Instituto Serzedello Corrêa

Inovação na Pesquisa de Jurisprudência Selecionada do TCU:

Expansão de Documentos usando Modelos de Linguagem

Leandro Carísio Fernandes

Edans Flávius de Oliveira Sandes

Coletânea de Pós-Graduação
Controle Governamental:
Tecnologias para Inovação

TRIBUNAL DE CONTAS DA UNIÃO

REPÚBLICA FEDERATIVA DO BRASIL

TRIBUNAL DE CONTAS DA UNIÃO

MINISTROS

Bruno Dantas (Presidente)
Vital do Rêgo Filho (Vice-Presidente)
Walton Alencar Rodrigues
Benjamin Zymler
João Augusto Ribeiro Nardes
Aroldo Cedraz de Oliveira
Jorge Antônio de Oliveira Francisco
Antonio Augusto Junho Anastasia
Jhonatan de Jesus

MINISTROS-SUBSTITUTOS

Augusto Sherman Cavalcanti
Marcos Bemquerer Costa
Weder de Oliveira

MINISTÉRIO PÚBLICO JUNTO AO TCU

Cristina Machado da Costa e Silva (Procuradora-Geral)
Lucas Furtado (Subprocurador-Geral)
Paulo Soares Bugarin (Subprocurador-Geral)
Marinus Eduardo de Vries Marsico (Procurador)
Júlio Marcelo de Oliveira (Procurador)
Sérgio Ricardo Costa Caribé (Procurador)
Rodrigo Medeiros de Lima (Procurador

Instituto Serzedello Corrêa
Escola Superior do Tribunal de Contas da União

Inovação na Pesquisa de Jurisprudência Selecionada do TCU:
Expansão de Documentos usando Modelos de Linguagem

Leandro Carísio Fernandes

Artigo de conclusão de curso submetido ao Instituto Serzedello Corrêa do Tribunal de Contas da União como requisito parcial para a obtenção do grau de especialista.

Orientador(a):
Prof. Edans Flávius de Oliveira Sandes

Banca examinadora:
Prof. Eric Hans Messias da Silva

Brasília | 2025

REFERÊNCIA BIBLIOGRÁFICA

Fernandes, Leandro Carísio. **Inovação na Pesquisa de Jurisprudência Selecionada do TCU:** Expansão de Documentos usando Modelos de Linguagem. 2025. Monografia (Especialização em Avaliação de Políticas Públicas) – Instituto Serzedello Corrêa, Escola Superior do Tribunal de Contas da União, Brasília DF. 44fl.

CESSÃO DE DIREITOS

NOME DO AUTOR: Leandro Carísio Fernandes
TÍTULO: Inovação na Pesquisa de Jurisprudência Selecionada do TCU: Expansão de Documentos usando Modelos de Linguagem
GRAU/ANO: Especialista/2025

_____

Leandro Carísio Fernandes
carisio@gmail.com

FICHA CATALOGRÁFICA

L131a Fernandes, Leandro Carísio

Inovação na Pesquisa de Jurisprudência Selecionada do TCU: Expansão de Documentos usando Modelos de Linguagem / Leandro Carísio Fernandes. – Brasília: ISC/TCU, 2025.
44 fl. (Artigo de Especialização)

1. Controle Governamental: Tecnologias para Inovação. 2. Tema 2. 3. Tema 3. I. Título.

CDU 02
CDD 020

Inovação na Pesquisa de Jurisprudência Selecionada do TCU: Expansão de Documentos usando Modelos de Linguagem

Leandro Carísio Fernandes

Trabalho de conclusão do curso de pós-graduação *lato sensu* em Controle Govenamental: Tecnologias para Inovação realizado pelo Instituto Serzedello Corrêa como requisito para a obtenção do título de especialista.

Brasília, 20 de janeiro de 2025.

