CEFET/RJ
DIRETORIA DE PESQUISA E PÓS-GRADUAÇÃO

LEXRANK GUIADO PARA RECUPERAÇÃO AVANÇADA EM ANÁLISE JURÍDICA

Fabio da Silva Gregorio

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Ciência da Computação, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ, como parte dos requisitos necessários à obtenção do grau de mestre.

Orientadores:
Eduardo Bezerra, D.Sc.

Rio de Janeiro,

Fevereiro de 2025

LexRank guiado para recuperação avançada em análise jurídica

Dissertação de Mestrado em Ciência da Computação, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, CEFET/RJ.

Fabio da Silva Gregorio

Aprovada por:

Presidente, Prof. Eduardo Bezerra da Silva, D.Sc. (CEFET/RJ)

Prof. Altigran Soares da Silva, D.Sc. (UFAM)

Prof.ª Kele Teixeira Belloze, D.Sc. (CEFET/RJ)

Prof. Gustavo Paiva Guedes e Silva, D.Sc. (CEFET/RJ)

Rio de Janeiro,

Fevereiro de 2025

Ficha catalográfica elaborada pela Biblioteca Central do CEFET/RJ

G821 Gregorio, Fabio da Silva
Lexrank guiado para recuperação avançada em análise jurídica / Fabio da Silva Gregorio. — 2025.
62f. : il. color. , enc.

Dissertação (Mestrado) Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, 2025.
Bibliografia : f. 58-62
Orientador: Eduardo Bezerra

1. Processamento de linguagem natural (Computação). 2. Recurso especial. 3. Procedimento sumário. I. Bezerra, Eduardo. (Orient.). II. Título.

CDD 006.312

Elaborada pela bibliotecária Tania Mello – CRB/7 nº 5507/04

DEDICATÓRIA

Dedico este trabalho à minha esposa, Scintilla,
por toda compreensão e companheirismo ao
longo desta jornada.

AGRADECIMENTOS

Primeiramente agradeço a Deus, criador de todas as coisas e fonte inesgotável de sabedoria.

Agradeço ao professor Eduardo Bezerra por toda orientação. Sua paciência e conhecimento tornou este trabalho possível.

Agradeço ao meu pai, Ivanil (in memoriam), à minha mãe, Lúcia, e ao meu irmão Leandro por me mostrar em todos os momentos o significado de família.

Agradeço aos amigos do TRF2, por todas as palavras de incentivo e colaboração com este trabalho.

Agradeço a João Batista da Silva (in memoriam), meu tio e amigo, pelo exemplo. Negro, de origem humilde. Desde cedo entendeu que o conhecimento é um bem precioso que devemos buscar a todo custo, e ele nos conduzirá a lugares inimagináveis. Sua dedicação aos estudos e seu profissionalismo o levaram a ser agraciado com Ordem do Mérito Judiciário do Distrito Federal e dos Territórios. Plantou boas sementes em sua jornada, seu legado não foi apagado e frutifica até os dias de hoje.

Por fim, agradeço ao CNPq, CAPES, FAPERJ e CEFET/RJ pelo financiamento parcial desta pesquisa.

RESUMO

LexRank guiado para recuperação avançada

em análise jurídica

Fabio da Silva Gregorio

Orientadores:
Eduardo Bezerra, D.Sc.

Resumo da Dissertação submetida ao Programa de Pós-graduação em Ciência da Computação do Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ como parte dos requisitos necessários à obtenção do grau de mestre.

A Constituição Brasileira prevê mecanismos para que cidadãos peticionem ao Judiciário, incluindo o chamado recurso especial. Esse tipo específico de recurso visa uniformizar a interpretação jurídica da legislação brasileira. O tratamento de recursos especiais é uma das tarefas diárias no Judiciário, regularmente apresentando demandas significativas em seus tribunais. Propomos um novo método chamado GLARE (LexRank guiado para recuperação avançada em análise jurídica), baseado em aprendizado de máquina não supervisionado, para auxiliar o analista jurídico a classificar um recurso especial em um tema de uma lista disponibilizada pelo Superior Tribunal de Justiça (STJ). Como parte desse método, propomos uma modificação do algoritmo LexRank baseado em grafos, que chamamos de Guided LexRank. Esse algoritmo gera o resumo de um recurso especial. O grau de similaridade entre o resumo gerado e diferentes temas é avaliado usando o algoritmo BM25. Como resultado, o método apresenta um ranking de temas mais adequados ao recurso especial analisado. O método proposto não requer rotulagem prévia do texto a ser avaliado e elimina a necessidade de grandes volumes de dados para treinar um modelo, como normalmente ocorre em modelos supervisionados. Avaliamos a eficácia do método aplicando-o a um corpus de recurso especial previamente classificado por especialistas humanos.

Palavras-chave:

Recurso especial, Processo judicial, Processamento de linguagem natural, Classificação de texto, Sumarização extrativa, Modelagem de tópicos, BERTopic, LexRank, BM25

Rio de Janeiro,

Fevereiro de 2025

ABSTRACT

LexRank guiado para recuperação avançada

em análise jurídica

Fabio da Silva Gregorio

Advisors:
Eduardo Bezerra, D.Sc.

Abstract of dissertation submitted to Programa de Pós-graduação em Ciência da Computação - Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ as partial fulfillment of the requirements for the degree of master.

The Brazilian Constitution provides mechanisms for citizens to petition the Judiciary, including the so-called special appeal. This specific type of appeal aims to standardize the legal interpretation of Brazilian legislation. The processing of special appeals is one of the daily tasks of the Judiciary, regularly presenting significant demands in its courts. We propose a new method called GLARE (Guided LexRank for Advanced Retrieval in Legal Analysis), based on unsupervised machine learning, to assist the legal analyst in classifying a special appeal on a topic from a list made available by the National Court of Brazil (STJ). As part of this method, we propose a modification of the graph-based LexRank algorithm, which we call Guided LexRank. This algorithm generates the summary of a special appeal. The degree of similarity between the generated summary and different topics is evaluated using the BM25 algorithm. As a result, the method presents a ranking of themes most appropriate to the special appeal analyzed. The proposed method does not require prior labeling of the text to be evaluated and eliminates the need for large volumes of data to train a model, as is typically the case in supervised models. We evaluate the effectiveness of the method by applying it to a special appeal corpus previously classified by human experts.

Key-words:

Special appeal, Legal Process, Natural Language Processing, Text Classification, Extractive Summarization, Topic Modeling, BERTopic, LexRank, BM25

Rio de Janeiro,

Fevereiro de 2025

Sumário

I Introdução **14**

I.1 Contextualização 14

I.2 Motivação 16

I.3 Objetivos 18

I.4 Metodologia 19

I.5 Estrutura da Dissertação 20

II Fundamentação Teórica **21**

II.1 Sistemas de recuperação de informação 21

II.1.1 Indexação 22

II.1.2 Similaridade 22

II.2 Agrupamentos 26

II.2.1 DBSCAN 28

II.2.2 HDBSCAN 28

II.3 Sumarização extrativa 29

II.3.1 Modelagem de tópicos 29

II.3.2 Algoritmos baseados em grafo 34

III Trabalhos Relacionados **37**

III.1 Seleção de publicações 37

III.2 Critérios de inclusão, exclusão e priorização 38

III.2.1 Critérios de inclusão 39

III.2.2 Critérios de exclusão 39

III.2.3 Critérios de priorização 39

III.3 Resultados 39

III.3.1 Publicações selecionadas 39

III.3.2 Comparação 42

IV Metodologia **43**

IV.1 Formulação do problema 43

IV.2 Documentos Jurídicos 44
IV.2.1 Recursos Especiais 44
IV.2.2 Temas Repetitivos 44
IV.3 Corpus e metadados associados 46
IV.4 Passos da metodologia 47
IV.4.1 Pré-processamento dos recursos especiais e temas 48
IV.4.2 Sumarização por meio do LexRank Guiado 49
IV.4.3 Cálculo de similaridades 53
IV.4.4 Avaliação do resultado 54

V Experimentos 56
V.1 Configurações experimentais 56
V.2 Comparando alternativas de pipeline 56
V.3 Impacto da sumarização de texto no desempenho 59
V.4 Comparação com a linha de base do Elasticsearch 62
V.5 Comparação com linhas de base de aprendizagem supervisionada 64
V.5.1 Avaliando a generalização do modelo em temas jurídicos não representados nos dados de treinamento 67

VI Conclusão 68
VI.1 Análise Retrospectiva 68
VI.2 Trabalhos futuros 70

Referências 71

A Recurso Especial 76

B Amostra de tema repetitivo submetido ao STJ 83

C Estrutura do Conjunto de Dados 85

D Resultados dos experimentos 87

Lista de Figuras

I.1 Série histórica CNJ 16

I.2 Série histórica vice-presidência do TRF2 17

II.1 Relacionamentos entre estruturas de dados. 23

II.2 Representação vetorial de documento. 24

II.3 Efeito da saturação no Term Frequency (TF)-Inverse Term Frequency (IDF) e no Best Match 25 (BM25). 27

II.4 Modelagem de tópicos guiada 33

II.5 Representação gráfica do grafo G 34

II.6 Grafo PageRank 35

III.1 Fluxo do processo de seleção dos artigos 38

IV.1 Esquema do método proposto 44

IV.2 Distribuição de temas no corpus 47

IV.3 Passos metodológicos da abordagem proposta. 49

IV.4 Grafo LexRank 50

IV.5 Diagrama do algoritmo LexRank Guiado 52

IV.6 Grafo LexRank guiado 53

IV.7 Relação entre recursos, resumos e temas 53

V.1 Análise da qualidade da classificação das sugestões 59

V.2 Impacto da sumarização no desempenho do modelo 60

V.3 Impacto da remoção de termos no desempenho 61

V.4 Comparação de remoção de termos com similaridade de cosseno 62

V.5 Solução adotada pelo Tribunal Regional Federal da 2ª Região (TRF2) 63

V.6 Distribuição de classes minoritárias 66

Lista de Tabelas

III.1 Resumo dos trabalhos relacionados 42

IV.1 Estatísticas do corpus de recursos especiais utilizado nesta pesquisa. 46
IV.2 Dados estatísticos sobre os temas repetitivos considerados nesta pesquisa. 47

V.1 Tratamento: Com a remoção de termos. Similaridade: BM25 57
V.2 Tratamento: Com remoção de termos. Similaridade: Cosseno 57
V.3 Tratamento: Sem remoção de termos. Similaridade: BM25 58
V.4 Tratamento: Sem remoção de termos. Similaridade: Cosseno 58
V.5 Análise comparativa de texto completo vs. texto resumido 61
V.6 Comparação de desempenho entre o Elasticsearch de base e o GLARE. 64
V.7 Desempenho: distribuição aleatória entre treino e teste 65
V.8 Desempenho: classes minoritárias no conjunto de teste 67
V.9 Desempenho: classes não vistas no treino 67

Lista de Abreviações

BM25	Best Match 25	18, 25, 26, 27, 48, 50, 51, 54, 69
CJF	Conselho Da Justiça Federal	17
CNJ	Conselho Nacional De Justiça	16, 17
CPC	Código De Processo Civil	15, 44, 48
CSJT	Conselho Superior Da Justiça Do Trabalho	17
CSV	Comma-separated Values	85
DOI	Digital Object Identifier	38
HDBSCAN	Hierarchical Density-Based Spatial Clustering Of Applications With Noise	32
IDF	Inverse Term Frequency	24, 25, 26, 27
LDA	Latent Dirichlet Allocation	30
MAP	Mean Average Precision	54
NDCG	Normalized Discounted Cumulative Gain	54, 55
NIST	National Institute Of Standards And Technology	25
NMF	Non-Negative Matrix Factorization	30
PNUD	Programa Das Nações Unidas Para O Desenvolvimento	17
PRISMA	Preferred Reporting Items For Systematic Reviews And Meta-Analyses	19, 37, 38, 39
SQL	Structured Query Language	21, 22
STJ	Superior Tribunal De Justiça	14, 15, 16, 17, 19, 43, 44, 46, 63, 85
SUS	Sistema Único De Saúde	15
TF	Term Frequency	23, 24, 25, 26, 27
TREC	Text REtrieval Conference	25
TRF2	Tribunal Regional Federal Da 2ª Região	16, 17, 19, 46, 62, 63
TST	Tribunal Superior Do Trabalho	17

Capítulo I Introdução

I.1 Contextualização

Como é possível ao ser humano compreender as informações contidas em um texto? Ou, ainda, como lhe é possível elencar trechos relevantes de um documento e elaborar um segundo documento que seja um resumo coerente do primeiro? O conhecimento tácito, adquirido pelas experiências de um indivíduo, proporciona um certo grau de automatismo na interpretação de textos corriqueiros e, ao mesmo tempo, oculta o complexo mecanismo da análise semântica de um texto.

Interpretação e elaboração de documentos textuais fazem parte de inúmeras atividades profissionais, e isto ocorre em diversos graus de complexidade. Dentre essas atividades, encontra-se o tratamento de processos jurídicos. No período compreendido entre o nascimento de um processo a partir de sua petição inicial e o seu encerramento, com sua baixa definitiva, inúmeros documentos são incorporados ao mesmo. Por diversos momentos, a tramitação de um processo está atrelada à elaboração de um documento, o qual é confeccionado após uma detalhada análise semântica de outros documentos já anexados ao processo e a análise de legislações relacionadas ao assunto em questão.

A Constituição Federal Brasileira de 1988 proporcionou diversos meios pelos quais um cidadão pode provocar uma ação do Poder Judiciário ao sentir-se lesado em seus direitos. Um desses meios é o chamado **recurso**, o qual por definição é o meio idôneo para impugnar uma decisão judicial visando o seu reexame, para tentar obter, na mesma relação processual, a reforma, a invalidação, o esclarecimento ou a integração do julgado. Em sentido geral, recurso é o poder de provocar o reexame de uma decisão, pela mesma autoridade judiciária, ou por outra hierarquicamente superior, visando a obter a sua reforma ou modificação [Con, 2023].

Um tipo específico de recurso, previsto no artigo 105 da Constituição, é o **recurso especial**. Este recurso cumpre a finalidade de uniformizar o entendimento jurídico sobre o regramento brasileiro. Tais recursos são julgados pelo STJ em situações de causas que foram decididas, em única ou última instância, pelos Tribunais Regionais Federais ou pelos tribunais dos Estados, do Distrito Federal e Territórios, quando a decisão recorrida: contrariar tratado ou lei federal, ou negar-lhes vigência; julgar válido ato de governo local contestado em face de lei federal; der a lei federal interpretação divergente da que lhe haja atribuído outro tribunal [Con, 2023].

Possuindo dimensões continentais e uma população estimada em mais de 207 milhões segundo o censo de 2022 [Cen, 2023], o Brasil, com seus problemas estruturais amplamente conhecidos, submete rotineiramente aos seus tribunais um quantitativo expressivo de demandas, as quais por vezes são muito semelhantes em seus conteúdos. Podemos citar como um problema recorrente o acionamento do poder judiciário para intervir no funcionamento do Sistema Único de Saúde (SUS) e garantir a um cidadão o direito a determinado tratamento ou medicamento.

Como forma de dar celeridade ao julgamento de processos e otimizar o fluxo de trabalho nos tribunais, alguns mecanismos foram instituídos. Um desses mecanismos é o de **recursos repetitivos** instituído no STJ com a Lei 11.672/2008. Com aplicação no julgamento de recursos especiais que tratem da mesma controvérsia jurídica[1], o sistema de recursos repetitivos estabelece que, para demandas semelhantes, sejam selecionados processos por amostragem com fim de que tal demanda seja julgada pelo STJ [CNJ, 2023].

Para cada controvérsia submetida ao STJ por meio de recurso especial, ocorre a definição e a divulgação de uma tese jurídica que deve ser aplicada a todos os recursos em que seja discutida idêntica questão de direito. Cabe aos tribunais de instâncias inferiores a observância do Código de Processo Civil (CPC) ao proceder a avaliação dos requisitos necessários para a admissibilidade de cada recurso especial solicitado, também lhes cabe o exame do CPC para aplicar o trâmite adequado ao recurso admitido.

Segundo o Código de Processo Civil, cabe ao presidente ou vice-presidente do tribunal recorrido: Negar seguimento ao recurso especial interposto contra acórdão[2] que esteja em conformidade com entendimento do STJ no regime de julgamento de recursos repetitivos; Encaminhar o processo ao seu órgão julgador para que haja um Juízo de Retratação se o acórdão recorrido divergir do entendimento do STJ no regime de julgamento de recursos repetitivos; Sobrestar o recurso que versar sobre controvérsia de caráter repetitivo ainda não decidida pelo STJ; E deverá selecionar o recurso representativo da controvérsia e remeter ao STJ, desde que o recurso ainda não tenha sido submetido ao regime de julgamento de recursos repetitivos e tenha sido refutado o juízo de retratação.

Observa-se que no tratamento do recurso especial, além do exame de admissibilidade para cumprimento de verificações formais, é imprescindível que seja feita a leitura na íntegra do conteúdo do recurso e, a partir do que for depreendido desta leitura, seja feita uma confrontação entre o fato apresentado no recurso e todos os temas presentes no sistema de recursos repetitivos do STJ, buscando encontrar uma relação de similaridade com algum tema existente na base de dados do sistema. Essa confrontação é necessária para ser dado o correto tratamento ao recurso, seja este o

[1]Desacordo ou debate existente em relação à interpretação, ou aplicação da lei em uma determinada situação.

[2]Decisão final do tribunal sobre o caso em questão, é elaborado por um grupo de juízes ou desembargadores que compõem o órgão colegiado.

encaminhamento ao STJ como exemplo amostral a ser julgado no sistema de recursos repetitivos, seja o sobrestamento do processo até que haja definição de uma tese sobre determinado tema, ou seja a devolução do processo ao órgão julgador pertencente ao tribunal para aplicação do entendimento já fixado em uma tese previamente divulgada pelo STJ.

O tratamento de recursos especiais é um típico caso do cotidiano no âmbito do poder judiciário, e segundo dados oficiais de 2021 foram mais de 50 mil recursos especiais submetidos ao STJ [de Justiça, 2021]. Trata-se de uma atividade complexa cuja tarefa central consiste em analisar na íntegra o conteúdo do recurso, identificar os pontos essenciais do que está sendo pleiteado e realizar uma comparação com os temas existentes no sistema de recursos repetitivos do STJ.

I.2 Motivação

O Conselho Nacional de Justiça (CNJ) mantém e atualiza periodicamente o painel de estatísticas do Poder Judiciário [Est]. No Brasil, segundo dados oficiais do CNJ, o Poder Judiciário finalizou o ano de 2020 com 75,4 milhões de processos em tramitação aguardando alguma solução definitiva (Figura I.1) [de Justiça, 2021].

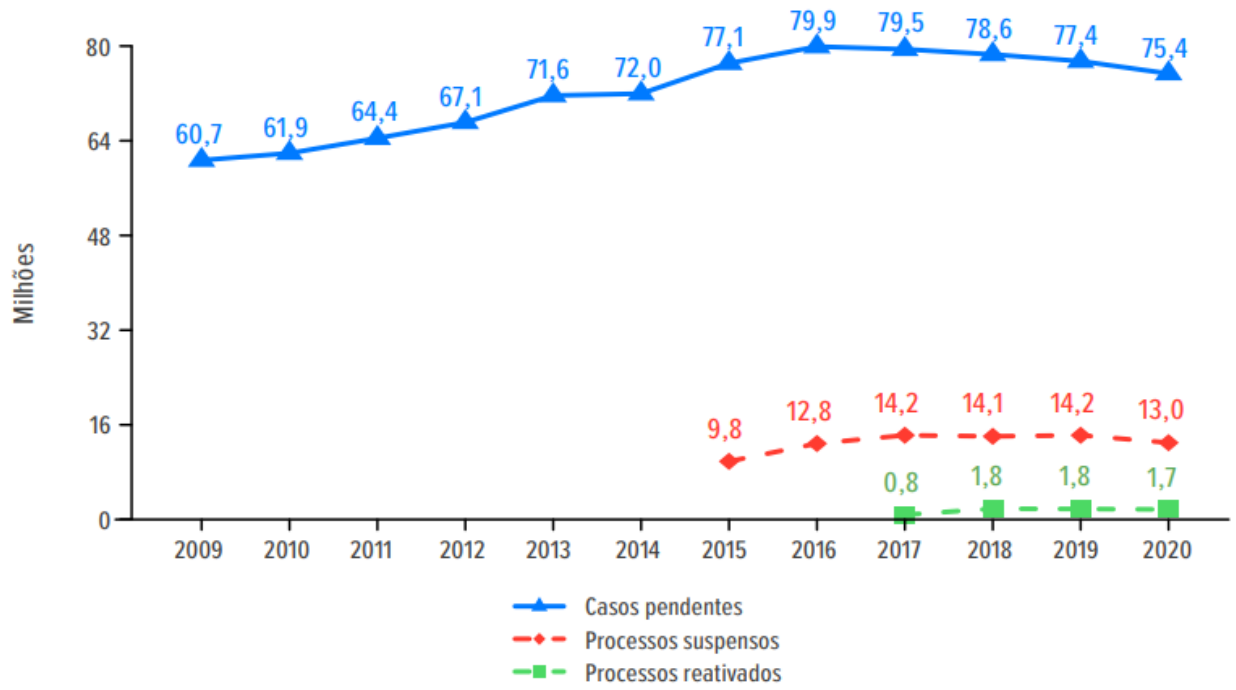


Figura I.1: Conselho Nacional de Justiça - Série histórica de processos pendentes de solução definitiva.

Segundo o Regimento interno do TRF2 [Reg], além de outras demandas, cabe ao vice-presidente do tribunal decidir sobre a admissibilidade de recurso especial, o qual é recebido e tratado por analistas judiciários lotados no gabinete da vice-presidência. Dados coletados no painel de estatísticas estão apresentados na Figura I.2. Verifica-se que em março de 2023, por exemplo, em média, o

tempo decorrido entre a petição inicial de um processo e sua análise pela vice-presidência do TRF2 era de 2952 dias.

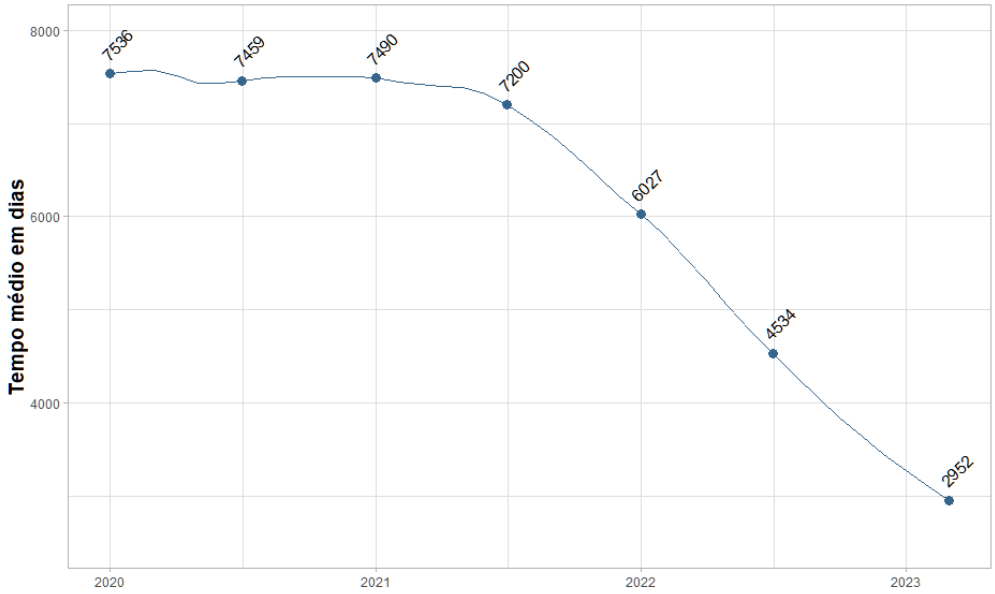


Figura I.2: Série histórica - Tempo médio dos processos que estão pendentes na vice-presidência do TRF2. Os valores assinalados na curva indicam há quantos dias, em média, um processo recebido na vice-presidência do TRF2 está tramitando no âmbito da Justiça Federal.

Analisando o gráfico da Figura I.2, é perceptível que há uma queda no tempo de tramitação após o ano de 2021. Essa queda pode ser explicada pelo lançamento do **Programa Justiça 4.0**. Esse programa foi desenvolvido pela parceria entre o CNJ e o Programa das Nações Unidas para o Desenvolvimento (PNUD) com apoio do Conselho da Justiça Federal (CJF), STJ, Tribunal Superior do Trabalho (TST) e Conselho Superior da Justiça do Trabalho (CSJT). O programa possui quatro eixos de atuação:

- **Inovação e Tecnologia -** Soluções disruptivas para transformar o Judiciário e melhorar a prestação de serviços a toda a sociedade.
- **Gestão de Informação e Políticas Judiciárias -** Formulação, implantação e monitoramento de políticas com base em evidências para fortalecer a promoção de direitos humanos.
- **Fortalecimento de Capacidades Institucionais do CNJ -** Transferência de conhecimento e soluções com foco na segurança jurídica, sustentabilidade de projetos e eficiência da prestação jurisdicional.
- **Prevenção e Combate à Corrupção e a Lavagem de Dinheiro e Recuperação de Ativos -** Robustecimento da atuação do Judiciário com melhor gestão de dados e informações e otimização da pesquisa de ativos.

Após a Constituição de 1988, houve uma ampliação de meios para os cidadãos buscarem seus direitos, resultando em aumento significativo no número de processos no Judiciário. Se por um lado o texto constitucional trouxe avanços na garantia de direitos do cidadão, por outro trouxe consigo o desafio de atender com celeridade as demandas da sociedade que foram em muito ampliadas. Ao analisarmos dados do TRF2, descobrimos que o tempo médio entre a admissão de um recurso especial pela vice-presidência e a sua decisão sobre o recurso é de 113,47 dias. A magnitude do trabalho desempenhado pelos órgãos da Justiça Federal, frente aos recursos disponíveis, foi o principal fator motivador desta pesquisa.

I.3 Objetivos

Esta dissertação tem como objetivo primário desenvolver um método eficaz, baseado em inteligência artificial e técnicas de processamento de linguagem natural, para auxiliar analistas jurídicos na tarefa de classificar recursos especiais em temas jurídicos específicos. Apresentamos o método GLARE (LexRank guiado para recuperação avançada em análise jurídica), baseado em aprendizado de máquina não supervisionado. O método engloba um algoritmo para sumarização de documentos, o qual também desenvolvemos no âmbito deste trabalho e o chamamos de LexRank guiado. Ainda como parte do método, é executada uma etapa de cálculo de similaridade com os temas, por meio do algoritmo BM25, fornecendo assim uma lista de temas mais prováveis para cada recurso especial apresentado. Os objetivos secundários são listados a seguir.

1. Construir um conjunto de dados rotulados de recursos especiais com seus respectivos temas, visando o reaproveitamento e a expansão do corpus para futuros trabalhos de processamento de linguagem natural no domínio jurídico brasileiro.

2. Analisar o impacto que a sumarização exerce na tarefa de classificar o documento em um tema.

3. Comparar o desempenho do método ao aplicar distintas formas de medição de similaridade.

4. Investigar de que forma o pré-processamento do texto afeta o desempenho do método proposto.

5. Validar o desempenho do método desenvolvido com um conjunto de dados de recursos especiais reais, analisando sua eficácia e identificando possíveis melhorias para ajustes futuros.

6. Confrontar o método desenvolvido com outras abordagens existentes, como classificadores tradicionais, para verificar as vantagens e limitações em termos de eficiência e precisão.

I.4 Metodologia

Para o desenvolvimento desta pesquisa, foram utilizados dois corpora distintos. O primeiro corpus é composto por 7.967 recursos especiais extraídos de processos jurídicos não sigilosos do TRF2, coletados no segundo semestre de 2023. Cada recurso foi classificado por um especialista em um tema. O segundo corpus consiste no conjunto textual de 190 temas dentre os disponibilizados pelo STJ[3]. Ambos os corpora foram estruturados de modo a permitir a análise integrada de recursos especiais e seus temas associados. Estatísticas detalhadas sobre os documentos e os temas estão apresentadas na Seção IV.3.

Para efetuar o levantamento bibliográfico que embasa a pesquisa proposta, foi adotada a metodologia Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA), amplamente reconhecida em revisões da literatura por sua robustez e rigor metodológico. Essa abordagem estruturou o processo de busca, aplicação de critérios de inclusão e exclusão, e priorização dos estudos relevantes, assegurando a identificação das contribuições mais significativas para o contexto da pesquisa. Os detalhes da aplicação do método, incluindo os termos de busca e critérios adotados, estão descritos no Capítulo III.

O método proposto nesta dissertação se inicia com a aplicação de uma técnica de sumarização desenvolvida no âmbito deste trabalho. Nessa etapa, um algoritmo é utilizado para extrair as informações mais relevantes do texto completo do recurso especial, gerando uma versão condensada que preserva o conteúdo essencial para a análise de similaridade. Esse processo de sumarização tem como foco destacar as partes do texto com maior relação com os temas a serem identificados, facilitando a classificação posterior. Em seguida, realiza-se uma avaliação de similaridade entre o texto sumarizado e os textos dos temas repetitivos previamente definidos, utilizando a métrica BM25, escolhida pela sua eficácia em problemas de recuperação de informação. Esse cálculo de similaridade é crucial para identificar quais temas possuem maior correlação com o conteúdo do recurso especial.

Com os resultados dessa avaliação, elabora-se uma lista ranqueada, onde os temas mais prováveis aparecem no topo. Esse ranqueamento permite que o analista jurídico visualize rapidamente os temas mais relevantes para o recurso especial analisado, facilitando a escolha ou verificação do tema correto, e reduzindo o esforço manual de busca. Por fim, realiza-se uma avaliação qualitativa dos resultados por meio de métricas apropriadas, verificando a adequação do modelo. Essa etapa final envolve a análise da relevância dos temas sugeridos em relação ao tema real do recurso especial, validando assim a eficácia do método proposto.

