CEFET/RJ
DIRETORIA DE PESQUISA E PÓS-GRADUAÇÃO

ANÁLISE DE PROCESSOS JUDICIAIS VIA PROCESSAMENTO DE LINGUAGEM NATURAL

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 2024

Análise de Processos Judiciais via
Processamento de Linguagem Natural

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. nome orientador, D.Sc. (orientador)

---

nome coorientador, D.Sc. (coorientador)

Membro 1, D.Sc.

Membro 2, D.Sc.

Rio de Janeiro,

Fevereiro de 2024

FICHA CATALOGRÁFICA A SER SOLICITADA NA BIBLIOTECA DO CEFET/RJ APÓS A REVISÃO FINAL DO TEXTO.

MAIS INFORMAÇÕES EM:
https://eic.cefet-rj.br/ppcic/index.php/ficha-catalografica-nada-consta/

DEDICATÓRIA

DEDICATÓRIA texto

AGRADECIMENTOS

Agradeço aos ...

Por fim, agradeço ao órgão de fomento, responsável por fomentar esta pesquisa.

RESUMO

Análise de Processos Judiciais via

Processamento de Linguagem Natural

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 vários mecanismos para os cidadãos acionarem o poder judiciário, incluindo o recurso, que é o direito de impugnar uma decisão judicial para reformar, invalidar, esclarecer ou integrar o julgado. O recurso especial, conforme definido no artigo 105 da Constituição, visa uniformizar o entendimento jurídico do direito brasileiro.Esse tipo de recurso é julgado pelo Superior Tribunal de Justiça (STJ) nos casos em que a decisão recorrida contrarie leis federais. O tratamento dos recursos especiais é uma tarefa diária na esfera do poder judiciário, o qual apresenta regularmente um quantitativo expressivo de demandas em seus tribunais. A adoção de ferramentas de inteligência artificial em tarefas repetitivas pode dar celeridade a tramitação de processos jurídicos e otimizar o uso de recursos humanos. Esta pesquisa tem como objetivo criar uma metodologia eficiente para classificar um recurso especial em um tema, para tal elaboramos uma metodologia baseada em sumarização extrativa. Para geração do resumo de um recurso especial, executamos comparações entre uma abordagem baseada no conceito da modelagem de tópicos e outra que cria sentenças por meio de algoritmo baseado em grafo. Buscamos também comparar duas abordagens distintas para avaliar a similaridade entre textos, em uma utilizamos a representação vetorial do texto e em outra utilizamos o próprio texto como dado de entrada para uma função de pontuação muito usada em sistemas de recuperação da informação. Nos experimentos realizados obtivemos resultados muito satisfatórios comparados com os dados de referência, provenientes da solução para classificação de recursos especiais adotada pelo Tribunal Regional Federal da 2ª Região (TRF2).

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 2024

ABSTRACT

Análise de Processos Judiciais via

Processamento de Linguagem Natural

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 for several mechanisms for citizens to activate the judiciary, including appeal, which is the right to challenge a judicial decision to reform, invalidate, clarify or integrate the judgment. The special appeal, as defined in article 105 of the Constitution, aims to standardize the legal understanding of Brazilian law. This type of appeal is judged by the Superior Tribunal de Justiça in cases where the appealed decision contravenes federal laws. The treatment of special resources is a daily task in the sphere of the judiciary, which regularly presents a significant number of demands in its courts. The adoption of artificial intelligence tools in repetitive tasks can speed up the processing of legal processes and optimize the use of human resources. This research aims to create an efficient methodology to classify a special resource in a topic. To this end, we developed a methodology based on extractive summarization. To generate the summary of a special resource, we performed comparisons between an approach based on the topic modeling concept and another that creates sentences using a graph-based algorithm. We also seek to compare two different approaches to evaluate the similarity between texts, in one we use the vector representation of the text and in the other we use the text itself as input data for a scoring function widely used in information retrieval systems. In the experiments carried out, we obtained very satisfactory results compared with reference data, coming from the solution for classifying special resources adopted by Tribunal Regional Federal da 2ª Região.

Key-words:

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

Rio de Janeiro,

Fevereiro de 2024

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

BERT	Bidirectional Encoder Representations From Transformers	32
BM25	Best Match 25	13, 14, 15, 36, 39, 42, 43, 44, 46, 52
CNJ	Conselho Nacional De Justiça	3
CNPJ	Cadastro Nacional De Pessoa Juridica	31
CPC	Código De Processo Civil	2, 5, 31
CPF	Cadastro De Pessoa Física	30
CSV	Comma-separated Values	67
HDBSCAN	Hierarchical Density-Based Spatial Clustering Of Applications With Noise	32, 33
HTML	Hypertext Markup Language	31
IDF	Inverse Term Frequency	12, 13, 14, 15
LDA	Latent Dirichlet Allocation	18
MAP	Mean Average Precision	38
NDCG	Normalized Discounted Cumulative Gain	38, 39, 46
NIST	National Institute Of Standards And Technology	13
NLTK	Natural Language Toolkit	30, 31, 34, 40
NMF	Non-Negative Matrix Factorization	18
PRISMA	Preferred Reporting Items For Systematic Reviews And Meta-Analyses	24, 25
SBERT	Sentence-BERT	31, 32, 34, 39, 52
SQL	Structured Query Language	9, 10
STF	Supremo Tribunal Federal	25
STJ	Superior Tribunal De Justiça	1, 2, 3, 5, 6, 23, 28, 51, 67
SUS	Sistema Único De Saúde	2
TF	Term Frequency	11, 12, 13, 14, 15
TREC	Text REtrieval Conference	13
TRF2	Tribunal Regional Federal Da 2ª Região	3, 6, 7, 27
UML	Unified Modeling Language	42

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 surgimento de um processo a partir de sua petição inicial e o seu encerramento, com sua baixa definitiva, inúmeros documentos são incorporados ao mesmo.

A incorporação de cada documento em um processo jurídico, e a ação de algum dos inúmeros agentes que interagem com o processo, sinaliza uma mudança no estado atual, provocando assim sua tramitação a partir da origem em direção ao seu encerramento. 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 é definido como sendo 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].

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].

Possuindo dimensões continentais e uma população estimada em mais de 207 milhões segundo o censo de 2022 [Cen], 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. Uma vez que a amostragem tenha sido enviada ao STJ os demais recursos especiais que tratem da mesma questão ficam paralisados nos tribunais de instâncias inferiores até que o tema da controvérsia seja julgado e um entendimento seja estabelecido [CNJ].

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.

[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.

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

I.2 Motivação

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.

O Conselho Nacional de Justiça (CNJ) mantém e atualiza periodicamente o painel de estatísticas do Poder Judiciário [Est]. Apesar de não estar disponível dados sobre o tempo médio para execução de tarefas, como por exemplo a análise de recursos especiais, é possível ter certa noção da magnitude de tempo analisando estatísticas que são intrinsecamente relacionadas ao tratamento de recurso especial. 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. É possível verificar no painel de estatísticas o tempo médio decorrido entre o início da ação judicial e a data corrente de processos que estão pendentes de julgamento pelo gabinete da vice-presidência. Os dados coletados no painel estão apresentados na Figura I.1, onde verifica-se que em março de 2023, por exemplo, a média de tempo em tramitação é de 2952 dias.

No Brasil, segundo dados oficiais do Conselho Nacional de Justiça, o Poder Judiciário finalizou o ano de 2020 com 75,4 milhões de processos em tramitação aguardando alguma solução definitiva (Figura I.2) [de Justiça, 2021].

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.

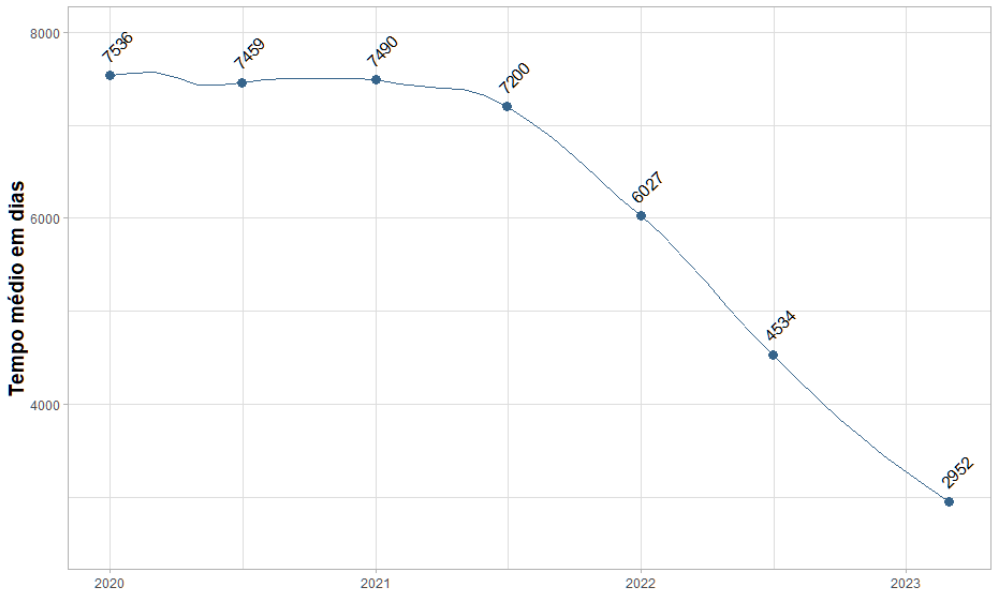


Figura I.1: Série histórica - Tempo médio do pendente na vice-presidência do TRF2

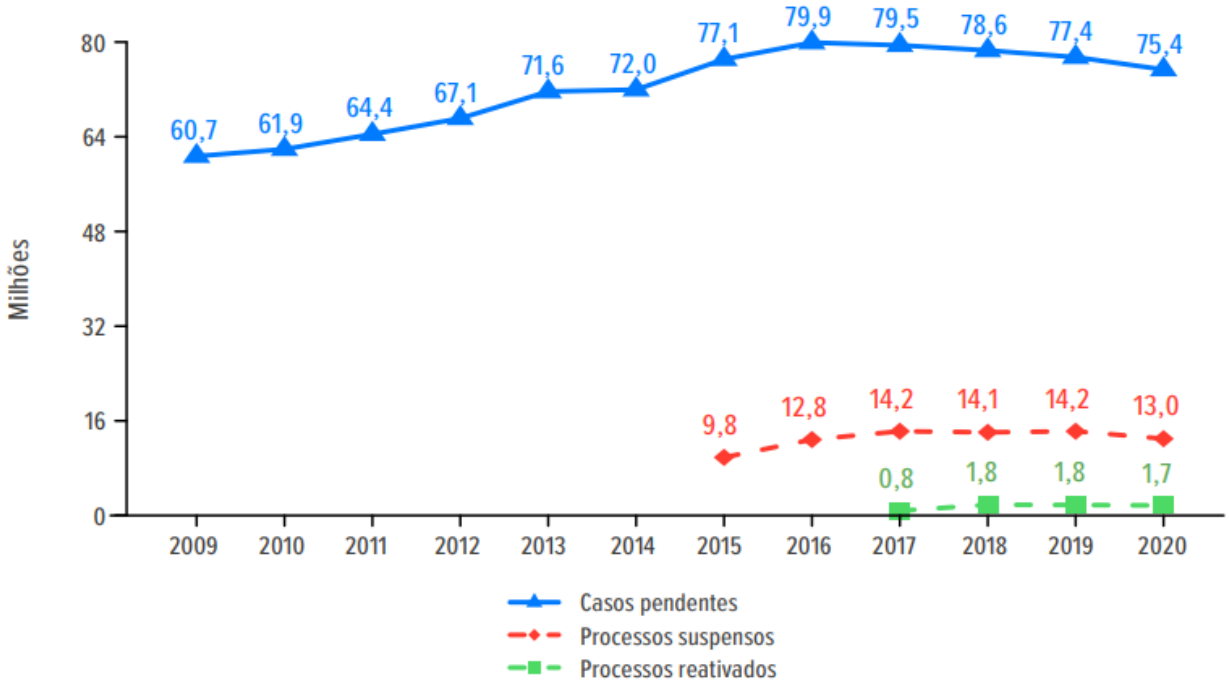


Figura I.2: Conselho Nacional de Justiça - Série histórica de casos pendentes

de busca distribuído, de código aberto, denominado Elasticsearch.

A magnitude do trabalho desempenhado pelos órgãos da Justiça Federal, frente aos recursos disponíveis, foi o principal fator motivador desta pesquisa.

I.3 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.

I.3.1 Recursos Especiais

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.

I.3.2 Temas

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.

I.4 Objetivos

Os objetivos a serem alcançados nesse projeto de dissertação são dois, conforme descrição a seguir.

- Classificar corretamente documentos pertencentes ao domínio jurídico por meio de aprendizado de máquina, tendo como base comparativa a solução atual adotada pelo TRF2 para classificação de recursos especiais;
- Criar um conjunto de dados formatado, que contenha textos de recursos especiais classificados, visando reaproveitamento em futuros trabalhos de processamento de linguagem natural no âmbito jurídico brasileiro no qual é muito escassa a disponibilização de dados.

I.5 Metodologia

Para o desenvolvimento deste trabalho, foram coletados dados de processos não sigilosos disponibilizados publicamente no Eproc do TRF2 [epr, b] e dados de temas repetitivos disponibilizados pelo STJ [STJ]. A partir das informações obtidas, foi estruturado um conjunto de dados com cerca de 8 mil registros e avaliado o desempenho da solução atual na classificação dos documentos, a qual servirá como base de comparação para a proposta deste trabalho.

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.
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.

Um esquema da solução adotada é apresentado na Figura I.3.

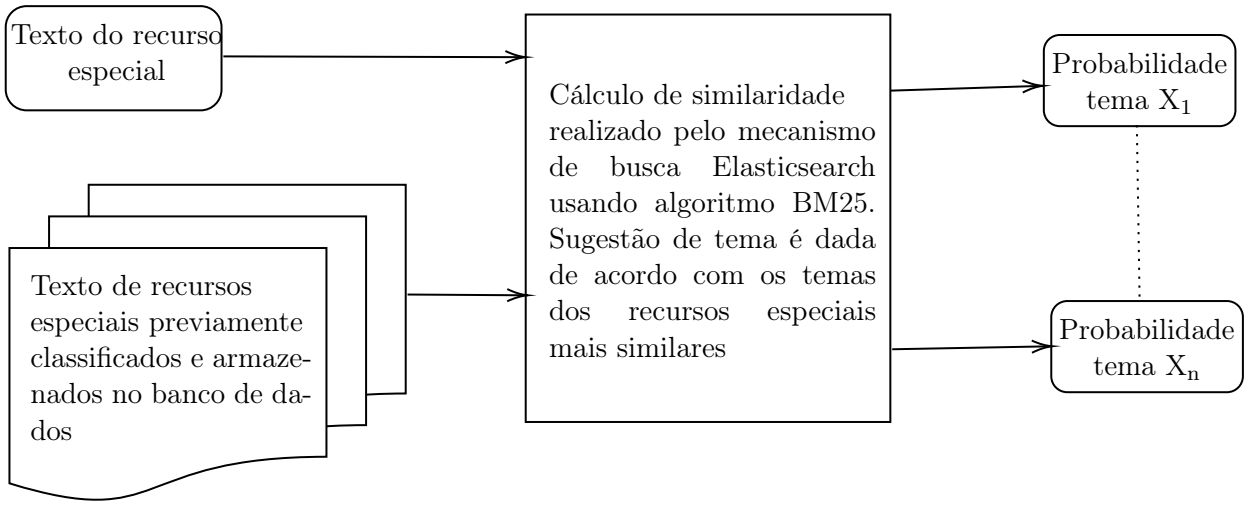


Figura I.3: Esquema atual de sugestões de temas para recursos especiais

Para fins de uma análise inicial da solução adotada pelo TRF2, foi extraída uma amostra para analisar o desempenho da solução atual na classificação de recursos que referem-se a um mesmo tema.

Observando o mapa de calor apresentado na figura I.4, constata-se que mesmo para recursos que são semelhantes , pois de fato pertencem ao mesmo tema jurídico, a solução atual não tem uma classificação uniforme. Como exemplo, no grupo de recursos pertencentes ao tema nº 1005 houve um percentual elevado de recursos para os quais nenhuma das sugestões dadas estava correta, embora para alguns recursos deste grupo foram dadas sugestões corretas como 1ª opção.

Um aspecto relevante na solução atual 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,

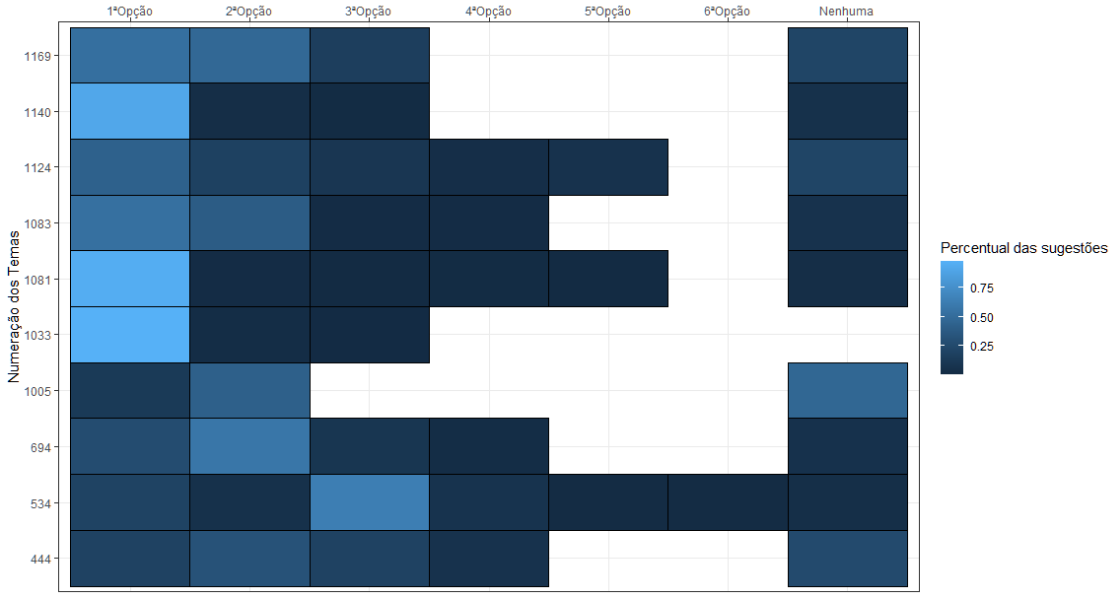


Figura I.4: Mapa de Calor - Desempenho na sugestão de alguns temas

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 poderia 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.