Banca Examinadora:

_____
Prof. Edans Flávius de Oliveira Sandes
Orientador
Tribunal de Contas da União

_____
Prof. Eric Hans Messias da Silva, MSc.
Avaliador
Tribunal de Contas da União

Resumo

Este estudo explora o uso de técnicas de expansão de documentos para aprimorar o sistema de busca de jurisprudência selecionada do Tribunal de Contas da União (TCU). Para isso, foram testadas quatro abordagens de expansão de documentos: 1 usando o método docT5query; 2 por meio de extração de sinônimos do enunciado; 3 através da reescrita do enunciado; e 4 combinando docT5query com a extração de sinônimos. Os experimentos foram realizados com o uso de *queries* reais e sintéticas, extraídas da base de jurisprudência selecionada do TCU, avaliando a eficácia do modelo BM25 com os documentos expandidos com as técnicas testadas. O estudo conclui que essas abordagens, se implementadas na pesquisa de jurisprudência selecionada do TCU, podem melhorar consideravelmente a sua qualidade. Os melhores resultados foram obtidos pela combinação do método docT5query com a extração de sinônimos, atingindo mais de 60% de aumento nas métricas $MRR@5$ e $nDCG@5$ quando testados com *queries* reais feitas por usuários.

Palavras-chave: Recuperação de Informação; docT5query; Sinônimos; Descasamento de Vocabulário; BM25.

Abstract

This study addresses the use of document expansion techniques to improve the selected jurisprudence search system of the Brazilian Federal Court of Accounts (TCU). Four document expansion techniques were tested: 1. using the docT5query method; 2. extracting synonyms from the document's summary; 3. rewriting the summary; and 4. combining docT5query with synonym extraction. Experiments were conducted using real and synthetic queries extracted from the TCU's selected jurisprudence system, evaluating the effectiveness of the BM25 model with the documents expanded using the proposed techniques. The study concludes that implementing these approaches in the TCU's selected jurisprudence search system can significantly enhance its quality. The best results were achieved by combining the docT5query method with synonym extraction, achieving over 60% improvements in $MRR@5$ and $nDCG@5$ *metrics when tested with real user-generated queries.*

Keywords*: Information Retrieval; docT5query; Synonyms; Vocabulary Mismatch; BM25.*

Lista de figuras

Lista de quadros

Lista de tabelas

Sumário

1. Introdução

O Tribunal de Contas da União (TCU) é um órgão essencial para o controle externo da Administração Pública. Em sua atuação, o TCU acumula uma vasta jurisprudência, com mais de 500.000 documentos que refletem a interpretação da legislação e a aplicação de princípios jurídicos em casos concretos.

Para facilitar a consulta de sua jurisprudência, o órgão disponibiliza a base de dados de jurisprudência selecionada[1], em que cada documento é estruturado ao redor de um enunciado elaborado por órgão especializado do Tribunal. Os enunciados buscam retratar o entendimento contido na deliberação da qual ele foi extraído.

Essa base facilita o acesso do público à jurisprudência do TCU, visto que condensa em pouco mais de 16 mil documentos decisões importantes do órgão. Assim, é uma base muito relevante para a tomada de decisões e para o apoio ao controle externo.

Além da base de dados, o Tribunal oferece também um sistema de pesquisa[2] que facilita ao cidadão a consulta a esses documentos. Entretanto, há limitações na ferramenta. Em uma recente análise, foram identificadas dificuldades no mecanismo de busca utilizado. Essas limitações chegam ao ponto de alguns usuários preferirem usar o Google à ferramenta disponibilizada, o que é um indicativo claro de que há espaço para melhorias da aplicação (Brasil, 2022).

A pesquisa disponibilizada é baseada em palavras-chave, e um dos problemas conhecidos é o descasamento de vocabulário, que ocorre quando o usuário utiliza, como termos de pesquisa, palavras diversas daquelas contidas nos documentos, mesmo que semanticamente próximas (Wang et al, 2024). Isso é algo comum quando pesquisas que possuem vocabulários muito específicos, como as pesquisas jurídicas, são disponibilizadas para um público amplo. Esse descompasso afeta a qualidade dos resultados obtidos, já que a busca é baseada em correspondências léxicas exatas.