[3]https://processo.stj.jus.br/repetitivos/temas_repetitivos/

I.5 Estrutura da Dissertação

Ao longo desta dissertação, os capítulos estão organizados de forma a apresentar a fundamentação teórica e prática do trabalho realizado, bem como os experimentos e resultados obtidos. No Capítulo II são abordados os conceitos fundamentais e teorias que sustentam o desenvolvimento do método proposto, fornecendo uma base teórica essencial para a compreensão do trabalho. O Capítulo III apresenta uma revisão dos trabalhos relacionados, discutindo abordagens existentes na literatura para tratamento de documentos jurídicos com técnicas de inteligência artificial. No Capítulo IV, é detalhada a metodologia desenvolvida para a solução do problema apresentado. Como parte desta metodologia apresentamos um algoritmo que desenvolvemos para a sumarização de documentos. Em seguida, no Capítulo V, são apresentados os experimentos realizados para validação do método. Por fim, o Capítulo VI traz as conclusões do trabalho, discutindo as principais contribuições, limitações e propostas para trabalhos futuros, com vistas ao avanço das aplicações de inteligência artificial no processamento de documentos jurídicos.

Capítulo II Fundamentação Teórica

Este capítulo apresenta alguns fundamentos necessários para o entendimento dos capítulos seguintes. A Seção II.1 trata de Sistemas de recuperação de informação, os quais visam atender a necessidade de informação do usuário demandada por consultas textuais. A Seção II.2 aborda algoritmos relacionados a agrupamento. Finalmente, a Seção II.3 apresenta os tipos de sumários e algumas técnicas usadas na geração dos mesmos.

II.1 Sistemas de recuperação de informação

Recuperação da informação é a ciência de procurar documentos ou informações em documentos. Esses documentos podem ser textuais ou multimídia, dados de áudio e vídeo que possuem palavras-chave descritivas associadas também podem ser alvos de busca [Silberschatz et al., 2011].

Os sistemas de recuperação de informação assemelham-se estruturalmente aos sistemas de banco de dados tradicionais no sentido de que os documentos são armazenados em um repositório, um índice estruturado é mantido e a consulta feita pelo usuário processa este índice para identificar o que lhe deve ser retornado. Apesar das semelhanças, existem algumas diferenças fundamentais entre esses sistemas [Zobel and Moffat, 2006] [Han et al., 2011]:

- Em sistemas de recuperação de informação assume-se que os dados não são estruturados, e não dispõem de qualquer esquema associado;
- Em sistemas de recuperação de informação as consultas são formadas principalmente por palavras-chave, as quais não possuem uma estrutura complexa ou sintaxe bem estabelecida como nas consultas em Structured Query Language (SQL).
- Em sistema de banco de dados uma correspondência é um registro que atende a uma condição lógica especificada, porém em sistemas de recuperação de informação, uma correspondência é um documento adequado à consulta de acordo com a heurística estatística e pode nem conter todos os termos da consulta.
- Os sistemas de banco de dados retornam todos os registros correspondentes. Os sistemas de recuperação de informação classificam as correspondências por sua similaridade estatística, e desta forma limita o que será apresentado ao usuário.

- Os sistemas de banco de dados atribuem uma chave de acesso exclusiva a cada registro e permitem a pesquisa nessa chave.

Os sistemas de recuperação de informação normalmente permitem expressões de consultas formadas por composições de palavras-chave e os conectivos lógicos **and**, **or** e **not**. Na recuperação de **texto completo**, todas as palavras em cada documento são consideradas palavras-chave. Em sua forma mais simples, se a consulta não apresentar conectivos, o sistema localiza e retorna todos os documentos que contêm todas as palavras-chave que constam na consulta. Sistemas mais sofisticados estimam a relevância dos documentos em resposta a determinada consulta, estabelecem um ranqueamento e apresentam os documentos na ordem da maior relevância estimada [Silberschatz et al., 2011].

II.1.1 Indexação

A partir da coleção de todos os documentos disponibilizados ao sistema de recuperação de informação, se origina um conjunto de dados denominado **corpus**, o qual é formado pela união dos termos existentes em cada um dos documentos. Por definição, índice é uma estrutura de dados que mapeia termos para os documentos que os contêm. O uso da indexação surge com a necessidade de se efetuar uma busca rápida e eficiente, que objetiva localizar os termos da consulta no corpus bem como identificar os documentos nos quais os termos estão contidos.

Dentre os tipos de implementação de índices existentes, pesquisas apontam como sendo mais eficiente o **índice de arquivo invertido**, ou **índice invertido** [Zobel and Moffat, 2006; Frakes and Baeza-Yates, 1992]. Conceitualmente o índice invertido pode ser visualizado como um agregado de dois tipos de estruturas de dados, **dicionário** e **lista invertida**. A cada termo K_i pertencente ao corpus, é associada uma lista S_i de identificadores de documentos que contêm K_i. O dicionário se relaciona com as diversas listas criadas, no sentido que o dicionário contem o ponteiro para o início da lista invertida de cada termo [Aggarwal, 2022a; Silberschatz et al., 2011]. A Figura II.1 esquematiza o relacionamento entre as estruturas de dados citadas.

II.1.2 Similaridade

Mais do que atender a uma consulta estritamente lógica, como nas consultas em SQL, dentro do contexto de sistema de recuperação um documento corresponde a uma necessidade de informação se o **usuário** perceber que ela é **relevante**. Porém, o conceito de relevância neste caso é inexato, e um documento pode ser relevante para uma necessidade de informação mesmo que não contenha nenhum dos termos de consulta ou irrelevante mesmo que contenha todos eles [Zobel and Moffat, 2006]. É necessário levar em consideração como a sinonímia, a homonímia e figuras de linguagem podem impactar na relevância de um documento. Sendo assim, uma questão inicial a ser analisada

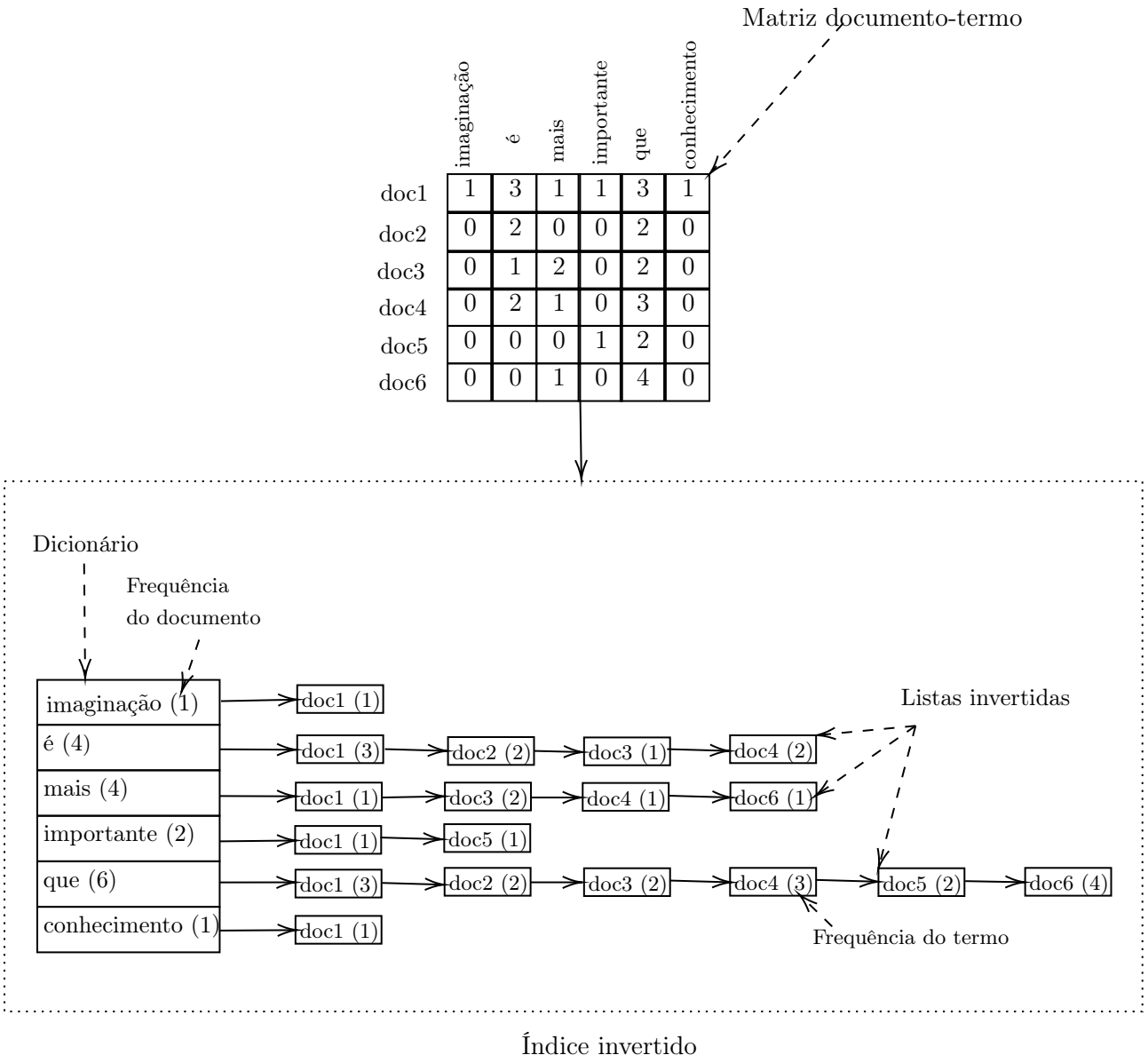


Figura II.1: Relacionamento entre estruturas de dados: Matriz documento-termo; Índice invertido; Dicionário e Lista invertida

é, dado um termo t de uma consulta qualquer, qual é a relevância de um documento D dado esse termo.

No contexto da recuperação de informação, a relevância de um documento é conhecida como **frequência de termo** (TF). Uma forma de medir $\mathrm{TF}(D,t)$, ou seja, a relevância de um documento D dado um termo t, é por meio da Equação II.1, na qual $n(D)$ indica o número de termos no documento e $n(D,t)$ indica o número de ocorrências do termo t no documento D [Silberschatz et al., 2011].

$$\mathrm{TF}(D,t) = \log\left(1 + \frac{n(D,t)}{n(D)}\right) \tag{II.1}$$

Considerando que uma consulta Q pode conter vários termos, a relevância de um documento para uma consulta com vários termos é estimada pela combinação das medidas de relevância do documento a cada termo. Supondo uma consulta com dois termos onde um "termo A" ocorre com mais frequência que o "termo B", um documento que contenha o "termo B" mas não o "termo A", deve ter uma **classificação** maior do que um documento que contenha o "termo A" mas não o "termo B". Para solucionar este problema, pesos são atribuídos aos termos usando a **frequência de documento inversa** (IDF), definida conforme a Equação II.2. Nessa equação, $n(t)$ indica o número de documentos que contêm o termo t e N a quantidade de documentos do corpus [Robertson and Jones, 1976].

$$\mathrm{IDF}(t) = \log\left(\frac{N}{n(t)}\right) \tag{II.2}$$

Uma técnica normalmente utilizada, que incorpora os pesos tanto da frequência do termo como da frequência do documento, é conhecida como TF-IDF. Nesta técnica, a **relevância** de um documento D dado um conjunto de termos de uma consulta Q é, então, definida conforme a Equação II.3 [Silberschatz et al., 2011].

$$\mathrm{r}(D, Q) = \sum_{t \in Q} \mathrm{TF}(D, t) \times \mathrm{IDF}(t) \tag{II.3}$$

Como forma de representação de documentos, Salton [Salton et al., 1975] propôs um modelo de espaço vetorial, considerando um espaço constituído de documentos D_i, cada um identificado por um ou mais termos de índice T_j. A Figura II.2 ilustra a representação do modelo proposto. Dados os vetores de representação de dois documentos, é possível calcular um coeficiente de similaridade s entre eles, $s(D_i, D_j)$ que reflete o grau de similaridade.

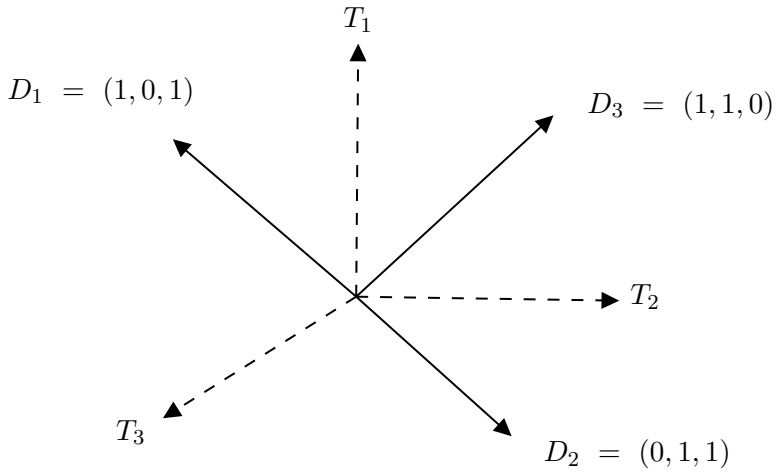


Figura II.2: Os documentos (D_1, D_2, D_3) estão representados em espaço vetorial tridimensional, onde cada dimensão refere-se a um termo que pode existir ou não documento representado.

Dentre inúmeras medidas de similaridade existentes, um ponto em comum é atender aos seguintes critérios:

- Menos peso é dado aos termos que aparecem em muitos documentos;
- Mais peso é dado a termos que aparecem muitas vezes em um documento;e
- Menos peso é dado a documentos que contêm muitos termos.

O objetivo é atribuir maior pontuação para documentos relevantes, favorecendo termos que parecem ser discriminatórios e reduzindo o impacto de termos que parecem ser distribuídos aleatoriamente [Zobel and Moffat, 2006]. Uma função de similaridade cosseno com normalização TFIDF atende aos critérios enumerados anteriormente, sendo o objetivo da função medir o cosseno do ângulo entre vetores multidimensionais que representem dois documentos. O cosseno entre o par de vetores não depende do tamanho dos mesmos mas somente do ângulo entre eles. Dado um par de vetores $X = (x_1, \ldots, x_i)$ e $Y = (y_1, \ldots, y_i)$, a similaridade cosseno é definida conforme a Equação II.4 [Aggarwal, 2022a], na qual x_i está relacionado à representação de X na i-ésima dimensão e y_i à representação de Y na i-ésima dimensão.

$$\text{Cosseno}(X, Y) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}} \tag{II.4}$$

Uma abordagem de processamento de linguagem natural puramente baseada no TF-IDF enquadra-se num modelo conceitual de representação textual conhecido como **bag-of-words** [bag]. Nesse modelo é desprezada a ordem na qual as palavras se apresentam no texto. Observa-se que, no TF-IDF, há uma computação numérica dos termos existentes nos documentos focada na frequência de ocorrência dos termos, não levando em consideração a ordem na qual os termos aparecem nos documentos.

Formulações de similaridade fundamentadas em princípios estatísticos possuem grande relevância conforme avaliações da Text REtrieval Conference (TREC) do National Institute of Standards and Technology (NIST), tendo grande destaque o sistema recuperação textual **Okapi** [Robertson et al., 1992] com função baseada no modelo probabilístico proposto por Robertson e Sparck Jones [Robertson and Jones, 1976], o qual, fundamentado no **Teorema de Bayes** estima a probabilidade de um documento D ser relevante para uma consulta Q.

Uma das instanciações mais conhecidas desse modelo probabilístico é o BM25 [Robertson and Zaragoza, 2009]. O BM25 (Equação II.5) expande o TF-IDF ao levar em consideração dois fatores que afetam a avaliação do conteúdo textual, são eles o **tamanho** de um documento e **frequência elevada** com que uma palavra aparece no texto. A **saturação** provocada por uma palavra que ocorre excessivamente, pode levar a uma avaliação equivocada do contexto de um documento, bem como afetar a avaliação de similaridade entre dois documentos.

$$\text{score}(D,Q) = \sum_{i=1}^{n} IDF(q_i) \times \frac{f(q_i,D) \times (k_1+1)}{f(q_i,D) + k_1 \times (1 - b + b \times \frac{|D|}{avgdl})} \tag{II.5}$$

Na equação acima, temos que:

- IDF(q_i) é a frequência de documento inversa do i-ésimo termo de consulta
- $f(q_i, D)$ é o número de vezes que q_i ocorre no documento D.
- $|D|$ é o comprimento do documento D em palavras.
- $avgdl$ é o comprimento médio de um documento pertencente ao corpus.
- k_1 e b são parâmetros livres.

Observe que $\frac{|D|}{avgdl}$ representa a razão entre o tamanho de um documento D e o tamanho médio de um documento do corpus. Esse fator, associado ao parâmetro b, pondera o impacto que o tamanho de um documento D exerce em uma consulta Q. Como exemplificação, supomos ser razoável que um termo $t \in Q$ com uma frequência igual a 1 em um documento D_1 com 10 palavras possua significância diferente em um documento D_2 de 100 palavras ainda que possua neste a mesma frequência 1.

O parâmetro k_1 está relacionado com a **saturação** da frequência de um termo em um documento. Busca-se impor um limite ao impacto que a frequência de um termo exerce em uma consulta, ou seja é um fator de ponderação para o componente $f(q_i, D)$ da equação. Por suposição, considere que para uma dada consulta Q seja relevante contabilizar a frequência de um termo até um valor máximo x desprezando qualquer ocorrência excedente. O parâmetro k_1 restringe o impacto da saturação de um termo num documento, limitando a curva da equação do BM25 a uma assíntota horizontal, diferentemente do que ocorre numa abordagem puramente TF-IDF (Figura II.3) [Plu].

II.2 Agrupamentos

O processo de gerar agrupamentos consiste em formar grupos a partir de um conjunto de objetos. Essa geração ocorre de modo que seja maximizada a semelhança entre objetos alocados a um mesmo grupo, e minimizada a semelhança entre objetos que pertençam a grupos distintos.

Em geral, os métodos de agrupamento podem ser classificados nas seguintes categorias [Han et al., 2011]:

- Métodos de particionamento - Dado um conjunto de n objetos, são construídas k grupos de dados, onde $k \leq n$. O algoritmo começa sua execução com um grupo inicial, de forma iterativa

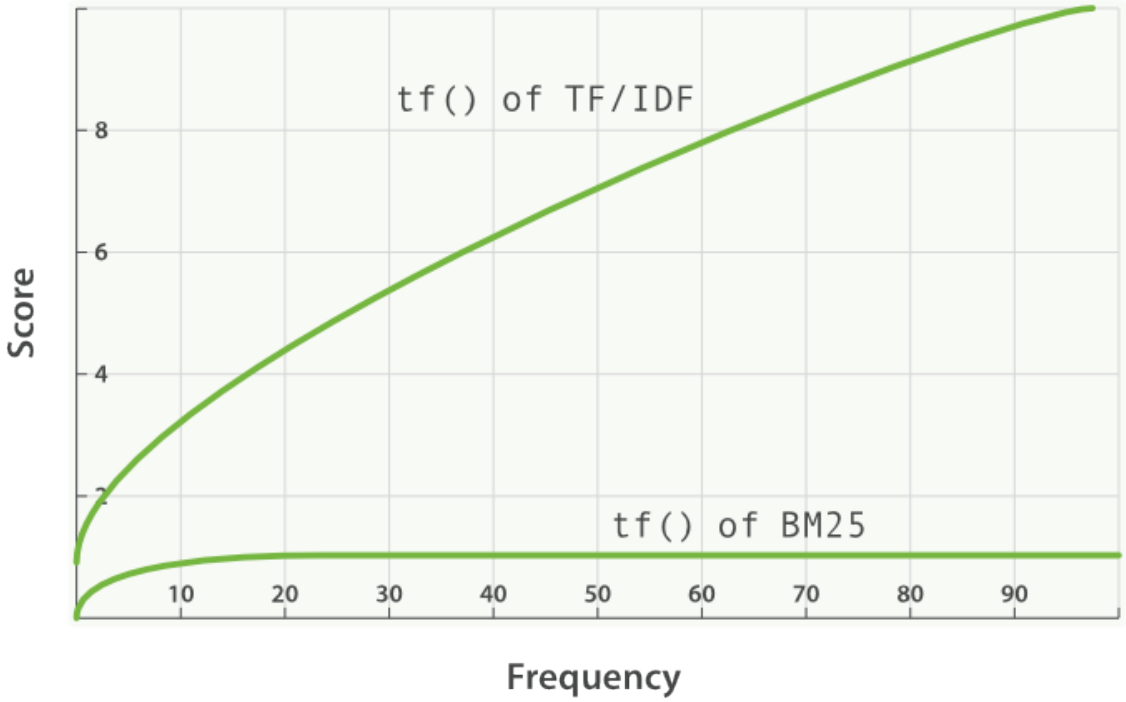


Figura II.3: Efeito da saturação no TF-IDF e no BM25. Fonte: `https://www.elastic.co/pt/blog/practical-bm25-part-2-the-bm25-algorithm-and-its-variables`

os objetos são movidos de uma grupo para outro na tentativa de encontrar um agrupamento ideal.

- Métodos hierárquicos - Dado um conjunto de objetos é feita uma decomposição hierárquica. Pode ser adotada um abordagem "bottom-up" na qual inicialmente cada grupo é formado por um único objeto e iterativamente vão ocorrendo junções até que seja atingida uma condição de parada estabelecida. Outra abordagem possível é a "top-down", na qual inicialmente existe um único grupo que contém todos os objetos, iterativamente são criados novos grupos e ocorre a distribuição dos objetos entre eles até que seja atingida uma condição de parada estabelecida.

- Métodos baseados em densidade - Considerando objetos distribuídos em um espaço de dados e para um dado valor de raio, os agrupamentos são formados identificando-se regiões de alta concentração de objetos em torno de pontos centrais, considerando o raio previamente estabelecido.

- Métodos baseado em grade - Formatam todo o espaço de dados em uma estrutura de grade com um número finito de células. Todas as operações de agrupamento são realizadas na estrutura da grade, sendo eficiente tratamento de dados espaciais. Apresentam como vantagem o processamento rápido, cujo tempo depende do número de células por dimensão e não da

quantidade de objetos. Podem ser integrados com outros métodos de agrupamento, como métodos baseados em densidade e métodos hierárquicos. Em geral a notação usada é a seguinte. Seja D um conjunto de dados de n objetos a serem agrupados. Um objeto é descrito por dm variáveis, onde cada variável também é chamada de atributo ou dimensão, e, portanto, também pode ser referido como um ponto em um espaço de dimensão dm.

II.2.1 DBSCAN

O DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*) é um algoritmo de agrupamento que pertence à família de métodos de agrupamento baseados em densidade. Esse algoritmo captura a percepção de que um grupo é um conjunto denso de pontos espacialmente próximos. A ideia é que, se um ponto específico pertencer a um grupo, ele deve estar próximo a muitos outros pontos nesse mesmo grupo.

A base conceitual do DBSCAN reside na caracterização de grupos como regiões de alta densidade em um espaço de características. Este método difere de técnicas tradicionais, pois não requer a especificação prévia do número de grupos desejados, proporcionando uma adaptabilidade intrínseca a estruturas de dados de natureza variada.

O DBSCAN possui dois hiperparâmetros, a saber, $\epsilon \in \Re$ e minPoints $\in Z$. No início de sua execução, o DBSCAN seleciona aleatoriamente um ponto p do conjunto de dados. Se houver mais do que minPoints pontos a uma distância até ϵ de p (incluindo o próprio ponto original), o DBSCAN considera todos eles como parte de um grupo. Em seguida, esse grupo é expandido verificando todos os novos pontos e verificando se eles também têm mais de pontos minPoints a uma distância de ϵ, aumentando o grupo de forma recursiva. Eventualmente, não haverá mais pontos para adicionar ao grupo. Nesse momento, o DBSCAN escolhe um novo ponto arbitrário e repete o processo.

Pode acontecer de o ponto p selecionado ter menos do que minPoints pontos em sua "bola" de raio ϵ e também não faça parte de nenhum outro grupo. Se for esse o caso, p é considerado um "ponto de ruído" que não pertence a nenhum grupo.

II.2.2 HDBSCAN

O algoritmo HDBSCAN (*Hierarchical Density-Based Spatial Clustering of Applications with Noise*) é uma extensão do DBSCAN [Campello et al., 2015]. O procedimento de agrupamento inicia-se pela construção de uma estrutura de dados em árvore e baseada na densidade, chamada de Árvore de Alcançabilidade. Esta árvore codifica informações sobre a conectividade entre pontos de dados, priorizando regiões de alta densidade e revelando potenciais estruturas hierárquicas.

A etapa subsequente compreende a extração de grupos da Árvore de Alcançabilidade, por meio de uma técnica de corte denominada Corte Hierárquico. Este processo resulta em uma hierarquia

de grupos, proporcionando uma visão granular das estruturas presentes nos dados.

O HDBSCAN é particularmente eficaz na identificação de grupos de densidades variáveis e formas complexas, além de apresentar robustez à presença de ruído e *outliers*. Sua capacidade de explorar a hierarquia densidade-conectividade o tornam uma escolha relevante em diversas aplicações, especialmente em contextos onde a estrutura dos dados é desconhecida ou altamente heterogênea. A complexidade computacional do HDBSCAN é mitigada por estratégias eficientes de indexação e algoritmos paralelos.

II.3 Sumarização extrativa

A sumarização extrativa consiste na aplicação de técnicas para extração de termos ou frases inteiras de um documento ou de um conjunto de documentos. Existem três variações principais em sistemas extrativos [Mehta and Majumder, 2019a], conforme descrição a seguir.

- Baseados em tópicos - Focam em atribuir importância relativa de palavras ou sentenças. O sistema pode atribuir pesos baseados na frequência dos termos. Com o armazenamento de informações sobre as frequências é possível selecionar as palavras ou sentenças mais importantes.

- Baseados em centralidade - Tentam identificar a sentença que melhor representa o conteúdo geral do documento. É estabelecido um ranqueamento, a sentença considerada mais importante é aquela cujo conteúdo é compartilhado com o maior número de outras sentenças do documento.

- Sumário geral - Trata a sumarização como um problema de otimização, tenta identificar qual subconjunto de sentenças melhor representa o documento. Não tem foco em sentenças individuais.

II.3.1 Modelagem de tópicos

Considere os conjuntos $A = \{$leão, onça, gato, tigre, jaguar$\}$ e $B = \{$porsche, ferrari, mercedes, jaguar$\}$. É perceptível que há conceitos ocultos (ou latentes) relacionados a estes conjuntos. O conjunto A refere-se a felinos e B a carros. Podemos dizer que textos que contêm elementos do conjunto A possuem uma relação semântica em torno de um tópico e os que contêm elementos do conjunto B se relacionam por um outro tópico.

Coleções de documentos podem ser representadas por matrizes de termos de documentos, como na Figura II.1. As palavras em um documento geralmente estarão predominantemente relacionadas a um tópico específico, o que causará correlações entre os atributos da matriz (termos). Estas

correlações podem ser aproveitadas para criar uma representação de baixa dimensão dos dados, e este processo é referido como redução de dimensionalidade.

Retomando o exemplo dos conjuntos A e B. Considere que em uma coleção, num certo grupo de documentos há o predomínio de palavras do conjunto A e em outro grupo a maioria das palavras pertence ao conjunto B. Intuitivamente, podemos expressar os documentos dessa coleção em função de certas características, neste caso felinos e carros.

Em resumo, um documento contendo a maioria das palavras do primeiro conjunto pode ser expresso como $(a, 0)$, um documento contendo a maioria das palavras do segundo conjunto pode ser expresso aproximadamente como $(0, b)$ e um documento contendo muitas palavras de ambos os conjuntos pode ser expresso como (c, d). Pode-se ver este novo conjunto de coordenadas como uma representação reduzida dos dados.

O que se busca é uma redução na dimensionalidade de representação de documentos, de modo que a maior parte do conhecimento semântico do corpus é retida e apenas o ruído é perdido. Como resultado, muitos algoritmos de recuperação da informação e mineração de texto mostram maior precisão quando a representação reduzida é usada no lugar da representação original. Pode-se dizer que no exemplo mencionado, foram extraídos os conceitos semânticos ocultos(ou latentes) "felinos" e "carros".

Esse procedimento de criar uma forma de representação resumida que explicita conceitos ocultos pode-se chamar de **modelagem de tópicos** [Aggarwal, 2022a]. A ideia central por trás da modelagem de tópicos é que cada documento pode ser visto como uma combinação de vários tópicos, e cada palavra em um documento contribui para um ou mais desses tópicos. Existem várias abordagens para implementar o conceito da modelagem de tópicos, e algumas das técnicas mais comuns incluem:

- Latent Dirichlet Allocation (LDA): Modelo probabilístico que assume que os documentos são uma mistura de tópicos e que as palavras em um documento são atribuídas a esses tópicos com uma certa probabilidade.
- Non-Negative Matrix Factorization (NMF): Técnica de álgebra linear que representa documentos como combinações lineares de tópicos, onde as entradas são não-negativas.
- Modelos Baseados em Redes Neurais: Usam modelos de linguagem baseados em redes neurais para a modelagem de tópicos, um exemplo é a técnica BERTopic [Grootendorst, 2022].

BERTopic

A biblioteca BERTopic [Grootendorst, 2022] é uma implementação Python projetada para facilitar a tarefa de agrupamento de tópicos em conjuntos de dados de texto. Essa biblioteca disponibiliza

um modelo de rede neural que implementa o conceito da modelagem de tópicos. Concretamente, é uma biblioteca que tem o propósito de identificar grupos de documentos de um corpus de entrada que compartilham tópicos semelhantes. Baseada na arquitetura BERT (Bidirectional Encoder Representations from Transformers), a BERTopic oferece uma abordagem não-supervisionada para a identificação de tópicos associados a documentos de um corpus.

O mecanismo subjacente da BERTopic opera por meio da projeção dos documentos de texto em um espaço latente de representações semânticas, onde a proximidade espacial entre os documentos reflete relações semânticas consideradas relevantes. Isso é alcançado por meio da extração de *embeddings* de texto utilizando modelos BERT pré-treinados, os quais são capazes de capturar nuances semânticas e contextuais.