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. "O gato caçou o rato" não possui o mesmo significado de "O rato caçou o gato", muito embora as duas sentenças apresentem exatamente as mesmas palavras, com as mesmas frequências. 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 pesquisa se dará 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 similaridade com o texto alvo $\mathcal{T}_{\dashv}$ minimizando perdas semânticas.

Dentro do domínio do Aprendizado de Máquina, o problema de classificar um recurso especial em um tema pode ser desmembrado em dois subproblemas. O primeiro, refere-se a sumarização do conteúdo do documento destacando seus pontos mais relevantes. O segundo problema refere-se a

avaliar a similaridade entre esse resumo criado e o conteúdo de um tema judicial, procedendo de fato a classificação do documento em um tema.

A sumarização pode ser dividida em dois tipos principais, a **sumarização extrativa** e o **resumo abstrato**. Na sumarização extrativa, um resumo é criado extraindo-se termos do documento original sem qualquer alteração.O resumo abstrato é construído com sentenças novas, mas em alguns casos podem ser aproveitadas sentenças do documento original, sendo ainda assim considerado um documento novo.

O procedimento metodológico consistirá em :

1. Aplicação de técnica para sumarização de um texto de recurso especial;
2. Avaliação da similaridade entre o **texto sumarizado** e os textos dos diversos **temas**;
3. Ranqueamento dos resultados de similaridade obtidos;
4. Medição qualitativa dos resultados gerados.

Pelo fato de estar disponível um conjunto de dados de recursos especiais que já foram rotulados em um tema, será possível mensurar o desempenho do procedimento adotado. Um esquema da solução proposta pode ser visto na Figura I.5.

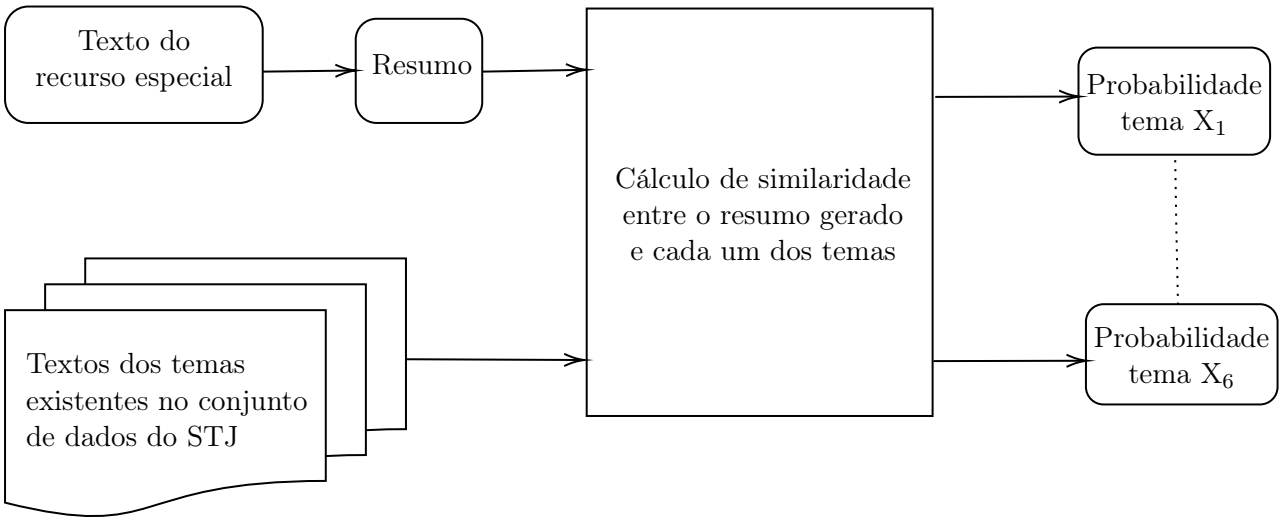


Figura I.5: Esquema proposto

Capítulo II Fundamentação Teórica

Este capítulo apresenta alguns fundamentos necessários para o entendimento dos capítulos seguintes. A primeira seção trata de Sistemas de recuperação de informação, os quais visam atender a necessidade de informação do usuário demandada por consultas textuais [Aggarwal, 2022]. A seção Sumarização de textos apresenta os tipos de sumários e algumas técnicas usadas na geração dos mesmos. A seção Modelos de embeddings apresenta tipos de vetores numéricos multidimensionais usados para representação de palavras com a preservação da semântica.

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 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, 2022] [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,

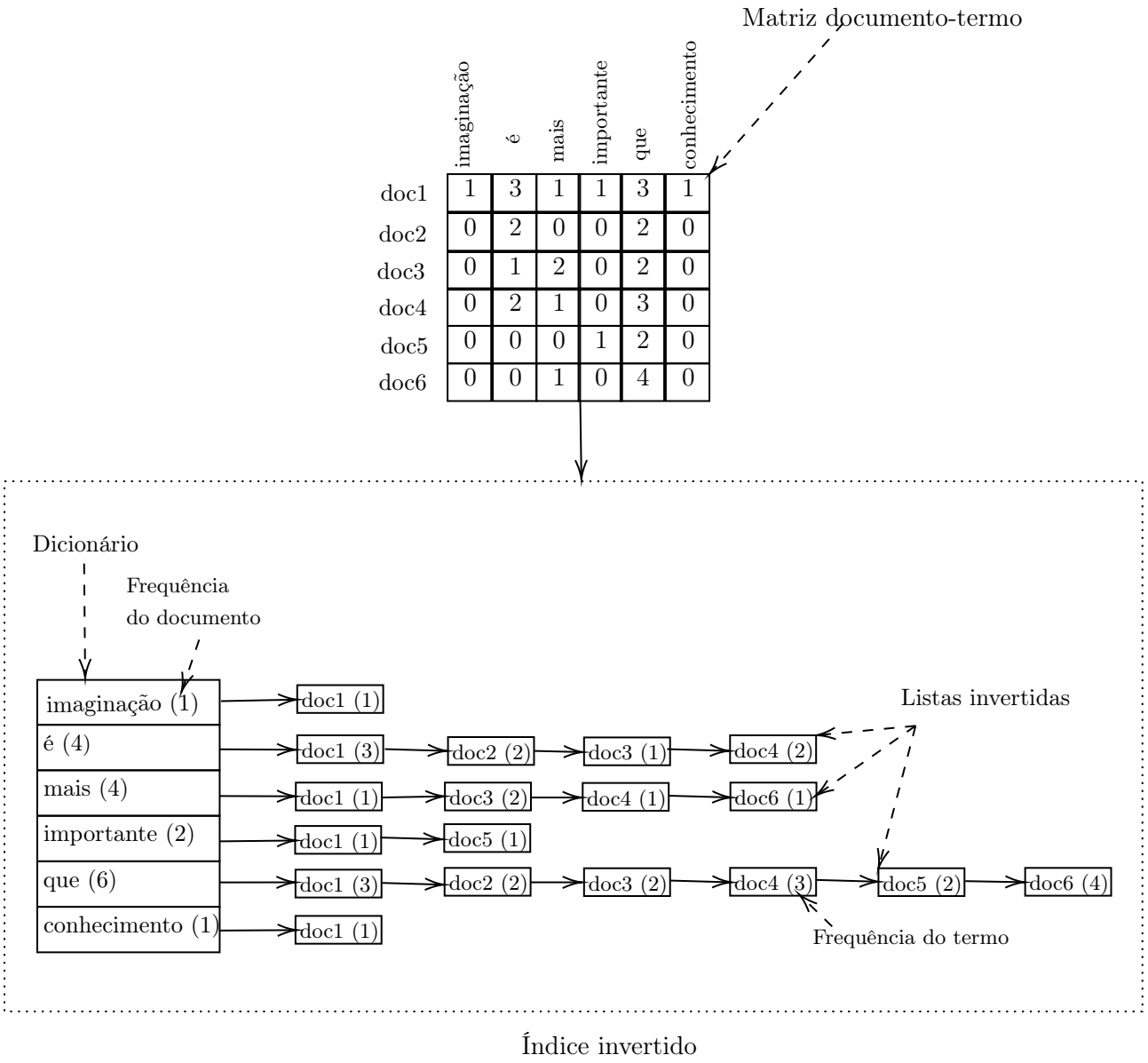


Figura II.1: Relacionamento das estruturas de dados: Matriz documento-termo; Índice invertido; Dicionário e Lista invertida

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 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 TF(d,t), ou seja, a relevância de um documento **d** dado um termo **t**, é

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

onde 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].

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].

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

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]

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

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. Um exemplo tridimensional pode ser visto na Figura II.2, contudo a representação pode ser estendida para t dimensões quando t diferentes termos estão presentes. 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 nos termos e pesos dos termos correspondentes.

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 aleatori-

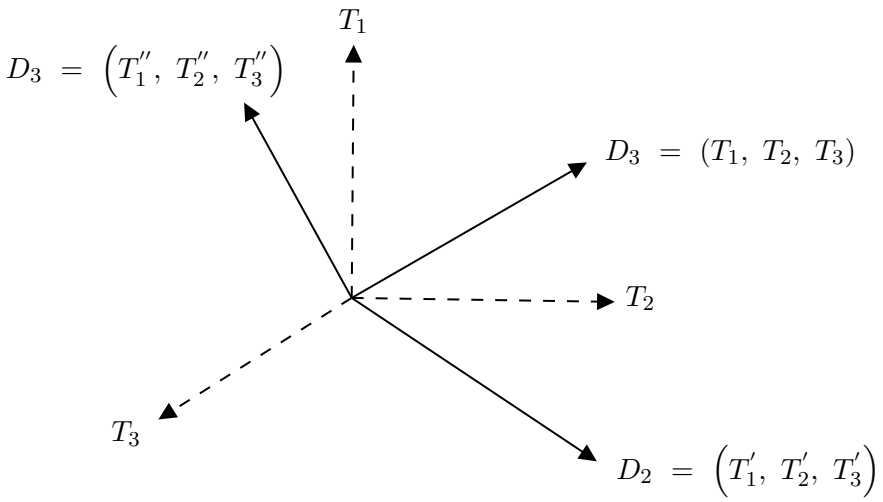


Figura II.2: Representação vetorial de um espaço de documentos

amente [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 = $(x_1 \ ... \ x_n)$ e Y = $(y_1 \ ... \ y_n)$, a similaridade cosseno é definida conforme a Equação II.4 [Aggarwal, 2022], onde x_n está relacionado à representação de X na n-ésima dimensão e y_n à representação de Y na n-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.

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

Onde: $IDF(q_i)$ 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 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 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 particionar um conjunto de objetos de dados em subconjuntos. De modo que neste processo seja maximizada a semelhança entre objetos de 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 partições de dados, onde cada partição representa um grupo e $k \leq n$. O algoritmo começa sua execução com uma partição inicial, de forma iterativa os objetos são movidos de uma partição para outra na tentativa de encontrar um particionamento ideal.

- Métodos hierárquicos - Dado um conjunto de objetos é feita uma decomposição hierárquica.

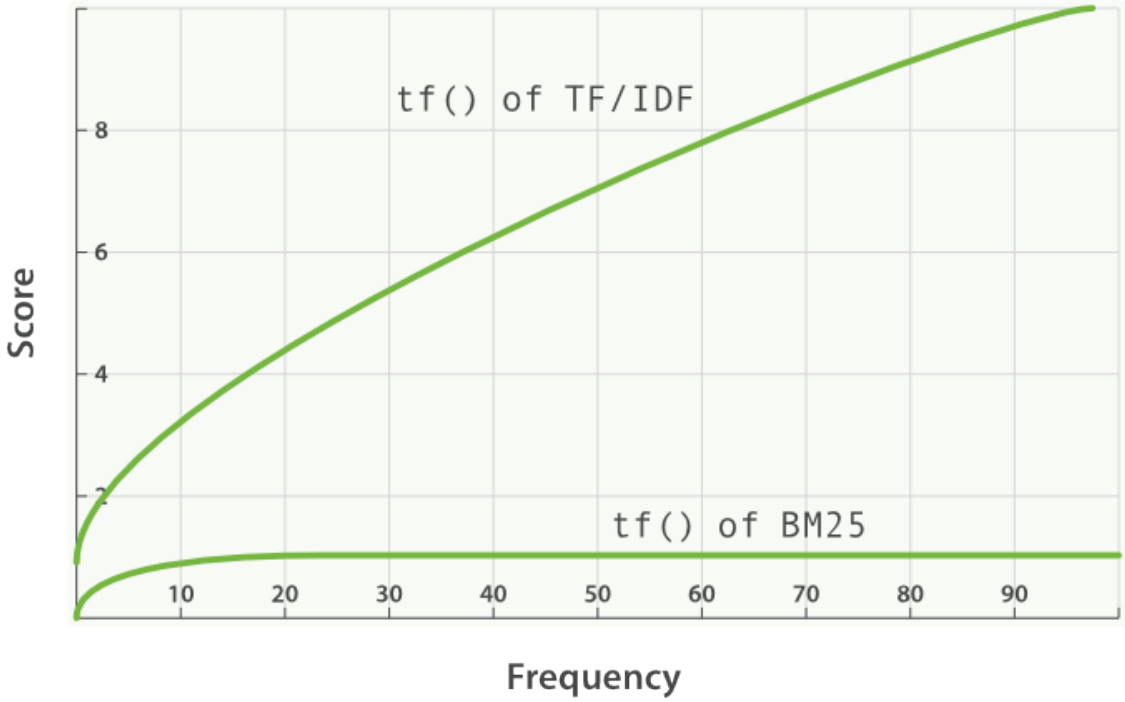


Figura II.3: Efeito da saturação no TF-IDF e no BM25

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 - Formata todo o espaço da dados em um estrutura de grade com um número finito de células

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 (*cluster*) 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 cluster. 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 árvore de alcançabilidade 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, 2019b]:

- 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, mercedez, 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, 2022]. A ideia central por trás da modelagem de tópicos é que cada documento pode ser visto como uma combinação 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.

O processo de agrupamento é executado por meio da aplicação de técnicas de agrupamento (*clustering*), o que permite a identificação de estruturas subjacentes e a alocação de documentos relacionados a tópicos similares. A BERTopic é configurada para usar o algoritmo 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.

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 :

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

pode ser vista na Figura II.4.

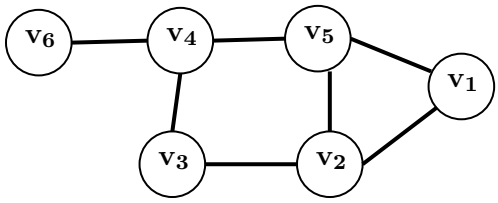


Figura II.4: Representação gráfica do grafo G

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 $nxnA(G) = [a_{ij}]$. Onde a_{ij} representa a relação entre os vértices v_i e v_j. A matriz de adjacência abaixo é uma outra forma de representar o grafo apresentado na Figura II.4

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

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.5). A ideia básica é que páginas com alta reputação estão ligadas a outras páginas que também possuem alta reputação [Aggarwal, 2022].

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

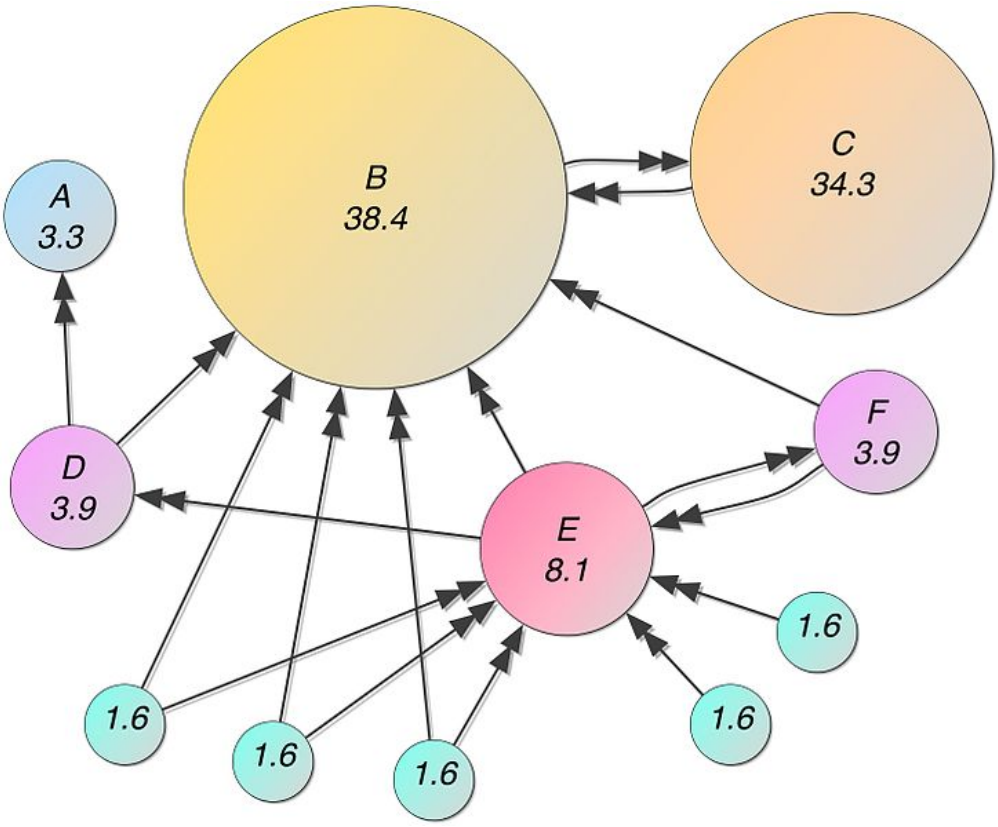


Figura II.5: O Pagerank é base para outros algoritmos baseados em grafo como o LexRank.Fonte: Mehta, Parth and Prasenjit Majumder. From Extractive to Abstractive Summarization: A Journey (2019, p.12) [Mehta and Majumder, 2019a]

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.

$$M_i, j = \frac{S_{ij}}{\sum_k S_{ik}} \tag{II.8}$$

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.