A literatura especializada apresenta algumas alternativas para mitigar esse problema (Lin, Nogueira e Yates, 2021). A proposta deste estudo concentra-se em analisar uma das possíveis técnicas na base de jurisprudência selecionada – a expansão de documentos (também chamada de enriquecimento de documentos). A estratégia envolve o enriquecimento dos documentos com novos textos, que podem tanto reforçar o peso de palavras já existentes no documento, como também incluir novos termos ou expressões, aumentando a probabilidade de que o sistema de busca recupere informações relevantes, mesmo que o usuário use em sua pesquisa palavras diferentes das contidas inicialmente nos documentos.

A escolha dessa abordagem é justificada pelo fato de exigir pouca alteração na arquitetura atual da pesquisa de jurisprudência selecionada, permitindo a sua evolução gradual.

---

[1] https://pesquisa.apps.tcu.gov.br/dados-abertos
[2] https://pesquisa.apps.tcu.gov.br/pesquisa/jurisprudencia-selecionada

Essas técnicas, embora recentes, já possuem bons resultados divulgados, especialmente para bases de dados disponíveis em língua inglesa. Uma questão pendente é saber se são métodos efetivos também em língua portuguesa e em que medida podem ser aplicados em bases de dados com contextos mais específicos, como as bases com documentos de origem legal ou jurisprudencial.

Assim, o objetivo deste trabalho é aplicar técnicas de expansão de documentos na pesquisa de jurisprudência selecionada para verificar possibilidades de melhoria em sua qualidade. Nesse contexto, são contribuições desse estudo: 1. a avaliação da qualidade da pesquisa atual de jurisprudência selecionada; 2. a aplicação de diferentes abordagens de expansão de documentos nesta base; 3. a aferição dessas abordagens, com a identificação da mais efetiva; e 4. o fornecimento de recomendações para melhorias na pesquisa de jurisprudência selecionada do TCU.

O artigo está estruturado em 5 seções. A seção seguinte apresenta uma revisão da literatura com os principais conceitos da área de recuperação de informação. A seção 3 descreve o método de estudo para aplicação de abordagens de expansão de documento na pesquisa de jurisprudência selecionada. Os resultados são discutidos na seção 4 e, por fim, as conclusões são apresentadas na última seção.

2. Revisão da Literatura

2.1. Conceitos de Recuperação de Informações

Recuperação de informações é uma área da ciência da computação que trabalha com a representação da informação, o seu armazenamento, a sua organização e o seu acesso para facilitar que o usuário encontre informações de seu interesse em um conjunto de documentos (Baeza-Yates e Ribeiro-Neto, 1999).

Um sistema de recuperação de informações é formado pelo processo de indexação e de consulta (Croft, Metzler e Strohman, 2009). No processo de indexação (

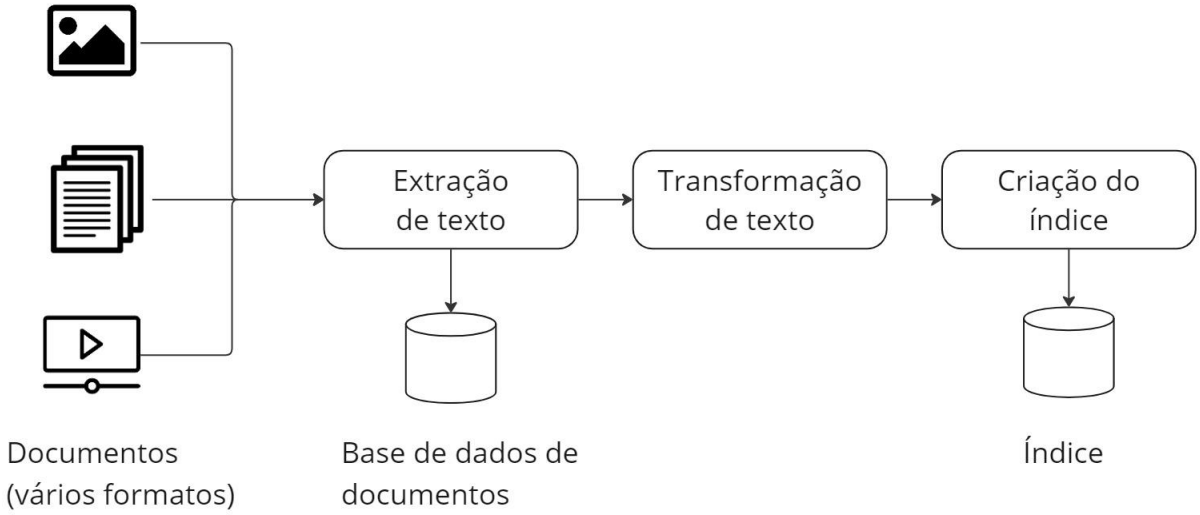
Figura 1), um conjunto de documentos é preparado para que possa ser consultado futuramente. A indexação tem as seguintes fases:

1. Extração de textos: Nesta etapa é feita a seleção e obtenção dos documentos que serão usados para consulta. Pode envolver um componente *crawler* para identificar e copiar documentos da Web, pode ser feita pelo acesso direto a bancos de dados ou sistemas de arquivos etc. Uma vez adquiridos, os documentos normalmente são armazenados para que possam ser retornados em uma consulta.

2. Transformação de textos: Além de serem armazenados, os documentos também são adaptados para serem inseridos em um índice para consulta. A transformação inclui o *parsing* e a limpeza do documento, extraindo apenas o conteúdo que será utilizado na busca.

3. Criação do índice: O índice é a estrutura de dados que armazena as informações que serão retornadas na pesquisa. Nesta etapa, deve-se criar um índice adequado para a aplicação. Para pesquisas textuais são utilizados principalmente dois índices, o índice invertido (para buscas baseadas em palavras-chave) e o índice vetorial (quando o documento e os termos de consulta são representados por vetores numéricos).

Figura 1: Processo de indexação



Fonte: Adaptado de Croft, Metzler e Strohman (2009).

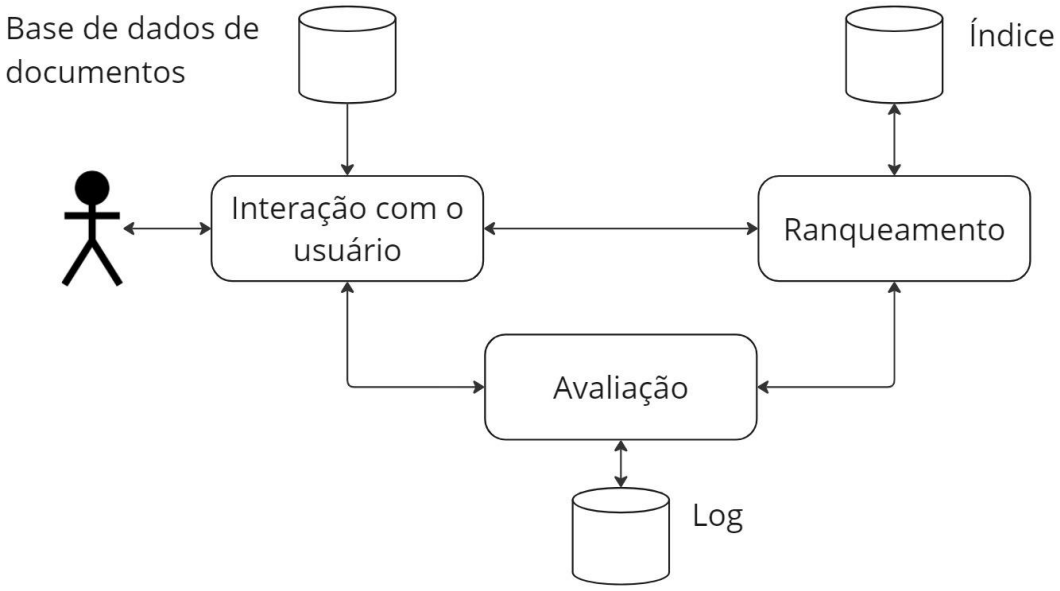
Após a indexação, o usuário poderá utilizar um motor de busca para encontrar uma informação. Essa interação do usuário com a ferramenta é feita no processo de consulta (Figura 2), que é composto por:

1. Interação com o usuário: Nesta etapa é disponibilizada uma interface para que o usuário encontre a sua necessidade de informação. O usuário traduz a sua necessidade de informação para um texto (*query*), que é utilizado na interação com o sistema.