São aplicadas técnicas de agrupamento que permitem a identificação de estruturas subjacentes e a alocação de documentos relacionados a tópicos similares. A BERTopic é configurada, por padrão, para usar o algoritmo de agrupamento HDBSCAN. Cada grupo identificado pelo HDBSCAN é considerado como um tópico. Para atribuir rótulos aos tópicos, o BERTopic utiliza as palavras mais representativas de cada grupo. Essas palavras são extraídas a partir das palavras mais frequentes e distintas dentro de cada grupo. Concretamente, a BERTopic computa uma variante da medida clássica TD-IDF denominada cTFIDF. Essa variante é computada por meio da Equação II.6.

$$w_{t,c} = \mathrm{tf}_{t,c} \times \log(1 \ + \ \frac{A}{\mathrm{tf}_t}) \tag{II.6}$$

Na Equação II.6, temos que:

- $\mathrm{tf}_{t,c}$ = frequência da palavra t no grupo c.
- tf_t = frequência da palavra t considerando todos os grupos.
- A = número médio de palavras por grupo.

O valor $w_{t,c}$ codifica a importância da palavra t dentro nos documentos do grupo c. Considerando um dos grupos gerados, as palavras são ordenadas (em ordem decrescente) com base na sua pontuação de relevância cTFIDF. Essa ordenação destaca as palavras que são distintivas para o grupo em questão. Uma vez ordenadas, as palavras mais relevantes são selecionadas como as palavras representativas do grupo. A quantidade de palavras selecionadas pode ser configurada pelo usuário ou seguir um critério predefinido.

BERTopic guiado

Considerando o contexto de atribuição de tema a um recurso especial, note que as palavras que compõem cada tema $t \in \mathcal{T}$ são conhecidas. Note também que o conjunto $\mathcal{T}$ possui todos os temas

possíveis para um recurso r.

A biblioteca BERTopic disponibiliza uma forma de **guiar** o procedimento de geração do tópico que representa um recurso r. Para isso, implementamos uma variação da abordagem BERTopic. Nessa variação, temos o intuito de influenciar as palavras candidatas a compor o tópico representativo de um recurso. A ideia é que, dentre todas as palavras possíveis, haja uma probabilidade maior de serem selecionadas para o tópico palavras que pertençam ao conjunto de temas $\mathcal{T}$.

O mecanismo da **modelagem de tópicos guiada**, esquematizado na Figura II.4, opera basicamente da seguinte forma:

- Criamos uma lista de **tópicos iniciais**, onde cada elemento da lista corresponde ao conjunto de palavras de um tema repetitivo pré-processado.
- Configuramos o hiperparâmetro **seed_topic_list** do modelo BERTopic para receber a lista de tópicos iniciais.
- O modelo cria representações vetoriais desses tópicos iniciais e de cada recurso.
- É criada uma matriz para armazenar informações de similaridade entre vetores de recursos e vetores de tópicos iniciais.
- Para cada vetor de recurso é verificado se é maior a similaridade com o vetor de algum tópico inicial ou se é maior com um vetor médio que representa todos os recursos. Caso seja maior com algum tópico, é registrada na matriz uma etiqueta referente a esse tópico caso contrário é registrado o valor -1 indicando que não há um tópico inicial com similaridade relevante em relação ao recurso.
- Os dados sobre recursos que têm similaridade com um tema comum, influenciam na composição de grupos com base em densidade feita pelo Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN)
- Uma matriz tf-idf é criada, e palavras que pertencem aos tópicos iniciais recebem um peso superior às demais. O modelo BERTopic por padrão aplica o peso **1,2**. Não é disponibilizado ao usuário do modelo uma interface para modificar esse valor de peso atribuído.
- Ao fim do processo, é aplicado o algoritmo cTFIDF(Seção II.3) para identificar as palavras que são parte de um tópico representativo de um grupo, contudo algumas palavras possuem uma probabilidade maior em compor um tópico devido ao fator de ponderação recebido.

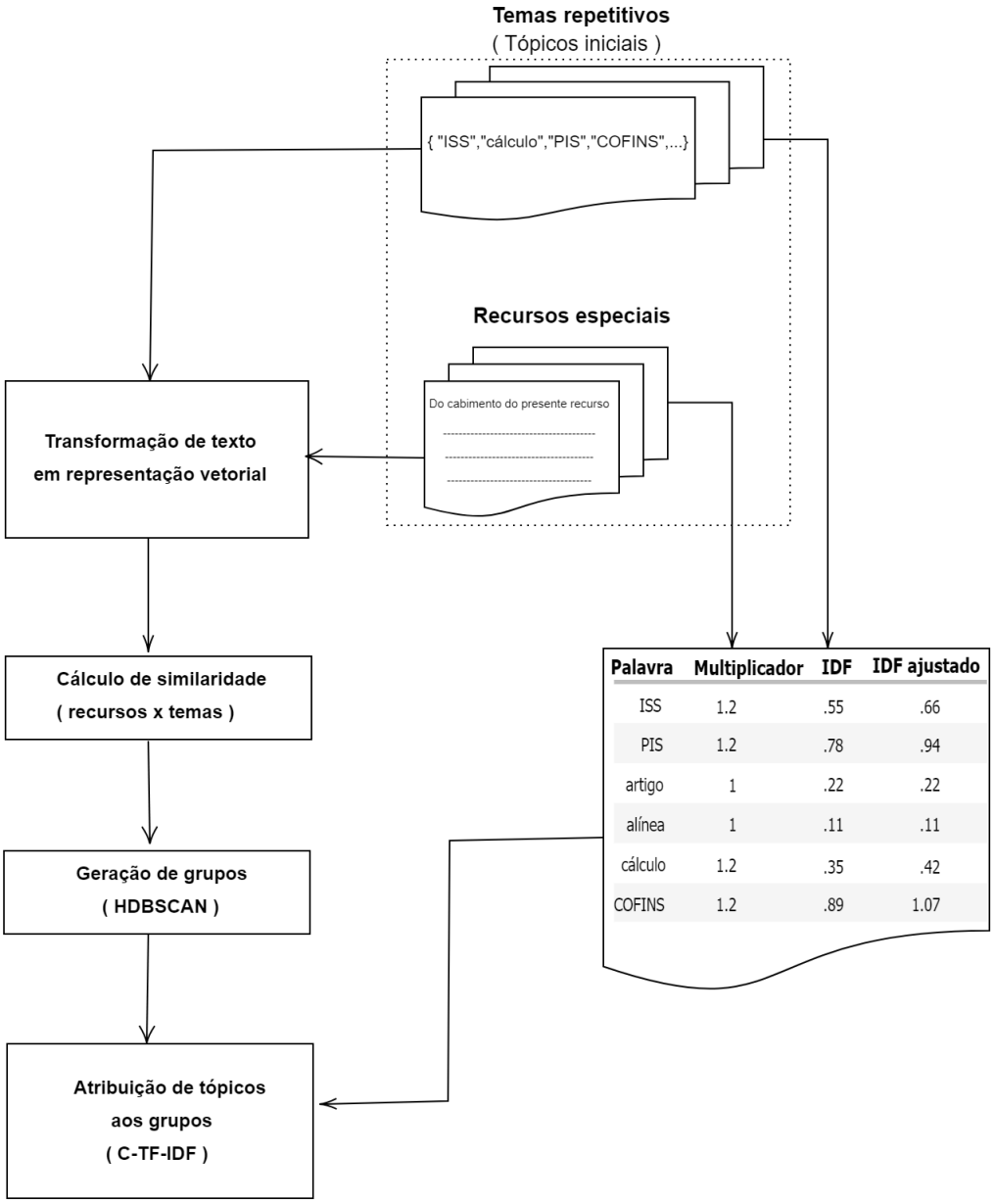


Figura II.4: Ilustração do funcionamento do BERTopic guiado. A técnica permite guiar a modelagem de tópicos mediante uma lista inicial de palavras (Tópicos iniciais). Essas palavras influenciam o processo de atribuição de tópicos, ajudando a direcionar a identificação de tópicos em textos conforme as preferências do usuário. Na figura, observa-se o fluxo geral: a representação vetorial de cada texto é gerada, os textos são agrupados com base em sua semelhanças semânticas, e, posteriormente, os tópicos são representados pelas palavras mais relevantes associadas a cada grupo, respeitando as diretrizes dos Tópicos iniciais.

II.3.2 Algoritmos baseados em grafo

Um grafo $G(V, E)$ é uma estrutura matemática que consiste em um conjunto não vazio de objetos denominados vértices $v \in V$, também chamados nós, e um conjunto de arestas $e \in E$ que conectam esses vértices. A teoria dos grafos possui aplicabilidade em diversas áreas do conhecimento, uma vez que os objetos de um grafo são abstrações que podem representar inúmeros tipos de entidades.

As arestas de um grafo podem apresentar atributos como direção e/ou pesos associados.Um grafo é uma entidade abstrata contudo é comum representá-lo graficamente. Uma representação gráfica para o grafo G abaixo pode ser vista na Figura II.5.

$$V = \{v_1, v_2, v_3, v_4, v_5, v_6\}, E = \{\{v_1, v_2\}, \{v_1, v_5\}, \{v_2, v_3\}, \{v_2, v_5\}, \{v_3, v_4\}, \{v_4, v_5\}, \{v_4, v_6\}\} \tag{II.7}$$

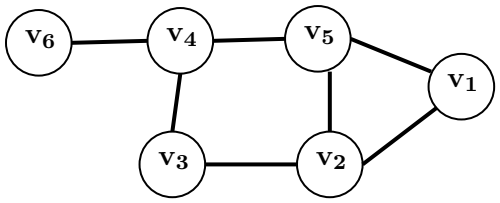


Figura II.5: Representação gráfica do grafo G

Uma forma comumente utilizada para representar um grafo é através da matriz de adjacência. Dado um grafo G com n vértices, podemos representá-lo como uma matriz $n \times n$ denotada por $A(G) = [a_{ij}]$, onde a_{ij} representa a relação entre os vértices v_i e v_j. Como exemplo, a matriz de adjacência abaixo é uma outra forma de representar o grafo apresentado na Figura II.5.

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

PageRank

O algoritmo **PageRank** usa a estrutura de **hiperlinks** das páginas **web** para criar um ranking com base em **reputação**. Tomando um grafo G como abstração, as páginas web são representadas como vértices $v \in V$ e os hiperlinks entre as páginas são as arestas $e \in E$. Um vértice recebe um peso (reputação) que está associado às relações com outros vértices (Figura II.6). A ideia básica é

que páginas com alta reputação estão ligadas a outras páginas que também possuem alta reputação [Aggarwal, 2022a].

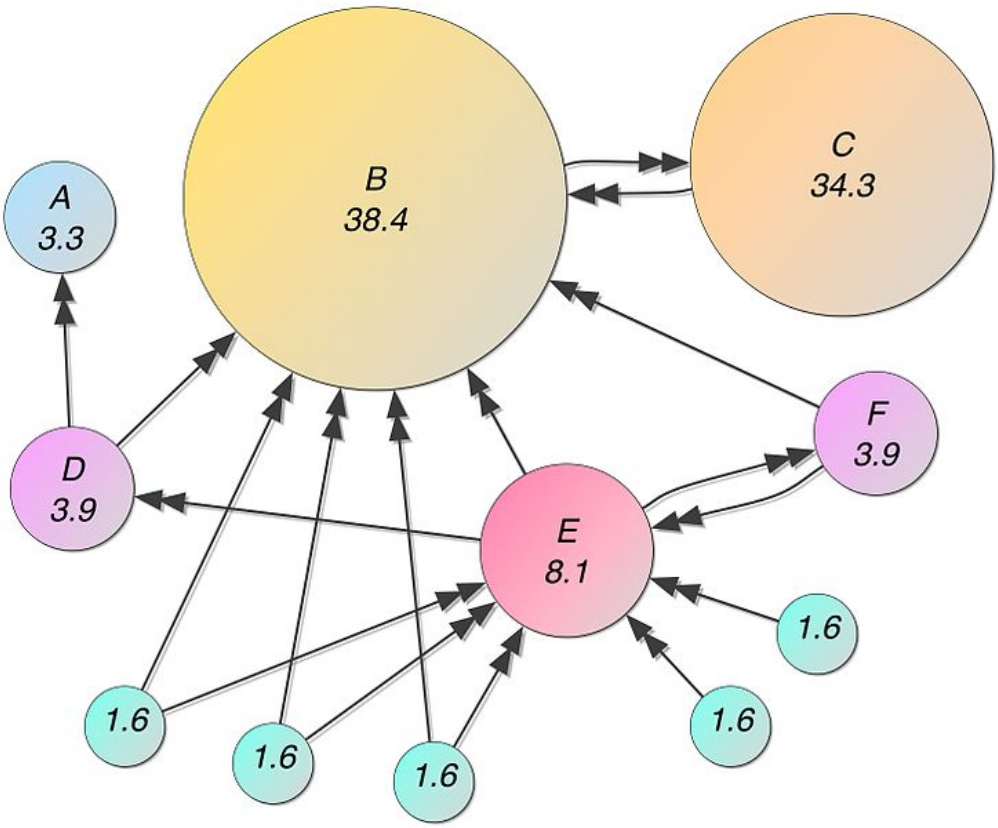


Figura II.6: Ilustração do algoritmo PageRank em um grafo ponderado. Os nós representam páginas web e estão associados a valores que indicam sua importância relativa. As arestas direcionadas simbolizam os hiperlinks entre as páginas, enquanto os pesos dos nós refletem as pontuações calculadas pelo algoritmo. A importância de cada página é determinada iterativamente, considerando tanto a quantidade quanto a qualidade dos links recebidos, bem como o número de links de saída das páginas que apontam para ela. Fonte: Mehta and Majumder [2019b].

LexRank

A ideia central do LexRank é representar o texto como um grafo, onde os nós representam as sentenças e as arestas representam a similaridade entre as sentenças. A similaridade entre duas sentenças pode ser medida usando alguma métrica de similaridade de texto, como a similaridade do cosseno. O algoritmo então calcula a **centralidade** de cada sentença no grafo, indicando sua importância relativa no documento [Erkan and Radev, 2004].

Na execução do algoritmo, os dados de similaridade entre sentenças são armazenados em uma **matriz de similaridade** S, onde S_{ij} contem o valor de similaridade entre as sentenças i e j. Os dados da linha da matriz são normalizados para garantir que cada linha some o valor 1. Isso é feito dividindo cada elemento da linha pelo somatório de todos os elementos naquela linha, gerando uma

matriz M normalizada.

$$M(i,j) = \frac{S(i,j)}{\sum_k S(i,k)} \tag{II.8}$$

A matriz M representa um grafo ponderado, onde cada nó é uma sentença e as arestas são ponderadas pela similaridade entre as sentenças correspondentes. O algoritmo de PageRank é aplicado ao grafo ponderado.

$$\mathrm{PR}(i) = (1-\delta) + \delta \times \sum_j (\frac{M_{ji}}{\sum_k M_{jk}} \times PR(j)) \tag{II.9}$$

Onde:

- PR(i) é a pontuação de PageRank para o nó (sentença)i
- δ é um fator de ponderação(normalmente entre 0.1 e 0.2

As pontuações de PageRank PR(i) são usadas para ranquear as sentenças. Sendo n um parâmetro definido pelo usuário do algoritmo, as n sentenças mais bem ranqueadas comporão o resumo do texto.

Capítulo III Trabalhos Relacionados

No cenário brasileiro, com dezenas de milhões de processos jurídicos em tramitação [de Justiça, 2021] e recursos escassos, o sistema judiciário mostra-se como um campo proeminente para aplicação de técnicas de inteligência artificial. Este projeto alinha-se com essa demanda ao analisar a aplicabilidade de técnicas de processamento de linguagem natural no tratamento de processos jurídicos. Diante do exposto, o objetivo desta revisão sistemática de literatura é investigar técnicas empregadas no processamento de documentos legais.

A estrutura do capítulo está organizada em três seções principais. Na Seção III.1, detalhamos as etapas do processo de busca e seleção das publicações; na Seção III.2, descrevemos os critérios aplicados para determinar a elegibilidade e priorização dos estudos analisados; e, finalmente, na Seção III.3, apresentamos os resultados da revisão, destacando os principais trabalhos selecionados e como eles se relacionam com a abordagem proposta nesta dissertação.

III.1 Seleção de publicações

Para realização desta revisão foi selecionada a base de dados Scopus. A escolha desta base foi devida a credibilidade e ao tamanho do acervo disponível. O tipo de documentos buscados ficou restrito a artigos efetivamente publicados em revistas científicas ou trabalhos apresentados em conferências. Não houve restrição temporal e os termos da busca foram combinados da seguinte forma:

- “legal” AND “document” AND “classification”
- “legal” AND “document” AND “topic” AND “modeling”
- “guided” AND “topic” AND “modeling”
- “lexrank” AND “document” AND “summarization”

A seleção dos artigos foi conduzida com base nas diretrizes do método PRISMA [Page et al., 2021]. A Figura III.1 detalha o fluxo da seleção dos artigos.

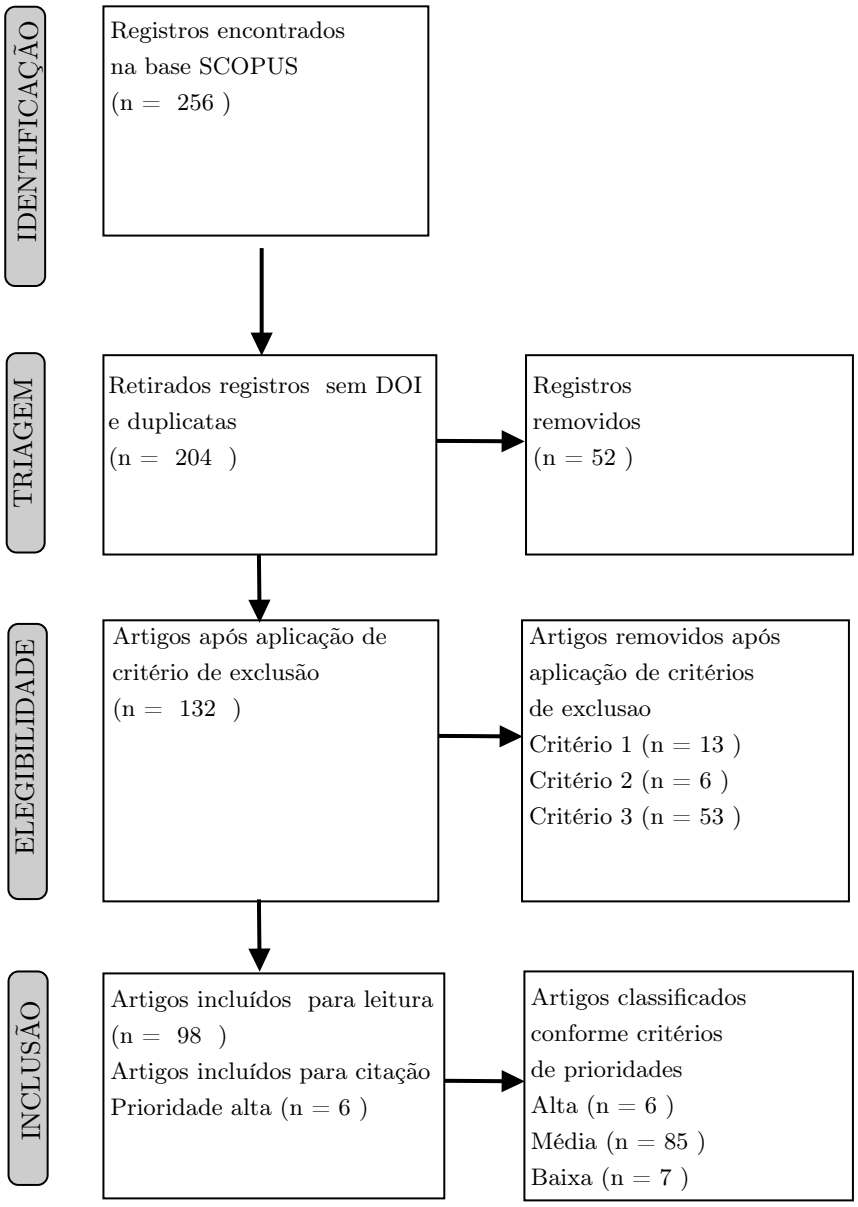


Figura III.1: Fluxo de seleção de artigos com base na metodologia PRISMA. O processo é dividido em quatro etapas principais: identificação, triagem, elegibilidade e inclusão. Na etapa de identificação, foram encontrados registros na base Scopus. Durante a triagem, registros sem DOI ou duplicados foram removidos. Na fase de elegibilidade, foram aplicados critérios de exclusão, resultando em um subconjunto de artigos considerados relevantes. Por fim, na etapa de inclusão, os artigos foram classificados com base em prioridades para citação e análise. Os números indicam a quantidade de registros processados em cada etapa.

III.2 Critérios de inclusão, exclusão e priorização

A aplicação da metodologia PRISMA envolveu a definição de critérios específicos de inclusão, exclusão e priorização dos estudos, com o objetivo de filtrar os trabalhos mais relevantes e representativos para o contexto desta dissertação.

III.2.1 Critérios de inclusão

Os critérios de inclusão possuem a função de direcionar o resultado ao assunto escolhido.

- **Critério 1 -** Trata de documentos jurídicos
- **Critério 2 -** Trata de modelagem de tópicos
- **Critério 3 -** Trata de sumarização de documentos

III.2.2 Critérios de exclusão

Os critérios de exclusão possuem a função de excluir resultados que não satisfazem restrições definidas para o trabalho.

- **Critério 1 -** Não relacionado a classificação de documentos
- **Critério 2 -** Trata de sentença em processo jurídico
- **Critério 3 -** Sem citação registrada em outra publicação

III.2.3 Critérios de priorização

Critérios que definem a relevância do artigo encontrado.

- **Prioridade alta -** Classificação de documento jurídico por modelagem de tópicos
- **Prioridade média -** Classificação de documento jurídico por método diverso da modelagem de tópicos
- **Prioridade baixa -** Sumariação extrativa de documento

III.3 Resultados

Aplicamos a metodologia PRISMA, conforme descrito na Figura III.1, e concluímos a pesquisa de publicações relacionadas na data de 14/08/2024. Apresentamos a seguir referências às publicações selecionadas com uso da metodologia.

III.3.1 Publicações selecionadas

Vianna et al. [2023] propõe uma solução baseada em técnicas de descoberta de tópicos para analisar, resumir e classificar documentos jurídicos. Eles usam conjuntos de dados do sistema jurídico brasileiro analisados por juristas para confirmar sua abordagem. O método proposto é dividido em três fases:

1. Filtragem — De acordo com critérios definidos por especialistas jurídicos, expressões regulares são usadas para selecionar um subconjunto de uma ampla coleção de documentos.

2. Pré-processamento — Remove pontuação, espaços em branco e outros termos definidos.

3. Descoberta — é a fase central do método em que quatro técnicas diferentes foram implementadas e avaliadas.

Este trabalho está relacionado ao nosso na medida em que também buscamos o tema intrínseco a um documento jurídico. No entanto, neste trabalho relacionado, especialistas humanos analisam os temas atribuídos aos documentos para atestar sua qualidade. Em nossa abordagem, calculamos a similaridade entre textos para atribuirmos o tema ao documento jurídico. Pelo fato de usarmos um conjunto de dados previamente rotulados por especialistas, podemos verificar a correta atribuição dos temas usando métricas conhecidas. Dessa forma, minimizamos qualquer viés que possa interferir na qualificação dos temas atribuídos. Outro ponto de diferenciação é que entre os conjuntos de documentos usados pelos autores, há um número considerável de documentos com tamanho médio de 220 palavras. Em comparação, o tamanho médio do corpus que usamos foi de aproximadamente quatro mil palavras. Ainda, como diferença, em nossos experimentos, buscamos comparar a modelagem BERTopic adotada por Vianna et al. [2023], com a sumarização baseada em grafos, sendo esta última o foco do nosso método.

Dalal et al. [2023] usa aprendizado de máquina para resumir documentos e parafraseá-los em uma linguagem mais direta no trabalho relacionado a precedentes legais indianos. O método proposto é uma combinação de duas técnicas de sumarização. LexRank [Erkan and Radev, 2004], um método baseado em grafo não supervisionado, é inicialmente aplicado ao documento para reduzi-lo a um comprimento correspondente a 35% do tamanho original. A segunda etapa usa o modelo de linguagem neural PEGASUS [Zhang et al., 2020] para sumarização de texto.

O PEGASUS é baseado na arquitetura Transformer [Vaswani et al., 2023]. Ele segue o padrão codificador-decodificador, onde o codificador processa o texto original para extrair representações contextuais enquanto o decodificador gera o resumo com base nessas representações. O codificador compreende várias camadas de codificação, enquanto o decodificador também tem várias camadas para gerar um resumo token por token. Uma camada linear é aplicada para mapear as saídas do decodificador para o vocabulário, permitindo a previsão das seguintes palavras no resumo. No método proposto por Dalal et al. [2023], o resumo extrativo gerado pelo modelo LexRank é dividido em frases e alimentado para o modelo PEGASUS. O modelo então deriva paráfrases para cada frase, tornando-a mais simples de entender. As frases são então mescladas para fornecer um resumo abstrato final.

Nosso trabalho está relacionado a Dalal et al. [2023], que usou o algoritmo LexRank baseado

em grafo para criar o resumo do documento legal. No entanto, como característica distintiva do nosso método, modificamos o algoritmo LexRank original, buscando uma maneira de orientá-lo no processo de sumarização e, assim, melhorar o resumo gerado.

Satwick Gupta et al. [2022] comparam técnicas de resumo e as aplicam a documentos legais. O método de trabalho abrange os estágios de Aquisição de Conjunto de Dados, Limpeza de Dados, Sumarização de Texto, Avaliação de Resumo e Análise Comparativa. No estágio de sumarização extrativa, as seguintes técnicas são comparadas: LexRank, TextRank, Redução, Luhn, Edmundson, Análise Semântica Latente e SumBasic. Os resultados mostram que abordagens baseadas em grafos, como LexRank e TextRank, têm melhor desempenho do que abordagens baseadas em frequência. Similarmente a Satwick Gupta et al. [2022], atestamos o bom desempenho da abordagem baseada em grafos, mas fazemos uma comparação com métodos que implementam modelagem de tópicos.

Vuong et al. [2022] propõe uma abordagem para resolver uma tarefa de jurisprudência apresentada na competição Legal Information Extraction and Entailment (COLIEE)[1]. A tarefa é encontrar casos jurídicos relevantes dado um caso de consulta de entrada. Os dados para a tarefa são baseados na jurisprudência do Tribunal Federal do Canadá.

Na primeira etapa, o objetivo é extrair do corpus os casos candidatos que apoiam a decisão do caso de consulta. Na segunda etapa da tarefa, dado um trio incluindo um caso de consulta, uma decisão do caso de consulta e uma lista de parágrafos candidatos dos documentos legais de suporte na fase anterior, esta fase visa identificar um ou mais parágrafos dos casos de documentos de suporte que implicam a decisão do caso de consulta.

Tanto a consulta quanto os documentos no corpus são textos com média de cerca de 3.000 palavras. Similarmente ao nosso trabalho, Vuong et al. [2022] reduz o texto longo em componentes semelhantes a parágrafos usando um algoritmo baseado em grafo para resolver a tarefa. Na abordagem Vuong et al. [2022], o resumo do texto rotula fracamente os dados avaliados. Os dados rotulados enviam as informações como entrada para um modelo BERT pré-treinado. Além do resultado produzido pelo modelo BERT, e similar ao que fazemos em nossa abordagem, Vuong et al. [2022] submete os textos ao algoritmo BM25 e usa o valor de saída do algoritmo como uma medida de similaridade.

Vuong et al. [2022] aplica uma pontuação combinada entre o resultado produzido pelo modelo BERT e BM25 para recuperar casos legais relevantes para um determinado caso de consulta. A abordagem proposta apresenta os documentos relevantes e os parágrafos que dão suporte à escolha.

Li et al. [2022] empregam técnica de modelagem de tópicos guiada na tarefa de classificação de documentos legais fornecidos pelo *California Innocence Project* e do conjunto de dados *20 Newsgroups*. O método proposto apresentou resultado superior ao ser comparado com os métodos

[1] `https://sites.ualberta.ca/~rabelo/COLIEE2020/`.

Fatoração de matriz não negativa semi-supervisionada (SSNMF), **Fatoração de matriz não negativa guiada (NMF guiada)** e **Supervisionado por tópico NMF**.

Conforme descrito no trabalho, o SSNMF é capaz de classificar diferentes documentos por meio de determinadas informações do rótulo, enquanto o NMF guiado pode orientar o conteúdo dos tópicos gerados por meio de palavras-semente a priori. O modelo proposto chamado **NMF Semi-Supervisionado Guiado (GSSNMF)**, é um modelo mais abrangente que pode aproveitar tanto informações de rótulos quanto palavras-semente importantes para melhorar o desempenho tanto na classificação multi-rótulos quanto na modelagem de tópicos.

III.3.2 Comparação

A Tabela III.1 apresenta uma visão geral das principais características dos trabalhos mencionados nesta seção. Ela sintetiza aspectos relevantes, como as técnicas empregadas, o uso de métodos baseados em grafos, a aplicação à classificação de documentos e quanto à abordagem comparativa de técnicas. A estrutura comparativa desta tabela visa auxiliar na contextualização da contribuição deste trabalho, permitindo destacar semelhanças e diferenças entre os trabalhos.

Tabela III.1: Visão geral das características dos trabalhos relacionados.