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

Onde:

- PR(i) é a pontuação de PageRank para o nó (sentença)i
- d 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 que atribuam a celeridade demandada pela sociedade. Iniciativas como o programa **Justiça 4.0**[1] buscam implementar uma transformação digital do Judiciário promovendo soluções que automatizam as atividades rotineiras dos tribunais.

Este projeto alinha-se com essa demanda ao analisar a aplicabilidade de técnicas de processamento de linguagem natural na classificação de processos jurídicos. Mais especificamente, buscamos avaliar a implementação do conceito da modelagem de tópicos na classificação de **recursos especiais** enviados ao STJ

Diante do exposto, o objetivo desta revisão sistemática de literatura é investigar técnicas empregadas no processamento de documentos legais. Tendo como foco principal implementações da modelagem de tópicos.

III.1 Seleção de artigos

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. As palavras-chave de buscas utilizadas foram: “legal”, “document”,“classification”,“guided”,“topic”, “modeling” e “survey” . O tipo de documentos buscados ficou restrito a artigos efetivamente publicados em revistas científicas, sendo excluídos trabalhos apresentados em conferências. A área de busca dos artigos foi ciência da computação. Não houve restrição temporal e os termos da busca foram combinados da seguinte forma:

- “legal” AND “document” AND “classification” AND “survey”
- “legal” AND “document” AND “classification”
- “legal” AND “document” AND “topic” AND “modeling”
- “guided” AND “topic” AND “modeling”

A seleção dos artigos foi conduzida com base nas diretrizes do método Preferred Reporting Items

[1] https://www.cnj.jus.br/tecnologia-da-informacao-e-comunicacao/justica-4-0/

for Systematic reviews and Meta-Analyses (PRISMA) [Page et al., 2021]. A Figura III.1 detalha o fluxo da seleção dos artigos.

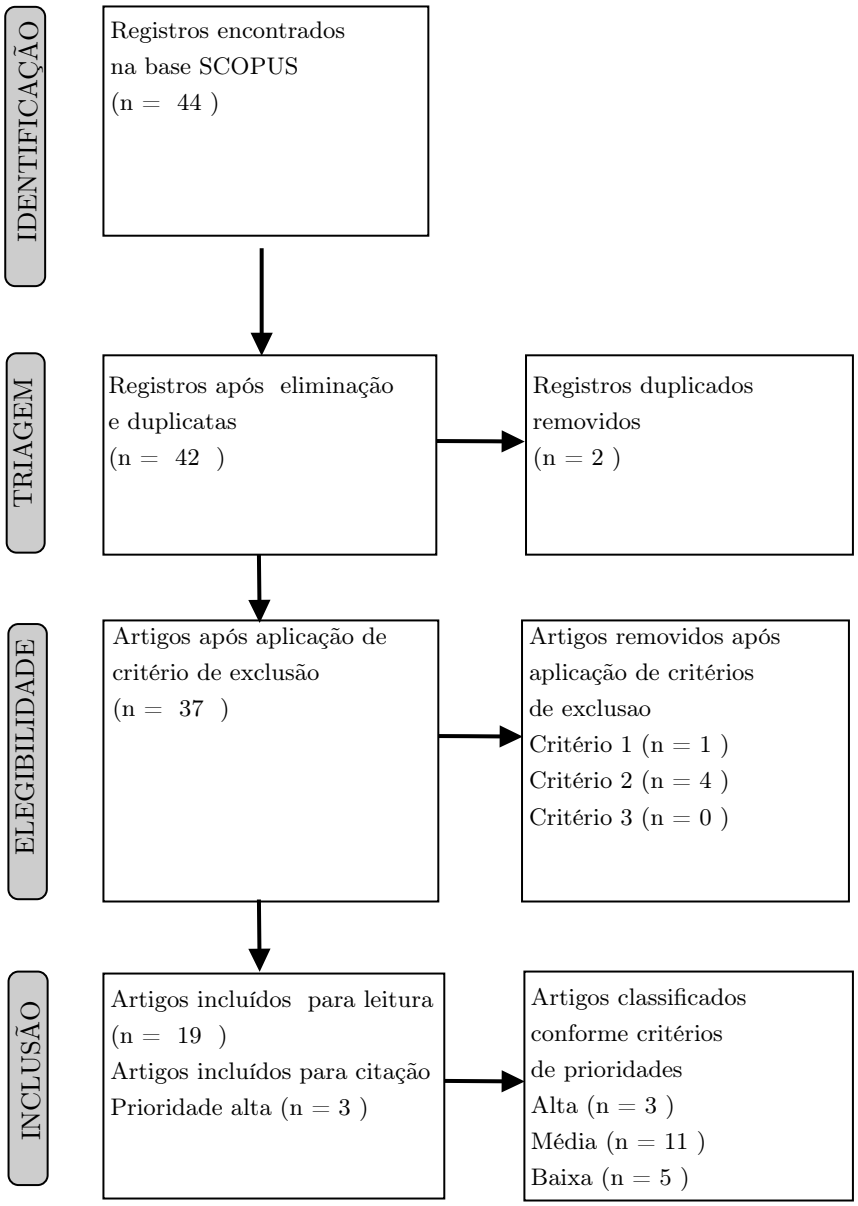


Figura III.1: Fluxo do processo de seleção dos artigos

III.2 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 classificação de documentos
- **Critério 3 -** Trata de modelagem de tópicos
- **Critério 4 -** Trata de sumarização de documentos

III.3 Critérios de exclusão

Os critérios de exclusão possuem a função de excluir resultados que não satisfazem ao objetivo do trabalho.

- **Critério 1 -** Não relacionado a classificação de documentos
- **Critério 2 -** Trata de sentença em processo jurídico
- **Critério 3 -** Método de aprendizado supervisionado

III.4 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.5 Resultados

Aplicamos a metodologia PRISMA, conforme descrito na Figura III.1, e procedemos a seleção de trabalhos relacionados na data de 26/11/2023. Selecionamos três artigos considerados mais relevantes, pelos critérios definidos, e os apresentamos nas seções a seguir.

III.5.1 LegalVis: Exploring and Inferring Precedent Citations in Legal Documents [Resck et al., 2023]

Precedente é uma decisão judicial tomada em um caso concreto, que pode servir como exemplo para outros julgamentos semelhantes. Na Constituição do Brasil está previsto o mecanismo da **Súmula Vinculante**, o qual permite ao Supremo Tribunal Federal (STF) consolidar entendimentos sobre questões judiciais estabelecendo precedentes que deverão ser seguidos.

Analisar documentos de processos judiciais e encontrar relação com súmulas vinculantes não é uma tarefa trivial. Resck et al. [2023], empregando técnicas de aprendizado de máquina, propuseram o **LegalVis**, um sistema de análise visual baseado na web projetado para apoiar a análise de documentos legais que citam ou poderiam potencialmente citar um precedente vinculante.

Na implementação da arquitetura do LegalVis são previstas duas etapas principais:

- Processo de aprendizagem - modelos de rede neural aprendem o que constitui uma citação de precedente.

- Identificação de potencial citação - modelos realizam a tarefa de identificar um citação potencial e empregam uma técnica de interpretabilidade para explicar a decisão tomada.

Um precedente vinculante pode abranger decisões relacionadas a diferentes assuntos, no LegalVis são utilizadas estratégias de modelagem de tópicos para agrupar documentos ligados ao mesmo precedente e identificar palavras relevantes que representem os grupos.

III.5.2 Guided Semi-Supervised Non-Negative Matrix Factorization [Li et al., 2022]

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 **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.5.3 A two-staged NLP-based framework for assessing the sentiments on Indian supreme court judgments [Gupta et al., 2023]

No estudo apresentado por Gupta et al. [2023] é proposto um modelo baseado em processamento de linguagem natural visando analisar a percepção pública sobre julgamentos realizados pela Suprema Corte indiana. O modelo é composto de dois estágios:

- Modelagem de tópicos - utiliza o algoritmo de aprendizado não supervisionado **Latent Dirichlet Allocation** para identificar tópicos relacionados aos julgamentos efetuados pela Suprema Corte.

- Análise de sentimentos - utiliza o modelo **Valence Aware Dictionary Sentiment Reasoner**[2], baseado em léxico e regras, para avaliar os sentimentos das pessoas sobre os tópicos relacionados aos julgamentos realizados pela Suprema Corte

[2]https://github.com/cjhutto/vaderSentiment

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. Na Seção IV.1, são descritos os dados usados deste trabalho. Na Seção IV.2 é apresentada uma representação formal do problema que abordamos nesta dissertação. Finalmente, a Seção IV.3 descreve em detalhes cada passo da metodologia proposta.

IV.1 Corpus e metadados associados

Estruturamos um corpus para o desenvolvimento da pesquisa proposta. O conteúdo desse corpus são documentos correspondentes a recursos especiais extraídos de processos jurídicos não sigilosos, no âmbito da justiça federal do Brasil. 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 susgestão 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.1 apresenta essa distribuição. Dados estatísticos sobre os temas são apresentados na tabela IV.2. Todos 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.

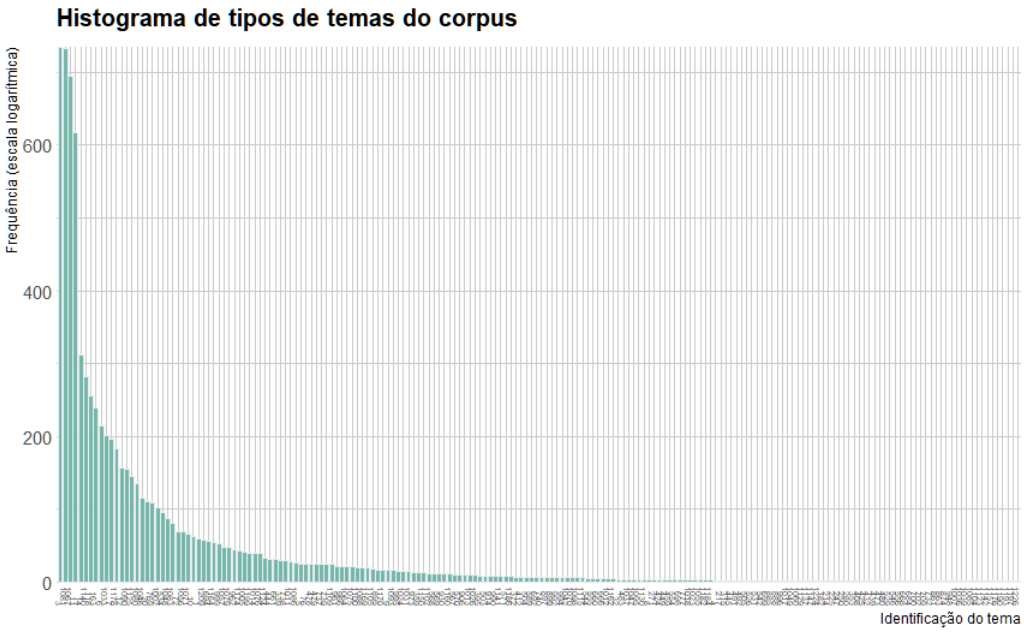


Figura IV.1: Distribuição de temas no corpus

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.

- 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.

IV.2 Formulação do problema

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 documentos representando temas repetitivos, que denotamos por $\mathcal{T}$. O problema que consideramos corresponde a identificar 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.

IV.3 Passos da metodologia

Nosso principal propósito neste trabalho é elaborar um procedimento que permita associar corretamente um dado recurso especial a um tema. Dessa forma, e seguindo a hipótese apresentada no Capítulo I, os passos da metodologia adotada nesta dissertação consistem nos listados a seguir. A Figura IV.2 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.

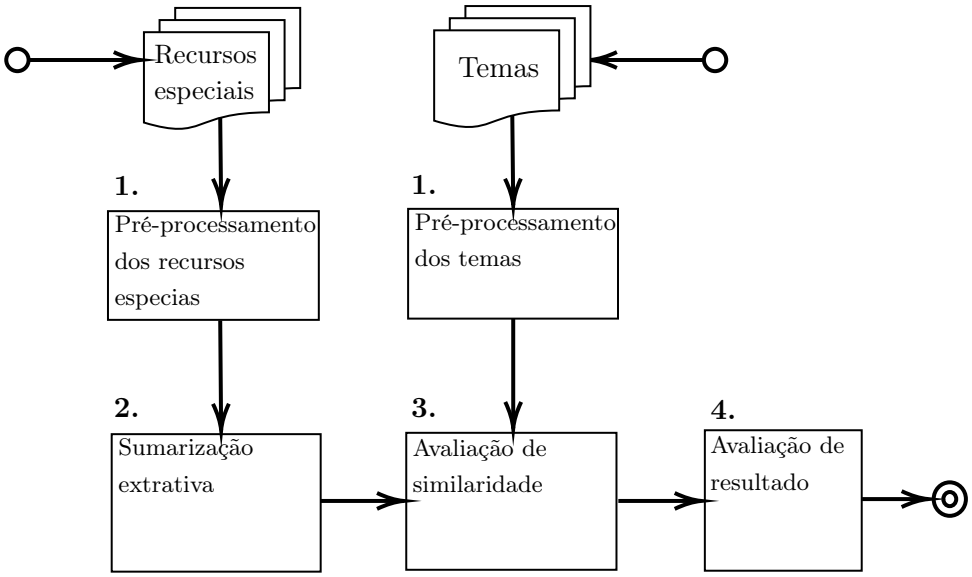


Figura IV.2: Passos da metodologia proposta.

IV.3.1 Pré-processamento dos recursos especiais e temas

Como descrito na Seção IV.1, o tamanho médio do texto de um recurso especial aproxima-se de 5 mil palavras. Neste universo de palavras buscamos identificar quais delas melhor representam a informação central de um recurso especial. Como etapa inicial para atingir esse objetivo, realizamos um pré-processamento no texto de cada recurso. Buscamos com isso eliminar palavras, sinais de pontuação e padrões numéricos irrelevantes para a tarefa em questão que acabam por gerar "ruído" na identificação da informação relevante.

Para eliminar palavras irrelevantes para o contexto do problema, fizemos uso da plataforma Natural Language Toolkit (NLTK)[1] para a linguagem Python. Por meio de uma função da NLTK, pudemos identificar e remover do texto palavras consideradas **stopwords** na língua portuguesa. Ou seja, removemos palavras consideradas comuns e que não apresentam grande importância para o significado do texto. Frequentemente são artigos, preposições e conjunções.

Pela natureza dos documentos contidos no corpus, vimos que seria pertinente remover palavras repetitivas que não agregariam para a análise em questão. Então removemos palavras tais como **tribunal, procuradoria, região, endereço, fone, fax, cep** e outras semelhantes.

Avaliamos também a necessidade de remover dos textos expressões que obedecessem a certos critérios. Desta forma utilizamos o módulo **re**[2] da biblioteca padrão da linguagem Python para tratar **expressões regulares**, assim identificamos e removemos algumas expressões, como por exemplo:

- $\backslash d\{3\}.\backslash d\{3\}.\backslash d\{3\}-\backslash d\{2\}$ - registros do Cadastro de Pessoa Física (CPF)

[1] https://www.nltk.org/search.html?q=stopwords
[2] https://docs.python.org/3/library/re.html

- $\backslash d\{2\}\backslash.\backslash d\{3\}\backslash.\backslash d\{3\}/\backslash d\{4\}-\backslash d\{2\}$ - registros do Cadastro Nacional de Pessoa Juridica (CNPJ)
- `\S+@\S+` - endereço de correio eletrônico

A fonte de dados utilizada na construção do corpus engloba alguns documentos escritos em Hypertext Markup Language (HTML). Ao analisarmos uma amostra do corpus constatamos que no lugar de alguns sinais de pontuação existiam elementos do tipo **entidades HTML**. Por exemplo:

- Onde deveriam existir "(aspas duplas) estava o elemento **"**
- Onde deveriam existir o sinal – (travessão) estava o elemento **–**

Para evitar que no processo de classificação dos recursos essas entidades HTML pudessem ser interpretadas como alguma palavras do texto, incluímos na lista de stopwords a serem removidas entidades HTML como **&ldquo**, **&rdquo**,**&lsquo**, **&rsquo**, **&bull**, **&middot**, **&sdot**, **&ndash** e outras.

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. Desta forma fizemos novamente uso do módulo **re** da biblioteca padrão da linguagem Python para localizarmos a primeira ocorrência da palavra **cabimento** no texto. De posse da localização dessa palavra (cabimento), eliminamos do texto todo conteúdo anterior a ela.