2. Ranqueamento: A *query* é enviada ao motor de busca, que acessa o índice de pesquisa e retorna os documentos mais similares relacionados à *query* do usuário. Em seguida, a *query* é pré-processada conforme cada caso (por exemplo, remoção de palavras comuns, conversão para formato vetorial etc). Por fim, os documentos são ordenados de acordo com algum critério de relevância.

3. Avaliação: A *query* e o resultado da busca podem ser enviados para uma etapa de avaliação. Essas relações podem ser salvas em um *log* de pesquisa para posteriormente serem utilizadas na evolução do sistema.

Figura 2: Processo de consulta



Fonte: Adaptado de Croft, Metzler e Strohman (2009).

2.2. Modelos de Recuperação de Informações

Formalmente um modelo de recuperação de informações é definido como uma quádrupla $\{\boldsymbol{D}, \boldsymbol{Q}, F, rel(Q, D)\}$, em que (Baeza-Yates e Ribeiro-Neto, 1999):

- $\boldsymbol{D}$ é o conjunto de documentos na coleção que será indexada;
- $\boldsymbol{Q}$ é o conjunto de necessidades de informações dos usuários representadas na forma de *queries*;
- F é o *framework* que modela a representação dos documentos, *queries* e como eles se relacionam;
- $rel(Q, D)$ é uma função de ranqueamento que associa um número real a um par formado por uma *query* Q e um documento D, onde $D \in \boldsymbol{D}$ e $Q \in \boldsymbol{Q}$. O número retornado pela função de ranqueamento indica o grau de relevância que D tem para Q. Na prática, os valores de $rel(Q, D)$ são utilizados apenas para a ordenação dos documentos retornados.

2.2.1. Avaliação de Sistemas de Recuperação de Informações

O aprimoramento e a evolução de sistemas de recuperação de informações dependem de uma forma de medir a qualidade dos resultados da pesquisa. No início dos anos 90, o NIST (*National Institute of Standards and Technology*) começou a promover as conferências TREC (*Text REtrieval Conference*), criadas para incentivar a pesquisa nesta área de conhecimento. Para isso, o instituto fornecia um conjunto de dados (*dataset*) para testes e procedimentos uniformes de avaliação de sistemas de busca (Baeza-Yates e Ribeiro-Neto, 1999). Com o tempo, foram disponibilizados

diversos *datasets*, a maioria em língua inglesa, que ainda são usados na avaliação de sistemas de recuperação de informação.

Para a avaliação de um sistema de pesquisa, são utilizados outros dois conjuntos de dados além do *dataset* com os documentos: 1. Um conjunto de *queries* padronizadas; e 2. Um conjunto de pares de critérios de relevância entre as *queries* padronizadas e os documentos (qrels – *query relevance judgments*). Com isso, a avaliação se dá comparando a ordenação de documentos retornados pelo sistema com a ordenação esperada no qrels.

Apesar de haver diversos *datasets* padronizados, eles normalmente estão disponíveis em língua inglesa. Há uma carência de *datasets* em língua portuguesa exclusivos para testes de sistemas de recuperação de informação. Devido à política de dados abertos da Administração Pública, há bases de dados em português, especialmente no campo legal e jurisprudencial. Entretanto, normalmente essas bases não trazem associado um qrels anotado, não sendo apropriados para a tarefa de testes de sistemas de recuperação de informações. Alguns exemplos de coleção de documentos sem essa anotação são as disponibilizadas por Junior et al (2023) e Siqueira et al (2024). Apesar disso, há esforços nessa área em português, como por exemplo o Quati (Bueno et al, 2024), que disponibiliza um *dataset* contendo trechos de texto curtos juntamente com o qrels.