Trabalho	**Técnica Principal**	**Uso de Grafos**	**Classificação de Documentos**	**Abordagem comparativa**
Vianna et al. [2023]	Descoberta de tópicos com BERTopic		x	
Dalal et al. [2023]	Sumarização com LexRank e PEGASUS	x		
Satwick Gupta et al. [2022]	Comparação de técnicas de sumarização	x		x
Vuong et al. [2022]	Recuperação de casos jurídicos com BM25 e BERT	x	x	
Li et al. [2022]	Modelagem de tópicos guiada (GSSNMF)		x	x
Este trabalho	Recuperação de casos jurídicos com BM25 e LexRank guiado	x	x	x

Capítulo IV Metodologia

Neste capítulo, descrevemos a metodologia proposta para resolver o problema abordado nesta dissertação, a saber, a identificação de temas relevantes para um recurso especial. A Seção IV.1 aborda a formulação do problema analisado nesta dissertação. Na Seção IV.2 são apresentadas características dos documentos jurídicos utilizados neste projeto. Na Seção IV.3, são descritos os dados usados deste trabalho. Finalmente, a Seção IV.4 descreve em detalhes cada passo da metodologia proposta.

IV.1 Formulação do problema

Em concordância com o que foi explicitado na Seção I.3, o problema a ser tratado nesta pesquisa refere-se ao trabalho de analisar o texto de um recurso especial e atribuí-lo a um tema específico dentre todos os temas existentes na base de dados do STJ.

No domínio do Aprendizado de Máquina, considerando a abordagem desta pesquisa, o problema de classificar um recurso especial em um tema pode ser desmembrado em dois subproblemas. O primeiro, refere-se à sumarização do conteúdo do documento destacando seus pontos mais relevantes. O segundo problema refere-se à avaliação da similaridade entre esse resumo criado e o conteúdo de um tema judicial, procedendo de fato a classificação do documento em um tema. Um esquema da solução proposta pode ser visto na Figura IV.1.

Aqui, apresentamos a formulação do problema para o qual propomos nossa metodologia. Considere um documento correspondente a um recurso especial e denotado por r. Considere também um conjunto de temas repetitivos, que denotamos por $\mathcal{T}$. O problema que consideramos corresponde à identificação de um subconjunto $\mathcal{T}' \subset \mathcal{T}$ de temas relacionados ao conteúdo do documento r. Além disso, para cada tema $t \in \mathcal{T}'$, desejamos produzir um valor σ_{rt} que indica o quão relacionados estão os documentos r e t e que permite ordenar os elementos em $\mathcal{T}' \subset \mathcal{T}$. Cabe ressaltar que um analista humano ao avaliar os elementos em $\mathcal{T}'$ selecionará no máximo um elemento que seja de fato o tema adequado ao recurso r.

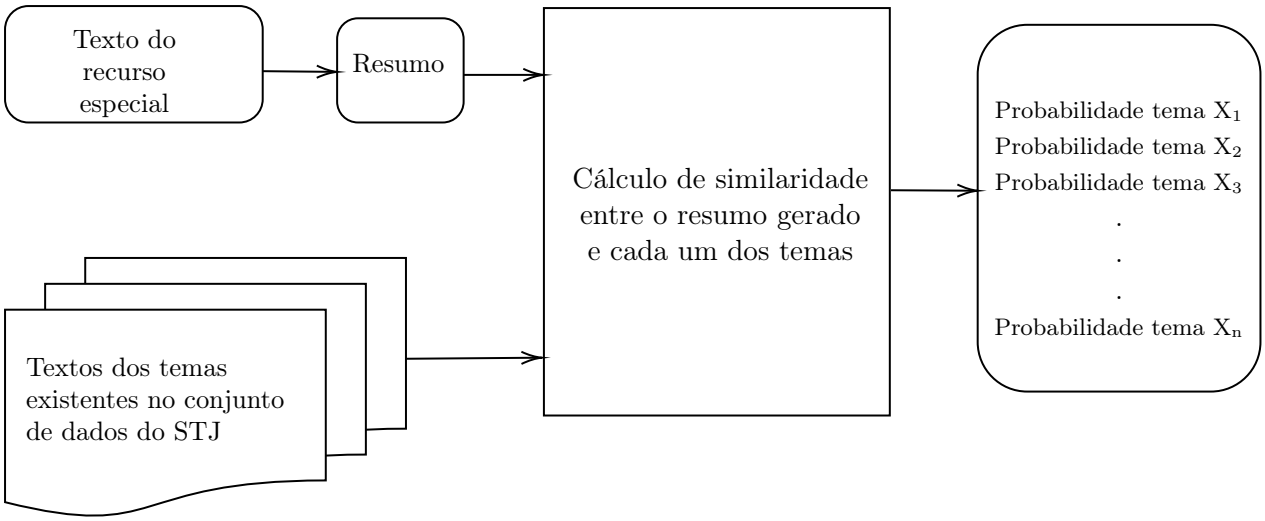


Figura IV.1: O esquema representa o método desenvolvido para a atribuição de tema ao recurso especial. O texto do recurso especial é sumarizado. A partir do resumo gerado é computada a similaridade com os temas. Como saída o método apresenta uma lista ordenada dos temas mais adequados ao recurso especial de entrada.

IV.2 Documentos Jurídicos

Fundamentalmente esta pesquisa lida com dois tipos de dados textuais, o recurso especial e o tema repetitivo. Apesar de não existir uma formalização da estrutura desses tipos de dados, eles apresentam certas particularidades no formato.

IV.2.1 Recursos Especiais

Conforme descrito na Seção I.1, o recurso especial é um meio de recorrer uma decisão judicial, previsto no artigo 105 da Constituição Federal. O julgamento deste tipo de recurso é de competência do STJ e visa uniformizar o entendimento jurídico sobre o regramento brasileiro.

Esse tipo de recurso é cabível quando a decisão recorrida: contrariar tratado ou lei federal, ou negar-lhes vigência; julgar válido ato de governo local contestado em face de lei federal; der a lei federal interpretação divergente da que lhe haja atribuído outro tribunal.

Embora não haja um modelo padrão específico para o recurso especial, o CPC é o guia que orienta a elaboração do documento e de certa forma indica as seções que devem existir no documento. O apêndice A apresenta um modelo de recurso especial comumente adotado.

IV.2.2 Temas Repetitivos

As informações dos temas repetitivos são disponibilizadas em formatos tabulares pelo STJ[STJ]. No apêndice B é apresentado um exemplo desses dados. O elemento denominado **"Questão submetida a julgamento"** contém o texto a ser analisado para o enquadramento de um recurso

especial em um tema.

Há três aspectos essenciais a destacar sobre temas repetitivos. Primeiro, cada tema repetitivo corresponde a um pequeno texto (cerca de 36 palavras por tema), tipicamente consistindo de um único parágrafo sem uma estrutura rigidamente definida. Esse aspecto é relevante porque quando classificamos um apelo especial em um tema repetitivo, lidamos com textos de uma ordem de magnitude muito diferente (um recurso especial tem cerca de 3.980 palavras).

Em segundo lugar, a lista de temas repetitivos é aumentada periodicamente devido às mudanças na legislação e novas demandas sociais. Esse aspecto influenciou diretamente nossa adoção de uma abordagem de aprendizado não supervisionado. Pois a escassez de dados de treinamento afeta diretamente o desempenho de uma abordagem de aprendizado supervisionado.

Por fim, pode haver temas muito semelhantes, como mostrado nos exemplos a seguir, em que ambos os temas se referem à mesma lei, ao mesmo artigo e ao prazo de prescrição:

- **Tema 568:** "Discute-se a sistemática para a contagem da prescrição intercorrente (prescrição após a propositura da ação) prevista no art. 40 e parágrafos da Lei da Execução Fiscal (Lei n. 6.830/80): quais são os obstáculos ao curso do prazo prescricional da prescrição prevista no art. 40, da LEF."

- **Tema 569:** "Discute-se a sistemática para a contagem da prescrição intercorrente (prescrição após a propositura da ação) prevista no art. 40 e parágrafos da Lei da Execução Fiscal (Lei n. 6.830/80): se a ausência de intimação da Fazenda Pública quanto ao despacho que determina a suspensão da execução fiscal (art. 40, § 1º) ilide a decretação da prescrição intercorrente."

Esse último aspecto influenciou nossa escolha de tratar o problema como uma recuperação de caso legal. A recuperação de caso legal se refere ao processo de identificar e recuperar casos legais relevantes de um banco de dados com base em uma consulta ou critério específico. Essa tarefa envolve pesquisar em vastas coleções de documentos legais para encontrar casos que sejam pertinentes a uma questão, questão ou contexto legal específico. Devido aos detalhes sutis que podem diferenciar um tema do outro, entendemos que é conveniente fornecer uma pequena lista de temas como uma saída do nosso método para o especialista humano poder escolher o tema apropriado em vez de oferecer um único tema.

As características dos recursos especiais e temas repetitivos exigem que o analista jurídico humano tenha boas habilidades de síntese. Ele terá que lidar com documentos que excedem quatro mil palavras cada e confrontá-los com textos de cerca de quarenta palavras.

IV.3 Corpus e metadados associados

Estruturamos um corpus para o desenvolvimento deste trabalho. O conteúdo desse corpus são documentos correspondentes a recursos especiais extraídos de processos jurídicos não sigilosos, no âmbito da justiça federal do Brasil. Esses documentos foram coletados durante o segundo semestre de 2023. Este corpus reúne informações de 7.967 recursos especiais (conceito descrito na Seção I.1). A Tabela IV.1 apresenta um sumário estatístico desse corpus.

Descrição	**Valor**
Número de documentos	7.967
Média de palavras por documento	4.672,48
Mediana de palavras por documento	3.980
Mínimo de palavras por documento	92
Máximo de palavras por documento	67.944
Tamanho total (em Mb)	262

Tabela IV.1: Estatísticas do corpus de recursos especiais utilizado nesta pesquisa.

Um recurso especial ao ser submetido para apreciação no TRF2 é analisado por um especialista humano (i.e., um analista judiciário). Como auxílio na análise do recurso, o especialista tem à disposição um mecanismo de busca (Elasticsearch) que lhe apresenta sugestões de temas para classificação do recurso. Cabe ao especialista aceitar ou não a sugestão dada pelo Elasticsearch e rotular o recurso especial com o tema correto.

Os recursos contidos no corpus utilizado estão distribuídos não uniformemente em 190 tipos de temas. O histograma da Figura IV.2 apresenta essa distribuição. Dados estatísticos sobre os temas são apresentados na Tabela IV.2. Todos os recursos possuem uma classificação real (i.e., a atribuição de um único tema) dada por um especialista.

Para cada recurso especial do corpus, além da classificação dada por um especialista, constam as sugestões de temas dadas pelo Elasticsearch, a probabilidade da relevância de cada sugestão dada pelo Elasticsearch, o texto do tema atribuído ao recurso e em alguns casos o texto da tese jurídica firmada pelo STJ. Mais formalmente, cada entrada nesse corpus pode ser representada como uma tupla (r, t, τ, S, t^*). Os elementos desta tupla são descritos abaixo.

- r é o conteúdo textual do recurso especial.
- t é o conteúdo textual do tema associado, conforme disponibilizado na base de dados do STJ. Este é o tema no qual um analista judiciário classificou o recurso especial r.
- τ é conteúdo textual da tese jurídica firmada pelo STJ após julgamento de recursos especiais repetitivos. No caso de conteúdo nulo, significa que até o momento da criação deste corpus, não havia uma tese firmada para t.

- S é o conjunto de sugestões apresentadas pela máquina de busca para o recurso r. Cada sugestão é um par $(t_i, \Pr(t_i))$, $1 \leq i \leq 6$, no qual t_i é o identificador do i-ésimo tema sugerido, e $\Pr(t_i)$ é a probabilidade correspondente (também gerada pela máquina de busca).
- t^* é o identificador do tema selecionado pelo especialista. Esse tema não necessariamente deve estar contido em S. Se não estiver, significa que nenhuma das sugestões dadas pela máquina de busca foi a correta.

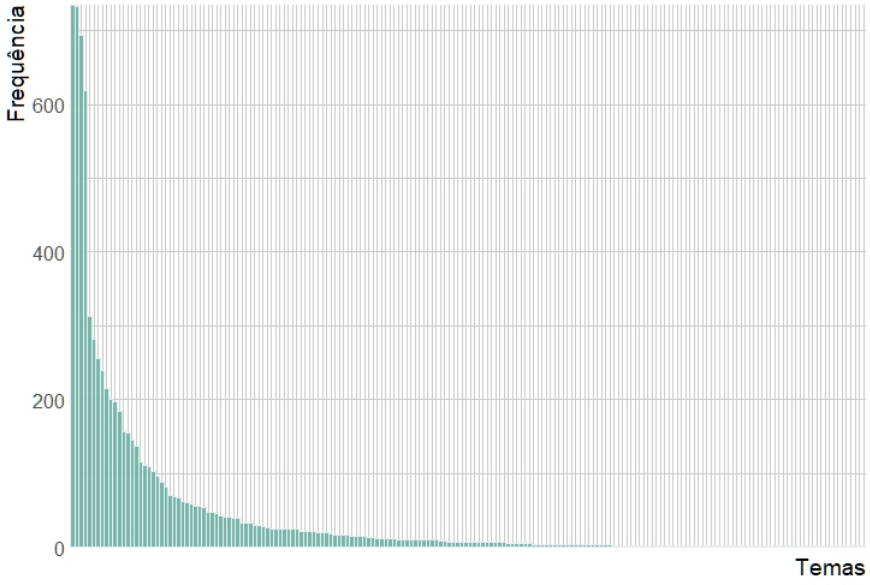


Figura IV.2: O histograma apresenta a distribuição dos temas no corpus analisado. Cada barra representa um tema específico, e sua altura indica a quantidade de documentos associados a ele. A figura evidencia a predominância de alguns temas, refletindo a repetitividade de determinados assuntos no conjunto de dados utilizado.

Descrição	**Valor**
Número de temas	190
Média de palavras por tema	43,3
Mediana de palavras por tema	36
Mínimo de palavras por tema	6
Máximo de palavras por tema	300
Tamanho total (em Kb)	56

Tabela IV.2: Dados estatísticos sobre os temas repetitivos considerados nesta pesquisa.

IV.4 Passos da metodologia

Sabe-se que num texto em linguagem natural, as palavras podem ser organizadas em sentenças, e o significado dessas sentenças depende do significado das palavras nelas contidas. Indo além, o

significado das sentenças também depende da ordem das palavras. A partir do estudo da linguagem, sabe-se que certas propriedades semânticas, por exemplo, a sinonímia, possibilitam a reescrita de um texto, modificando palavras, e até mesmo reduzindo o tamanho, sem contundo alterar-lhe o sentido [Cançado, 2005].

Diante do exposto, o desenvolvimento desta metodologia está baseado na hipótese de que dado um texto qualquer $\mathcal{T}$ composto de z palavras, é possível sintetizá-lo em um agrupamento $\mathcal{T}'$ de x palavras, onde $x < z$, de modo que $\mathcal{T}'$ preserve a informação essencial de $\mathcal{T}$. Uma questão a ser respondida, com a redução da dimensão do texto original $\mathcal{T}$, é se resulta ou não em ganho para a avaliação de similaridade com um texto alvo $\mathcal{T}_{\dashv}$. O que se espera, por hipótese, é que ao remover palavras irrelevantes para o entendimento da informação central do texto $\mathcal{T}$, gerando $\mathcal{T}'$, ocorra favorecimento na avaliação de similaridade com o texto alvo $\mathcal{T}_{\dashv}$ minimizando perdas semânticas.

O método que propomos está ancorado na sumarização textual por meio do algoritmo LexRank guiado em combinação com o algoritmo BM25 para computação de similaridade. Utilizamos o acrônimo GLARE (Guided LexRank for Advanced Retrieval in Legal Analysis) para denominar nosso método. Os passos do método estão nas seções seguintes. A Figura IV.3 apresenta um diagrama esquemático dos passos definidos. Nas próximas seções deste capítulo, descrevemos os detalhes de cada um desses passos.

1. Processar o conteúdo textual de cada recurso r e cada tema t, bem como gerar suas respectivas representações vetoriais.

2. Gerar representação textual que sintetize o conteúdo do recurso r e representação vetorial desse resumo. Da mesma forma gerar representação vetorial para o texto de cada tema $t \in \mathcal{T}$.

3. Efetuar comparação aos pares, entre representação textual do recurso e cada tema, produzindo o valor $\sigma_{r't}$ associado ao nível de similaridade entre os elementos comparados.

4. Produzir e ordenar o conjunto $\mathcal{T}' \subset \mathcal{T}$.

5. Avaliar os resultados.

IV.4.1 Pré-processamento dos recursos especiais e temas

Um recurso especial não é um tipo estruturado de texto definido por alguma norma. Porém, o seu conteúdo deve apresentar elementos mínimos exigidos pelo CPC. O trecho de um recurso especial onde se concentra a informação essencial do que está sendo pleiteado inicia na da Seção comumente intitulada **DO CABIMENTO DO PRESENTE RECURSO ESPECIAL**, conforme exemplo do Apêndice A.

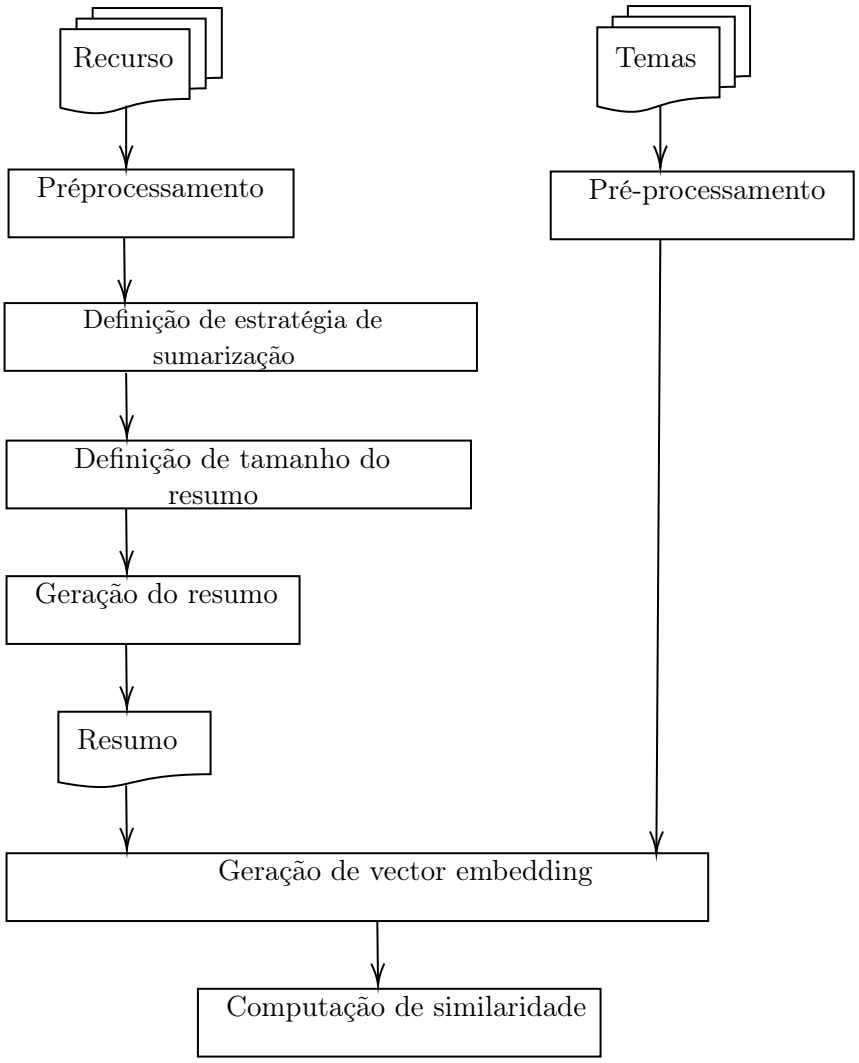


Figura IV.3: Esta figura descreve, de forma sequencial, os passos metodológicos seguidos na abordagem desenvolvida. As etapas incluem a seleção de documentos, pré-processamento, sumarização, vetorização e computação de similaridade, fornecendo uma visão integrada e estruturada do fluxo aplicado.

Uma vez identificado o núcleo do documento, nesta etapa do método, selecionamos os parágrafos com as informações essenciais do documento. Em seguida, aplicamos um tratamento de remoção para eliminar palavras, sinais de pontuação e padrões numéricos irrelevantes para a tarefa em questão, que geram "ruído" na identificação de informações relevantes. Em geral, artigos, preposições e conjunções. Devido à natureza dos textos contidos no corpus, removemos endereços e números de documentos que nomeiam pessoas físicas e jurídicas.

IV.4.2 Sumarização por meio do LexRank Guiado

Adotamos uma abordagem de sumarização extrativa para evitar possíveis alucinações. Dessa forma, ao trabalhar com termos e estruturas que estão, de fato, no texto original, buscamos preservar sua semântica e evitar perdas na próxima etapa do método.

Conforme descrito na Seção II.3.2, o LexRank representa o texto como um grafo ponderado,

onde os nós representam sentenças e as arestas representam a similaridade entre as sentenças. Os pesos associados aos nós estão relacionados ao grau de centralidade de cada sentença(Figura IV.4). Considerando o contexto deste trabalho, quando enviamos o texto de um apelo especial para o LexRank, o algoritmo gera um resumo, que é formado pelo conjunto de sentenças que melhor representam o apelo. Devemos considerar que uma sentença não incluída no resumo ainda pode ser importante para classificar o apelo especial em um tema. O LexRank pode ignorar uma frase porque calcula a centralidade de cada frase sem levar em conta nenhum fator externo além do próprio texto.

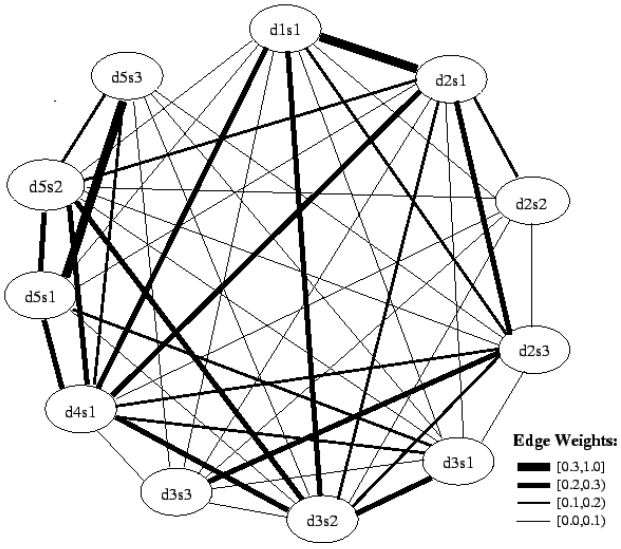


Figura IV.4: O grafo ponderado representa a sumarização de um texto com a aplicação do algoritmo LexRank. Os nós representam as sentenças do texto. As espessuras das arestas refletem o grau de centralidade das sentenças. Fonte: Erkan and Radev [2004].

Otterbacher et al. [2009] apresenta uma versão do algoritmo LexRank que considera o viés com base em um tópico de interesse. É proposta uma abordagem para introduzir o viés que parte da equação geral do LexRank e adiciona um termo de prioridade com base em uma função de peso. De acordo com a similaridade que uma dada frase no texto tem com um tópico previamente definido, a frase receberá um peso que influenciará na composição do resumo final do texto.

Propomos uma mudança no algoritmo LexRank para orientar a tarefa de sumarização, chamamos esse novo algoritmo de LexRank guiado. Com a mudança proposta, a decisão de uma frase compor o resumo do texto é resultado de uma combinação de dois fatores. Um dos fatores é interno ao texto, ou seja, o grau de centralidade da frase calculado no algoritmo LexRank original. O outro fator é externo ao texto, calculado de acordo com a similaridade entre cada frase a ser resumida e os textos dos temas deste trabalho. Conforme detalhado na Seção II.1.2, o BM25 é um modelo probabilístico que permite a computação de similaridade de documentos com base na frequência dos termos, ajustado pelo comprimento do texto. No LexRank guiado, o valor de similaridade do valor externo é computado por meio do BM25.

Dessa forma, no LexRank guiado, para cada sentença s, os seguintes passos são realizados:

1. Calcular a centralidade de grau γ_s, no documento D, da mesma forma que é calculada no LexRank.

2. Enviar cada sentença s como entrada para o algoritmo BM25 visando cada tema $t \in T$. Armazenar a maior pontuação encontrada ao executar o BM25, σ_{st}.

3. Calcular a pontuação combinada usando dois fatores de ponderação (ver Equação IV.1).

Propomos uma abordagem combinada para calcular a pontuação, usando a Equação IV.1. As sentenças selecionadas para compor o resumo do texto serão aquelas com os maiores valores de pontuação combinada. Nessa equação, α e β são parâmetros arbitrários. Os pesos α e β foram ajustados empiricamente em nossos experimentos preliminares para maximizar o desempenho nos dados de validação. Após essa etapa preliminar, os valores $\alpha = 1$ e $\beta = 3$ foram fixados e aplicados consistentemente em todos os experimentos subsequentes. O algoritmo do LexRank guiado está esquematizado na Figura IV.5.

$$\text{Pontuação combinada} = \alpha \times \gamma_s + \beta \times \sigma_{st} \tag{IV.1}$$

Ao combinar fatores internos e externos para a seleção de sentenças, espera-se que o LexRank guiado apresente resultados superiores, na atribuição temática, em relação ao LexRank clássico. Essa hipótese é corroborada pelos experimentos apresentados no Capítulo V.

Nossa abordagem de sumarização proposta, LexRank guiado, difere do Biased Lexrank Otterbacher et al. [2009] em dois aspectos. Primeiro, o LexRank guiado usa um algoritmo diferente para combinar fatores para pesar a relevância de uma frase. A segunda diferença é que nosso método usa o algoritmo BM25 para avaliar a importância das frases no texto e frases externas ao texto. A Figura IV.6 ilustra a aplicação do LexRank guiado na sumarização de um texto.

Ao término desta etapa de **sumarização extrativa**, temos disponível um resumo como representação para cada recurso. Como exemplificado na Figura IV.7, extraímos as sentenças mais significativas do recurso r e compomos um resumo r' a partir dessas. Na etapa seguinte, efetuaremos a avaliação de similaridade entre pares r' e cada $t \in \mathcal{T}$, a fim de gerarmos os valores σ_{rt}. Desta forma, σ_{rt} representará a similaridade entre um recurso r e um tema $t \in \mathcal{T}$.

É importante ratificar que estamos em um contexto de recuperação da informação, ou seja, para uma dada consulta(recurso) feita por um usuário deve ser retornada uma lista de conteúdos(temas) ordenada por ordem de relevância. Deste modo, temos para cada recurso r uma lista $\mathcal{L}$ de tuplas (r_i, t_i, σ_{rt}), onde:

- r_i - é o identificador do recurso

- t - é o identificador do tema

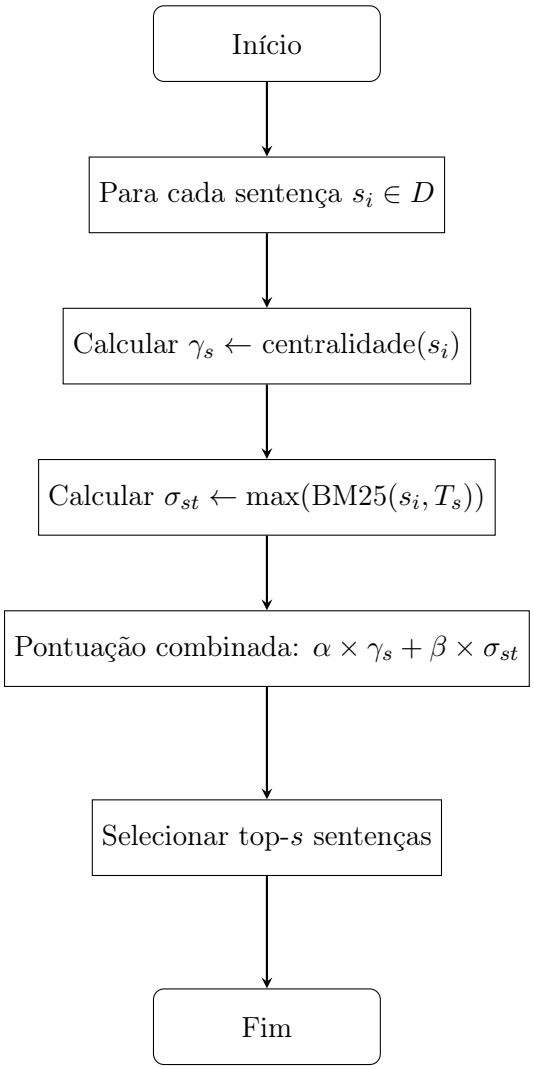


Figura IV.5: O algoritmo começa com a análise de cada sentença s_i pertencente ao documento D. Em seguida, calcula-se a centralidade γ_s da sentença, conforme o LexRank clássico. Após isso, é calculada a similaridade σ_{st} de cada sentença com os temas utilizando o BM25. A pontuação combinada de cada sentença é calculada com base na centralidade e na similaridade com os temas, ponderada pelos parâmetros α e β. Por fim, as top-s sentenças com maior pontuação combinada são selecionadas para formar o resumo final.

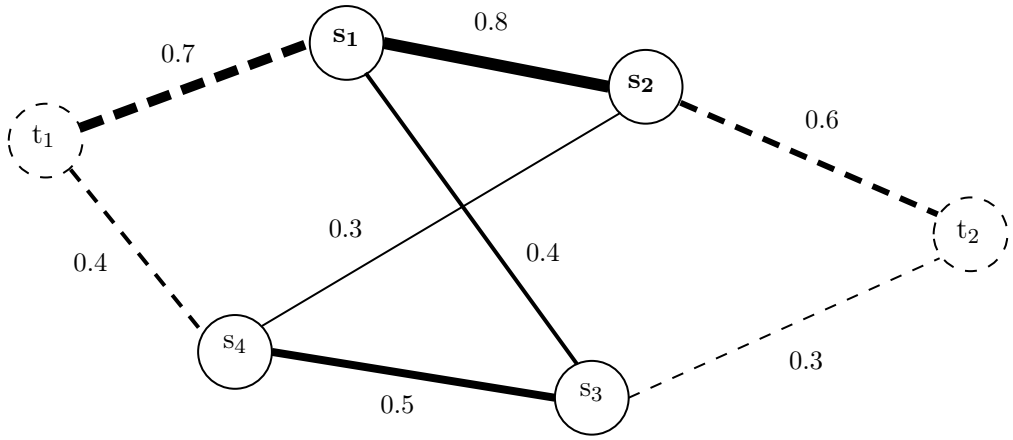


Figura IV.6: O grafo mostra as sentenças do texto original (s_1, s_2, s_3, s_4) como nós, conectados por arestas ponderadas. A espessura das arestas reflete a pontuação combinada, derivada da centralidade interna γ_s e da similaridade externa σ_{st} com os temas(t_1, t_2)

.