No Apêndice B, encontra-se uma amostra de texto de um tema repetitivo. Observa-se que é um texto sucinto com tamanho em torno de 30 palavras. No entanto, mais de 30 por cento são artigos e preposições. Nesta etapa de pré-processamento de temas fizemos uso da plataforma NLTK para identificar e possibilitar remoção de palavras consideradas **stopwords** da lingua portuguesa, predominantemente artigos, preposições e conjunções. O objetivo é contribuir para uma melhor avaliação de similaridade numa etapa posterior.

Com a execução de experimentos iniciais, identificamos procedimentos repetitivos que impactaram significativamente o tempo de processamento. Deste modo, refatoramos a implementação do experimento e optamos por salvar em arquivo uma estrutura de dados contendo cada texto processado e suas respectivas representações vetoriais.Desta forma, os dados foram serializados (i.e., convertidos em um fluxo de bytes) e salvos em um arquivo para, numa etapa posterior, sofrerem o processo inverso, proporcionando otimização na realização dos experimentos computacionais.

Para gerar uma representação vetorial do texto, fizemos uso do framework Sentence-BERT (SBERT) [Reimers and Gurevych, 2019] que representa o estado da arte na transformação de textos e imagens em vetores, **embeddings**. O SBERT é derivado do Bidirectional Encoder Representations

from Transformers (BERT), o qual é um modelo de rede neural pré-treinado e baseado na arquitetura de rede neural conhecida como Transformers [Vaswani et al., 2017]. Para um conteúdo textual de entrada o SBERT gera na saída um vetor denso de dimensão fixa com tamanho 512, o qual incorpora as relações semânticas do texto.

IV.3.2 Sumarização extrativa

Nesta etapa, buscamos criar uma representação textual resumida do documento de entrada, contudo preservando a semântica do texto. Ao fim da etapa geramos uma representação vetorial do resumo textual obtido, desta forma temos duas formas de representação, uma textual e outra vetorial. A partir dessas representações será avaliada a similaridade com os temas na etapa seguinte. Cabe ressaltar que o tratamento aplicado na etapa anterior, de certa forma, cria um representação sintética do texto original. Porém, intensificando a síntese do texto, fizemos uso de duas abordagens distintas para gerar uma nova representação. Em uma abordagem buscamos, encontrar um conjunto de palavras que expressem **tópicos** relacionados ao texto. Em outra abordagem, buscamos identificar **sentenças** relevantes do texto. Realizamos a descrição dessas duas abordagens nas próximas seções.

Extração de tópicos

É razoável supor que recursos especiais rotulados sob um mesmo tema possuem algum grau de similaridade semântica. Nesta abordagem, por meio de aprendizado não supervisionado, buscamos agrupar documentos que possuem conteúdo semântico similar. Por fim, buscamos explicitar através de tópicos qual é esse conteúdo.

Para esta tarefa fizemos uso de um módulo de geração de tópicos implementado com a biblioteca BERTopic (Seção II.3). Como parâmetros de entrada para o módulo, foram adicionados os textos dos recursos com as respectivas representações vetorias calculadas e armazenadas no passo anterior(Seção IV.3.1. Também configuramos o módulo BERTopic para utilizar o algoritmo Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) (veja descrição na Seção II.2.2) para organizar os recursos especiais (representados como vetores) em grupos.

No decorrer de experimentos preliminares, através da função `get_topic_info` disponibilizada no BERTopic, pudemos identificar que os tópicos estavam sendo gerados com erros ortográficos. Tomamos a decisão de alterar a configuração padrão do BERTopic, a qual utiliza o modelo SBERT **all-MiniLM-L6-v2**[3], e passamos a utilizar o modelo multilíngue do SBERT **distiluse-base-multilingual-cased-v1**[4].

[3]https://maartengr.github.io/BERTopic/api/bertopic.html
[4]https://www.sbert.net/docs/pretrained_models.html

Como resultado da aplicação deste passo, obtemos as palavras que melhor representam cada grupo. Esse conjunto de palavras é o tópico que representa cada grupo. Uma vez identificado o tópico de cada grupo, podemos avaliar como seus membros (recursos especiais) se relacionam com os temas.

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. Dessa forma, buscamos 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 IV.3, 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 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.

Extração de sentenças

Nesta abordagem de sumarização fizemos uso do algoritmo LexRank (Seção II.3) utilizando uma implementação disponível para a linguagem Python[5].

Cada recurso r submetido é fragmentado em sentenças $s_i \in S$ por meio de uma função da plataforma NLTK, onde s_i é o identificador da i-ésima sentença do recurso r. Cada sentença é transformada numa representação vetorial V_i por meio de um modelo SBERT, onde V_i é o identificador do vetor que representa a i-ésima sentença do recurso r. É feito um cálculo de similaridade por cosseno (Seção II) entre pares de vetores V_i e os valores são armazenados e passados como parâmetros de entrada para o LexRank.

O algoritmo computa o grau de **centralidade** de cada sentença $s_{ri} \in S$ avaliando a quantidade de outras sentenças com as quais s_i possui similaridade relevante, um parâmetro " threshold" do algoritmo determina um limiar de relevância para o valor de similaridade.

No término da execução do LexRank, considerando um parâmetro "size" inerente ao algoritmo, um resumo com as sentenças mais relevantes do texto é obtido. Novamente, os dados foram serializados (i.e., convertidos em um fluxo de bytes) e salvos em um arquivo para numa etapa posterior sofrerem o processo inverso, proporcionando otimização no processamento do experimento.

LexRank guiado

Conforme descrito na Seção II.3, o algoritmo LexRank se propõe a gerar um resumo extraindo as sentenças mais relevantes de um texto. Em essência o algoritmo realiza o cômputo de similaridade entre todas as sentenças do texto, os valores computados determinam o **grau de centralidade** de cada sentença. As sentenças com grau de centralidade mais elevado compõem o resumo do texto.

Considerando o contexto deste trabalho, ao submetermos o texto de um recurso especial ao LexRank o algoritmo nos dá como saída um resumo, o qual é formado pelo conjunto das sentenças que melhor representam o recurso. Estejamos atentos ao detalhe da possibilidade de uma sentença que não faz parte do resumo ser importante para a tarefa de classificar o recurso especial em um tema. Isso pode ocorrer pois o LexRank calcula o grau de centralidade de cada sentença sem considerar qualquer fator externo ao texto avaliado.

Propomos uma alteração no algoritmo LexRank de modo a guiá-lo na tarefa de classificação. Com a alteração proposta, a decisão para que uma sentença componha o resumo do texto é resultado da combinação de dois fatores. Um dos fatores é interno ao texto, ou seja, o grau de centralidade da sentença calculado no algoritmo original. O outro fator é externo ao texto, calculado em função

[5]https://pypi.org/project/lexrank/

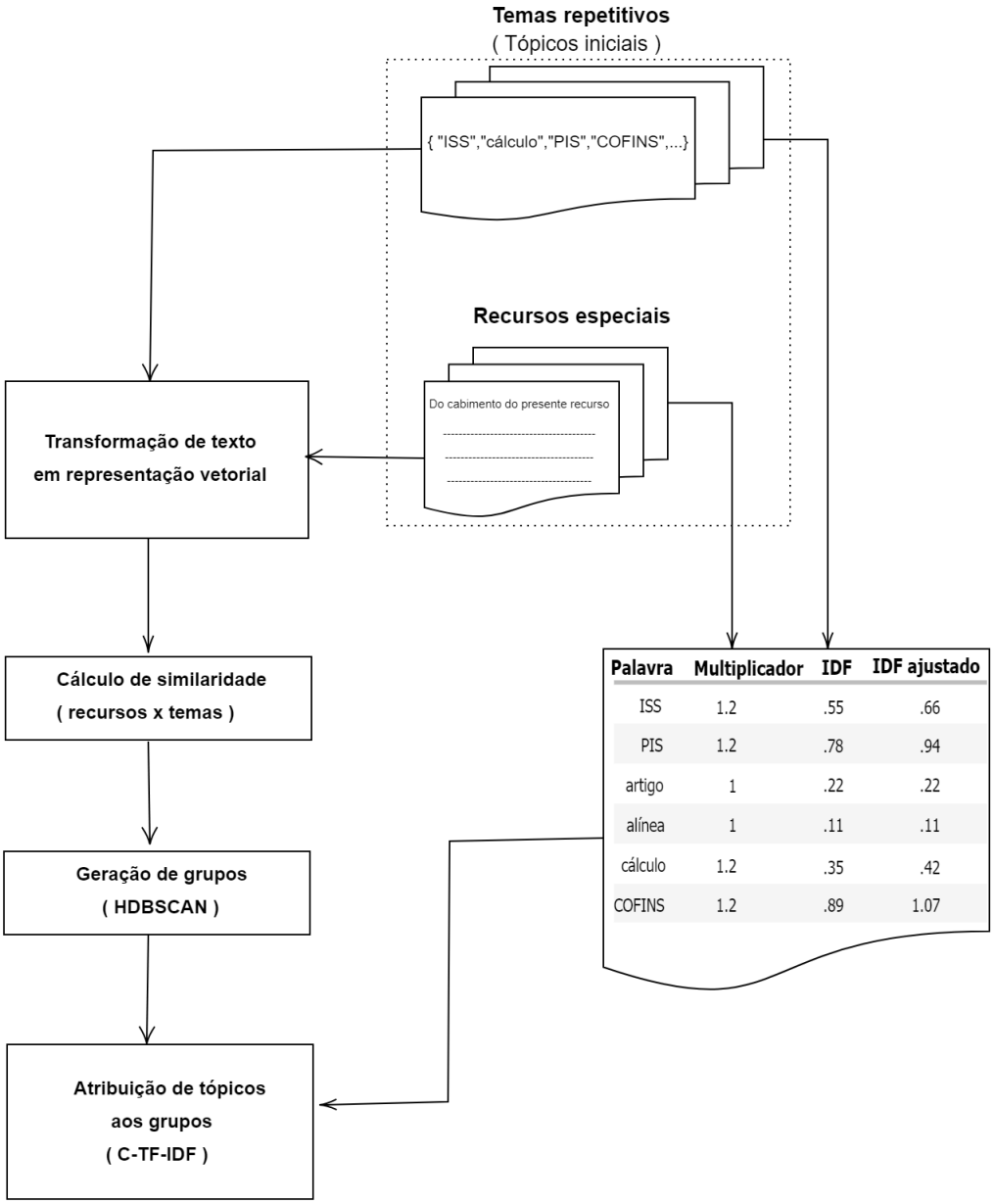


Figura IV.3: Modelagem de tópicos guiada. [Grootendorst, 2022]

da similaridade entre cada sentença do texto a ser resumido e os temas. Desta forma, no LexRank guiado, para cada sentença s:

- Computamos o grau de centralidade γ_s, conforme cálculo original do algoritmo.
- Submetemos s como entrada para o algoritmo BM25 tendo como alvo cada tema $t \in \mathcal{T}$. Identificamos e armazenamos o maior score encontrado na execução do BM25, σ_{st}.
- Calculamos o score combinado usando dois fatores de ponderação (Equação IV.1).

O score combinado é calculado da seguinte forma:

$$\text{Score combinado} = \alpha * \gamma_s + \beta * \sigma_{st} \tag{IV.1}$$

Onde α e β são parâmetros arbitrários. Desta forma, as sentenças selecionadas para compor o resumo do texto serão aquelas com os maiores valores de score combinado. A quantidade de sentenças selecionadas, e consequentemente o tamanho do resumo, permanece limitada pelo parâmetro "size" do algoritmo original.

IV.3.3 Avaliação de similaridade

Nesta etapa temos por objetivo obter um valor numérico que indique o grau de similaridade entre um **recurso** e um **tema**. Destacamos que após o processamento executado nas etapas anteriores foram geradas novas formas de representação para um mesmo recurso.

Inicialmente fizemos uso da representação que mais se aproxima do texto original do recurso r. Seja r' a representação do recurso r obtida como saída da etapa **Pré-processamento dos recursos especiais e temas**. Computamos a similaridade entre pares r' e cada $t \in \mathcal{T}$ e geramos os valores $\sigma_{r'}$. Desta forma, $\sigma_{r'}$ representa a similaridade entre um recurso r e um tema $t \in \mathcal{T}$.

Ao término da etapa **Sumarização extrativa**, temos disponíveis duas novas representações para cada recurso:

- Resumo por tópico - Entendemos por tópico um conjunto de palavras sem qualquer conectivo sintático ligando-as. Considerando a abordagem **BERTopic** exemplificada na Figura IV.4, cada recurso $r \in \mathcal{R}$ é incluído em um subconjunto $\mathcal{R}' \subset \mathcal{R}$. Cada subconjunto possui um e somente um tópico que o identifica. Desta forma, dado um recurso r conhecemos o tópico r' que o representa. Efetuamos então a avaliação de 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 $t \in \mathcal{T}$.
- Resumo por sentença - Na abordagem **LexRank**, como exemplificado na Figura IV.5, extraímos as sentenças mais significativas do recurso r e compomos um resumo r' a partir

dessas.Efetuamos então a avaliação de 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 $t \in \mathcal{T}$.

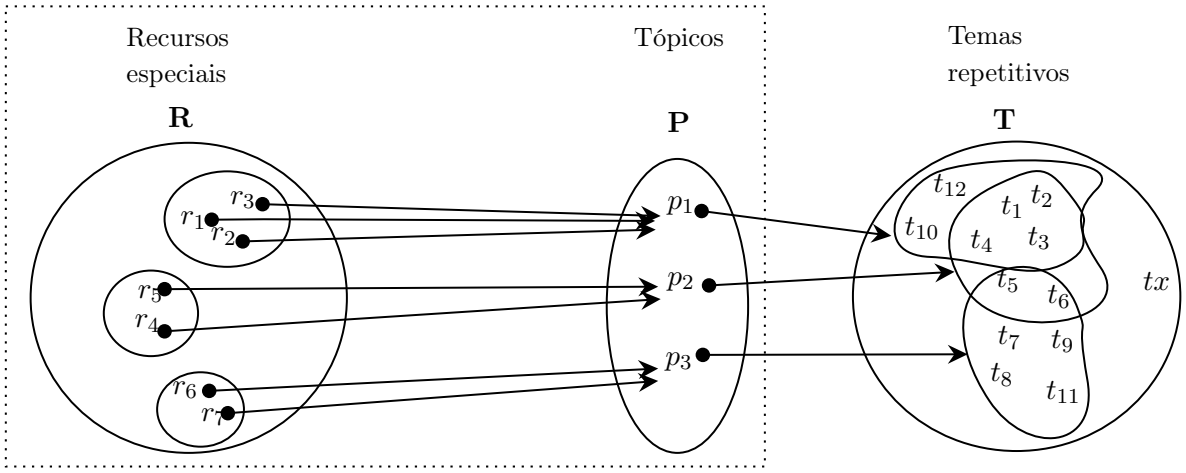


Figura IV.4: Bertopic - Relação entre recursos, tópicos e temas

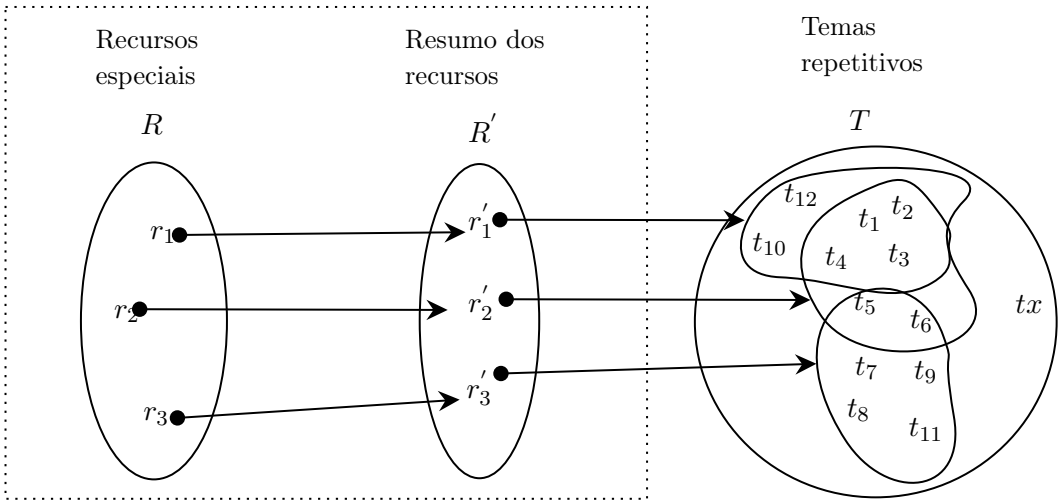


Figura IV.5: LexRank - Relação entre recursos, resumos e temas

Geração de lista de temas

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. Em ambas as abordagens descritas anteriormente, computamos um valor numérico σ_{rt} que representa a similaridade entre um recurso r e um tema $t \in \mathcal{T}$. Deste modo, obtemos 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
- 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 passando a conter os n temas que melhor estão relacionados ao recurso avaliado.

IV.3.4 Avaliação de resultado

Conforme descrito na Seção IV.1, 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 humanoAo 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.

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

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

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

F1-score

Calculada pela média harmônica entre precision e recall.

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

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.

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

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

$$\text{NDCG at k} = \frac{\text{DCG at k}}{\text{IDCG at k}} \quad \text{(IV.9)}$$

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

IV.4 Parâmetros de variação

Na metodologia adotada existem elementos que, pela natureza do problema tratado,possuem potencial de impactar significativamente o resultado esperado. Por conta disto, adotamos alguns parâmetros como de livre variação e computamos os resultados.

IV.4.1 Similaridade

Fizemos uso de duas abordagens distintas para computar a similaridade entre tópicos e temas repetitivos:

- Aplicação de um modelo SBERT para converter r' e cada t em vetores densos. Cálculo de similaridade por cosseno entre os vetores, gerando σ_{rt} para cada par avaliado;
- Uso das palavras do tópico r' como palavras-chaves de uma consulta. As palavras-chaves servem como parâmetro de entrada para o algoritmo BM25 [Trotman et al., 2014] o qual terá como alvo cada $t \in \mathcal{T}$, gerando σ_{rt} para cada par avaliado;

IV.4.2 Remoção de termos

Em relação ao pré-processamento do texto efetuamos duas abordagens distintas. Numa abordagem executamos todos os passos da metodologia exceto o de pré-processamento e computamos as métricas qualitativas. Em outra abordagem buscamos identificar e remover elementos do texto.