De posse de um *dataset* de avaliação e do qrels, é necessário utilizar métricas para medir aspectos da qualidade da pesquisa. Há várias métricas que podem ser usadas, sendo as duas mais comuns a precisão e a revocação (*recall*), definidas como (Croft, Metzler e Strohman, 2009; Manning, Eaghavan e Schùtze, 2008; Mitra e Craswell, 2018):

$$P = \frac{|\boldsymbol{D}_{relevant} \cap \boldsymbol{D}_{retrieved}|}{|\boldsymbol{D}_{retrieved}|}$$

$$R = \frac{|\boldsymbol{D}_{relevant} \cap \boldsymbol{D}_{retrieved}|}{|\boldsymbol{D}_{relevant}|}$$

Em que P e R são a precisão e a revocação do sistema quando uma *query* Q é pesquisada, e $\boldsymbol{D}_{relevant}$ e $\boldsymbol{D}_{retrieved}$ são, respectivamente, o conjunto de documentos relevantes referentes aquela *query* e o conjunto de documentos retornados pelo sistema de busca.

Em sistemas de busca em Web, é comum analisar apenas os primeiros k resultados. Dessa forma, fala-se em precisão em k ($P@k$) e revocação em k ($P@k$). Nesse caso, considera-se que $\boldsymbol{D}_{retrieved}$ se refere aos k primeiros documentos retornados.

Apesar de serem as métricas mais conhecidas e utilizadas, a precisão e a revocação não consideram a posição dos elementos nas listas. Para contornar esse problema, é necessário utilizar métricas específicas para esse fim. Nessa categoria se destacam as métricas $RR@k$ (*reciprocal rank*) e $nDCG@k$ (*normalized discounted cumulative gain*):

$$RR@k = \frac{1}{rank}$$

$$nDCG@k = \frac{DCG@k}{IDCG@k} = \frac{\sum_{i=1}^{k} \frac{rel_i}{log_2(i+1)}}{IDCG@k}$$

No conjunto de equações acima, o $RR@k$ e o $nDCG@k$ são calculados observando apenas os k primeiros documentos retornados; $rank$ é a posição do primeiro documento relevante retornado; $DCG@k$ (*discounted cumulative gain*) é uma medida que considera a posição relativa dos documentos retornados; rel_i é o escore de relevância do i'ésimo documento retornado; e $IDCG@k$ (*ideal discounted cumulative gain*) é o valor do $DCG@k$ ideal, ou seja, calculado considerando que os k primeiros documentos retornados são os k documentos mais relevantes, em ordem de relevância.

A avaliação de um sistema de buscas é feita consultando as *queries* do qrels e informando as métricas obtidas para cada *query* (por exemplo, através de um histograma) ou uma medida central para todas elas, normalmente a média. Nesse caso, é utilizado o nome $MRR@k$ (*Mean Reciprocal Rank*) para a média dos $RR@k$ e, para as outras métricas, mantém-se a nomenclatura.

2.3. Modelos de Recuperação de Informações Tradicionais

Modelos tradicionais consideram que cada documento é formado por um conjunto de palavras que são indexadas (*index terms*, termos de indexação ou palavras-chave). Esse conjunto pode incluir todas as palavras do documento ou apenas parte delas (por exemplo, alguns campos específicos). As palavras-chave podem ser pré-processadas antes de serem indexadas para facilitar o processo de consulta.

A importância dada a um termo de indexação varia com o modelo. Após indexar todos os documentos da base, há um total de V termos únicos indexados, o que representa o vocabulário da base de dados. Assim, pode-se associar pesos distintos para cada termo indexado de um documento. Formalmente, seja t_i para $i \in [1, V]$ os termos de pesquisa presentes no vocabulário e D algum documento indexado. Se o termo t_i estiver presente no documento D, pode-se definir algum peso $w_{D,i} > 0$ que representará a relevância de t_i em D. Caso o termo não esteja presente neste documento, $w_{D,i} = 0$. Dessa forma, o documento é definido como $D = [w_{D,1}, \dots, w_{D,V}]$, um vetor com V elementos.