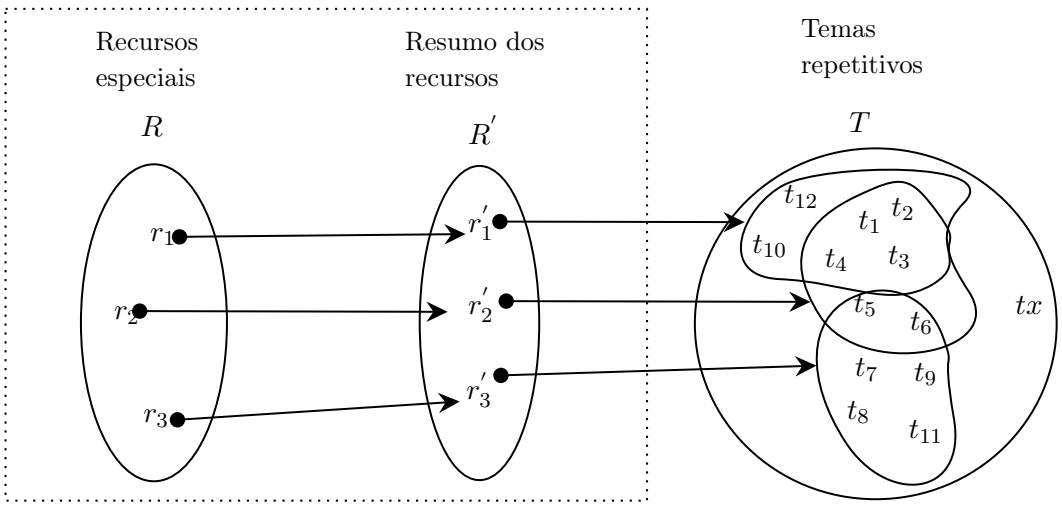


Figura IV.7: A figura apresenta a relação entre os recursos especiais, os resumos gerados e os temas associados. O diagrama demonstra como os recursos são reduzidos a resumos por meio da aplicação do algoritmo LexRank guiado. Esses resumos, em seguida, são comparados aos textos representativos dos temas para determinar a classificação final.

- σ_{rt} - é o grau de similaridade entre o recurso e o tema

A lista $\mathcal{L}$ é ordenada pelo elemento σ_{rt}. Conforme parâmetro n definido pelo usuário, $\mathcal{L}$ é limitada ao tamanho n, contendo os n temas que melhor estão relacionados ao recurso avaliado.

IV.4.3 Cálculo de similaridades

Nesta etapa temos por objetivo obter um valor numérico que indique o grau de similaridade entre um **recurso** e um **tema**. É prática comum quantificar a similaridade semântica entre textos aplicando similaridade de cosseno. No entanto, Steck et al. [2024] alerta contra o uso cego desta técnica. Usando observações empíricas, Steck et al. [2024] demonstra que a similaridade de cosseno de incorporações aprendidas pode produzir resultados arbitrários. Resultados de similaridade dife-

rentes podem ser obtidos no mesmo conjunto de dados. Esses resultados discrepantes são devidos às várias formas de treinamento que o modelo de linguagem grande (LLM) usado para gerar a incorporação vetorial do texto pode sofrer.

Tendo em vista os resultados apresentados por Steck et al. [2024], consideramos o uso do BM25 como alternativa ao uso da similaridade por cosseno. Desta foram, usamos o resumo gerado na etapa anterior como parâmetro de entrada para o algoritmo BM25(II.1.2, visando o texto do tema como alvo.

Seja r' a representação do recurso r obtida como saída da etapa de **sumarização**, computamos a similaridade entre pares r' e cada $t \in \mathcal{T}$ e geramos os valores σ_{rt}. Desta forma, σ_{rt} representa a similaridade entre um recurso r e um tema $t \in \mathcal{T}$.

IV.4.4 Avaliação do resultado

Conforme descrito na Seção IV.3, todos os recursos que fazem parte do conjunto de dados possuem uma classificação dada por um analista humano. De posse da lista de temas relevantes e da classificação do recurso dada por um analista humano. Ao longo das iterações sobre os passos, computamos as métricas para avaliar o desempenho obtido. Fizemos uso de algumas das métricas mas utilizadas no contexto de recuperação da informação: Recall, Mean Average Precision (MAP), F1-score e Normalized Discounted Cumulative Gain (NDCG).

Recall

$$\text{Recall} = \frac{\text{Itens relevantes no ranking de k posições}}{\text{Total de itens relevantes}} \tag{IV.2}$$

Considerando que um recurso só é classificado por um analista humano em um único tema, o Recall torna-se a métrica mais relevante a ser observada neste contexto, pois, por definição, avalia a proporção com que um tema relevante para o usuário é localizado no ranking de k posições.

MAP

Considerando uma lista $\mathcal{L}$ contendo k elementos recomendados em uma consulta, a precisão média é calculada considerando a posição em que um item realmente relevante encontra-se na lista. O MAP é a média dessas precisões médias para todos os itens.

$$\text{Precision at k} = \frac{\text{Quantidade de itens relevantes até a posição k}}{k} \tag{IV.3}$$

$$\text{Average Precision} = \frac{\sum_{k=1}^{k}(\text{Precision at k} \times \text{Relevância do item em k})}{\text{Número total de itens relevantes}} \tag{IV.4}$$

$$MAP = \frac{\sum_{q=1}^{Q} \text{Average Precision para a consulta q}}{Q} \tag{IV.5}$$

F1-score

Calculada pela média harmônica entre precision e recall.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{IV.6}$$

NDCG

Semelhante ao Precision at K, porém leva em consideração a relevância dos itens e a posição em que estão no ranking. O ganho acumulado é normalizado para lidar com diferentes escalas de relevância.

$$\text{DCG at k} = \sum_{i=1}^{k} \frac{2^{\text{rel}_i} - 1}{\log_2(i+1)} \text{onde, rel}_i \text{ é o grau de relevância do elemento na posição } i \tag{IV.7}$$

$$\text{IDCG at k é o DCG ideal, onde os itens estão ordenados por relevância.} \tag{IV.8}$$

$$\text{NDCG at k} = \frac{\text{DCG at k}}{\text{IDCG at k}} \tag{IV.9}$$

$$NDCG = \frac{\sum_{q=1}^{Q} \text{NDCG para a consulta q}}{Q} \tag{IV.10}$$

Capítulo V Experimentos

Neste capítulo apresentamos uma análise experimental detalhada do desempenho e da eficácia do método proposto. Cada seção deste capítulo aborda aspectos específicos da avaliação experimental, desde a configuração do ambiente computacional até comparações detalhadas com métodos existentes. A Seção V.1 detalha a configuração experimental, incluindo o ambiente computacional e as bibliotecas Python usadas para garantir a reprodutibilidade. A Seção V.2 avalia o modelo proposto variando técnicas de pré-processamento, métodos de sumarização, comprimentos de resumo e abordagens de cálculo de similaridade. A Seção V.3 explora o impacto do processo de sumarização de texto, confirmando que, mesmo com a redução da dimensionalidade, a sumarização melhora a precisão em comparação ao uso do texto completo. A Seção V.4 compara nosso método com a linha de base do Elasticsearch. A Seção V.5 compara nosso método com modelos supervisionados, mostrando que o GLARE é competitivo.

V.1 Configurações experimentais

Os experimentos foram executados em um servidor AMD EPYC 7452 32-Core 2.35GHz 52GB RAM. Os experimentos foram conduzidos usando um conjunto de bibliotecas Python com versões específicas. As bibliotecas utilizadas incluem **argparse** versão 1.1, **BERTopic** 0.16, **csv** 1.0, **LexRank** 0.1.0, **nltk** 3.6.7, **numpy** 1.22.4, **pandas** 2.1.4, **rank_bm25** 0.2.2, **rank_eval** 0.1.3, **re** 2.2.1, **sentence_transformers** 2.2.2 e **torch** 1.12.1. Essas versões foram selecionadas para se alinharem à metodologia e garantir que os resultados experimentais fossem confiáveis e reproduzíveis.

V.2 Comparando alternativas de pipeline

Realizamos uma série de experimentos comparativos para avaliar o desempenho da nossa abordagem de recuperação de casos legais. Esses experimentos visavam avaliar o impacto de vários parâmetros e estratégias na eficácia do método. Ao variar sistematicamente técnicas de pré-processamento, tipos de representação de texto, tamanhos de resumo e métodos de cálculo de similaridade, buscamos identificar a configuração ideal para melhorar a precisão e a relevância. Esta seção detalha a configuração experimental e os parâmetros específicos testados para entender sua influência no desempenho do modelo. Em resumo, os parâmetros que variamos em nossos experimentos compa-

rativos consistem no seguinte:

- Pré-processamento — Adotamos a remoção ou retenção de termos do recurso especial como um parâmetro de decisão.
- Tipo de representação — Na etapa seguinte ao pré-processamento, escolhemos o tipo de representação que será adotado posteriormente no fluxo. Poderíamos manter o texto ou gerar um resumo extrativo usando as estratégias BERTopic, BERTopic guiado, LexRank ou LexRank guiado.
- Tamanho do resumo — Fazemos variações nos tamanhos dos resumos gerados usando um parâmetro passado para o algoritmo.
- Cálculo de similaridade — Nesta etapa, possuímos o texto/resumo e seu embedding vetorial. Dessa forma, podemos avaliar a similaridade do apelo com o tema de duas maneiras. Uma maneira é usar o texto/resumo como uma consulta de entrada para o algoritmo BM25 visando o texto do tema. Outra maneira é calcular a similaridade do cosseno com o embedding vetorial do apelo e dos temas.

Para cada texto de recurso especial enviado como entrada em nosso método, a saída foi uma lista de 6 temas sugeridos considerados mais apropriados para o texto. Os melhores resultados são condensados nas Tabelas V.1, V.2, V.3 e V.4 de acordo com os parâmetros de variação.

Tabela V.1: Tratamento: Com a remoção de termos. Similaridade: BM25

Técnica de sumarização	**recall@6**	**F1-score**	**map@6**	**ndcg@6**
GLARE (LexRank guiado—15 frases)	**0.7575**	**0.6268**	**0.5345**	**0.5902**
LexRank—60 frases	0.7215	0.6014	0.5157	0.5672
BERTopic—55 palavras	0.5696	0.4552	0.3791	0.4266
BERTopic guiado—60 palavras	0.5551	0.4525	0.3819	0.4253

Tabela V.2: Tratamento: Com remoção de termos. Similaridade: Cosseno

Técnica de sumarização	**recall@6**	**F1-score**	**map@6**	**ndcg@6**
LexRank guiado—10 frases	0.4704	**0.3858**	0.3269	**0.3626**
LexRank—20 frases	0.3543	0.2700	0.2181	0.2519
BERTopic—45 palavras	**0.4792**	0.3700	**0.5157**	0.3454
BERTopic guiado—45 palavras	0.4612	0.3552	0.2887	0.3322

O melhor resultado obtido entre todos os experimentos foi um recall@6 igual a 0,75750. Em resumo, aplicando o método que propomos, GLARE, que integra o LexRank guiado para sumarização e o BM25 para avaliação de similaridade, nosso método foi capaz de sugerir o tema correto para o apelo para aproximadamente 76% dos registros em nosso conjunto de dados.

Tabela V.3: Tratamento: Sem remoção de termos. Similaridade: BM25

Técnica de sumarização	**recall@6**	**F1-score**	**map@6**	**ndcg@6**
LexRank—50 frases	0.7302	0.5852	0.4883	0.5484
GLARE (LexRank guiado—30 frases)	**0.7545**	**0.6361**	**0.5498**	**0.6009**
BERTopic—25 palavras	0.5506	0.4452	0.3736	0.4180
BERTopic guiado—60 palavras	0.5526	0.4500	0.3796	0.4228

Tabela V.4: Tratamento: Sem remoção de termos. Similaridade: Cosseno

Técnica de sumarização	**recall@6**	**F1-score**	**map@6**	**ndcg@6**
LexRank—30 frases	0.4176	0.3036	0.2385	0.2827
LexRank guiado—10 frases	0.3798	0.2804	0.2223	0.2610
BERTopic—45 palavras	**0.4638**	**0.3605**	**0.2949**	**0.3373**
BERTopic guiado—45 palavras	0.4616	0.3566	0.2905	0.3333

A diferença de desempenho entre o método LexRank guiado e o BERTopic ao usar similaridade de cosseno pode ser atribuída à natureza dos resumos gerados e à maneira como a similaridade é calculada. Enquanto o Guide LexRank, que é baseado em gráficos, prioriza a estrutura e a importância das frases, ele não captura similaridades semânticas contextuais de forma tão eficaz, resultando em desempenho inferior em comparação ao BERTopic quando avaliado com similaridade de cosseno. No entanto, quando o BM25 é usado para avaliação, que enfatiza a correspondência exata de termos, o Guide LexRank supera significativamente o BERTopic, destacando sua eficácia na preservação de termos-chave e estrutura textual relevante.

Para entender melhor os resultados, conduzimos mais experimentos para avaliar o comportamento de cada estratégia de geração de resumo variando o tamanho do resumo gerado. Também avaliamos o comportamento de cada estratégia de geração de resumo removendo termos considerados irrelevantes no texto. O objetivo era encontrar pontos de saturação no desempenho das estratégias de sumarização e a relação com os parâmetros de variação dos experimentos.

Os melhores resultados estão no quadrante superior direito no gráfico de dispersão mostrado na Figura V.1. Esta região representa os resultados com o maior valor de recall associado a uma classificação de sugestões de melhor qualidade, medido pelas métricas MAP e NDCG. Pelo que é mostrado na Figura V.1, os melhores resultados foram alcançados com o algoritmo que propusemos, LexRank guiado.

Com base nos resultados mostrados na Figura V.2, o desempenho do método depende da combinação adequada da estratégia de sumarização e do método de avaliação de similaridade. O desempenho superior ficou claro ao usar o algoritmo BM25 para avaliar a similaridade entre o resumo do recurso especial e os temas repetitivos em vez de usar a medição de similaridade do cosseno.

Com base nos resultados apresentados na Figura V.3, não podemos relatar diretamente a remoção de termos do texto com desempenho aprimorado do método. No entanto, o desempenho

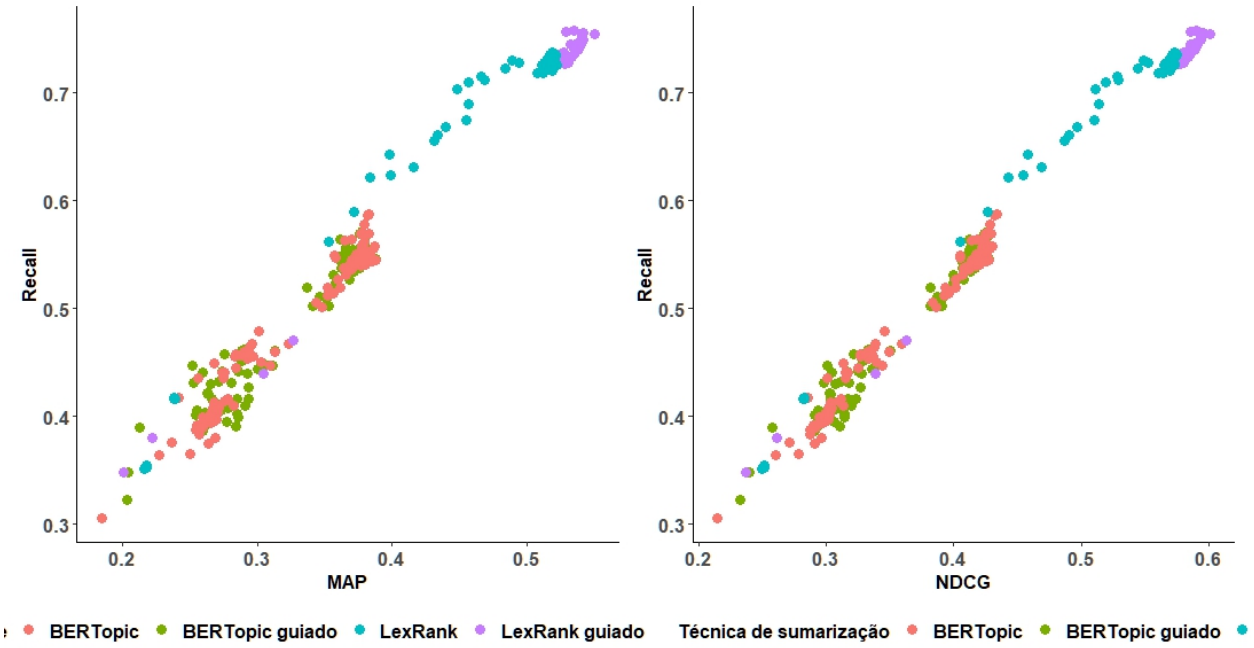


Figura V.1: A qualidade da classificação das sugestões, dada por cada técnica de sumarização, pode ser medida pelas métricas MAP e NDCG.

superior ao combinar a avaliação de similaridade usando o algoritmo BM25 com técnicas de sumarização baseadas em gráficos é notável.

Os resultados apresentados na Figura V.4 destacam o efeito da remoção de termos, especialmente em técnicas de sumarização baseadas em gráficos. A técnica guiada LexRank se beneficia da remoção de termos irrelevantes pela aplicação de similaridade de cosseno. O efeito oposto ocorre quando se realiza a mesma remoção combinada com a técnica tradicional LexRank. Também é digno de nota o desempenho constante em técnicas baseadas em gráficos, mesmo com variação substancial no número de frases no resumo do texto.

Ao usar o algoritmo BM25 para gerar uma medida de similaridade, buscamos uma alternativa à medida tradicional de similaridade de cosseno. Conforme descrito na Seção IV.4.3, os parâmetros e dados de treinamento usados por um modelo que gera a representação vetorial de um texto afetam diretamente o cálculo de similaridade de cosseno. Com base nos resultados obtidos nos experimentos, confirmamos a viabilidade e superioridade do BM25, sobre cosseno, na tarefa realizada. Verificamos a eficácia do método proposto para encontrar o tema de um apelo especial, fornecendo o texto do apelo e uma lista de temas repetitivos. A combinação do algoritmo baseado em gráfico que propomos, o LexRank guiado, com o algoritmo BM25 produziu ganhos efetivos sobre outras abordagens que analisamos.

V.3 Impacto da sumarização de texto no desempenho

Aggarwal [2022b] afirma que algum processo de redução de dimensionalidade pode alterar a

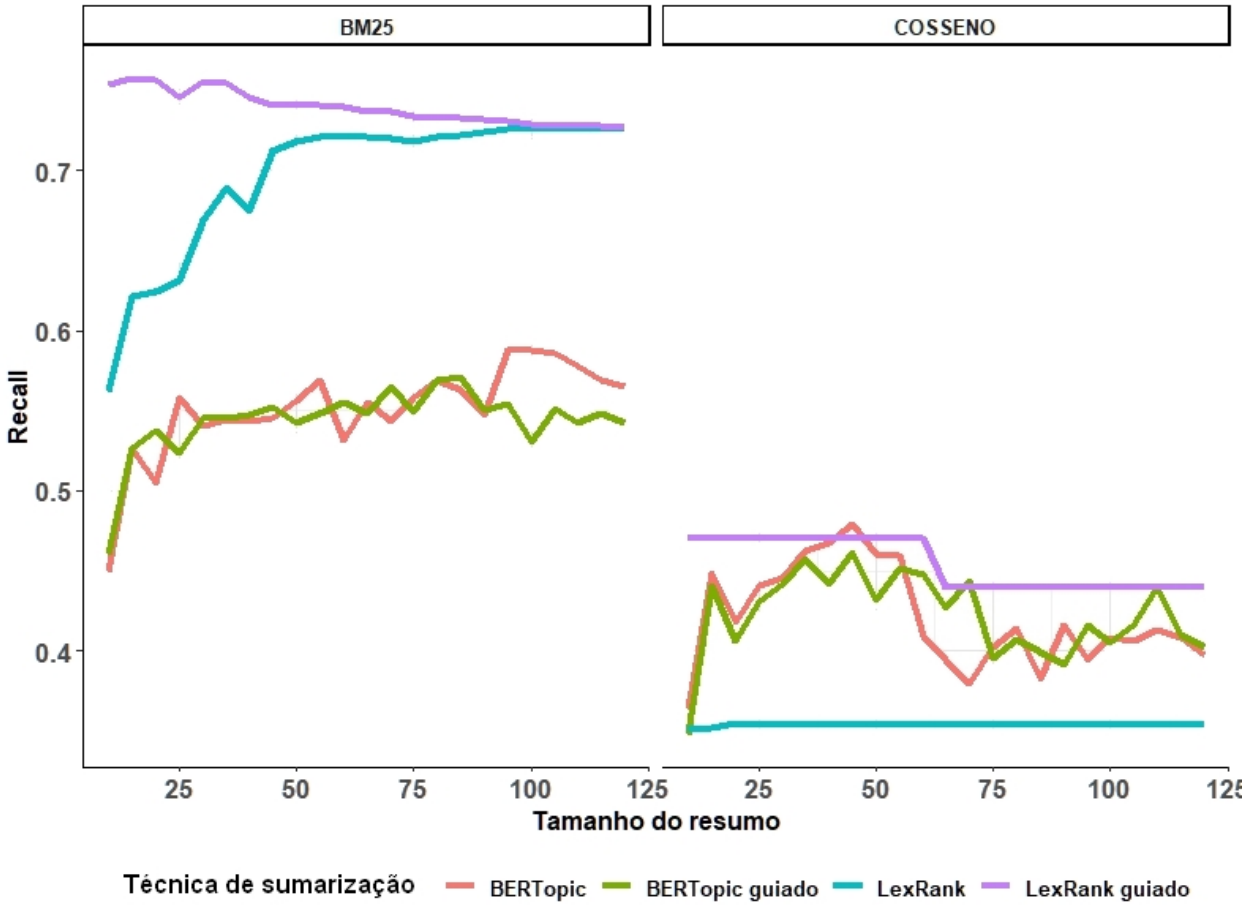


Figura V.2: O desempenho do modelo foi avaliado variando-se a quantidade de frases na composição do resumo. Outro aspecto considerado foi a métrica utilizada para avaliar a similaridade entre o resumo do recurso especial e o tema, seja pelo BM25 ou pelo cosseno.

representação de um documento. Embora possa parecer que a redução da dimensionalidade perde informação, muitas vezes é possível escolher uma dimensionalidade de representação na qual o conhecimento semântico no corpus é mantido e a maior parte do ruído é reduzida.

O objetivo da etapa de sumarização proposta em nosso método (Seção IV.4.2) é reduzir o ruído para melhorar a qualidade da representação e, consequentemente, facilitar o processo de classificação do apelo especial em um tema. Em todos os experimentos, obtivemos um recall maior que zero, mostrando que usar o resumo do texto efetivamente para classificar um apelo especial em um tema é eficaz. No entanto, também queríamos avaliar se havia um ganho ao usar o resumo em vez do texto completo. Para avaliar esse aspecto, incluímos dois casos em nossos experimentos nos quais usamos o texto completo, conforme descrito abaixo.

- Avaliação da similaridade entre o texto completo do apelo especial e o texto de cada tema usando o BM25.
- Avaliação de similaridade entre a representação vetorial do texto completo do recurso especial e a representação vetorial de cada tema usando cosseno.

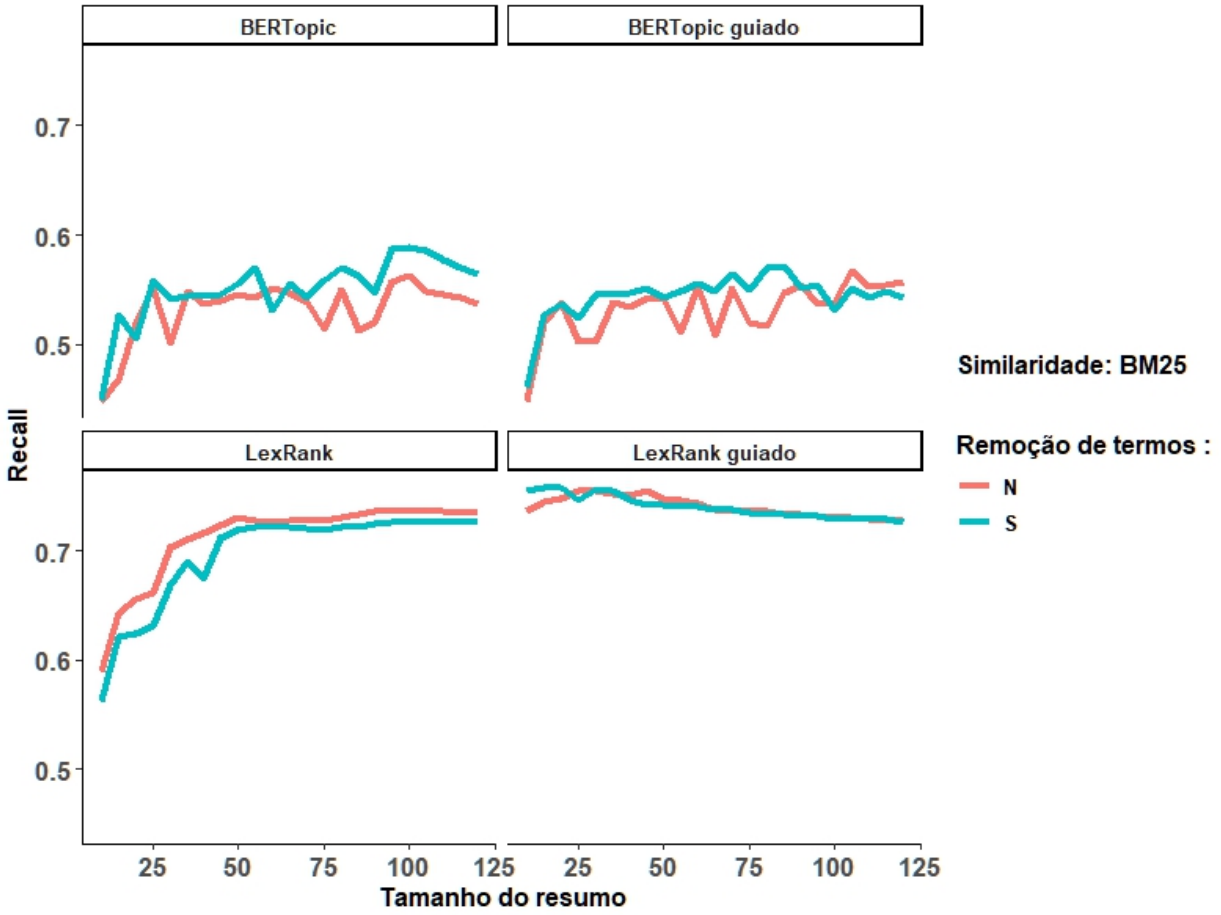


Figura V.3: Para cada técnica de sumarização aplicada, o desempenho foi avaliado sob o aspecto da remoção de termos considerados irrelevantes no texto original do recurso especial. Outros aspectos considerados foram a similaridade medida pelo BM25 e o número de frases contidas no resumo após a aplicação das técnicas de sumarização.

A tabela V.5 apresenta a análise comparativa do texto completo versus texto resumido. Nesta tabela, usamos os seguintes mnemônicos para representar cada alternativa: FT (ou seja, texto completo é usado), GLARE15 (ou seja, LexRank guiado- resumo com 15 frases), FTE (ou seja, representação vetorial de texto completo) e GLR10 (ou seja, LexRank guiado- resumo com 10 frases-representação vetorial). De acordo com esses resultados, concluímos que é possível suprimir termos de um texto e ainda obter um resultado superior ao resultado obtido usando o texto completo.