Fizemos uso de expressões regulares adequadas ao texto tratado e também usamos uma função da plataforma NLTK[6], novamente computamos as métricas e avaliamos se houve ganho.

IV.4.3 Tamanho do sumário

Referente à etapa de sumarização extrativa, queremos avaliar se há algum impacto significativo no cômputo de similaridade entre recursos e temas variando-se o tamanho do resumo obtido. Para isto efetuamos alterações paramétricas em ambas as abordagens:

- **Bertopic** - Variamos um hiperparâmetro do modelo BERTopic chamado **top_n_words**. Este parâmetro define a quantidade de palavras que compõem o tópico gerado pelo modelo.
- **Lexrank** - Variamos no algoritmo o parâmetro que define a quantidade de sentenças que compõem o resumo do texto.

[6]https://www.nltk.org/

Capítulo V Experimentos

Os experimentos foram executados sobre o corpus de 7.967 documentos apresentado no Capítulo IV, cujo resumo estatístico encontra-se na Tabela IV.1. No corpus constam as sugestões de temas dadas pela máquina de busca ao especialista humano. Esse conjunto de sugestões , intrínseco no corpus, constitui a linha de base usada como referência para nossos experimentos. Neste capítulo fornecemos informações sobre o hardware e configurações de software usados na condução dos experimentos (Seção V.1). Apresentamos os resultados dos estudos de ablação (Seção V.2). Apresentamos também uma síntese dos resultados obtidos e a comparamos com o resultado da linha de base (Seção V.3).

V.1 Configurações

Os experimentos foram executados em um servidor AMD EPYC 7452 32-Core 2.35GHz 52GB RAM. Na tabela V.1 constam as versões das bibliotecas python utilizadas nos experimentos.

Tabela V.1: Versão das bibliotecas

Biblioteca	**versão**
argparse	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
torch	1.12.1

V.2 Estudos de ablação

Conduzimos estudos de ablação em conformidade com os parâmetros de variação apresentados na Seção IV.4. Em resumo, os parâmetros de variação consistem em:

- Pré-processamento - Adotamos como um parâmetro de decisão a remoção ou permanência de termos do recurso especial.

- Tipo de representação - Na etapa seguinte ao pré-processamento escolhemos o tipo de representação que seria adotada adiante no fluxo. Como opção poderíamos manter o texto ou gerar um resumo extrativo utilizando as estratégias BERTopic, BERTopic guiado, LexRank ou LexRank guiado.

- Tamanho do resumo - Em caso de resumo, executamos variações nos tamanhos dos resumos gerados mediante parâmetro passado ao algoritmo.

- Cálculo de similaridade - Nesta etapa temos a posse do texto/resumo juntamente com sua representação vetorial. Desta forma podemos avaliar a similaridade do recurso com o tema de duas formas. Uma forma é utilizando o texto/resumo como consulta de entrada para o algoritmo BM25 tendo como alvo o texto do tema. Outra forma é , de posse da representação vetorial de recurso e dos temas podemos executar um cálculo de similaridade por cosseno.

Considerando as combinações possíveis, executamos um total de 180 experimentos distintos, cuja íntegra dos resultados consta no Apêndice D. Na figura V.1 demonstramos o fluxo de processamento efetuado nos experimentos.

Na etapa de geração de resumo do texto temos quatro abordagens distintas que executam o mesmo tipo de tarefa. Visando a otimização do código e prevendo futuras variações dos algoritmos, fizemos o uso padrão de projeto "Strategy" [Gamma et al., 1995] o qual é bem adequado a este contexto. O padrão Strategy define uma família de algoritmos, encapsula as distintas implementações de cada um deles e os torna intercambiáveis. Um diagrama de classes Unified Modeling Language (UML) do padrão implementado é apresentado na Figura V.2

Os dados gerais do Apêndice D foram condensados e distribuídos nas Tabelas V.2, V.3, V.4 e V.5. Nessas tabelas são apresentados o melhor desempenho por tipo de representação de texto, considerando se houve ou não remoção de termos irrelevantes e se o tipo de avaliação de similaridade foi por cosseno ou pelo algoritmo BM25 . Os tipos de representação podem ser o próprio texto ou um resumo, considerando que o resumo pode ser por tópicos nas abordagens BERTopic/BERTopic guiado e por sentenças nas abordagens LexRank/LexRank guiado.

Tabela V.2: Melhor resultado por tipo de representação

Tipo de representação	**recall@6**	**f1-score**	**map@6**	**ndcg@6**
Texto	0,71457	0,60415	0,52329	0,57144
Resumo - LexRank guiado - 15 sentenças	**0,75750**	**0,62679**	**0,53455**	**0,59018**
Resumo - LexRank - 60 sentenças	0,72148	0,60145	0,51567	0,56725
Resumo - BERTopic - 55 palavras	0,5696	0,45524	0,37912	0,42662
Resumo - BERTopic guiado - 60 palavras	0,55517	0,45252	0,3819	0,42535

Tratamento: Com remoção de termos | Similaridade: BM25

Nas Figuras V.3 e V.4 buscamos representar o impacto no desempenho de cada abordagem mediante a variação de tamanho do resumo gerado. Comparando os dois gráficos observamos

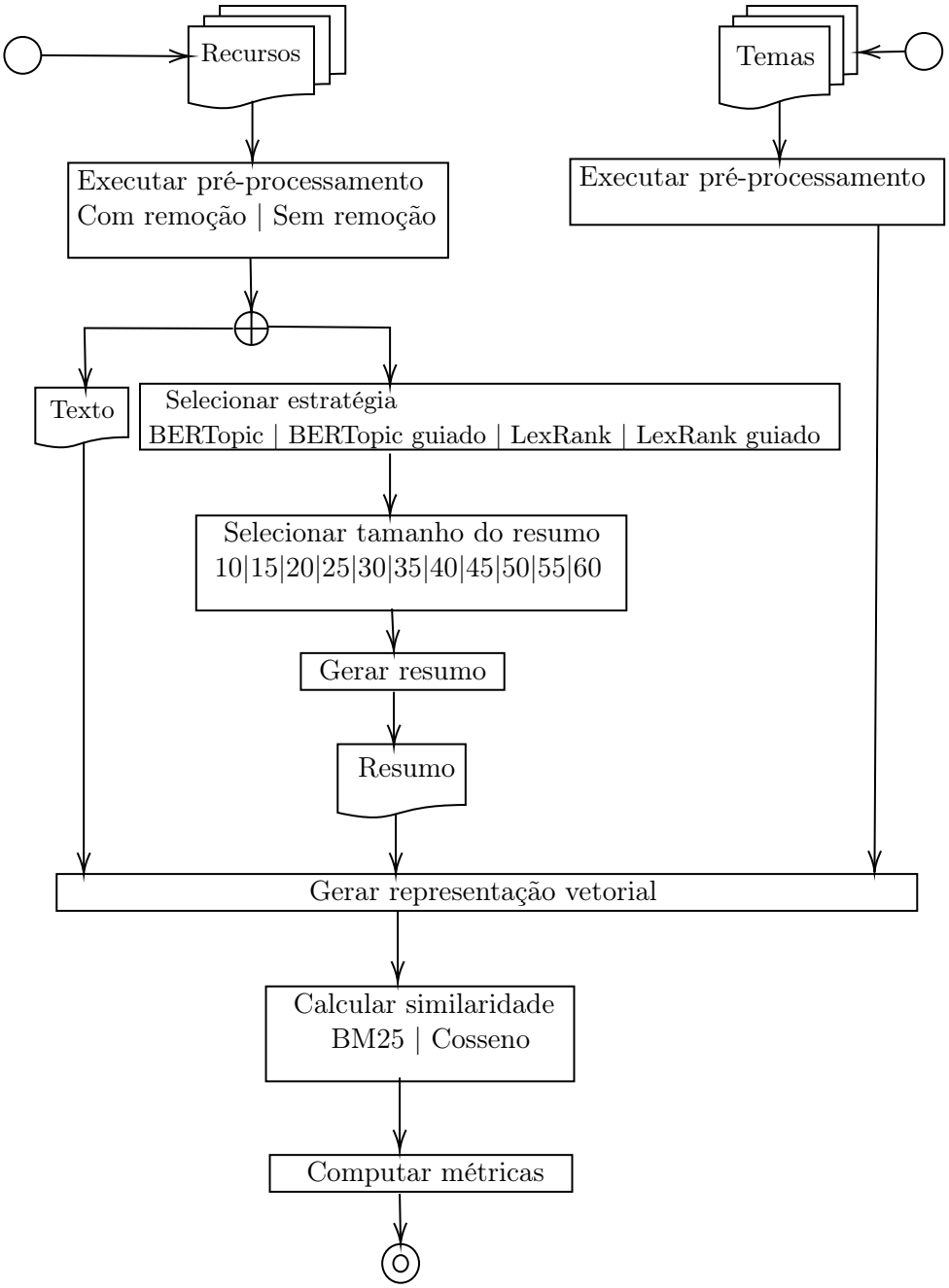


Figura V.1: Fluxo de processamento dos experimentos

Tabela V.3: Melhor resultado por tipo de representação

Tipo de representação	**recall@6**	**f1-score**	**map@6**	**ndcg@6**
Texto	0,13631	0,10726	0,08842	0,09989
Resumo - LexRank guiado - 10 sentenças	0,47044	**0,38577**	0,32693	**0,36259**
Resumo - LexRank - 20 sentenças	0,35434	0,27001	0,21815	0,25191
Resumo - BERTopic - 45 palavras	**0,47923**	0,37009	**0,51567**	0,34546
Resumo - Bertopic guiado - 45 palavras	0,46128	0,35516	0,28873	0,33222

Tratamento: Com remoção de termos | Similaridade: Cosseno

também o impacto causado pelo tipo de **similaridade** adotado:

- **BM25** - Na Figura V.3 observamos que a combinação LexRank/BM25 possui grande vantagem, obtendo incrementos sucessivos no recall até atingir um ponto limite no qual o tamanho

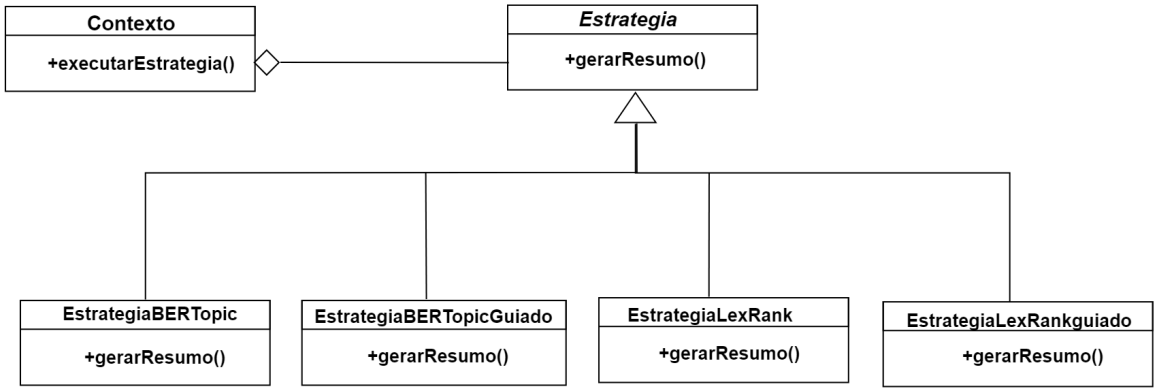


Figura V.2: Diagrama de classes do padrão de projeto implementado

Tabela V.4: Melhor resultado por tipo de representação

Tipo de representação	**recall@6**	**f1-score**	**map@6**	**ndcg@6**
Texto	0,72813	0,59771	0,51337	0,56385
Resumo - LexRank - 50 sentenças	0,73026	0,58523	0,48827	0,54842
Resumo - LexRank guiado - 30 sentenças	**0,75449**	**0,63608**	**0,54980**	**0,60090**
Resumo - BERTopic - 25 palavras	0,55065	0,44520	0,37365	0,41796
Resumo - BERTopic guiado - 60 palavras	0,55265	0,45007	0,37961	0,42284

Tratamento: Sem remoção de termos | Similaridade: BM25

Tabela V.5: Melhor resultado por tipo de representação

Tipo de representação	**recall@6**	**f1-score**	**map@6**	**ndcg@6**
Texto	0,14372	0,10870	0,08740	0,10128
Resumo - LexRank - 30 sentenças	0,41760	0,30358	0,23846	0,28272
Resumo - LexRank guiado - 10 sentenças	0,37982	0,28043	0,22227	0,26108
Resumo - BERTopic - 45 palavras	**0,46379**	**0,36052**	**0,29487**	**0,33728**
Resumo - BERTopic guiado - 45 palavras	0,46165	0,35662	0,29052	0,33335

Tratamento: Sem remoção de termos | Similaridade: Cosseno

do resumo é em torno de 60 sentenças.

- **Cosseno** - Na Figura V.4 observamos que as abordagens BERTopic e BERTopic guiado possuem desempenhos semelhantes, atingindo seus melhores valores de recall com tópicos de tamanho em torno de 50 palavras e apresentando uma tendência de queda no recall conforme aumenta-se o tamanho do resumo após essa dimensão. Uma observação relevante é que na combinação LexRank/cosseno praticamente não há variação no recall com o incremento no tamanho dos resumos gerados.

Nas Figuras V.5 e V.6 buscamos demonstrar o impacto no desempenho das abordagens mediante o tratamento na fase de pré-processamento do texto (Seção IV.3.1). Destaca-se o fato da semelhança

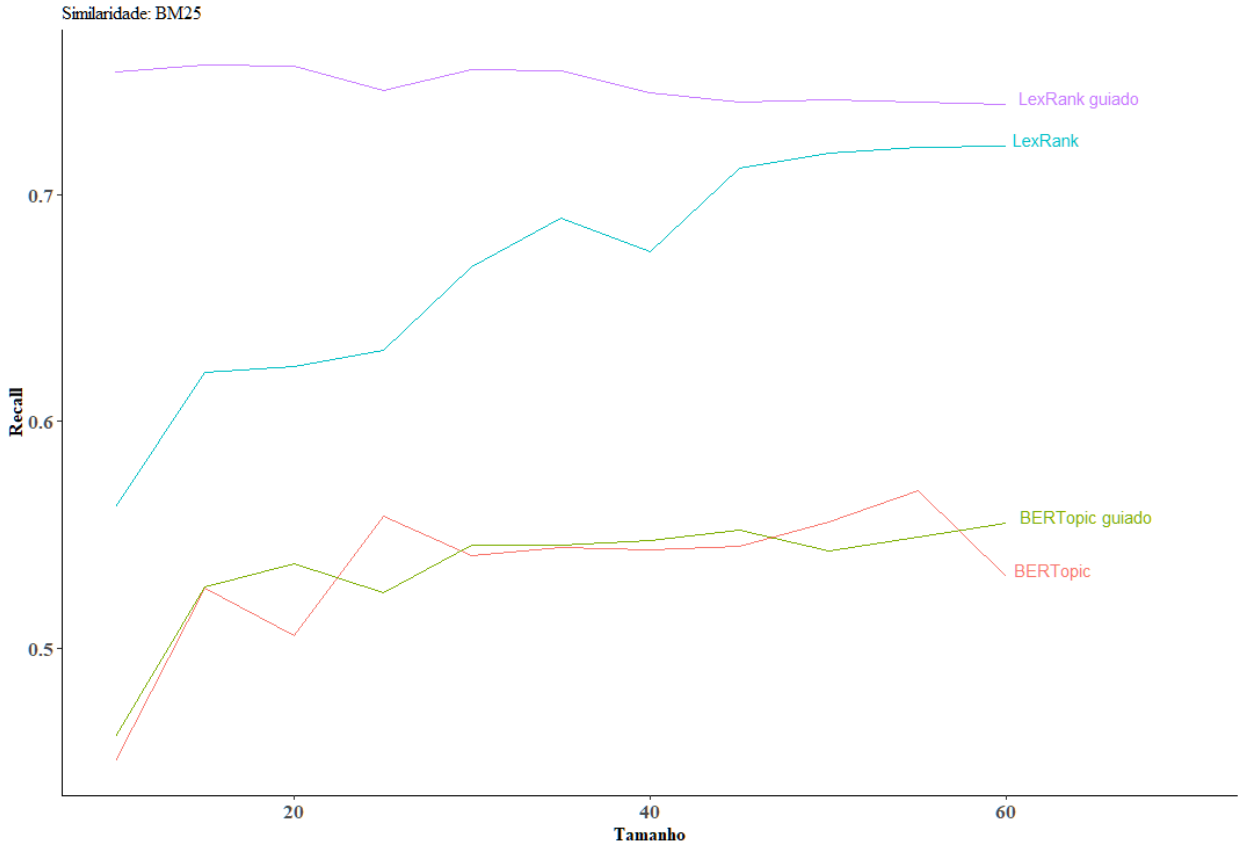


Figura V.3: Desempenho com variação de tamanho do resumo | Similaridade: bm25

de desempenho com ou sem remoção de termos considerados irrelevantes.

V.3 Síntese dos resultados

Conforme elucidado na Seção IV.3.4, para o contexto do problema avaliado o recall apresenta-se como a métrica mais importante. Desta forma o recall foi tratado como o fator preponderante para distinguir o desempenho entre os resultados obtidos.