De forma semelhante, a *query* do usuário também é vista como um conjunto de palavras que são pré-processadas antes de serem pesquisadas nos documentos. Esse pré-processamento que é realizado nas *queries* e nos documentos serve, entre outras coisas, para remover palavras que não agregam informação ao texto, reduzindo a complexidade do documento e, consequentemente, da operação de pesquisa.

A forma como é feita a pesquisa, ou seja, como é realizado o cálculo de similaridade entre uma *query* e um documento, define o tipo de modelo, que pode ser classificado em três categorias: booleanos, de espaço vetorial, ou probabilísticos.

2.3.1. Modelos booleanos

No modelo booleano, a pesquisa considera a teoria de conjuntos e a álgebra booleana. Os documentos retornados são todos aqueles que satisfazem completamente a expressão descrita pela *query* (Baeza-Yates e Ribeiro-Neto, 1999; Croft, Metzler e Strohman, 2009).

Nessa abordagem os pesos são binários ($w_{D,i} \in \{0, 1\}$) e indicam apenas se a palavra está presente ou ausente no documento. A *query* é uma expressão booleana tradicional descrita com operadores booleanos (AND, OR, NOT) e cabe ao motor de busca apenas aplicá-la em cada documento e retornar aqueles que satisfazem aos critérios. O conjunto de documentos resultante é não-ordenado (Baeza-Yates e Ribeiro-Neto, 1999; Croft, Metzler e Strohman, 2009; Wang et al, 2024).

2.3.2. Modelo de Espaço Vetorial

Esse modelo reconhece que o uso de pesos binários para representar os documentos é um fator limitante e propõe um *framework* onde a equivalência parcial da *query* é possível. Os valores de $w_{D,i}$ são não-binários e podem ser usados para calcular o grau de similaridade entre uma *query* e um documento (Baeza-Yates e Ribeiro-Neto, 1999).

A query Q é representada por $Q = [w_{Q,1}, \dots, w_{Q,V}]$, e, um documento da base como $D = [w_{D,1}, \dots, w_{D,V}]$. São dois vetores V-dimensionais, onde V é a quantidade de palavras únicas indexadas na base de dados, ou seja, o espaço vetorial é o espaço do vocabulário do *corpus*. O cálculo da similaridade é feito através de uma medida de proximidade entre esses vetores (Salton, Wong e Yang, 1975), como a similaridade de cosseno:

$$rel(Q, D) = \frac{Q.D}{|Q|.|D|}$$

Diferentemente do modelo booleano, nesse caso não é necessária uma equivalência exata de todos os termos da *query* no documento. As contribuições de cada termo são acumuladas e basta um dos termos da *query* estar presente no documento para que $rel(Q, D) \neq 0$.

A efetividade desse método reside no cálculo dos pesos $w_{D,i}$. O método mais conhecido dessa categoria é o TF-IDF, que ajusta o peso de acordo com a frequência do termo de pesquisa no documento (TF – *term frequency*) e com o inverso da frequência com que o mesmo termo aparece em todos os documentos (IDF – *inverse document frequency*). No método TF-IDF, quanto mais vezes um termo de pesquisa aparece no documento (TF), maior a sua relevância, mas isso é balanceado com a sua raridade em todos os documentos indexados.

O TF-IDF tem as vantagens de ser um método de fácil implementação e possuir bom desempenho computacional e de resultados. Além disso, ao contrário dos métodos booleanos, ele já ordena a lista de documentos retornados.

2.3.3. Modelos Probabilísticos

Esses modelos calculam os pesos usando conceitos probabilísticos. O método mais conhecido dessa categoria e o mais utilizado é o BM25 (*Best Match* 25). Proposto na década de 90 (Robertson et al, 1995), o BM25, assim como o TF-IDF, considera a correspondência exata com os termos de pesquisa da *query*, mas não exige que todos eles estejam presentes no documento. Cada termo da *query* presente no documento contribui com um escore de relevância que é ponderado pela frequência da palavra no documento e pelo inverso da probabilidade com que essa palavra aparece na coleção de documentos.