Tabela V.5: Análise comparativa de texto completo vs. texto resumido

Representação	**similaridade**	**recall@6**	**F1-score**	**map@6**	**ndcg@6**
FT	BM25	0.7146	0.6041	0.5233	0.5714
GLARE15	BM25	**0.7575**	**0.6268**	**0.5345**	**0.5902**
FTE	Cosseno	0.1363	0.1073	0.0884	0.0999
GLR10	Cosseno	0.4704	0.3858	0.3269	0.3626

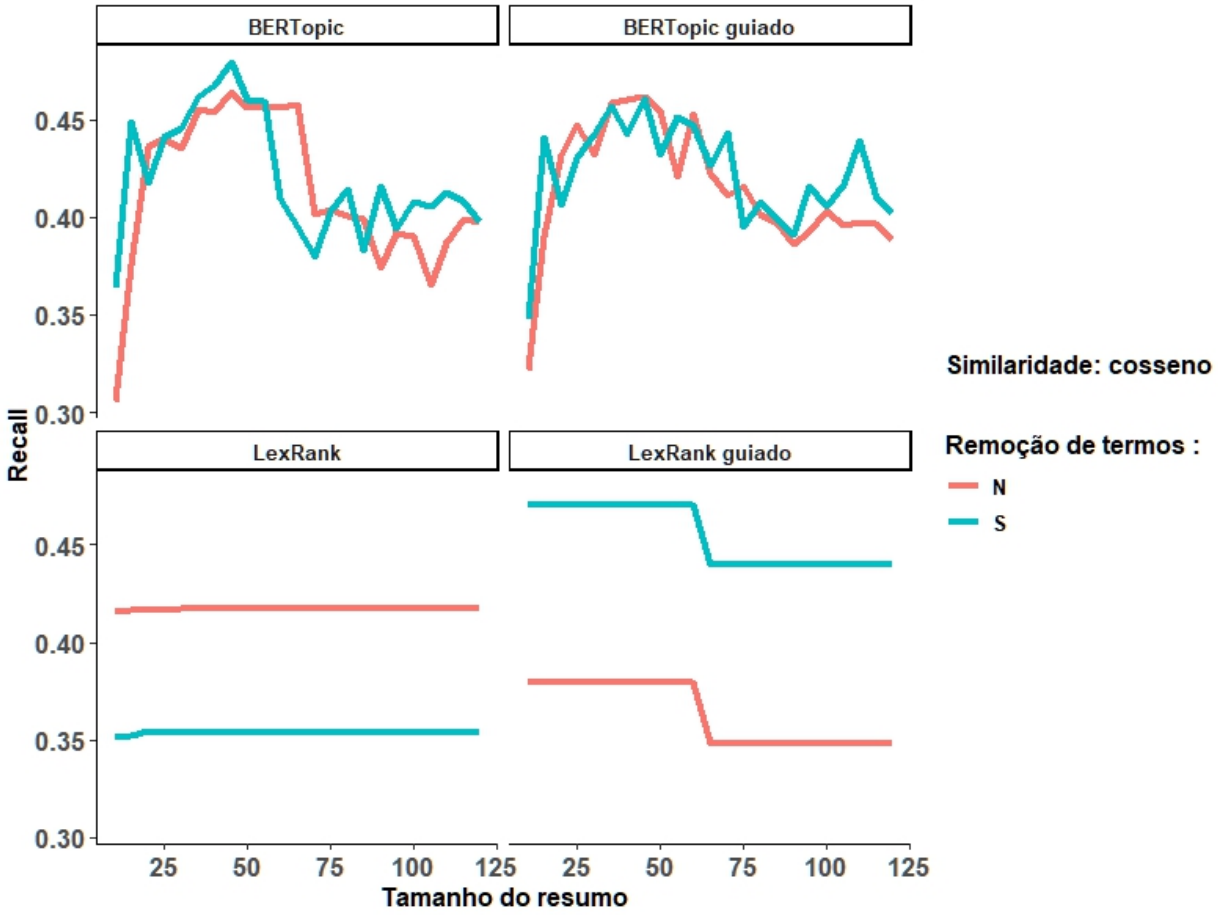


Figura V.4: Para cada técnica de sumarização aplicada, o desempenho foi avaliado sob o aspecto da remoção de termos considerados irrelevantes no texto original do recurso especial. Outros aspectos considerados foram a similaridade medida pelo cosseno e o número de frases contidas no resumo após a aplicação das técnicas de sumarização.

V.4 Comparação com a linha de base do Elasticsearch

Conforme mencionado anteriormente na Seção IV.3, como auxílio no tratamento de recursos especiais, o TRF2 adotou uma solução baseada em inteligência artificial para sugestão de tema a um recurso especial. A solução é o mecanismo de busca distribuído, de código aberto, denominado Elasticsearch. Trata-se de uma solução integrada entre o sistema processual eletrônico - Eproc [epr], utilizado no TRF2.

Em uma visão macro, a solução adotada pelo TRF2 apresenta os seguintes passos ao sugerir temas para um recurso novo:

1. Um novo recurso é submetido ao sistema.

2. O mecanismo de busca compara o novo recurso com outros que já existem na base de dados histórica, os quais já foram previamente classificados por um analista humano.

3. O mecanismo de busca gera uma lista ordenada de recursos históricos com base na semelhança

que possuem com o recurso novo.

4. O sistema sugere ao analista humano um conjunto de temas repetitivos, em conformidade com os temas dos recursos contidos na lista gerada anteriormente.

De posse da lista de temas gerada pelo sistema, o analista humano atribui um tema ao recurso especial. Cabe ressaltar que a lista é uma **sugestão**, a decisão sobre o tema cabe ao analista o qual pode inclusive discordar de todas as sugestões da lista. Um esquema da solução adotada é apresentado na Figura V.5.

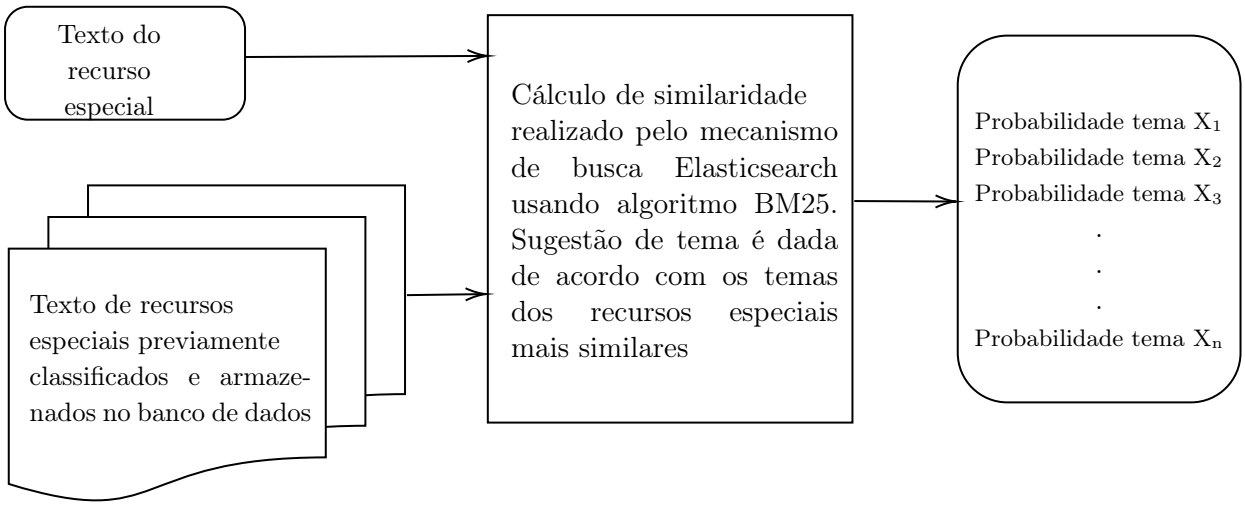


Figura V.5: O esquema ilustra a solução adotada pelo TRF2 para atribuir temas aos recursos especiais. Dado um recurso especial de entrada, é computada a similaridade com recursos existentes na base de dados que foram previamente classificados em um tema. De acordo com os valores de similaridade obtidos, os temas mais adequados são sugeridos para o novo recurso especial.

Um aspecto relevante nesta solução é a dependência de recursos especiais semelhantes na base de dados do TRF2. De fato, caso seja submetido um novo recurso especial que poderia ser perfeitamente enquadrado em um tema já divulgado na base de dados de temas repetitivos do STJ, mas que não haja um recurso anterior semelhante na base de dados do TRF2, a solução atual, a priori, é incapaz de dar uma sugestão correta. Esse empecilho pode ser superado com uma solução baseada numa arquitetura na qual fosse prevista a comparação direta entre o conteúdo do recurso especial e a base de temas repetitivos.

A solução do TRF2, utilizando o Elasticsearch, foi identificada como um dos modelos de linha de base para comparação neste estudo. O desempenho do método proposto, GLARE, foi avaliado em relação a esta linha de base para avaliar sua eficácia na classificação de recursos especiais. Os resultados, detalhados na Tabela V.6, destacam as melhorias significativas alcançadas com o GLARE. Nosso método demonstrou desempenho superior em todas as métricas avaliadas. Esses resultados ressaltam a eficácia da abordagem proposta na identificação de temas relevantes para recursos especiais, superando o desempenho da linha de base do Elasticsearch.

Tabela V.6: Comparação de desempenho entre o Elasticsearch de base e o GLARE.

Representação	recall@6	F1-score	map@6	ndcg@6
GLARE	**0.7575**	**0.6268**	**0.5345**	**0.5902**
Baseline-Elasticsearch	0.3491	0.3084	0.3275	0.3182

V.5 Comparação com linhas de base de aprendizagem supervisionada

Estruturamos o problema de atribuir uma lista classificada de temas a um determinado recurso como um problema de classificação supervisionada, tratando os temas como rótulos. Cada recurso no corpus é associado a um tema específico, e esses temas servem como rótulos de destino para os modelos de classificação. Ao enquadrar o problema dessa forma, podemos treinar modelos supervisionados para aprender a relação entre o conteúdo dos documentos e seus temas correspondentes.

Para obter uma classificação dos temas, semelhante à abordagem não supervisionada, exploramos as probabilidades geradas pelos modelos supervisionados. Usamos os modelos XGBoost [Chen and Guestrin, 2016] e Logistic Regression [Pedregosa et al., 2011], ambos configurados em um esquema one-vs-rest (OvR), onde cada modelo aprende independentemente a distinguir um tema de todos os outros. Após o treinamento, os modelos retornam uma probabilidade associada a cada tema para cada documento. Essas probabilidades são então classificadas em ordem decrescente para produzir uma classificação, fornecendo uma saída consistente com a proposta do nosso método não supervisionado.

Validação Cruzada

A técnica de validação cruzada divide os dados rotulados em k segmentos iguais. Um dos k segmentos é usado para teste, e os segmentos restantes $(k-1)$ são usados para treinamento. Este processo é repetido k vezes usando cada um dos k segmentos como o conjunto de teste [Aggarwal, 2022b]. Em ambos os modelos supervisionados, usamos uma técnica de validação cruzada estratificada de dez vezes, garantindo que cada dobra tivesse a mesma proporção de classes que o conjunto de dados original.

Reamostragem de Classes Minoritárias

Conforme mostrado na Seção IV.3, o corpus usado em nosso projeto apresenta desequilíbrio entre as classes. Para abordar o desequilíbrio de classes, conduzimos dois experimentos com os modelos supervisionados: um sem aplicar nenhuma técnica de balanceamento e o outro usando um método de reamostragem disponível na biblioteca scikit-learn. A técnica de reamostragem aumenta artificialmente o número de exemplos de classes minoritárias no conjunto de treinamento, equalizando o número de exemplos entre as classes. A aplicação dessa técnica nos ajudou a avaliar

o impacto do balanceamento no desempenho do modelo.

Treinamento e calibração do modelo

Usamos os classificadores XGBoost e Logistic Regression em um esquema one-vs-rest (OvR), permitindo a classificação multiclasse. As saídas de probabilidade dos modelos foram calibradas usando um recurso disponível no Sci-Kit Learn com o método sigmoide, com base em Zadrozny and Elkan [2002].

Com as técnicas usadas, buscamos construir modelos robustos que pudessem servir como uma linha de base confiável para comparação com nossa abordagem não supervisionada proposta. As métricas na Tabela V.7 comparam nosso modelo com os modelos supervisionados. Os dados relacionados aos modelos supervisionados são apresentados no formato *média aritmética* ± *desvio padrão*, pois se referem ao resultado obtido pela aplicação da validação cruzada em 10 dobras.

Método	**Reamostragem**	**recall@6**	**F1-score**	**map@6**	**ndcg@6**
LogReg	Sim	0.9494±0.0050	0.6835±0.0089	0.1329±0.0007	0.9114±0.0054
LogReg	Não	**0.9543±0.0056**	0.6712±0.0082	0.1365±0.0008	**0.9242±0.0043**
XGBoost	Sim	0.8712±0.0103	0.7215±0.0100	0.1313±0.0013	0.8553±0.0080
XGBoost	Não	0.9065±0.0086	**0.7447±0.0090**	0.1364±0.0014	0.8861±0.8861
GLARE	n/a	0.7575	0.6268	**0.5345**	0.5902

Tabela V.7: Desempenho comparativo: algoritmos não supervisionados *vs* supervisionados (XGBoost: Extreme Gradient Boosting. LogReg: Logistic Regression).

Desempenho em classes com poucos exemplos

Conforme relatado na Seção IV.2, é comum que, para determinados temas, o banco de dados de um tribunal possua poucos recursos especiais rotulados. Em certos cenários, há a possibilidade de que o recurso recentemente submetido ao tribunal apresente uma questão inédita ao Judiciário. Essas são situações importantes com as quais o perito humano lida ao analisar um recurso especial. Portanto, são as premissas sobre as quais desenvolvemos nosso método para classificar o apelo especial de um tema.

Para avaliar o desempenho dos modelos supervisionados em classes com poucos exemplos, conduzimos um experimento no qual extraímos dois subconjuntos disjuntos do corpus original obedecendo às seguintes regras:

- O conjunto de treinamento do modelo tem 80% de o corpus original.
- O conjunto de teste do modelo tem 20% do corpus original.
- O conjunto de teste compreende principalmente classes com a menor representação no corpus original.

- O conjunto de treinamento tem a garantia de ter pelo menos um exemplo de cada classe do corpus original.
- Respeitando as regras impostas, a inserção de dados nos conjuntos de treinamento e teste ocorre aleatoriamente.

O histograma na Figura V.6 mostra a frequência dessas classes minoritárias no conjunto de treinamento do modelo supervisionado.

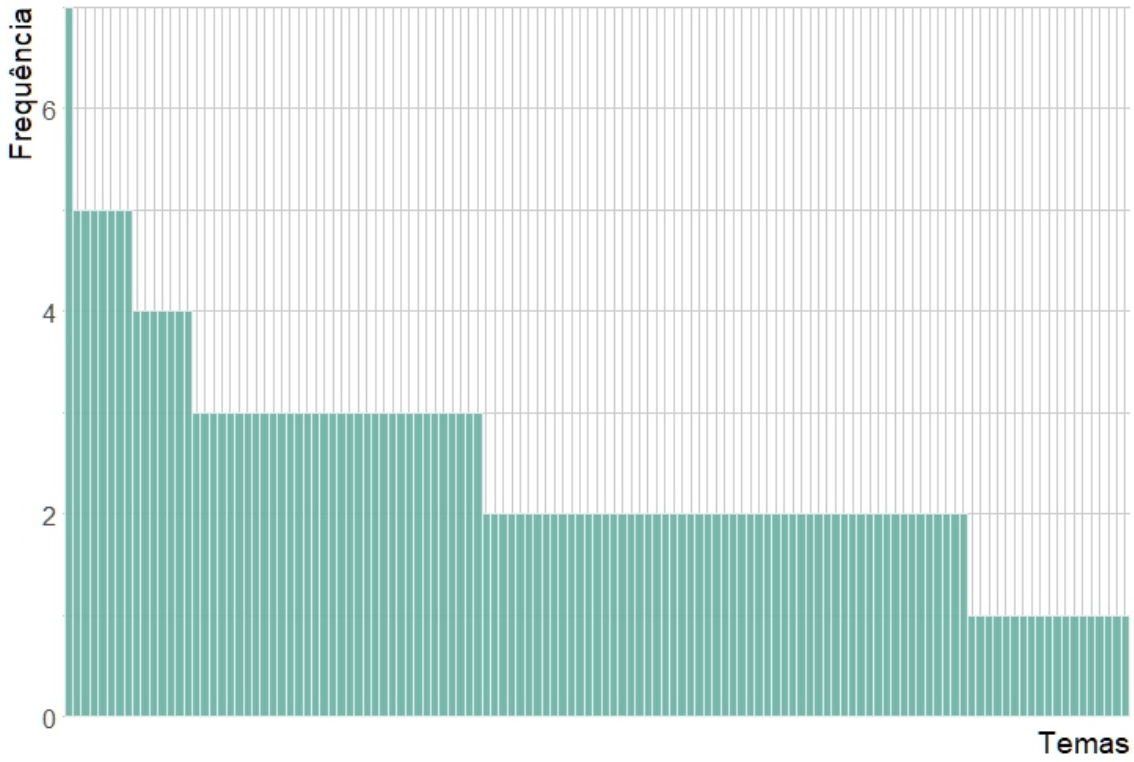


Figura V.6: O histograma apresenta a distribuição dos temas no conjunto de treinamento dos modelos supervisionados, nos experimentos com classes minoritárias. Cada barra representa um tema específico, e sua altura indica a quantidade de documentos associados a ele. A figura evidencia a baixa ocorrência de alguns temas no conjunto de dados utilizado.

A melhor configuração de hiperparâmetros para o modelo foi identificada aplicando validação cruzada ao conjunto de treinamento. Depois disso, o modelo foi treinado em todo o conjunto de treinamento (Seção V.5) . Este modelo final foi aplicado ao conjunto de teste previamente separado (20% dos dados do corpus original). Como o conjunto de teste não foi usado em nenhum momento durante o treinamento ou validação, isso nos fornece uma avaliação imparcial do desempenho do modelo. Durante este experimento, vimos que os modelos supervisionados tiveram um desempenho ruim. Este resultado foi particularmente surpreendente quando comparado ao desempenho do nosso

método não supervisionado em execução no mesmo conjunto de dados. Nosso modelo fez atribuição de tema correta em 72% dos casos, como visto na Tabela V.8.

Tabela V.8: Desempenho comparativo: algoritmo não supervisionado *vs* supervisionados com classes minoritárias no conjunto de teste (XGBoost: Extreme Gradient Boosting. LogReg: Logistic Regression).

Método	**Reamostragem**	**recall@6**	**F1-score**	**map@6**	**ndcg@6**
LogReg	Sim	0.5442	0.3903	0.0696	0.4818
LogReg	Não	0.6064	0.3223	0.0646	0.4920
XGBoost	Sim	0.2241	0.2323	0.0322	0.2075
XGBoost	Não	0.2931	0.2412	0.0375	0.2548
GLARE	n/a	**0.7200**	**0.6059**	**0.5230**	**0.5717**

V.5.1 Avaliando a generalização do modelo em temas jurídicos não representados nos dados de treinamento

Em outro experimento, examinamos um cenário extremo em que documentos de apelação especial no conjunto de teste se alinham com tópicos sem documentos no conjunto de treinamento. Por meio desse experimento, replicamos a situação em que um especialista humano encontra uma apelação especial que não se assemelha a nenhum conteúdo em um banco de dados histórico. Mesmo sem outro documento de referência previamente classificado, o especialista deve ser capaz de analisar o documento e classificá-lo em um tópico da lista de tópicos repetitivos (Tabela V.9).

Tabela V.9: Desempenho comparativo: algoritmo não supervisionado *vs* supervisionados, com classes sem representantes no conjunto de treinamento (XGBoost: Extreme Gradient Boosting. LogReg: Logistic Regression).

Método	**Reamostragem**	**recall@6**	**F1-score**	**map@6**	**ndcg@6**
LogReg	Sim	0.0000	0.0000	0.0000	0.0000
LogReg	Não	0.0000	0.0000	0.0000	0.0000
XGBoost	Sim	0.0000	0.0000	0.0000	0.0000
XGBoost	Não	0.0000	0.0000	0.0000	0.0000
GLARE	n/a	**0.6903**	**0.5917**	**0.5177**	**0.5598**

Os resultados mostraram que os modelos supervisionados tiveram dificuldades com a generalização. Eles não conseguiram fazer avaliações corretas em classes para as quais não foram treinados. Ficou evidente que um volume substancial de dados rotulados é essencial para o modelo supervisionado aprender padrões representativos. Em contraste, nosso modelo não supervisionado provou ser eficaz no cenário analisado. Em aproximadamente 70% dos casos, o modelo identificou corretamente o tema do recurso especial.

Capítulo VI Conclusão

Neste capítulo apresentamos a conclusão do trabalho e discutimos possíveis trabalhos futuros. Na Seção VI.1 procedemos uma breve análise retrospectiva e analisamos os resultados alcançados. Na Seção VI.2 apresentamos algumas sugestões de possíveis trabalhos futuros.

Todos os experimentos descritos neste artigo foram desenvolvidos com a linguagem de programação Python 3.10. Para fins de reprodutibilidade, nosso código-fonte está disponível publicamente em `https://github.com/AILAB-CEFET-RJ/r2t`. Os procedimentos de instalação e de replicação dos experimentos estão descritos no próprio repositório.

Os conjuntos de dados que foram construídos no contexto desta dissertação estão publicamente disponíveis em formato CSV. Eles podem ser baixados no Zenodo, um repositório de dados online de código aberto: `http://doi.org/10.5281/zenodo.13696090`.

VI.1 Análise Retrospectiva

Conforme descrito na Seção I.3, o objetivo primário deste trabalho é desenvolver um método eficaz, baseado em inteligência artificial e técnicas de processamento de linguagem natural, para auxiliar analistas jurídicos na tarefa de classificar recursos especiais em temas jurídicos específicos. Por meio dos resultados experimentais apresentados no Capítulo V, confirmamos a eficácia do nosso método, especialmente em um cenário onde há poucos exemplos de documentos para uma determinada classe, conforme demonstrado na Seção V.5.

Em concordância para atingimento do objetivo primário definimos alguns objetivos secundários (Seção I.3). Em alinhamento com os objetivos secundários construímos o conjunto de dados detalhado na Seção IV.3, o qual foi utilizado ao longo dos experimentos. Um ponto de destaque na execução dos experimentos foi o ganho obtido ao gerarmos a representação vetorial dos textos, serializá-la e salvá-la em arquivo para uso posterior.

Elaboramos o método GLARE ancorado na hipótese de que dado um texto qualquer seria possível gerar um resumo que preservasse informação essencial desse texto (Seção I.4). Comprovamos a veracidade da hipótese verificando nos experimentos a correta classificação dos recursos especiais por meio dos resumos dos textos. Quantitativamente a métrica do *recall* comprovou a veracidade da hipótese. Em 100% dos experimentos houve resultado positivo, ou seja, obtivemos um *recall*

superior a zero. Os resultados confirmaram ser possível gerar um resumo que preserve o conteúdo essencial do texto e a partir desse resumo realizar uma classificação.

Ainda relacionado a hipótese de gerar um resumo preservando a semântica do texto, tínhamos um objetivo secundário que era analisar o impacto que a sumarização exerce na tarefa de classificar o documento em um tema. Para avaliação deste objetivo, incluímos em nossos experimentos dois casos nos quais utilizamos o texto integral com suas respectivas representações vetoriais, desta forma foi possível realizar comparações com os casos nos quais geramos resumos para efetuar a classificação em um tema. Conforme os resultados apresentados no Apêndice D, sintetizados na Seção V.3 , observamos resultados superiores ao utilizarmos os resumos textuais em comparação com o uso do texto integral.

Um outro objetivo secundário era comparar distintas formas de medição da similaridade. Tradicionalmente utiliza-se a similaridade de cosseno, contudo, mostramos ser viável o uso do algoritmo BM25 para analisar a similaridade entre textos. Nosso método, GLARE, combina o Guided LexRank para sumarização e o BM25 para avaliação de similaridade, oferecendo uma solução não supervisionada e eficaz. Os gráficos e dados tabulares apresentados no Capítulo V apresentam essa comparação entre similaridade por cosseno e BM25.

Relacionado ao tratamento dos recursos especiais, tínhamos um objetivo secundário neste trabalho que era o de investigar de que forma o pré-processamento do texto afetava o desempenho do método GLARE. Analisando os resultados apresentados na Seção V.2, não identificamos um ganho de desempenho especificamente atribuído ao pré-processamento.

O conjunto de dados utilizado cooperou para atingirmos um objetivo secundário de validar o método GLARE com um conjunto de dados de recursos especiais reais. O método se mostrou eficaz. Porém, ao analisarmos os resultados apresentados na Seção V.2 identificamos espaço para melhorias especialmente na etapa de pré-processamento textual.

Ainda como um dos objetivos secundários, buscamos confrontar o método GLARE com outras abordagens existentes.Ao compararmos nosso método com modelos supervisionados ele mostrou-se competitivo, conforme pode ser visto na Seção V.5. Nosso método apresentou um resultado excepcional, em relação aos modelos supervisionados de base, especialmente nos cenários onde há poucos dados rotulados para determinada classe.

Nosso método não considera nenhuma estrutura linguística específica e pode ser aplicado em contextos além da atribuição de tema de recursos especiais e até mesmo do domínio legal. Como é uma abordagem de aprendizado de máquina não supervisionado, não requer dados para treinamento como o aprendizado supervisionado. Provou ser uma solução eficaz e versátil, adequada para outros problemas não explorados neste projeto.

VI.2 Trabalhos futuros

Considerando que o desempenho do nosso método foi muito maior do que a linha de base em cenários com classes minoritárias, uma direção promissora para trabalhos futuros seria desenvolver um sistema híbrido que combinasse aprendizado supervisionado e não supervisionado, criando um sistema mais robusto e eficiente.

Além disso, trabalhos futuros poderiam explorar a integração de Large Language Models (LLMs) como parte do pipeline de pré-processamento em nosso método. LLMs, como GPT-4, demonstraram potencial significativo no aprimoramento de tarefas de pré-processamento de dados, incluindo normalização de texto, detecção de erros e imputação de dados [Zhang et al., 2023]. Ao alavancar LLMs, é possível melhorar a qualidade e a consistência dos dados de entrada, levando a um desempenho ainda melhor e a uma aplicabilidade mais ampla do nosso método em vários domínios legais e não legais.

Referências

Constituição. https://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm, 2023. (Accessed on 02/23/2023).

Censo 2022 | ibge. https://www.ibge.gov.br/estatisticas/sociais/populacao/22827-censo-demografico-2022.html?edicao=35938&t=resultados, 2023. (Acessado em 04/23/2023).

Cnj serviço: Saiba a diferença entre repercussão geral e recurso repetitivo - portal cnj. https://www.cnj.jus.br/cnj-servico-saiba-a-diferenca-entre-repercussao-geral-e-recursos-repetitivos/, 2023. (Acessado em 04/23/2023).

Conselho Nacional de Justiça. Justiça em números 2021 / conselho nacional de justiça. – brasília: Cnj,2021. https://www.cnj.jus.br/wp-content/uploads/2021/09/relatorio-justica-em-numeros2021-12.pdf, 2021. (Acessado em 07/02/2022).

Estatísticas e painéis de gestão - portal cnj. https://www.cnj.jus.br/programas-e-acoes/estatistica/. (Accessed on 06/13/2023).

Regimento interno do trf2. https://static.trf2.jus.br/nas-internet/documento/institucional/publicacoes/trf2-regimento-interno-2023-03-07.pdf. (Accessed on 06/13/2023).

A. Silberschatz, S. Sundarshan, and H.F. Korth. *Sistema de Banco de Dados*. Elsevier, 2011. ISBN 9788535239898.

Justin Zobel and Alistair Moffat. Inverted files for text search engines. *ACM Computing Surveys*, 38(2), July 2006. ISSN 0360-0300. doi: 10.1145/1132956.1132959. URL http://doi.acm.org/10.1145/1132956.1132959.

J. Han, M. Kamber, and J. Pei. *Data Mining: Concepts and Techniques*. The Morgan Kaufmann Series in Data Management Systems. Elsevier Science, 2011. ISBN 9780123814807. URL https://books.google.com.br/books?id=pQws07tdpjoC.

William B. Frakes and Ricardo A. Baeza-Yates, editors. *Information Retrieval: Data Structures & Algorithms*. Prentice-Hall, 1992. ISBN 0-13-463837-9.

Charu C. Aggarwal. *Machine Learning for Text, Second Edition*. Springer, 2022a. ISBN 978-3-030-96622-5. doi: 10.1007/978-3-030-96623-2. URL https://doi.org/10.1007/978-3-030-96623-2.

S.E. Robertson and K. Sparck Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(3):129 – 146, 1976. doi: 10.1002/asi.4630270302. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-0016958419&doi=10.1002%2fasi.4630270302&partnerID=40&md5=ea439b59d6110a1c6cde459d4ef88882. Cited by: 1327.

G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, nov 1975. ISSN 0001-0782. doi: 10.1145/361219.361220. URL https://doi-org.ez108.periodicos.capes.gov.br/10.1145/361219.361220.

A statistical approach to machine translation. https://dl.acm.org/doi/pdf/10.5555/92858.92860. (Accessed on 06/05/2023).

Stephen E. Robertson, Steve Walker, Micheline Hancock-Beaulieu, Aarron Gull, and Marianna Lau. Okapi at TREC. In Donna K. Harman, editor, *Proceedings of The First Text REtrieval Conference, TREC 1992, Gaithersburg, Maryland, USA, November 4-6, 1992*, volume 500-207 of *NIST Special Publication*, pages 21–30. National Institute of Standards and Technology (NIST), 1992. URL http://trec.nist.gov/pubs/trec1/papers/02.txt.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389, apr 2009. ISSN 1554-0669. doi: 10.1561/1500000019. URL https://doi.org/10.1561/1500000019.

Pluggable similarity algorithms | elasticsearch: The definitive guide [2.x] | elastic. https://www.elastic.co/guide/en/elasticsearch/guide/current/pluggable-similarites.html#bm25-saturation. (Accessed on 01/06/2024).

Ricardo J. G. B. Campello, Davoud Moulavi, Arthur Zimek, and Jörg Sander. Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Trans. Knowl. Discov. Data*, 10(1), jul 2015. ISSN 1556-4681. doi: 10.1145/2733381. URL https://doi-org.ez108.periodicos.capes.gov.br/10.1145/2733381.

Parth Mehta and Prasenjit Majumder. *From Extractive to Abstractive Summarization: A Journey*. 01 2019a. ISBN 978-981-13-8933-7. doi: 10.1007/978-981-13-8934-4.

Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.

P. Mehta and P. Majumder. *From Extractive to Abstractive Summarization: A Journey*. Springer Nature Singapore, 2019b. ISBN 9789811389344.

G. Erkan and D. R. Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479, December 2004. ISSN 1076-9757. doi: 10.1613/jair.1523. URL `http://dx.doi.org/10.1613/jair.1523`.

Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, Roger Chou, Julie Glanville, Jeremy M Grimshaw, Asbjørn Hróbjartsson, Manoj M Lalu, Tianjing Li, Elizabeth W Loder, Evan Mayo-Wilson, Steve McDonald, Luke A McGuinness, Lesley A Stewart, James Thomas, Andrea C Tricco, Vivian A Welch, Penny Whiting, and David Moher. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372, 2021. doi: 10.1136/bmj.n71. URL `https://www.bmj.com/content/372/bmj.n71`.

Daniela Vianna, Edleno Silva de Moura, and Altigran Soares da Silva. A topic discovery approach for unsupervised organization of legal document collections. *Artificial Intelligence and Law*, pages 1–30, 2023.