O gráfico da Figura V.7 apresenta o desempenho da linha de base. Observa-se no gráfico que em 65,1% dos casos nenhuma das 6 sugestões dadas pela solução de base foi aceita pelo especialista humano como correta, e em **34,9%** dos casos, alguma das 6 sugestões dadas pelo sistema foi considerada correta segundo análise do especialista. Apresentando um **recall@6** igual a **0,34906** (Tabela V.6).

Dentre os 180 experimentos realizados, **175 apresentaram desempenho superior ao baseline e 5 apresentaram desempenho inferior**(vide Apêndice D). No melhor resultado alcançado nos experimentos(Tabela V.7) foi obtido um recall@6 igual a **0,75750** o que representa um **incremento de 117,01%** sobre o recall do baseline. Considerando todo o conjunto de experimentos, a **mediana** do recall@6 apresentou um incremento de **34,78%** sobre o recall do baseline.

Embora, para este trabalho, consideremos o recall como a métrica mais importante, nas Fi-

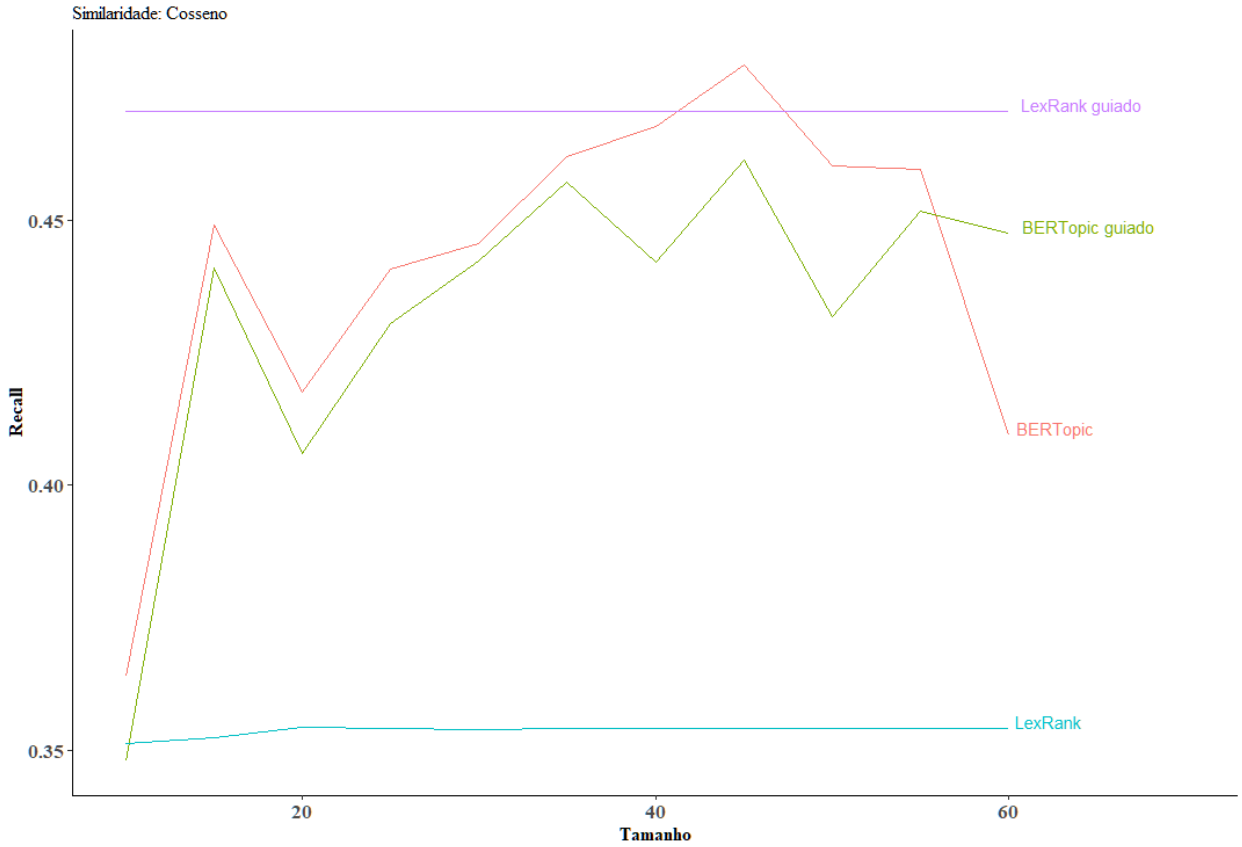


Figura V.4: Desempenho com variação de tamanho do resumo | Similaridade: cosseno

Tabela V.6: Métricas baseline

	recall@6	map@6	f1-score	ndcg@6
Elasticsearch	**0.34906**	0.30838	0.32746	0.31823

Tabela V.7: Melhor resultado alcançado considerando métrica recall@6

	recall@6	map@6	f1-score	ndcg@6
Sumarização: LexRank guiado Tamanho : 15 sentenças Tratamento : Com remoção de termos Similaridade : BM25	**0.75750**	0.53455	0.62679	0.59018

guras V.8 e V.9 buscamos também representar em um diagrama de dispersão as métricas F1 e NDCG. Essas métricas representam um fator de qualidade das listas de sugestões de temas para cada assunto. A métrica F1(vide Seção IV.6) implicitamente contém a precisão de acerto da lista e a NDCG(vide Seção IV.3.4 indica a posição que o resultado correto é indicado na lista de sugestão de temas. Desta forma, nos gráficos mencionados os melhores resultados são os posicionados no quadrante superior direito.Observa-se que nesta região há predominância dos resultados obtidos com a aplicação do LexRank. Em ambos os gráficos sinalizamos através de uma linha horizontal pontilhada o recall@6 da linha de base, buscando evidenciar os resultados alcançados nos experimentos em relação à linha.

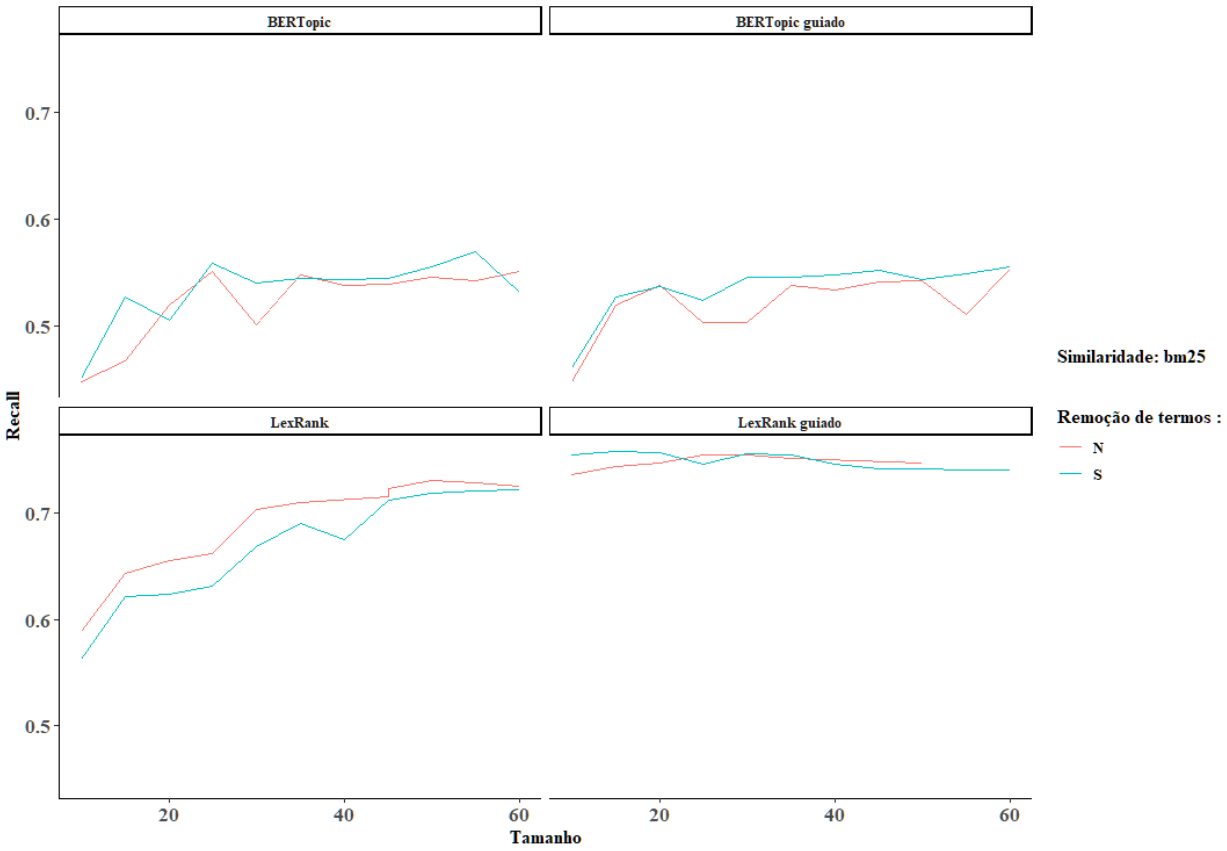


Figura V.5: Desempenho com variação de tratamento dos termos do texto

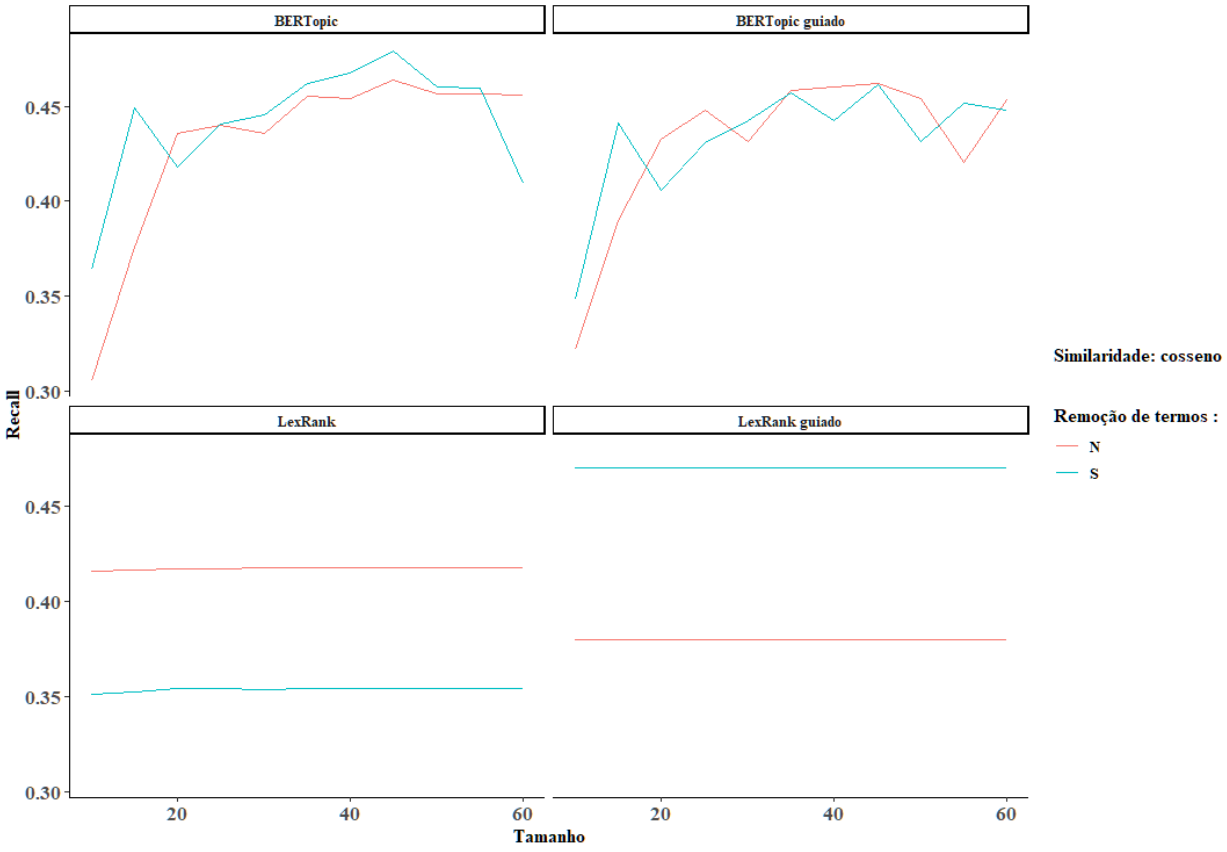


Figura V.6: Desempenho com variação de tratamento dos termos do texto

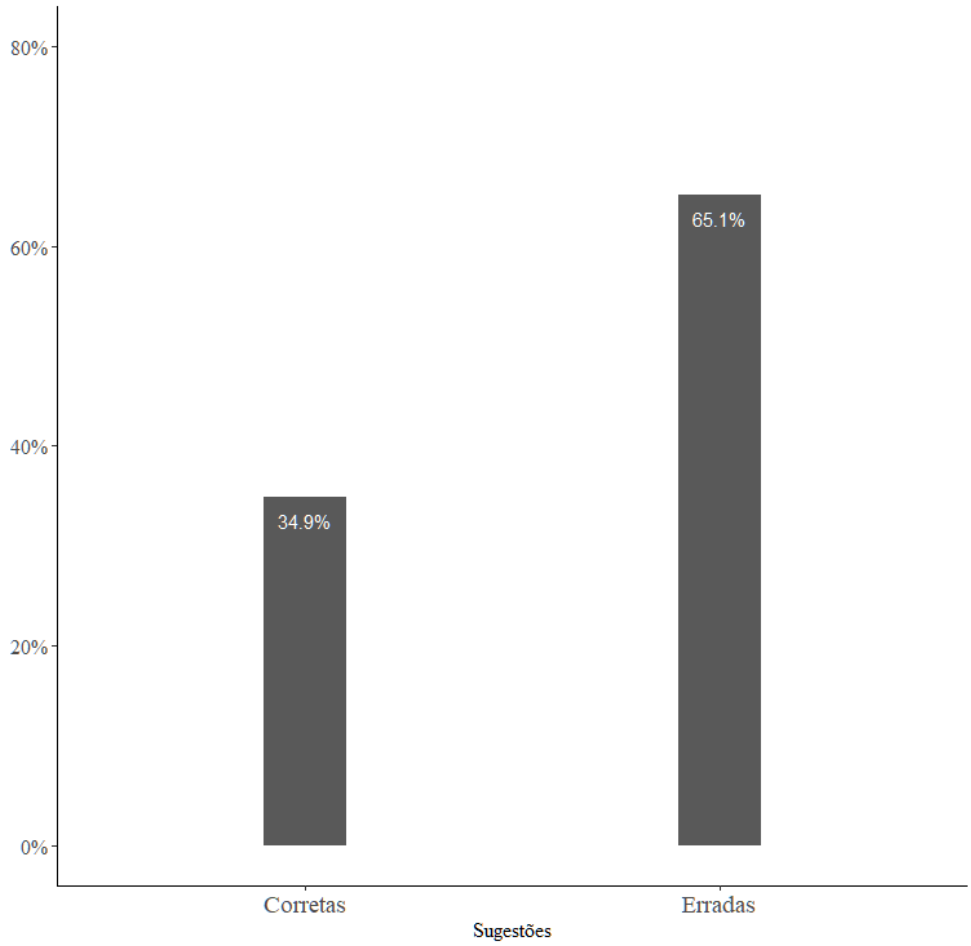


Figura V.7: Desempenho na classificação de recurso pelo Elasticsearch(baseline) - Lista com 6 sugestões de temas

