho. Mineração de texto: a clusterização aplicada em artigos científicos de Química, por meio do modelo Cassiopeia. *Lumen et Virtus*, v. 15, n. 42, p. 7482–7505, 2024. DOI: 10.56238/levv15n42-070. Disponível em: https://periodicos.newsciencepubl.com/LEV/article/view/1712

BERWANGER, Giordano Pydd. *Sumarização multi-documento para o português com modelos transformer*. 2022. 64 f. Trabalho de Conclusão de Curso (Engenharia de Computação) – Universidade Tecnológica Federal do Paraná, Toledo, 2022.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, v. 3, p. 993–1022, 2003.

CROFT, W. B.; METZLER, D.; STROHMAN, T. *Search Engines: Information Retrieval in Practice*. Addison-Wesley, 2009.

DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: *Proceedings of NAACL–HLT*. 2019.

ELHADI, M. Extractive Summarization Using Structural Syntax, Term Expansion and Refinement. *International Journal of Intelligence Science*, v. 7, p. 55–71, 2017. DOI: 10.4236/ijis.2017.73004.

FRIEDMAN, M. (1937). The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. Journal of the American Statistical Association, 32(200), 675–701. https://doi.org/10.2307/2279372

GAMBHIR, M.; GUPTA, V. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*, 2017.

GUELPELI, Marcus Vinícius Carvalho. *Cassiopeia: Um modelo de agrupamento de textos baseado em sumarização*. 2012. Tese (Doutorado em Computação) – Universidade Federal Fluminense, Niterói, 2012.

KENDALL, M.G. and BABINGTON Smith, B. (1939) The Problem of m Rankings. The Annals of Mathematical Statistics, 10, 275-287. http://dx.doi.org/10.1214/aoms/1177732186

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

MANI, I. *Automatic Summarization*. Cambridge: MIT Press, 2001. (Natural Language Processing; v. 3).

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

MIHALCEA, R.; TARAU, P. TextRank: Bringing order into texts. In: *Proceedings of EMNLP*. 2004.

MIKOLOV, T. et al. Distributed Representations of Words and Phrases and their Compositionality. In: *Advances in Neural Information Processing Systems*. 2013.

NENKOVA, A.; McKEOWN, K. A survey of single-document summarization. *Foundations and Trends in Information Retrieval*, 2012.

PARDO, T. A. S.; RINO, L. H. M. A sumarização automática de textos: principais características e metodologias. In: *Anais do XXIII Congresso da Sociedade Brasileira de Computação*. Vol. VIII: III Jornada de Minicursos de Inteligência Artificial, 2002.

REIMERS, N.; GUREVYCH, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: *EMNLP/IJCNLP – System Demonstrations*. 2019.

ROCHA, V. J. C.; GUELPELI, M. V. C. PragmaSUM: Automatic text summarizer based on user profile. *International Journal of Current Research*, v. 9, n. 7, p. 53935–53942, 2017.

ROSS, S. M. *A First Course in Probability*. 10. ed. Pearson, 2014.

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

SUMY Developers. *Sumy (v. 0.11.0)*. Disponível em: https://pypi.org/project/sumy/