Sarthak Dalal, Amit Singhal, and Brejesh Lall. Lexrank and pegasus transformer for summarization of legal documents. In Dilip Singh Sisodia, Lalit Garg, Ram Bilas Pachori, and M. Tanveer, editors, *Machine Intelligence Techniques for Data Analysis and Signal Processing*, pages 569–577, Singapore, 2023. Springer Nature Singapore. ISBN 978-981-99-0085-5.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

M. N. Satwick Gupta, N. L. Siva Narayana, V. Sai Charan, Kunam Balaram Reddy, Malaya Dutta Borah, and Deepali Jain. Extractive summarization of indian legal documents. In Ripon Patgiri, Sivaji Bandyopadhyay, Malaya Dutta Borah, and Valentina Emilia Balas, editors, *Edge Analytics*, pages 629–638, Singapore, 2022. Springer Singapore. ISBN 978-981-19-0019-8.

Yen Thi-Hai Vuong, Quan Minh Bui, Ha-Thanh Nguyen, Thi-Thu-Trang Nguyen, Vu Tran, Xuan-Hieu Phan, Ken Satoh, and Le-Minh Nguyen. Sm-bert-cr: a deep learning approach for case law retrieval with supporting model. *Artif. Intell. Law*, 31(3):601–628, aug 2022. ISSN 0924-8463.

Pengyu Li, Christine Tseng, Yaxuan Zheng, Joyce A. Chew, Longxiu Huang, Benjamin Jarman, and Deanna Needell. Guided semi-supervised non-negative matrix factorization. *Algorithms*, 15(5), 2022. doi: 10.3390/a15050136. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85129443693&doi=10.3390%2fa15050136&partnerID=40&md5=d3939cbd3d22c9e3c531e50357a332a7. Cited by: 2; All Open Access, Gold Open Access.

Stj - precedentes qualificados. https://processo.stj.jus.br/repetitivos/temas_repetitivos/. (Accessed on 06/05/2023).

Márcia Cançado. Manual de semântica. *Belo Horizonte: Editora UFMG*, 2005.

Jahna Otterbacher, Gunes Erkan, and Dragomir R. Radev. Biased lexrank: Passage retrieval using random walks with question-based priors. *Information Processing and Management*, 45(1):42 – 54, 2009. doi: 10.1016/j.ipm.2008.06.004. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-55949118940&doi=10.1016%2fj.ipm.2008.06.004&partnerID=40&md5=e1dd3e6c435b45ef7fb58daa85323487. Cited by: 73.

Harald Steck, Chaitanya Ekanadham, and Nathan Kallus. Is cosine-similarity of embeddings really about similarity? In *Companion Proceedings of the ACM on Web Conference 2024*, WWW '24, page 887–890, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400701726.

Charu C. Aggarwal. *Machine Learning for Text, Second Edition.* Springer, Gewerbestrasse, 2022b. ISBN 978-3-030-96622-5.

eproc | trf 2. https://portaleproc.trf2.jus.br/. (Accessed on 06/05/2023).

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16. ACM, aug 2016.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, page 694–699, New York, NY, USA, 2002. Association for Computing Machinery. ISBN 158113567X.

Haochen Zhang, Yuyang Dong, Chuan Xiao, and Masafumi Oyamada. Large language models as data preprocessors, 2023.

Apêndice A Recurso Especial

EXCELENTÍSSIMO SENHOR DOUTOR DESEMBARGADOR PRESIDENTE DO EGRÉGIO TRIBUNAL DE JUSTIÇA DO ESTADO DE XXXXXXXXXX

Embargos de Declaração autuado sob nº XXXXXXXXXXXXX

Recorrente: XXXXXXXXXXXXX

Recorrido: XXXXXXXXXXXXXX

Recorrido: XXXXXXXXXXXXXX

XXXXXXXXXXX, (Prenome), já qualificado nos autos, de número em epígrafe, que move em face de ato do PXXXXXXXXXXXX, pessoa jurídica de direito público, já qualificado nos autos, e em face do PRESIDENTE DA ASSOCIAÇÃO DE COMERCIANTES LOCAIS (ACSC-XX), pessoa jurídica de direito privado, também já qualificado nos autos, vem, respeitosamente e tempestivamente, por seu advogado e bastante procurador que esta subscreve, procuração em anexo (Doc), com escritório profissional situado na (Rua), (número), (bairro), (CEP), (Cidade), (Estado), endereço eletrônico: xxxxxxxx@gmail.com, nos termos do artigo 105, inciso III, alínea a, e artigo 1.029 e seguintes do Código de Processo Civil, interpor o presente **RECURSO ESPECIAL** contra respeitável Acórdão proferido nas fls._____, que negou ser impossível a devolução dos valores recebidos irregularmente pela associação em face dos recorridos qualificados nos autos, interposto nos Embargos de Declaração da ação em epígrafe.

Requer seja recebido e processado o presente recurso, intimando-se a parte contrária para que ofereça, dentro do prazo legal, as contrarrazões e, após seja o recurso admitido e encaminhado com as inclusas razões ao Colendo Superior Tribunal de Justiça.

Termos em que,

Pede deferimento.

XXXXXX, XX de XXXX de XXXX.

Nome do Advogado

OAB/XX XXX.XXX

RAZÕES DE RECURSO ESPECIAL

Embargos de Declaração autuado sob nº XXXXXXXXXXXXX

Recorrente: XXXXXXXXXX

Recorrido: XXXXXXXXXX

Recorrido: XXXXXXXXXX

Colenda Turma,

Ínclitos Julgadores

XXXXXXX , não se conformando com o respeitável Acórdão de fls._____, vem, respeitosamente, apresentar as razões do presente Recurso Especial.

I. A) - DO CABIMENTO DO PRESENTE RECURSO ESPECIAL

(NCPC, ART. 1.029, inc. II)

Lado outro, disciplina o artigo 105, III, a, da Constituição Federal, que é da competência, exclusiva, do Superior Tribunal de Justiça, apreciar Recurso Especial, fundado em decisão proferida em última ou única instância, quando a mesma contrariar lei federal ou negar-lhe vigência.

A hipótese se ajusta aos ditames supra, e, nessa esteira, converge ao exame deste Recurso Especial.

a) Contrariar tratado ou lei federal, ou negar-lhes vigência:

É cabível salientar que o direito de recorrer decorre de um princípio denominado duplo grau de jurisdição, que é um dos desdobramentos do chamado Devido Processo Legal. Esse princípio não está expressamente previsto no nosso Direito Constitucional, mas decorre da lógica do sistema, que prevê competências recursais a cada um dos órgãos do Poder Judiciário na Constituição. Por outro lado, é expresso no Pacto de San José da Costa Rica, a principal Carta de Direitos Humanos do nosso continente1. Decerto, então, inexistir o óbice contido na Súmula nº. 7 do STJ. Acrescente-se que a decisão guerreada, como dito, contrariou os princípios da proporcionalidade e razoabilidade.

b) Julgar válido ato de governo local contestado em face de lei federal conforme redação dada pela Emenda Constitucional nº 45, de 2004:

"julgar válido ato de governo local contestado em face de lei federal".

c) Der a lei federal interpretação divergente da que lhe seja atribuída em outro tribunal.

O objetivo do presente Recurso Especial é a correta aplicação da Lei da Ação Popular – Lei Federal nº 4.717, de 29 de junho de 1965 –, que não foi corretamente aplicada em seu art. 11, pois foi negado o pedido de devolução dos valores recebidos indevidamente em razão da ilegal concessão do serviço de "Zona Azul" à Associação de Comerciantes de São Caetano, pelo Decreto nº 01/2019, sem a realização de licitação prévia

I. B) DA TEMPESTIVIDADE DO PRESENTE REsp

A Recorrente fora intimada da decisão guerreada por meio do Diário da Justiça nº ___. Esse circulou no dia xx de xxxx de xxxx (xxxxxx).

Diante disso, mostra-se tempestiva a interposição em espécie, *ex vi* do artigo 1.003 § 5º do Código de Ritos.

I. C) DO PREPARO

Essas se encontram acompanhadas do devido preparo (custas e guias de porte de remessa e retorno), uma vez que o processo é físico (CPC, art. 1.007, *caput* c/c § 3º).

"O sistema constitucional brasileiro não permite o controle normativo abstrato de leis municipais, quando contestadas em face da Constituição Federal. A fiscalização de constitucionalidade das leis e atos municipais, nos casos em que estes venham a ser questionados em face da Carta da Republica, somente se legitima em sede de controle incidental (método difuso). Desse modo, inexiste, no ordenamento positivo brasileiro, a ação direta de inconstitucionalidade de lei municipal, quando impugnada 'in abstracto' em face da Constituição Federal. Doutrina. Precedentes do Supremo Tribunal Federal."

I. D) DO PREQUESTIONAMENTO

Art . 1025. Consideram-se incluídos no acórdão os elementos que o embargante suscitou, para fins de pré-questionamento, ainda que os embargos de declaração sejam inadmitidos ou rejeitados, caso o tribunal superior considere existentes erro, omissão, contradição ou obscuridade.

I.E) DA SÍNTESE DOS FATOS

XXXXXXXXX, comerciante e cidadão de XXXXXXXX, no estado de XXXXXXXX, ficou inconformado com uma decisão do prefeito de sua cidade, XXXXXXXX, o qual, por meio do Decreto Municipal nº 01/2019, transferiu a cobrança do serviço de estacionamento em locais públicos, denominado "Zona Azul", sem a devida licitação (modalidade imposta aos entes públicos para realizar a denominada concessão de serviços públicos), para uma associação de comerciantes locais (ACSC-XX), cujo presidente, Sr. XXXXXX, é seu amigo pessoal e principal colaborador de sua campanha eleitoral.

XXXXXXXXX com a finalidade de propor uma ação para anular a transferência indevida do serviço público para essa associação, em razão do descumprimento de mandamentos constitucionais que exigem a realização de licitação para a concessão de serviços públicos e da imoralidade da medida de beneficiar os seus conhecidos.

A Ação Popular foi promovida, a qual foi julgada improcedente pelo juiz de primeiro grau e, em sede de apelação, foi julgada procedente.

Contudo, o Tribunal anulou o Decreto nº 01/2019, como foi requerido, mas nada foi dito a respeito da devolução do dinheiro pago à associação durante a época em que explorou ilegalmente o serviço público de "Zona Azul" no município.

Após ter opostos Embargos de Declaração da decisão do Tribunal de Justiça, os desembargadores acolheram o recurso e complementaram a decisão.

No entanto, eles julgaram ser impossível a devolução dos valores recebidos irregularmente pela associação ré, em razão de a Lei da Ação Popular prever apenas a anulação do ato ilegal realizado,

não a reparação dos danos decorrentes desse ato. O inconformismo das partes sucumbentes e a possibilidade de que sejam cometidos erros judiciais são as razões para que exista o direito de recorrer das decisões judiciais.

II – DA FUNDAMENTAÇÃO JURÍDICA

Tem-se que o pleito dos recorrentes encontra substancial amparo em nossa Carta Magna, conforme dispõe o artigo 105, inciso III, da Constituição Federal Brasileira:

Art. 105. Compete ao Superior Tribunal de Justiça:

(...)

III - julgar, em recurso especial, as causas decididas, em única ou última instância, pelos Tribunais Regionais Federais ou pelos tribunais dos Estados, do Distrito Federal e Territórios, quando a decisão recorrida:

Consubstanciando ainda mais nesse sentido, o artigo 1.029 e seguintes, do Código de Processo Civil, estabelece as disposições gerais para seu procedimento:

Art. 1.029. O recurso extraordinário e o recurso especial, nos casos previstos na Constituição Federal , serão interpostos perante o presidente ou o vice-presidente do tribunal recorrido, em petições distintas que conterão:

I - a exposição do fato e do direito;

II - a demonstração do cabimento do recurso interposto;

III - as razões do pedido de reforma ou de invalidação da decisão recorrida.

III - DAS RAZÕES RECURSAIS

O intuito da presente demanda visa a verificação correta da aplicação da Lei Federal nº 4.717/65, em razão da especialidade da norma que não foi corretamente aplicada, negando o pedido de devolução dos valores recebidos indevidamente em razão da ilegal concessão do serviço de "Zona Azul" à Associação de Comerciantes de São Caetano, pelo Decreto nº 01/2019.

Salienta-se pelo exposto, que o presente recurso não requer o reexame de prova, repita-se, mas sim, e tão somente, que seja verificada a correta aplicação do dispositivo legal da Lei da Ação Popular, consubstanciado no artigo 105, inciso III, alínea a, da Constituição Federal Brasileira:

Art. 105. Compete ao Superior Tribunal de Justiça:

(...)

III - julgar, em recurso especial, as causas decididas, em única ou última instância, pelos Tribunais Regionais Federais ou pelos tribunais dos Estados, do Distrito Federal e Territórios, quando a decisão recorrida:

a) contrariar tratado ou lei federal, ou negar-lhes vigência;

O Acórdão proferido pelo Egrégio Tribunal de Justiça do Estado de Pernambuco, julgou ser impossível a devolução dos valores recebidos irregularmente pela associação, em razão de a Lei

da Ação Popular prever apenas a anulação do ato ilegal realizado, e não a reparação dos danos decorrentes desse ato.

Ora, nobres Julgadores, a decisão supramencionada contraria lei federal, nega-lhe vigência e dá interpretação divergente da que lhe haja atribuído outro tribunal, pois o artigo 11, da Lei nº 4.717/65, é de claro entendimento ao afirmar que:

Art. 11. A sentença que, julgando procedente a ação popular, decretar a invalidade do ato impugnado, condenará ao pagamento de perdas e danos os responsáveis pela sua prática e os beneficiários dele, ressalvada a ação regressiva contra os funcionários causadores de dano, quando incorrerem em culpa.

Assim, conforme dispõe o artigo 93, inciso IX, da Constituição Federal Brasileira, o qual faz menção à obrigatoriedade da exposição das motivações judiciais, para alcançar o fim a que se destina no Direito, é essencial que a tutela jurisdicional seja prestada de forma clara e fundamentada, sob pena de nulidade.

Art. 93. Lei complementar, de iniciativa do Supremo Tribunal Federal, disporá sobre o Estatuto da Magistratura, observados os seguintes princípios

(...)

IX todos os julgamentos dos órgãos do Poder Judiciário serão públicos, e ***fundamentadas todas as decisões, sob pena de nulidade****, podendo a lei limitar a presença, em determinados atos, às próprias partes e a seus advogados, ou somente a estes, em casos nos quais a preservação do direito à intimidade do interessado no sigilo não prejudique o interesse público à informação; (Redação dada pela Emenda Constitucional nº 45, de 2004)*

Vale frisar que a decisão embargada, como se demonstrou nos fatos, também deixou de acolher o pedido de devolução dos valores recebidos indevidamente em razão da ilegal concessão do serviço de "Zona Azul" à Associação de Comerciantes de São Caetano, por meio do Decreto nº 01/2019, sem a realização de licitação prévia.

O Código de Processo Civil, conforme estabelece o artigo 1.022, parágrafo único, inciso II, dispõe:

Art. 1.022. Cabem embargos de declaração contra qualquer decisão judicial para:

(...)

Parágrafo único. Considera-se omissa a decisão que:

(...)

II - incorra em qualquer das condutas descritas no art. 489, § 1º.

Assim, consubstanciando, dispõe o artigo 489, § 1º, inciso IV, do Código de Processo Civil, a seguinte redação:

Art. 489. São elementos essenciais da sentença:

(...)

§ 1º Não se considera fundamentada qualquer decisão judicial, seja ela interlocutória, sentença ou acórdão, que:

(...)

IV - não enfrentar todos os argumentos deduzidos no processo capazes de, em tese, infirmar a conclusão adotada pelo julgador;

Desta forma, por meio de todo o exposto, demonstra-se que, ainda tendo decisões favoráveis à parte recorrente, devido à falta da aplicação de tais detrimentos legais e essenciais para seu julgamento, o direito da mesma encontra-se fragilizado.

IV - DO DIREITO

II.A) Da ofensa aos artigos 1.022, II art. 489 parágrafo 1º, IV art. 373, I e art. 1.013 e incisos, todos de NCPC/2015;

Não tendo sido acolhidos os embargos de declaração, acabou se por infringir os art. 1.022, II e 489, parágrafo 1º, IV do NCPC/2015, que assim estão dispostos:

Art. 1.022 Cabem Embargos de Declaração contra qualquer decisão judicial para:

I – esclarecer obscuridade ou eliminar contradição;

II – suprir omissão de ponto ou questão sobre o qual devia se pronunciar o juiz de ofício ou a requerimento.

III – corrigir erro material

Nesse compasso, examinando-se esses acórdãos, constata-se similitude fática entre a decisão recorrida e o aresto apontado como paradigma. Por isso, revelam teses diversas na interpretação do mesmo dispositivo legal.

Por tudo isso, merece, há de ser conhecido este recurso especial também pela alínea c, do Texto Maior.

V – DOS PEDIDOS

Como se vê todos os dispositivos da Lei Federal e decisão paradigma acima transcritos, resta cabalmente demonstrada a violação de dispositivos de lei federal e a divergência jurisdicional acerca da interpretação dos dispositivos legais violados.

FACE AO EXPOSTO, e tendo sido atendidos todos os requisitos de admissibilidade recursal, **requer a recorrente**:

a) seja recebido, processado e admitido o presente Recurso Especial;

b) seja intimada a recorrida, para, querendo apresentar sua resposta, no prazo previsto em lei;

c) seja juntados os comprovantes das custas do despacho de admissibilidade e da interposição de recurso em instância inferior;

d) sejam juntados os acórdãos e certidões em anexo, para fim de fazer prova da divergên-

cia/dissídio jurisdicional na forma do art. 1.029 parágrafo 1º do NCPC.

e) seja dado provimento ao presente Recurso Especial, determinando-se a NULIDADE do acórdão por falta de fundamentação, não apreciação de todos os argumentos e erro na valoração das provas, reconhecendo-se a prescrição anual, tudo com base nos fundamentos acima aludidos, por ser matéria de DIREITO e JUSTIÇA.

Respeitosamente, pede deferimento.

Cidade, 00 de XXXX de XXXX.

Advogado – OAB/XX xxxxxx

Apêndice B Amostra de tema repetitivo submetido ao STJ

TEMA REPETITIVO 1031	
Questão submetida a julgamento	**Possibilidade de reconhecimento da especialidade da atividade de vigilante, exercida após a edição da Lei 9.032/1995 e do Decreto 2.172/1997, com ou sem o uso de arma de fogo.**
Tese Firmada	É possível o reconhecimento da especialidade da atividade de Vigilante, mesmo após EC 103/2019, com ou sem o uso de arma de fogo, em data posterior à Lei 9.032/1995 e ao Decreto 2.172/1997, desde que haja a comprovação da efetiva nocividade da atividade, por qualquer meio de prova até 5.3.1997, momento em que se passa a exigir apresentação de laudo técnico ou elemento material equivalente, para comprovar a permanente, não ocasional nem intermitente, exposição à atividade nociva, que coloque em risco a integridade física do Segurado.
Anotações NUGEPNAC	Dados parcialmente recuperados via sistema Athos e Projeto Accordes.Afetação na sessão eletrônica iniciada em 25/9/2019 e finalizada em 1/10/2019 (Primeira Seção).Vide Controvérsia n. 133/STJ. REsp n. 1831377/PR sobrestado pelo Tema 1.209/STF (decisão da Vice-Presidência do STJ de 9/2/2022). Tema 1.209/STJ sobrestado.Decisão da Vice-Presidência do STJ, publicada no DJe de 1/2/2022, no Resp n. 1.830.508/RS, nos seguintes termos: "Por meio de ofício encaminhado a todos os tribunais, o Supremo Tribunal Federal recomendou que, nos feitos representativos de controvérsia, ainda que se vislumbre questão infraconstitucional, o recurso extraordinário seja admitido de forma a permitir o pronunciamento da Suprema Corte sobre a existência, ou não, de matéria constitucional no caso e, eventualmente, de repercussão geral. Assim, diante da relevância da matéria debatida e considerando que o aresto recorrido foi proferido sob o rito dos arts. 1.036 e seguintes do Código de Processo Civil, entende-se ser o caso de remessa do apelo extremo ao Pretório Excelso, na qualidade de representativo de controvérsia. Diante do exposto, com fulcro no art. 1.036, § 1º, do Código de Processo Civil, admite-se o presente recurso extraordinário."Os recursos especiais n. 1.813.371/SP e 1.831.377/PR também tiveram seus recursos extraordinários recebidos na qualidade de representativo de controvérsia. Vide acórdão proferido na Pet n. 10.679/RN, relator Ministro Napoleão Nunes Maia Filho, DJe de 22/5/2019.
Informações Complementares	Há determinação de suspensão do processamento de todos os processos pendentes, individuais ou coletivos, que versem acerca da questão delimitada e tramitem no território nacional (acórdão publicado no DJe de 21/10/2019).
Repercussão Geral	Tema 1209/STF - Reconhecimento da atividade de vigilante como especial, com fundamento na exposição ao perigo, seja em período anterior ou posterior à promulgação da Emenda Constitucional 103/2019.

Apêndice C Estrutura do Conjunto de Dados

O conjunto de dados, armazenados em um arquivo Comma-separated values (CSV), consiste em 7.967 registros referentes a recursos especiais. Os recursos estão distribuídos não uniformemente em 190 tipos de temas. Todos possuem uma classificação real dada por um especialista. O detalhamento dos metadados encontra-se a seguir.

- **Recurso:** Dado textual que armazena todo o conteúdo do recurso especial.
- **Tema:** Dado textual que armazena o conteúdo do tema, conforme disponibilizado na base de dados do STJ. Este é o tema que um especialista classificou o recurso especial submetido.
- **Tese:** Dado textual que armazena o conteúdo da tese jurídica firmada pelo STJ após julgamento de recursos especiais repetitivos. No caso de conteúdo nulo, significa que até o momento desta pesquisa não havia uma tese firmada sobre o tema relacionado.
- **Num_tema_cadastrado:** Dado numérico do tipo inteiro que representa o número de tema, conforme disponibilizado na base de dados do STJ. Este é o tema que um especialista classificou o recurso especial submetido.
- **Sugerido_1:** Este é o número de tema dado como primeira sugestão pela máquina de busca Elasticsearch.
- **Sugerido_2:** Este é o número de tema dado como segunda sugestão dada pela máquina de busca Elasticsearch.
- **Sugerido_3:** Este é o número de tema dado como terceira sugestão dada pela máquina de busca Elasticsearch.
- **Sugerido_4:** Este é o número de tema dado como quarta sugestão dada pela máquina de busca Elasticsearch.
- **Sugerido_5:** Este é o número de tema dado como quinta sugestão dada pela máquina de busca Elasticsearch.
- **Sugerido_6:** Este é o número de tema dado como sexta sugestão dada pela máquina de busca Elasticsearch.

- **Probabilidade_1:** Dado numérico . Representa a probabilidade da primeira sugestão dada ser a correta, segundo o Elasticsearch.
- **Probabilidade_2:** Dado numérico . Representa a probabilidade da segunda sugestão dada ser a correta, segundo o Elasticsearch.
- **Probabilidade_3:** Dado numérico . Representa a probabilidade da terceira sugestão dada ser a correta, segundo o Elasticsearch.
- **Probabilidade_4:** Dado numérico . Representa a probabilidade da quarta sugestão dada ser a correta, segundo o Elasticsearch.
- **Probabilidade_5:** Dado numérico . Representa a probabilidade da quinta sugestão dada ser a correta, segundo o Elasticsearch.
- **Probabilidade_6:** Dado numérico . Representa a probabilidade da sexta sugestão dada ser a correta, segundo o Elasticsearch.
- **Sugestao_adotada:** Dado numérico do tipo inteiro. Representa qual sugestão dada pelo Elasticsearch era de fato a correta segundo o especialista. Exemplo: número 1 , significa que a primeira sugestão dada realmente era a correta. No caso do número ser zero, significa que nenhuma das sugestões dadas pelo Elasticsearch era a correta.

Apêndice D Resultados dos experimentos

Resultados dos experimentos realizados estão apresentados a seguir neste apêndice. Para cada experimento está indicada a **origem** dos dados submetidos para computação de métricas. A **origem** é determinada pela combinação dos parâmetros descritos no Capítulo V.

origem	recall	f1-score	map	ndcg
bertopic_palavras_10_bm25	0,4473	0,3660	0,3096	0,3443
bertopic_palavras_10_cosseno	0,3054	0,2304	0,1849	0,2148
bertopic_palavras_10_remocao_bm25	0,4501	0,3620	0,3027	0,3398
bertopic_palavras_10_remocao_cosseno	0,3639	0,2801	0,2277	0,2608
bertopic_palavras_15_bm25	0,4674	0,3824	0,3236	0,3594
bertopic_palavras_15_cosseno	0,3759	0,2904	0,2366	0,2712
bertopic_palavras_15_remocao_bm25	0,5265	0,4272	0,3594	0,4014
bertopic_palavras_15_remocao_cosseno	0,4490	0,3358	0,2682	0,3134
bertopic_palavras_20_bm25	0,5195	0,4265	0,3617	0,4014
bertopic_palavras_20_cosseno	0,4357	0,3374	0,2753	0,3154
bertopic_palavras_20_remocao_bm25	0,5052	0,4088	0,3433	0,3837
bertopic_palavras_20_remocao_cosseno	0,4176	0,3065	0,2421	0,2856
bertopic_palavras_25_bm25	0,5506	0,4452	0,3737	0,4180
bertopic_palavras_25_cosseno	0,4402	0,3375	0,2737	0,3155
bertopic_palavras_25_remocao_bm25	0,5584	0,4508	0,3779	0,4230
bertopic_palavras_25_remocao_cosseno	0,4408	0,3395	0,2761	0,3173
bertopic_palavras_30_bm25	0,5013	0,4108	0,3479	0,3863
bertopic_palavras_30_cosseno	0,4354	0,3224	0,2559	0,3005
bertopic_palavras_30_remocao_bm25	0,5406	0,4461	0,3797	0,4201
bertopic_palavras_30_remocao_cosseno	0,4455	0,3474	0,2847	0,3251
bertopic_palavras_35_bm25	0,5479	0,4446	0,3741	0,4179
bertopic_palavras_35_cosseno	0,4554	0,3494	0,2834	0,3267
bertopic_palavras_35_remocao_bm25	0,5446	0,4418	0,3716	0,4150
bertopic_palavras_35_remocao_cosseno	0,4620	0,3593	0,2940	0,3363
bertopic_palavras_40_bm25	0,5375	0,4340	0,3639	0,4076
bertopic_palavras_40_cosseno	0,4539	356,0000	0,2929	0,3335
bertopic_palavras_40_remocao_bm25	0,5435	0,4511	0,3856	0,4252
bertopic_palavras_40_remocao_cosseno	0,4677	0,3625	0,2959	0,3389
bertopic_palavras_45_bm25	0,5393	0,4438	0,3770	0,4178
bertopic_palavras_45_cosseno	0,4638	0,3605	0,2949	0,3373
bertopic_palavras_45_remocao_bm25	0,5450	0,4497	0,3828	0,4235
bertopic_palavras_45_remocao_cosseno	0,4792	0,3701	0,3014	0,3455
bertopic_palavras_50_bm25	0,5452	0,4531	0,3876	0,4274
bertopic_palavras_50_cosseno	0,4563	0,3588	0,2957	0,3361
bertopic_palavras_50_remocao_bm25	0,5558	0,4507	0,3791	0,4232
bertopic_palavras_50_remocao_cosseno	0,4603	0,3561	0,2904	0,3325
bertopic_palavras_55_bm25	0,5420	0,4405	0,3710	0,4140
bertopic_palavras_55_cosseno	0,4566	0,3531	0,2879	0,3302
bertopic_palavras_55_remocao_bm25	0,5696	0,4552	0,3791	0,4266
bertopic_palavras_55_remocao_cosseno	0,4595	0,3726	0,3134	0,3494
bertopic_palavras_60_bm25	0,5506	0,4476	0,3771	0,4206

origem	recall	f1-score	map	ndcg
bertopic_palavras_60_cosseno	0,4558	0,3598	0,2972	0,3369
bertopic_palavras_60_remocao_bm25	0,5313	0,4330	0,3653	0,4068
bertopic_palavras_60_remocao_cosseno	0,4094	0,3343	0,2824	0,3133
bertopic_palavras_65_bm25	0,5473	0,4445	0,3742	0,4177
bertopic_palavras_65_cosseno	0,4576	0,3504	0,2839	0,3269
bertopic_palavras_65_remocao_bm25	0,5554	0,4552	0,3856	0,4280
bertopic_palavras_65_remocao_cosseno	0,3948	0,3190	0,2676	0,2988
bertopic_palavras_70_bm25	0,5382	0,4359	0,3662	0,4094
bertopic_palavras_70_cosseno	0,4018	0,3227	0,2696	0,3023
bertopic_palavras_70_remocao_bm25	0,5434	0,4429	0,3738	0,4162
bertopic_palavras_70_remocao_cosseno	0,3798	0,3147	0,2686	0,2959
bertopic_palavras_75_bm25	0,5137	0,4207	0,3562	0,3956
bertopic_palavras_75_cosseno	0,4039	0,3239	0,2703	0,3028
bertopic_palavras_75_remocao_bm25	0,5582	0,4570	0,3869	0,4297
bertopic_palavras_75_remocao_cosseno	0,4027	0,3199	0,2653	0,2988
bertopic_palavras_80_bm25	0,5505	0,4520	0,3834	0,4254
bertopic_palavras_80_cosseno	0,4009	0,3203	0,2666	0,2998
bertopic_palavras_80_remocao_bm25	0,5695	0,4578	0,3828	0,4294
bertopic_palavras_80_remocao_cosseno	0,4142	0,3312	0,2759	0,3099
bertopic_palavras_85_bm25	0,5122	0,4176	0,3524	0,3923
bertopic_palavras_85_cosseno	0,3986	0,3191	0,2660	0,2983
bertopic_palavras_85_remocao_bm25	0,5632	0,4428	0,3648	0,4137
bertopic_palavras_85_remocao_cosseno	0,3828	0,3075	0,2569	0,2875
bertopic_palavras_90_bm25	0,5198	0,4196	0,3518	0,3937
bertopic_palavras_90_cosseno	0,3743	0,3097	0,2641	0,2911
bertopic_palavras_90_remocao_bm25	0,5473	0,4327	0,3578	0,4050
bertopic_palavras_90_remocao_cosseno	0,4157	0,3333	0,2781	0,3112
bertopic_palavras_95_bm25	0,5565	0,4498	0,3773	0,4221
bertopic_palavras_95_cosseno	0,3919	0,3095	0,2558	0,2897
bertopic_palavras_95_remocao_bm25	0,5874	0,4631	0,3823	0,4334
bertopic_palavras_95_remocao_cosseno	0,3943	0,3172	0,2654	0,2970
bertopic_palavras_100_bm25	0,5624	453,0000	0,3792	0,4248
bertopic_palavras_100_cosseno	0,3899	0,3116	0,2595	0,2918
bertopic_palavras_100_remocao_bm25	0,5881	0,4634	0,3823	0,4336
bertopic_palavras_100_remocao_cosseno	0,4077	0,3259	0,2714	0,3048
bertopic_palavras_105_bm25	0,5490	0,4331	0,3576	0,4047
bertopic_palavras_105_cosseno	0,3654	0,2970	0,2502	0,2785
bertopic_palavras_105_remocao_bm25	0,5862	0,4622	0,3815	0,4322
bertopic_palavras_105_remocao_cosseno	0,4058	0,3235	0,2690	0,3025
bertopic_palavras_110_bm25	0,5455	0,4411	0,3703	0,4139
bertopic_palavras_110_cosseno	0,3872	0,3072	0,2545	0,2875