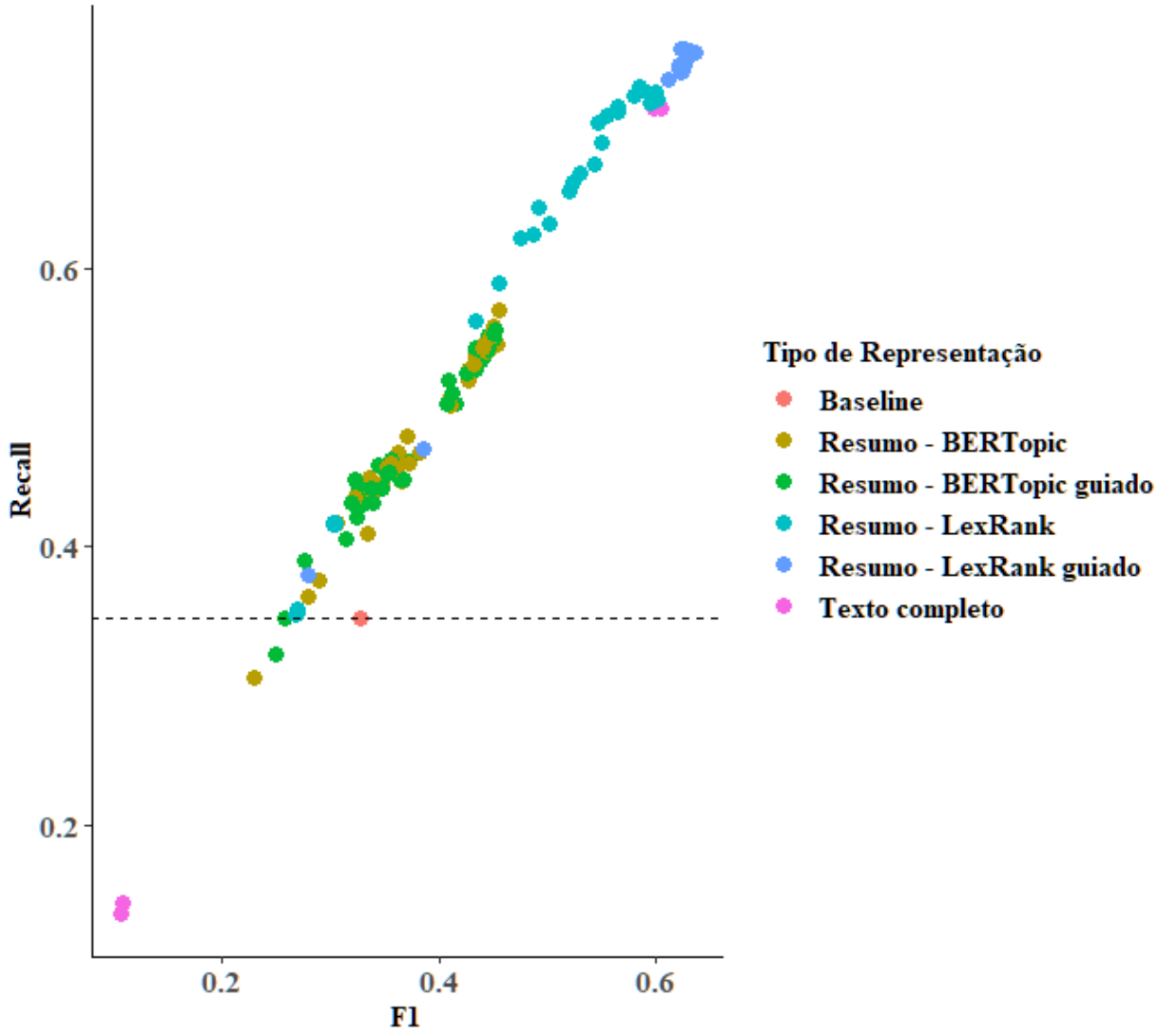


Figura V.8: Desempenho por tipo de representação

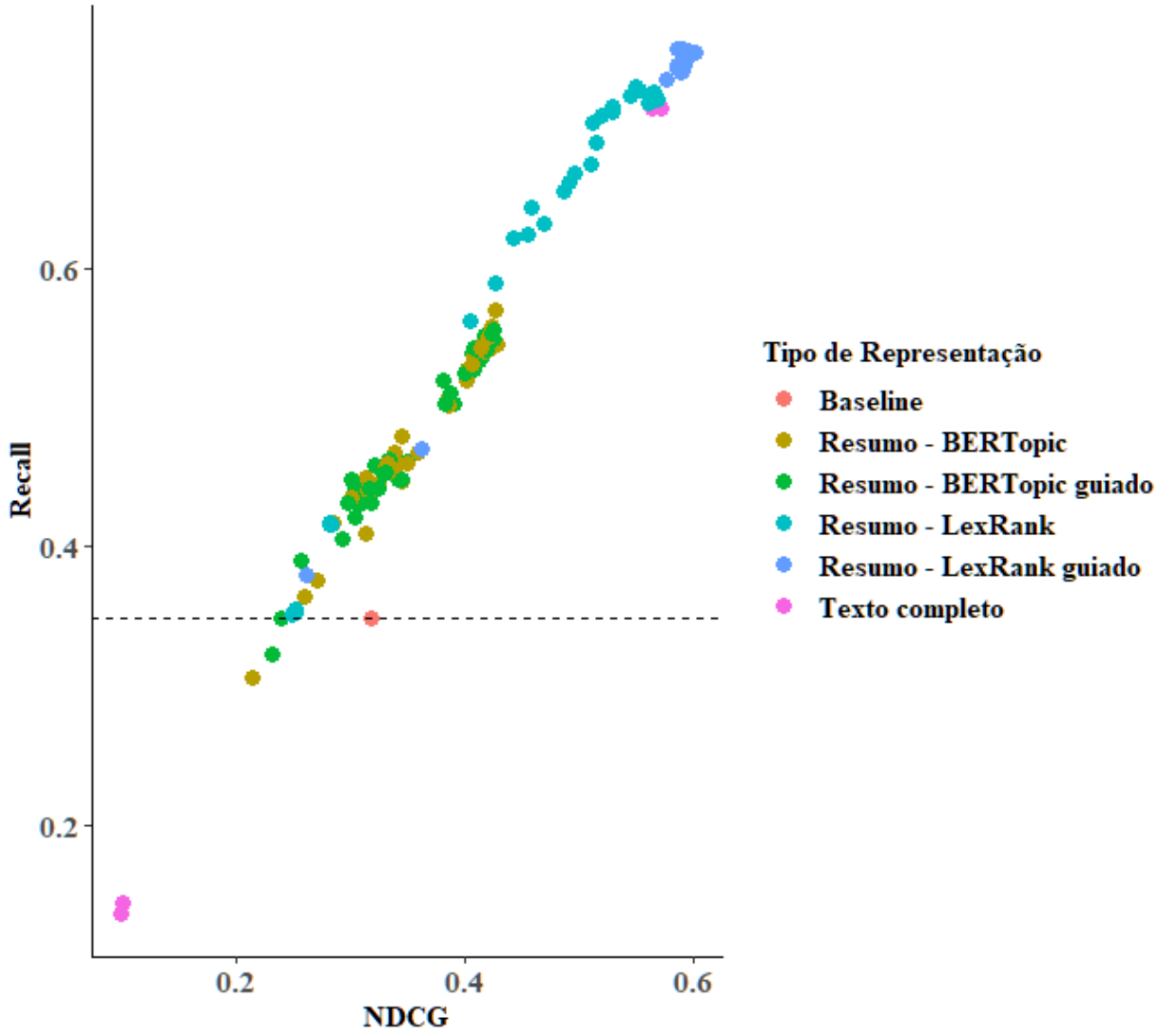


Figura V.9: Desempenho por tipo de representação

Capítulo VI Conclusões

Neste capítulo procedemos uma análise retrospectiva sobre o trabalho desenvolvido e apresentamos algumas conclusões obtidas.

VI.1 Análise Retrospectiva

Conforme descrição do Capítulo I, buscamos com este trabalho uma solução para o problema de classificar um recurso especial em um tema repetitivo definido pelo STJ. Derivado deste problema, temos como objetivo geral desta pesquisa demonstrar que a tarefa de classificação de um recurso especial em um tema é mais eficiente comparando-se diretamente o texto de um recurso com o texto de um tema. Sendo assim, o ponto fundamental que diferencia nossa abordagem daquela usada como referência é que na solução de referência (Seção I.5) um recurso especial novo é comparado com diversos recursos especiais, previamente classificados, existentes na base histórica de dados. A partir da similaridade com os recursos da base histórica um tema é sugerido para o recurso novo. Em nossa abordagem, um recurso especial novo é comparado diretamente com vários temas buscando encontrar qual é o tema mais adequado para classificá-lo.

Através dos resultados apresentados no Capítulo V, vemos que em 97% dos casos o resultado de nossos experimentos foram superiores aos de referência. Esses resultados apontam para um ganho substancial ao se comparar diretamente o texto de um recurso novo com os textos de cada tema repetitivo. Isso indica também que, embora dois recursos especiais compartilhem muitas palavras iguais eles podem diferir em seu principal conteúdo semântico.

Elaboramos nossa metodologia ancorada na hipótese de que dado um texto qualquer seria possível gerar um resumo que preservasse informação essencial desse texto (Seção I.5). A forma adotada para avaliar o atingimento deste objetivo específico, e comprovar a veracidade da hipótese, foi verificar se em algum experimento utilizando o resumo do texto conseguimos classificar corretamente um recurso especial conforme a classificação dada por um especialista humano. Quantitativamente a métrica do recall mediu se este objetivo foi alcançado ou não. Em 100% dos experimentos houve resultado positivo ou seja, obtivemos um recall superior a zero. Estes resultados sugerem que é 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 outro objetivo específico que era gerar um resumo que nos proporcionasse um ganho na tarefa de classificação em relação ao uso do texto integral. Para avaliação deste aspecto, 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 efetuara classificação em um tema. Conforme os resultados apresentados no Apêndice D, os quais são exibidos graficamente no Capítulo V, verificamos que é possível suprimir termos de um texto e ainda assim obter um resultado na classificação que seja superior ao resultado obtido utilizando um texto integral.

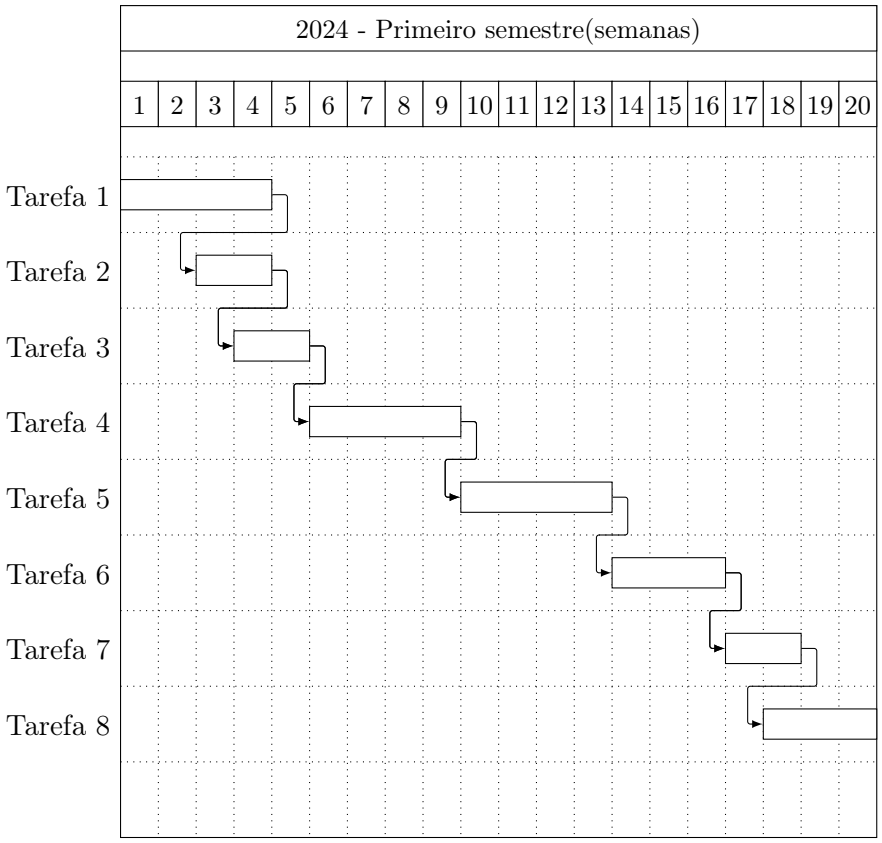
Consideramos o fato da quantidade de dados para os experimentos não ser expressiva, em comparação com a quantidade de dados normalmente utilizada em tarefas de processamento de linguagem natural. Apesar da escassez de dados, o fato de termos uma linha de base utilizando os mesmos dados tornam válidos os experimentos que realizamos.

Uma dificuldade inicial para a geração de resumo extrativo foi encontrar uma ferramenta que não restringisse o tamanho do texto a ser processado, a combinação dos modelos SBERT e BERTopic atenderam bem a este propósito. Buscamos implementar uma variação na abordagem BERTopic tradicional buscando **guiar** o processo de obtenção de tópicos, porém os resultados ficaram aquém do esperado. Um ponto a ser investigado é se ao variar o parâmetro de ponderação, fixo por padrão no código do BERTopic guiado, conseguiremos resultados superiores aos da abordagem BERTopic tradicional.

Um ponto de destaque na execução dos experimentos foi o ganho obtido ao gerar a representação vetorial dos textos e salvá-la em arquivo para uso posterior. A medição do tempo ganho com esta otimização é uma tarefa importante a ser realizada.

Avaliando os resultados dos experimentos, é nítida a superioridade no desempenho ao se adotar o algoritmo LexRank para extração de sentenças combinado com o BM25 para computação de similaridade. Também ficou comprovado nos experimentos que a modificação que implementamos no LexRank, gerando uma variação guiada do algoritmo, redundou em ganhos na tarefa de classificação superando a bordagem com o LexRank tradicional. Um ponto a ser investigado é obtenção de uma proporção ótima entre os fatores que compõem o score combinado.

VI.2 Plano para conclusão



Tarefa 1 → Elaborar e submeter artigo

Tarefa 2 → Analisar recursos classificados incorretamente pelos algoritmos

Tarefa 3 → Implementar medição dos tempos de execução dos algoritmos

Tarefa 4 → Efetuar alteração em hiperparâmetros do LexRank guiado e avaliar resultados

Tarefa 5 → Efetuar alteração em hiperparâmetros do BERTopic guiado e avaliar resultados

Tarefa 6 → Elaborar versão final do texto da pesquisa

Tarefa 7 → Preparar apresentação de defesa da dissertação

Tarefa 8 → Defesa da dissertação

Referências

Lucas E. Resck, Jean R. Ponciano, Luis Gustavo Nonato, and Jorge Poco. Legalvis: Exploring and inferring precedent citations in legal documents. *IEEE Transactions on Visualization and Computer Graphics*, 29(6):3105 – 3120, 2023. doi: 10.1109/TVCG.2022.3152450. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85125329439&doi=10.1109%2fTVCG.2022.3152450&partnerID=40&md5=6086793e9402f3a5b8c3e0e71ba613fe. Cited by: 3; All Open Access, Green Open Access.

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.

Isha Gupta, Indranath Chatterjee, and Neha Gupta. A two-staged nlp-based framework for assessing the sentiments on indian supreme court judgments. *International Journal of Information Technology (Singapore)*, 15(4):2273 – 2282, 2023. doi: 10.1007/s41870-023-01273-z. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85153731773&doi=10.1007%2fs41870-023-01273-z&partnerID=40&md5=34472b9d96632d98ec20b5d5a199b9d6. Cited by: 1; All Open Access, Bronze Open Access, Green Open Access.

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

Constituição. https://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm. (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. (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/. (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).

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

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

:: eproc - consulta processual - busca de processo ::. https://eproc.trf2.jus.br/eproc/externo_controlador.php?acao=processo_consulta_publica, b. (Accessed on 06/05/2023).

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

Charu C. Aggarwal. *Machine Learning for Text, Second Edition*. Springer, 2022. 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.

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.

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 2019b. 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.

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`.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, abs/1908.10084, 2019. URL `http://arxiv.org/abs/1908.10084`.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf`.

Andrew Trotman, Antti Puurula, and Blake Burgess. Improvements to bm25 and language models examined. In *Proceedings of the 19th Australasian Document Computing Symposium*, ADCS '14, page 58–65, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450330008. doi: 10.1145/2682862.2682863. URL `https://doi-org.ez108.periodicos.capes.gov.br/10.1145/2682862.2682863`.

E. Gamma, R. Helm, R. Johnson, and J. Vlissides. *Design Patterns: Elements of Reusable Object-Oriented Software*. Addison-Wesley professional computing series. Addison-Wesley, 1995. ISBN 9783827330437.

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 8.187 registros referentes a recursos especiais. Os recursos estão distribuídos não uniformemente em 191 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 arredondado para inteiro. Representa a probabilidade da primeira sugestão dada ser a correta, segundo o Elasticsearch.
- **Probabilidade_2:** Dado numérico arredondado para inteiro. Representa a probabilidade da segunda sugestão dada ser a correta, segundo o Elasticsearch.
- **Probabilidade_3:** Dado numérico arredondado para inteiro. Representa a probabilidade da terceira sugestão dada ser a correta, segundo o Elasticsearch.
- **Probabilidade_4:** Dado numérico arredondado para inteiro. Representa a probabilidade da quarta sugestão dada ser a correta, segundo o Elasticsearch.
- **Probabilidade_5:** Dado numérico arredondado para inteiro. Representa a probabilidade da quinta sugestão dada ser a correta, segundo o Elasticsearch.
- **Probabilidade_6:** Dado numérico arredondado para inteiro. 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 Elastisearch era a correta.

Apêndice D Resultados dos experimentos

Os dados de todos os 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 (Figura V.1).

origem	recall	f1-score	map	ndcg	mrr
lexrank_sentencas_50_bm25	0,73026	0,58523	0,48827	0,54842	0,48827
lexrank_sentencas_55_bm25	0,72813	0,58856	0,49389	0,55247	0,49389
lexrank_sentencas_60_bm25	0,72574	0,59958	0,51078	0,56436	0,51078
lexrank_sentencas_45_bm25	0,72323	0,57969	0,48369	0,54366	0,48369
lexrank_sentencas_60_remocao_bm25	0,72148	0,60145	0,51567	0,56725	0,51567
lexrank_sentencas_55_remocao_bm25	0,72123	0,60019	0,51394	0,56591	0,51394
lexrank_sentencas_50_remocao_bm25	0,71846	0,59487	0,50756	0,56035	0,50756
lexrank_sentencas_40_bm25	0,71570	0,56446	0,46599	0,52829	0,46599
texto_bm25	0,71520	0,59771	0,51337	0,56385	0,51337
texto_remocao_bm25	0,71457	0,60415	0,52329	0,57144	0,52329
lexrank_sentencas_45_remocao_bm25	0,71181	0,56483	0,46816	0,52863	0,46816
lexrank_sentencas_35_bm25	0,70980	0,55538	0,45614	0,51859	0,45614
lexrank_sentencas_30_bm25	0,70365	0,54730	0,44780	0,51117	0,44780
lexrank_sentencas_35_remocao_bm25	0,68997	0,54922	0,45617	0,51388	0,45617
lexrank_sentencas_40_remocao_bm25	0,67491	0,54348	0,45489	0,50962	0,45489
lexrank_sentencas_30_remocao_bm25	0,66838	0,53013	0,43927	0,49620	0,43927
lexrank_sentencas_25_bm25	0,66148	0,52404	0,43389	0,49050	0,43389
lexrank_sentencas_20_bm25	0,65545	0,51995	0,43088	0,48651	0,43088
lexrank_sentencas_15_bm25	0,64340	0,49161	0,39777	0,45829	0,39777
lexrank_sentencas_25_remocao_bm25	0,63173	0,50119	0,41536	0,46897	0,41536
lexrank_sentencas_20_remocao_bm25	0,62420	0,48668	0,39881	0,45442	0,39881
lexrank_sentencas_15_remocao_bm25	0,62181	0,47468	0,38385	0,44229	0,38385
lexrank_sentencas_10_bm25	0,58931	0,45606	0,37196	0,42605	0,37196
bertopic_palavras_55_remocao_bm25	0,56960	0,45524	0,37912	0,42662	0,37912
lexrank_sentencas_10_remocao_bm25	0,56245	0,43348	0,35263	0,40491	0,35263
bertopic_palavras_25_remocao_bm25	0,55843	0,45078	0,37793	0,42300	0,37793
bertopic_palavras_50_remocao_bm25	0,55579	0,45073	0,37907	0,42319	0,37907
bertopic_guiado_palavras_60_remocao_bm25	0,55517	0,45252	0,38190	0,42535	0,38190
bertopic_guiado_palavras_60_bm25	0,55265	0,45007	0,37961	0,42284	0,37961
bertopic_guiado_palavras_45_remocao_bm25	0,55190	0,44485	0,37258	0,41751	0,37258
bertopic_palavras_25_bm25	0,55065	0,44520	0,37365	0,41796	0,37365
bertopic_palavras_60_bm25	0,55065	0,44761	0,37706	0,42060	0,37706
bertopic_guiado_palavras_55_remocao_bm25	0,54876	0,45090	0,38266	0,42422	0,38266
bertopic_palavras_35_bm25	0,54789	0,44464	0,37414	0,41788	0,37414
bertopic_guiado_palavras_40_remocao_bm25	0,54738	0,44812	0,37933	0,42115	0,37933
bertopic_guiado_palavras_30_remocao_bm25	0,54563	0,44532	0,37616	0,41872	0,37616
bertopic_guiado_palavras_35_remocao_bm25	0,54563	0,45338	0,38781	0,42747	0,38781
bertopic_palavras_50_bm25	0,54525	0,45313	0,38764	0,42739	0,38764
bertopic_palavras_45_remocao_bm25	0,54500	0,44973	0,38282	0,42350	0,38282
bertopic_palavras_35_remocao_bm25	0,54462	0,44175	0,37156	0,41503	0,37156
bertopic_palavras_40_remocao_bm25	0,54349	0,45111	0,38557	0,42522	0,38557
bertopic_guiado_palavras_50_remocao_bm25	0,54286	0,43455	0,36227	0,40727	0,36227
bertopic_guiado_palavras_50_bm25	0,54249	0,44507	0,37731	0,41876	0,37731
bertopic_palavras_55_bm25	0,54199	0,44049	0,37101	0,41401	0,37101
bertopic_guiado_palavras_45_bm25	0,54111	0,44479	0,37758	0,41884	0,37758
bertopic_palavras_30_remocao_bm25	0,54060	0,44611	0,37973	0,42010	0,37973
bertopic_palavras_45_bm25	0,53935	0,44376	0,37696	0,41783	0,37696
bertopic_guiado_palavras_20_bm25	0,53797	0,43304	0,36236	0,40633	0,36236
bertopic_guiado_palavras_35_bm25	0,53772	0,43917	0,37115	0,41321	0,37115
bertopic_palavras_40_bm25	0,53747	0,43398	0,36391	0,40763	0,36391
bertopic_guiado_palavras_20_remocao_bm25	0,53734	0,44227	0,37579	0,41618	0,37579

origem	recall	f1-score	map	ndcg	mrr
bertopic_guiado_palavras_40_bm25	0,53395	0,43844	0,37191	0,41283	0,37191
bertopic_palavras_60_remocao_bm25	0,53132	0,43295	0,36532	0,40685	0,36532
bertopic_guiado_palavras_15_remocao_bm25	0,52692	0,43334	0,36798	0,40792	0,36798
bertopic_palavras_15_remocao_bm25	0,52655	0,42722	0,35942	0,40144	0,35942
bertopic_guiado_palavras_25_remocao_bm25	0,52416	0,42575	0,35845	0,39961	0,35845
bertopic_guiado_palavras_15_bm25	0,51952	0,40846	0,33652	0,38166	0,33652
bertopic_palavras_20_bm25	0,51952	0,42650	0,36173	0,40142	0,36173
bertopic_guiado_palavras_55_bm25	0,51073	0,41266	0,34618	0,38725	0,34618
bertopic_palavras_20_remocao_bm25	0,50521	0,40881	0,34330	0,38372	0,34330
bertopic_guiado_palavras_30_bm25	0,50270	0,40668	0,34146	0,38188	0,34146
bertopic_guiado_palavras_25_bm25	0,50257	0,41487	0,35323	0,39054	0,35323
bertopic_palavras_30_bm25	0,50132	0,41078	0,34794	0,38634	0,34794
bertopic_palavras_45_remocao_cosseno	0,47923	0,37009	0,30144	0,34546	0,30144
bertopic_palavras_40_remocao_cosseno	0,46768	0,36250	0,29595	0,33893	0,29595
bertopic_palavras_15_bm25	0,46743	0,38241	0,32357	0,35936	0,32357
bertopic_palavras_45_cosseno	0,46379	0,36052	0,29487	0,33728	0,29487
bertopic_palavras_35_remocao_cosseno	0,46203	0,35934	0,29400	0,33626	0,29400
bertopic_guiado_palavras_45_cosseno	0,46165	0,35662	0,29052	0,33335	0,29052
bertopic_guiado_palavras_45_remocao_cosseno	0,46128	0,35516	0,28873	0,33222	0,28873
bertopic_guiado_palavras_10_remocao_bm25	0,46090	0,37282	0,31300	0,35016	0,31300
bertopic_palavras_50_remocao_cosseno	0,46027	0,35613	0,29042	0,33250	0,29042
bertopic_guiado_palavras_40_cosseno	0,46015	0,35601	0,29030	0,33287	0,29030
bertopic_palavras_55_remocao_cosseno	0,45952	0,37263	0,31337	0,34937	0,31337
bertopic_guiado_palavras_35_cosseno	0,45827	0,34444	0,27592	0,32153	0,27592
bertopic_guiado_palavras_35_remocao_cosseno	0,45714	0,35850	0,29488	0,33568	0,29488
bertopic_palavras_55_cosseno	0,45663	0,35313	0,28788	0,33016	0,28788
bertopic_palavras_50_cosseno	0,45626	0,35883	0,29569	0,33609	0,29569
bertopic_palavras_60_cosseno	0,45575	0,35977	0,29718	0,33690	0,29718
bertopic_palavras_35_cosseno	0,45538	0,34936	0,28339	0,32671	0,28339
bertopic_guiado_palavras_50_cosseno	0,45425	0,35537	0,29184	0,33289	0,29184
bertopic_palavras_40_cosseno	0,45387	0,35600	0,29286	0,33350	0,29286
bertopic_guiado_palavras_60_cosseno	0,45312	0,35432	0,29089	0,33124	0,29089
bertopic_guiado_palavras_55_remocao_cosseno	0,45161	0,35194	0,28831	0,32908	0,28831
bertopic_palavras_10_remocao_bm25	0,45011	0,36197	0,30270	0,33979	0,30270
bertopic_palavras_15_remocao_cosseno	0,44898	0,33582	0,26823	0,31340	0,26823
bertopic_guiado_palavras_10_bm25	0,44785	0,36347	0,30584	0,34122	0,30584
bertopic_guiado_palavras_25_cosseno	0,44760	0,32240	0,25194	0,30061	0,25194
bertopic_guiado_palavras_60_remocao_cosseno	0,44760	0,36701	0,31102	0,34437	0,31102
bertopic_palavras_10_bm25	0,44735	0,36595	0,30962	0,34430	0,30962
bertopic_palavras_30_remocao_cosseno	0,44546	0,34737	0,28468	0,32510	0,28468
bertopic_guiado_palavras_30_remocao_cosseno	0,44220	0,34722	0,28583	0,32526	0,28583
bertopic_guiado_palavras_40_remocao_cosseno	0,44207	0,33837	0,27408	0,31633	0,27408
bertopic_guiado_palavras_15_remocao_cosseno	0,44094	0,32664	0,25940	0,30467	0,25940
bertopic_palavras_25_remocao_cosseno	0,44082	0,33950	0,27606	0,31730	0,27606
bertopic_palavras_25_cosseno	0,44019	0,33752	0,27369	0,31547	0,27369

origem	recall	f1-score	map	ndcg	mrr
bertopic_palavras_20_cosseno	0,43567	0,33741	0,27531	0,31540	0,27531
bertopic_palavras_30_cosseno	0,43542	0,32239	0,25595	0,30048	0,25595
bertopic_guiado_palavras_20_cosseno	0,43228	0,33350	0,27147	0,31189	0,27147
bertopic_guiado_palavras_30_cosseno	0,43166	0,31922	0,25325	0,29795	0,25325
bertopic_guiado_palavras_50_remocao_cosseno	0,43166	0,34014	0,28064	0,31832	0,28064
bertopic_guiado_palavras_25_remocao_cosseno	0,43053	0,32874	0,26588	0,30717	0,26588
bertopic_guiado_palavras_55_cosseno	0,42061	0,32462	0,26430	0,30353	0,26430
bertopic_palavras_20_remocao_cosseno	0,41760	0,30654	0,24214	0,28557	0,24214
lexrank_sentencas_30_cosseno	0,41760	0,30358	0,23846	0,28272	0,23846
lexrank_sentencas_35_cosseno	0,41760	0,30373	0,23865	0,28287	0,23865
lexrank_sentencas_40_cosseno	0,41760	0,30373	0,23865	0,28287	0,23865
lexrank_sentencas_45_cosseno	0,41760	0,30373	0,23865	0,28287	0,23865
lexrank_sentencas_50_cosseno	0,41760	0,30373	0,23865	0,28287	0,23865
lexrank_sentencas_55_cosseno	0,41760	0,30373	0,23865	0,28287	0,23865
lexrank_sentencas_60_cosseno	0,41760	0,30373	0,23865	0,28287	0,23865
lexrank_sentencas_20_cosseno	0,41710	0,30354	0,23859	0,28269	0,23859
lexrank_sentencas_25_cosseno	0,41684	0,30339	0,23849	0,28255	0,23849
lexrank_sentencas_15_cosseno	0,41622	0,30384	0,23925	0,28296	0,23925
lexrank_sentencas_10_cosseno	0,41584	0,30310	0,23845	0,28227	0,23845
bertopic_palavras_60_remocao_cosseno	0,40944	0,33425	0,28239	0,31325	0,28239
bertopic_guiado_palavras_20_remocao_cosseno	0,40592	0,31376	0,25570	0,29364	0,25570
bertopic_guiado_palavras_15_cosseno	0,38986	0,27586	0,21345	0,25755	0,21345
bertopic_palavras_15_cosseno	0,37593	0,29044	0,23663	0,27121	0,23663
bertopic_palavras_10_remocao_cosseno	0,36388	0,28014	0,22774	0,26083	0,22774
lexrank_sentencas_20_remocao_cosseno	0,35434	0,27004	0,21815	0,25191	0,21815
lexrank_sentencas_35_remocao_cosseno	0,35409	0,26983	0,21797	0,25172	0,21797
lexrank_sentencas_40_remocao_cosseno	0,35409	0,26984	0,21797	0,25172	0,21797
lexrank_sentencas_45_remocao_cosseno	0,35409	0,26984	0,21797	0,25172	0,21797
lexrank_sentencas_50_remocao_cosseno	0,35409	0,26984	0,21797	0,25172	0,21797
lexrank_sentencas_55_remocao_cosseno	0,35409	0,26984	0,21797	0,25172	0,21797
lexrank_sentencas_60_remocao_cosseno	0,35409	0,26984	0,21797	0,25172	0,21797
lexrank_sentencas_25_remocao_cosseno	0,35396	0,26975	0,21791	0,25164	0,21791
lexrank_sentencas_30_remocao_cosseno	0,35383	0,26972	0,21791	0,25161	0,21791
lexrank_sentencas_15_remocao_cosseno	0,35233	0,26975	0,21854	0,25170	0,21854
lexrank_sentencas_10_remocao_cosseno	0,35132	0,26773	0,21627	0,24970	0,21627
elasticsearch_bm25	0,34906	0,32746	0,30838	0,31824	0,30838
bertopic_guiado_palavras_10_remocao_cosseno	0,34794	0,25772	0,20465	0,23994	0,20465
bertopic_guiado_palavras_10_cosseno	0,32208	0,24945	0,20355	0,23247	0,20355
bertopic_palavras_10_cosseno	0,30538	0,23038	0,18495	0,21478	0,18495
texto_cosseno	0,14372	0,10870	0,08740	0,10128	0,08740
texto_remocao_cosseno	0,13631	0,10726	0,08842	0,09989	0,08842

origem	recall	f1-score	map	ndcg	mrr
lexrank_guiado_sentencas_50_bm25	0,74683	0,62581	0,53854	0,59079	0,53854
lexrank_guiado_sentencas_55_bm25	0,74558	0,62109	0,53222	0,58547	0,53222
lexrank_guiado_sentencas_60_bm25	0,74256	0,62083	0,53339	0,58560	0,53339
lexrank_guiado_sentencas_45_bm25	0,75399	0,62958	0,54041	0,59385	0,54041
lexrank_guiado_sentencas_40_bm25	0,74972	0,62843	0,54092	0,59326	0,54092
lexrank_guiado_sentencas_50_cosseno	0,37982	0,28043	0,22227	0,26108	0,22227
lexrank_guiado_sentencas_55_cosseno	0,37982	0,28043	0,22227	0,26108	0,22227
lexrank_guiado_sentencas_60_cosseno	0,37982	0,28043	0,22227	0,26108	0,22227
lexrank_guiado_sentencas_45_cosseno	0,37982	0,28043	0,22227	0,26108	0,22227
lexrank_guiado_sentencas_40_cosseno	0,37982	0,28043	0,22227	0,26108	0,22227
lexrank_guiado_sentencas_10_bm25	0,73591	0,61106	0,52243	0,57586	0,52243
lexrank_guiado_sentencas_15_bm25	0,74419	0,62124	0,53315	0,58599	0,53315
lexrank_guiado_sentencas_20_bm25	0,74708	0,62615	0,53892	0,59101	0,53892
lexrank_guiado_sentencas_25_bm25	0,75424	0,62969	0,54045	0,59387	0,54045
lexrank_guiado_sentencas_30_bm25	0,75449	0,63608	0,54980	0,60090	0,54980
lexrank_guiado_sentencas_35_bm25	0,75198	0,62901	0,54061	0,59355	0,54061
lexrank_guiado_sentencas_10_cosseno	0,37982	0,28043	0,22227	0,26108	0,22227
lexrank_guiado_sentencas_15_cosseno	0,37982	0,28043	0,22227	0,26108	0,22227
lexrank_guiado_sentencas_20_cosseno	0,37982	0,28043	0,22227	0,26108	0,22227
lexrank_guiado_sentencas_25_cosseno	0,37982	0,28043	0,22227	0,26108	0,22227
lexrank_guiado_sentencas_30_cosseno	0,37982	0,28043	0,22227	0,26108	0,22227
lexrank_guiado_sentencas_35_cosseno	0,37982	0,28043	0,22227	0,26108	0,22227
lexrank_guiado_sentencas_10_remocao_bm25	0,75424	0,62951	0,54018	0,59362	0,54018
lexrank_guiado_sentencas_15_remocao_bm25	0,75750	0,62679	0,53455	0,59018	0,53455
lexrank_guiado_sentencas_20_remocao_bm25	0,75712	0,62274	0,52887	0,58575	0,52887
lexrank_guiado_sentencas_25_remocao_cosseno	0,47044	0,38577	0,32693	0,36259	0,32693
lexrank_guiado_sentencas_30_remocao_cosseno	0,47044	0,38577	0,32693	0,36259	0,32693
lexrank_guiado_sentencas_35_remocao_bm25	0,75524	0,63064	0,54133	0,59485	0,54133
lexrank_guiado_sentencas_10_remocao_cosseno	0,47044	0,38577	0,32693	0,36259	0,32693
lexrank_guiado_sentencas_15_remocao_cosseno	0,47044	0,38577	0,32693	0,36259	0,32693
lexrank_guiado_sentencas_20_remocao_cosseno	0,47044	0,38577	0,32693	0,36259	0,32693
lexrank_guiado_sentencas_25_remocao_bm25	0,74633	0,62479	0,53730	0,58981	0,53730
lexrank_guiado_sentencas_30_remocao_bm25	0,75574	0,63048	0,54084	0,59461	0,54084
lexrank_guiado_sentencas_35_remocao_cosseno	0,47044	0,38577	0,32693	0,36259	0,32693
lexrank_guiado_sentencas_40_remocao_bm25	0,74545	0,62573	0,53914	0,59090	0,53914
lexrank_guiado_sentencas_45_remocao_bm25	0,74131	0,62331	0,53772	0,58887	0,53772
lexrank_guiado_sentencas_50_remocao_bm25	0,74194	0,62369	0,53795	0,58920	0,53795
lexrank_guiado_sentencas_55_remocao_bm25	0,74093	0,62270	0,53700	0,58820	0,53700
lexrank_guiado_sentencas_60_remocao_bm25	0,74005	0,62223	0,53677	0,58779	0,53677
lexrank_guiado_sentencas_40_remocao_cosseno	0,47044	0,38577	0,32693	0,36259	0,32693
lexrank_guiado_sentencas_45_remocao_cosseno	0,47044	0,38577	0,32693	0,36259	0,32693
lexrank_guiado_sentencas_50_remocao_cosseno	0,47044	0,38577	0,32693	0,36259	0,32693
lexrank_guiado_sentencas_55_remocao_cosseno	0,47044	0,38577	0,32693	0,36259	0,32693
lexrank_guiado_sentencas_60_remocao_cosseno	0,47044	0,38577	0,32693	0,36259	0,32693