bertopic_palavras_110_remocao_bm25	0,5776	0,4581	0,3795	0,4286
bertopic_palavras_110_remocao_cosseno	0,4128	0,3252	0,2683	0,3035
bertopic_palavras_115_bm25	0,5434	0,4445	0,3761	0,4178
bertopic_palavras_115_cosseno	0,3989	0,3148	0,2599	0,2943
bertopic_palavras_115_remocao_bm25	0,5696	0,4534	0,3765	0,4245
bertopic_palavras_115_remocao_cosseno	0,4078	0,3232	0,2677	0,3020
bertopic_palavras_120_bm25	0,5367	0,4381	0,3701	0,4117
bertopic_palavras_120_cosseno	0,3971	0,3173	0,2642	0,2972
bertopic_palavras_120_remocao_bm25	0,5647	0,4472	0,3701	0,4185
bertopic_palavras_120_remocao_cosseno	0,3973	0,3217	0,2702	0,3017
bertopic_guiado_palavras_10_bm25	0,4478	0,3635	0,3058	0,3412
bertopic_guiado_palavras_10_cosseno	0,3221	0,2495	0,2036	0,2325
bertopic_guiado_palavras_10_remocao_bm25	0,4609	0,3728	0,3130	0,3502
bertopic_guiado_palavras_10_remocao_cosseno	0,3479	0,2577	0,2047	0,2399
bertopic_guiado_palavras_15_bm25	0,5195	0,4085	0,3365	0,3817
bertopic_guiado_palavras_15_cosseno	0,3899	0,2759	0,2134	0,2576
bertopic_guiado_palavras_15_remocao_bm25	0,5269	0,4333	0,3680	0,4079
bertopic_guiado_palavras_15_remocao_cosseno	0,4409	0,3266	0,2594	0,3047
bertopic_guiado_palavras_20_bm25	0,5380	0,4330	0,3624	0,4063
bertopic_guiado_palavras_20_cosseno	0,4323	0,3335	0,2715	0,3119
bertopic_guiado_palavras_20_remocao_bm25	0,5373	0,4423	0,3758	0,4162
bertopic_guiado_palavras_20_remocao_cosseno	0,4059	0,3138	0,2557	0,2936
bertopic_guiado_palavras_25_bm25	0,5026	0,4149	0,3532	0,3905
bertopic_guiado_palavras_25_cosseno	0,4476	0,3224	0,2519	0,3006
bertopic_guiado_palavras_25_remocao_bm25	0,5242	0,4258	0,3584	0,3996
bertopic_guiado_palavras_25_remocao_cosseno	0,4305	0,3287	0,2659	0,3072
bertopic_guiado_palavras_30_bm25	0,5027	0,4067	0,3415	0,3819
bertopic_guiado_palavras_30_cosseno	0,4317	0,3192	0,2533	0,2979
bertopic_guiado_palavras_30_remocao_bm25	0,5456	0,4453	0,3762	0,4187
bertopic_guiado_palavras_30_remocao_cosseno	0,4422	0,3472	0,2858	0,3253
bertopic_guiado_palavras_35_bm25	0,5377	0,4392	0,3711	0,4132
bertopic_guiado_palavras_35_cosseno	0,4583	0,3444	0,2759	0,3215
bertopic_guiado_palavras_35_remocao_bm25	0,5456	0,4534	0,3878	0,4275
bertopic_guiado_palavras_35_remocao_cosseno	0,4571	0,3585	0,2949	0,3357
bertopic_guiado_palavras_40_bm25	0,5340	0,4384	0,3719	0,4128
bertopic_guiado_palavras_40_cosseno	0,4601	0,3560	0,2903	0,3329
bertopic_guiado_palavras_40_remocao_bm25	0,5474	0,4481	0,3793	0,4212
bertopic_guiado_palavras_40_remocao_cosseno	0,4421	0,3384	0,2741	0,3163
bertopic_guiado_palavras_45_bm25	0,5411	0,4448	0,3776	0,4188
bertopic_guiado_palavras_45_cosseno	0,4617	0,3566	0,2905	0,3333
bertopic_guiado_palavras_45_remocao_bm25	0,5519	0,4449	0,3726	0,4175

bertopic_guiado_palavras_45_remocao_cosseno	0,4613	0,3552	0,2887	0,3322
bertopic_guiado_palavras_50_bm25	0,5425	0,4451	0,3773	0,4188
bertopic_guiado_palavras_50_cosseno	0,4542	0,3554	0,2918	0,3329
bertopic_guiado_palavras_50_remocao_bm25	0,5429	0,4346	0,3623	0,4073
bertopic_guiado_palavras_50_remocao_cosseno	0,4317	0,3401	0,2806	0,3183
bertopic_guiado_palavras_55_bm25	0,5107	0,4127	0,3462	0,3873
bertopic_guiado_palavras_55_cosseno	0,4206	0,3246	0,2643	0,3035
bertopic_guiado_palavras_55_remocao_bm25	0,5488	0,4509	0,3827	0,4242
bertopic_guiado_palavras_55_remocao_cosseno	0,4516	0,3519	0,2883	0,3291
bertopic_guiado_palavras_60_bm25	0,5527	0,4501	0,3796	0,4228
bertopic_guiado_palavras_60_cosseno	0,4531	0,3543	0,2909	0,3312
bertopic_guiado_palavras_60_remocao_bm25	0,5552	0,4525	0,3819	0,4254
bertopic_guiado_palavras_60_remocao_cosseno	0,4476	0,3670	0,3110	0,3444
bertopic_guiado_palavras_65_bm25	0,5087	0,4157	0,3514	0,3907
bertopic_guiado_palavras_65_cosseno	0,4220	0,3239	0,2628	0,3024
bertopic_guiado_palavras_65_remocao_bm25	0,5480	0,4374	0,3639	0,4095
bertopic_guiado_palavras_65_remocao_cosseno	0,4265	0,3479	0,2938	0,3266
bertopic_guiado_palavras_70_bm25	0,5519	0,4508	0,3810	0,4238
bertopic_guiado_palavras_70_cosseno	0,4112	0,3275	0,2721	0,3069
bertopic_guiado_palavras_70_remocao_bm25	0,5647	0,4405	0,3611	0,4118
bertopic_guiado_palavras_70_remocao_cosseno	0,4435	0,3584	0,3008	0,3357
bertopic_guiado_palavras_75_bm25	0,5195	0,4226	0,3561	0,3970
bertopic_guiado_palavras_75_cosseno	0,4162	0,3249	0,2664	0,3035
bertopic_guiado_palavras_75_remocao_bm25	0,5498	0,4375	0,3633	0,4096
bertopic_guiado_palavras_75_remocao_cosseno	0,3953	0,3258	0,2771	0,3059
bertopic_guiado_palavras_80_bm25	0,5171	0,4226	0,3573	0,3971
bertopic_guiado_palavras_80_cosseno	0,4014	0,3118	0,2548	0,2912
bertopic_guiado_palavras_80_remocao_bm25	0,5692	0,4527	0,3758	0,4240
bertopic_guiado_palavras_80_remocao_cosseno	0,4076	0,3309	0,2785	0,3102
bertopic_guiado_palavras_85_bm25	0,5471	0,4488	0,3804	0,4220
bertopic_guiado_palavras_85_cosseno	0,3975	0,3165	0,2629	0,2962
bertopic_guiado_palavras_85_remocao_bm25	0,5712	0,4547	0,3776	0,4261
bertopic_guiado_palavras_85_remocao_cosseno	0,3995	0,3334	0,2861	0,3140
bertopic_guiado_palavras_90_bm25	0,5543	0,4470	0,3745	0,4191
bertopic_guiado_palavras_90_cosseno	0,3863	0,3107	0,2598	0,2913
bertopic_guiado_palavras_90_remocao_bm25	0,5508	0,4437	0,3715	0,4163
bertopic_guiado_palavras_90_remocao_cosseno	0,3911	0,3293	0,2843	0,3106
bertopic_guiado_palavras_95_bm25	0,5376	0,4395	0,3717	0,4131
bertopic_guiado_palavras_95_cosseno	0,3926	0,3150	0,2631	0,2954
bertopic_guiado_palavras_95_remocao_bm25	0,5548	0,4406	0,3654	0,4126
bertopic_guiado_palavras_95_remocao_cosseno	0,4157	0,3441	0,2935	0,3235

bertopic_guiado_palavras_100_bm25	0,5385	0,4402	0,3722	0,4139
bertopic_guiado_palavras_100_cosseno	0,4030	0,3169	0,2612	0,2964
bertopic_guiado_palavras_100_remocao_bm25	0,5308	0,4264	0,3563	0,3995
bertopic_guiado_palavras_100_remocao_cosseno	0,4055	0,3265	0,2733	0,3058
bertopic_guiado_palavras_105_bm25	0,5668	0,4549	0,3798	0,4263
bertopic_guiado_palavras_105_cosseno	0,3959	0,3151	0,2617	0,2950
bertopic_guiado_palavras_105_remocao_bm25	0,5517	0,4406	0,3667	0,4129
bertopic_guiado_palavras_105_remocao_cosseno	0,4163	0,3386	0,2853	0,3173
bertopic_guiado_palavras_110_bm25	0,5524	0,4449	0,3725	0,4173
bertopic_guiado_palavras_110_cosseno	0,3971	0,3197	0,2675	0,2998
bertopic_guiado_palavras_110_remocao_bm25	0,5429	0,4388	0,3682	0,4117
bertopic_guiado_palavras_110_remocao_cosseno	0,4396	0,3515	0,2929	0,3281
bertopic_guiado_palavras_115_bm25	0,5544	0,4443	0,3707	0,4164
bertopic_guiado_palavras_115_cosseno	0,3963	0,3172	0,2644	0,2972
bertopic_guiado_palavras_115_remocao_bm25	0,5483	0,4389	0,3660	0,4113
bertopic_guiado_palavras_115_remocao_cosseno	0,4097	0,3403	0,2910	0,3201
bertopic_guiado_palavras_120_bm25	0,5568	0,4411	0,3653	0,4128
bertopic_guiado_palavras_120_cosseno	0,3884	0,3112	0,2596	0,2917
bertopic_guiado_palavras_120_remocao_bm25	0,5424	0,4344	0,3623	0,4071
bertopic_guiado_palavras_120_remocao_cosseno	0,4025	0,3340	0,2853	0,3142
lexrank_sentencas_10_bm25	0,5893	0,4561	0,3720	0,4260
lexrank_sentencas_10_cosseno	0,4158	0,3031	0,2385	0,2823
lexrank_sentencas_10_remocao_bm25	0,5624	0,4335	0,3526	0,4049
lexrank_sentencas_10_remocao_cosseno	0,3513	0,2677	0,2163	0,2497
lexrank_sentencas_15_bm25	0,6434	0,4916	0,3978	0,4583
lexrank_sentencas_15_cosseno	0,4162	0,3038	0,2392	0,2830
lexrank_sentencas_15_remocao_bm25	0,6218	0,4747	0,3839	0,4423
lexrank_sentencas_15_remocao_cosseno	0,3523	0,2698	0,2185	0,2517
lexrank_sentencas_20_bm25	0,6555	0,5200	0,4309	0,4865
lexrank_sentencas_20_cosseno	0,4171	0,3035	0,2386	0,2827
lexrank_sentencas_20_remocao_bm25	0,6242	0,4867	0,3988	0,4544
lexrank_sentencas_20_remocao_cosseno	0,3543	0,2700	0,2181	0,2519
lexrank_sentencas_25_bm25	0,6615	0,5240	0,4339	0,4905
lexrank_sentencas_25_cosseno	0,4168	0,3034	0,2385	0,2825
lexrank_sentencas_25_remocao_bm25	0,6317	0,5012	0,4154	0,4690
lexrank_sentencas_25_remocao_cosseno	0,3540	0,2698	0,2179	0,2516
lexrank_sentencas_30_bm25	0,7037	0,5473	0,4478	0,5112
lexrank_sentencas_30_cosseno	0,4176	0,3036	0,2385	0,2827
lexrank_sentencas_30_remocao_bm25	0,6684	0,5301	0,4393	0,4962
lexrank_sentencas_30_remocao_cosseno	0,3538	0,2697	0,2179	0,2516
lexrank_sentencas_35_bm25	0,7098	0,5554	0,4561	0,5186

lexrank_sentencas_35_cosseno	0,4176	0,3037	0,2387	0,2829
lexrank_sentencas_35_remocao_bm25	0,6900	0,5492	0,4562	0,5139
lexrank_sentencas_35_remocao_cosseno	0,3541	0,2698	0,2180	0,2517
lexrank_sentencas_40_bm25	0,7157	0,5645	0,4660	0,5283
lexrank_sentencas_40_cosseno	0,4176	0,3037	0,2387	0,2829
lexrank_sentencas_40_remocao_bm25	0,6749	0,5435	0,4549	0,5096
lexrank_sentencas_40_remocao_cosseno	0,3541	0,2698	0,2180	0,2517
lexrank_sentencas_45_bm25	0,7232	0,5797	0,4837	0,5437
lexrank_sentencas_45_cosseno	0,4176	0,3037	0,2387	0,2829
lexrank_sentencas_45_remocao_bm25	0,7118	0,5648	0,4682	0,5286
lexrank_sentencas_45_remocao_cosseno	0,3541	0,2698	0,2180	0,2517
lexrank_sentencas_50_bm25	0,7303	0,5852	0,4883	0,5484
lexrank_sentencas_50_cosseno	0,4176	0,3037	0,2387	0,2829
lexrank_sentencas_50_remocao_bm25	0,7185	0,5949	0,5076	0,5603
lexrank_sentencas_50_remocao_cosseno	0,3541	0,2698	0,2180	0,2517
lexrank_sentencas_55_bm25	0,7281	0,5886	0,4939	0,5525
lexrank_sentencas_55_cosseno	0,4176	0,3037	0,2387	0,2829
lexrank_sentencas_55_remocao_bm25	0,7212	0,6002	0,5139	0,5659
lexrank_sentencas_55_remocao_cosseno	0,3541	0,2698	0,2180	0,2517
lexrank_sentencas_60_bm25	0,7257	0,5996	0,5108	0,5644
lexrank_sentencas_60_cosseno	0,4176	0,3037	0,2387	0,2829
lexrank_sentencas_60_remocao_bm25	0,7215	0,6015	0,5157	0,5672
lexrank_sentencas_60_remocao_cosseno	0,3541	0,2698	0,2180	0,2517
lexrank_sentencas_65_bm25	0,7274	0,6011	0,5122	0,5659
lexrank_sentencas_65_cosseno	0,4176	0,3037	0,2387	0,2829
lexrank_sentencas_65_remocao_bm25	0,7215	0,6009	0,5149	0,5667
lexrank_sentencas_65_remocao_cosseno	0,3541	0,2698	0,2180	0,2517
lexrank_sentencas_70_bm25	0,7271	0,6022	0,5139	0,5672
lexrank_sentencas_70_cosseno	0,4176	0,3037	0,2387	0,2829
lexrank_sentencas_70_remocao_bm25	0,7201	0,6026	0,5180	0,5688
lexrank_sentencas_70_remocao_cosseno	0,3541	0,2698	0,2180	0,2517
lexrank_sentencas_75_bm25	0,7281	0,6017	0,5127	0,5665
lexrank_sentencas_75_cosseno	0,4176	0,3037	0,2387	0,2829
lexrank_sentencas_75_remocao_bm25	0,7187	0,5978	0,5116	0,5635
lexrank_sentencas_75_remocao_cosseno	0,3541	0,2698	0,2180	0,2517
lexrank_sentencas_80_bm25	0,7299	0,6032	0,5140	0,5679
lexrank_sentencas_80_cosseno	0,4176	0,3037	0,2387	0,2829
lexrank_sentencas_80_remocao_bm25	0,7210	0,5994	0,5129	0,5650
lexrank_sentencas_80_remocao_cosseno	0,3541	0,2698	0,2180	0,2517
lexrank_sentencas_85_bm25	0,7329	0,6053	0,5155	0,5697
lexrank_sentencas_85_cosseno	0,4176	0,3037	0,2387	0,2829

lexrank_sentencas_85_remocao_bm25	0,7224	0,6005	0,5138	0,5660
lexrank_sentencas_85_remocao_cosseno	0,3541	0,2698	0,2180	0,2517
lexrank_sentencas_90_bm25	0,7357	0,6071	0,5168	0,5713
lexrank_sentencas_90_cosseno	0,4176	0,3037	0,2387	0,2829
lexrank_sentencas_90_remocao_bm25	0,7239	0,6023	0,5156	0,5677
lexrank_sentencas_90_remocao_cosseno	0,3541	0,2698	0,2180	0,2517
lexrank_sentencas_95_bm25	0,7373	0,6089	0,5186	0,5731
lexrank_sentencas_95_cosseno	0,4176	0,3037	0,2387	0,2829
lexrank_sentencas_95_remocao_bm25	0,7265	0,6047	0,5179	0,5700
lexrank_sentencas_95_remocao_cosseno	0,3541	0,2698	0,2180	0,2517
lexrank_sentencas_100_bm25	0,7365	0,6092	0,5194	0,5735
lexrank_sentencas_100_cosseno	0,4176	0,3037	0,2387	0,2829
lexrank_sentencas_100_remocao_bm25	0,7272	0,6056	0,5188	0,5709
lexrank_sentencas_100_remocao_cosseno	0,3541	0,2698	0,2180	0,2517
lexrank_sentencas_105_bm25	0,7373	0,6089	0,5186	0,5731
lexrank_sentencas_105_cosseno	0,4176	0,3037	0,2387	0,2829
lexrank_sentencas_105_remocao_bm25	0,7264	0,6062	0,5202	0,5718
lexrank_sentencas_105_remocao_cosseno	0,3541	0,2698	0,2180	0,2517
lexrank_sentencas_110_bm25	0,7355	0,6089	0,5195	0,5734
lexrank_sentencas_110_cosseno	0,4176	0,3037	0,2387	0,2829
lexrank_sentencas_110_remocao_bm25	0,7266	0,6069	0,5210	0,5724
lexrank_sentencas_110_remocao_cosseno	0,3541	0,2698	0,2180	0,2517
lexrank_sentencas_115_bm25	0,7343	0,6093	0,5207	0,5740
lexrank_sentencas_115_cosseno	0,4176	0,3037	0,2387	0,2829
lexrank_sentencas_115_remocao_bm25	0,7264	0,6070	0,5213	0,5726
lexrank_sentencas_115_remocao_cosseno	0,3541	0,2698	0,2180	0,2517
lexrank_sentencas_120_bm25	0,7350	0,6097	0,5209	0,5743
lexrank_sentencas_120_cosseno	0,4176	0,3037	0,2387	0,2829
lexrank_sentencas_120_remocao_bm25	0,7271	0,6070	0,5209	0,5725
lexrank_sentencas_120_remocao_cosseno	0,3541	0,2698	0,2180	0,2517
lexrank_guiado_sentencas_10_bm25	0,7359	0,6111	0,5224	0,5759
lexrank_guiado_sentencas_10_cosseno	0,3798	0,2804	0,2223	0,2611
lexrank_guiado_sentencas_10_remocao_bm25	0,7542	0,6295	0,5402	0,5936
lexrank_guiado_sentencas_10_remocao_cosseno	0,4704	0,3858	0,3269	0,3626
lexrank_guiado_sentencas_15_bm25	0,7442	0,6212	0,5331	0,5860
lexrank_guiado_sentencas_15_cosseno	0,3798	0,2804	0,2223	0,2611
lexrank_guiado_sentencas_15_remocao_bm25	0,7575	0,6268	0,5346	0,5902
lexrank_guiado_sentencas_15_remocao_cosseno	0,4704	0,3858	0,3269	0,3626
lexrank_guiado_sentencas_20_bm25	0,7471	0,6262	0,5389	0,5910
lexrank_guiado_sentencas_20_cosseno	0,3798	0,2804	0,2223	0,2611
lexrank_guiado_sentencas_20_remocao_bm25	0,7571	0,6227	0,5289	0,5858

lexrank_guiado_sentencas_20_remocao_cosseno	0,4704	0,3858	0,3269	0,3626
lexrank_guiado_sentencas_25_bm25	0,7542	0,6297	0,5404	0,5939
lexrank_guiado_sentencas_25_cosseno	0,3798	0,2804	0,2223	0,2611
lexrank_guiado_sentencas_25_remocao_bm25	0,7463	0,6248	0,5373	0,5898
lexrank_guiado_sentencas_25_remocao_cosseno	0,4704	0,3858	0,3269	0,3626
lexrank_guiado_sentencas_30_bm25	0,7545	0,6361	0,5498	0,6009
lexrank_guiado_sentencas_30_cosseno	0,3798	0,2804	0,2223	0,2611
lexrank_guiado_sentencas_30_remocao_bm25	0,7557	0,6305	0,5408	0,5946
lexrank_guiado_sentencas_30_remocao_cosseno	0,4704	0,3858	0,3269	0,3626
lexrank_guiado_sentencas_35_bm25	0,7520	0,6290	0,5406	0,5935
lexrank_guiado_sentencas_35_cosseno	0,3798	0,2804	0,2223	0,2611
lexrank_guiado_sentencas_35_remocao_bm25	0,7552	0,6306	0,5413	0,5948
lexrank_guiado_sentencas_35_remocao_cosseno	0,4704	0,3858	0,3269	0,3626
lexrank_guiado_sentencas_40_bm25	0,7497	0,6284	0,5409	0,5933
lexrank_guiado_sentencas_40_cosseno	0,3798	0,2804	0,2223	0,2611
lexrank_guiado_sentencas_40_remocao_bm25	0,7454	0,6257	0,5391	0,5909
lexrank_guiado_sentencas_40_remocao_cosseno	0,4704	0,3858	0,3269	0,3626
lexrank_guiado_sentencas_45_bm25	0,7540	0,6296	0,5404	0,5939
lexrank_guiado_sentencas_45_cosseno	0,3798	0,2804	0,2223	0,2611
lexrank_guiado_sentencas_45_remocao_bm25	0,7413	0,6233	0,5377	0,5889
lexrank_guiado_sentencas_45_remocao_cosseno	0,4704	0,3858	0,3269	0,3626
lexrank_guiado_sentencas_50_bm25	0,7468	0,6258	0,5385	0,5908
lexrank_guiado_sentencas_50_cosseno	0,3798	0,2804	0,2223	0,2611
lexrank_guiado_sentencas_50_remocao_bm25	0,7419	0,6237	0,5380	0,5892
lexrank_guiado_sentencas_50_remocao_cosseno	0,4704	0,3858	0,3269	0,3626
lexrank_guiado_sentencas_55_bm25	0,7456	0,6211	0,5322	0,5855
lexrank_guiado_sentencas_55_cosseno	0,3798	0,2804	0,2223	0,2611
lexrank_guiado_sentencas_55_remocao_bm25	0,7409	0,6227	0,5370	0,5882
lexrank_guiado_sentencas_55_remocao_cosseno	0,4704	0,3858	0,3269	0,3626
lexrank_guiado_sentencas_60_bm25	0,7426	0,6208	0,5334	0,5856
lexrank_guiado_sentencas_60_cosseno	0,3798	0,2804	0,2223	0,2611
lexrank_guiado_sentencas_60_remocao_bm25	0,7401	0,6222	0,5368	0,5878
lexrank_guiado_sentencas_60_remocao_cosseno	0,4704	0,3858	0,3269	0,3626
lexrank_guiado_sentencas_65_bm25	0,7375	0,6146	0,5268	0,5795
lexrank_guiado_sentencas_65_cosseno	0,3484	0,2548	0,2008	0,2372
lexrank_guiado_sentencas_65_remocao_bm25	0,7372	0,6199	0,5349	0,5856
lexrank_guiado_sentencas_65_remocao_cosseno	0,4398	0,3601	0,3048	0,3385
lexrank_guiado_sentencas_70_bm25	0,7360	0,6147	0,5276	0,5797
lexrank_guiado_sentencas_70_cosseno	0,3484	0,2548	0,2008	0,2372
lexrank_guiado_sentencas_70_remocao_bm25	0,7368	0,6197	0,5348	0,5855
lexrank_guiado_sentencas_70_remocao_cosseno	0,4398	0,3601	0,3048	0,3385

lexrank_guiado_sentencas_75_bm25	0,7358	0,6143	0,5272	0,5793
lexrank_guiado_sentencas_75_cosseno	0,3484	0,2548	0,2008	0,2372
lexrank_guiado_sentencas_75_remocao_bm25	0,7344	0,6177	0,5330	0,5836
lexrank_guiado_sentencas_75_remocao_cosseno	0,4398	0,3601	0,3048	0,3385
lexrank_guiado_sentencas_80_bm25	0,7365	0,6140	0,5265	0,5789
lexrank_guiado_sentencas_80_cosseno	0,3484	0,2548	0,2008	0,2372
lexrank_guiado_sentencas_80_remocao_bm25	0,7329	0,6167	0,5323	0,5827
lexrank_guiado_sentencas_80_remocao_cosseno	0,4398	0,3601	0,3048	0,3385
lexrank_guiado_sentencas_85_bm25	0,7325	0,6110	0,5240	0,5761
lexrank_guiado_sentencas_85_cosseno	0,3484	0,2548	0,2008	0,2372
lexrank_guiado_sentencas_85_remocao_bm25	0,7329	0,6171	0,5328	0,5831
lexrank_guiado_sentencas_85_remocao_cosseno	0,4398	0,3601	0,3048	0,3385
lexrank_guiado_sentencas_90_bm25	0,7330	0,6112	0,5241	0,5763
lexrank_guiado_sentencas_90_cosseno	0,3484	0,2548	0,2008	0,2372
lexrank_guiado_sentencas_90_remocao_bm25	0,7323	0,6159	0,5314	0,5819
lexrank_guiado_sentencas_90_remocao_cosseno	0,4398	0,3601	0,3048	0,3385
lexrank_guiado_sentencas_95_bm25	0,7316	0,6101	0,5232	0,5752
lexrank_guiado_sentencas_95_cosseno	0,3484	0,2548	0,2008	0,2372
lexrank_guiado_sentencas_95_remocao_bm25	0,7311	0,6150	0,5306	0,5810
lexrank_guiado_sentencas_95_remocao_cosseno	0,4398	0,3601	0,3048	0,3385
lexrank_guiado_sentencas_100_bm25	0,7309	0,6093	0,5223	0,5744
lexrank_guiado_sentencas_100_cosseno	0,3484	0,2548	0,2008	0,2372
lexrank_guiado_sentencas_100_remocao_bm25	0,7293	0,6136	0,5296	0,5798
lexrank_guiado_sentencas_100_remocao_cosseno	0,4398	0,3601	0,3048	0,3385
lexrank_guiado_sentencas_105_bm25	0,7298	0,6085	0,5218	0,5737
lexrank_guiado_sentencas_105_cosseno	0,3484	0,2548	0,2008	0,2372
lexrank_guiado_sentencas_105_remocao_bm25	0,7285	0,6135	0,5298	0,5798
lexrank_guiado_sentencas_105_remocao_cosseno	0,4398	0,3601	0,3048	0,3385
lexrank_guiado_sentencas_110_bm25	0,7284	0,6077	0,5213	0,5730
lexrank_guiado_sentencas_110_cosseno	0,3484	0,2548	0,2008	0,2372
lexrank_guiado_sentencas_110_remocao_bm25	0,7286	0,6126	0,5284	0,5787
lexrank_guiado_sentencas_110_remocao_cosseno	0,4398	0,3601	0,3048	0,3385
lexrank_guiado_sentencas_115_bm25	0,7275	0,6072	0,5211	0,5726
lexrank_guiado_sentencas_115_cosseno	0,3484	0,2548	0,2008	0,2372
lexrank_guiado_sentencas_115_remocao_bm25	0,7283	0,6116	0,5271	0,5777
lexrank_guiado_sentencas_115_remocao_cosseno	0,4398	0,3601	0,3048	0,3385
lexrank_guiado_sentencas_120_bm25	0,7276	0,6069	0,5205	0,5722
lexrank_guiado_sentencas_120_cosseno	0,3484	0,2548	0,2008	0,2372
lexrank_guiado_sentencas_120_remocao_bm25	0,7266	0,6111	0,5272	0,5774
lexrank_guiado_sentencas_120_remocao_cosseno	0,4398	0,3601	0,3048	0,3385