Devido à sua simplicidade e efetividade, ainda é muito utilizado para estabelecer parâmetros iniciais (*baselines*) de comparação (Thakur et al, 2021). O modelo é também utilizado como ponto partida para pesquisas acadêmicas e em ambientes de produção de sistemas de recuperação de informações comerciais (Lin, Nogueira e Yates, 2021). Além disso, mesmo quando sistemas de busca mais complexos são utilizados, é comum que o BM25 seja usado no primeiro estágio de ranqueamento para fazer uma filtragem inicial dos documentos.

Considerando que a query Q seja formada pelas palavras $q_1, \dots, q_n$ (ou seja, composta por n termos de pesquisa), o BM25 calcula a função de ranqueamento $rel(Q, D)$ como:

$$rel(Q, D) = \sum_{i=1}^{n} IDF(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \left(1 - b + b \frac{|D|}{avgdl}\right)}$$

$$IDF(q_i) = ln\left(\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} + 1\right)$$

Em que $f(q_i, D)$ é o número de vezes que a palavra q_i aparece no documento D, $|D|$ é o tamanho do documento D em palavras e $avgdl$ é o tamanho médio de um documento da coleção $\boldsymbol{D}$. k_1 e b são parâmetros que podem ser usados para configurar o método. $IDF(q_i)$ é um peso relacionado ao inverso da frequência da palavra q_i nos documentos, $n(q_i)$ é o número de documentos que contém a palavra q_i e N é o total de documentos em $\boldsymbol{D}$.

A implementação do TF-IDF e do BM25 utiliza as mesmas estruturas de dados e conceitos similares. Entretanto, devido à teoria utilizada para se chegar à equação do BM25, ele é tido como um modelo probabilístico (Lin, Nogueira e Yates, 2021).

2.3.4. Problemas de abordagens léxicas

Modelos de recuperação de informação tradicionais são léxicos, ou seja, baseiam-se em uma correspondência exata entre as palavras do documento e as usadas na *query*. A *query* e o documento são representados por estruturas que dependem apenas das palavras, sem importar o seu significado ou contexto.

Um dos problemas de abordagens léxicas é o descasamento de vocabulário (*vocabulary mismatch*), que ocorre quando o usuário usa termos na *query* que, embora sejam semanticamente iguais ou similares aos usados no documento, são diferentes lexicamente (Wang et al, 2024). O efeito disso é que o escore de relevância calculado por métodos tradicionais pode ser menor do que o que seria esperado se a semântica fosse considerada. Dependendo do caso, isso pode fazer com que documentos relevantes não sejam encontrados.

Há três formas de tratar esse problema (Lin, Nogueira e Yates, 2021):

1. expandir as *queries*: nessa situação, a *query* é enriquecida com a inserção de novos termos. Isso pode ser feito usando um dicionário léxico de palavras (sinônimos) ou por meio de métodos mais elaborados. A técnica atua no processo de consulta e é facilmente testável, pois exige apenas a alteração da *query*.

2. expandir os documentos: é similar à técnica de expansão de *queries*, mas atua nos documentos antes de serem indexados (Nogueira et al, 2019). Tradicionalmente a expansão de *queries* tem sido mais utilizada em detrimento da expansão de documentos porque estes são mais difíceis de testar, visto que envolve a reindexação de toda a base de dados. Além disso, com métodos tradicionais, a expansão de *queries* tende a gerar resultados melhores do que a expansão de documentos (Nogueira et al, 2019).

3. ir além da correspondência exata de termos: nesse caso, as representações da *query* e do documento deixam o espaço vocabular e passam para o espaço semântico (Lin, Nogueira e Yates, 2021).

As expansões de *queries* e de documentos possuem a vantagem de que é possível manter o motor de busca, pois basta adicionar um novo elemento para enriquecer um texto no momento da busca (no caso da expansão de *queries*) ou no momento da indexação (no caso da expansão de documentos). Por outro lado, elas não incorporam o significado das palavras, o que só é feito a partir da alteração de pesquisa léxica para semântica. Essa alteração implica a troca do motor de busca e da escolha de métodos para representação vetorial de textos, o que usualmente é feito com modelos de redes neurais (Bengio et al, 2003).

Embora a diferença entre o conteúdo do documento e o texto da consulta seja o problema mais significativo nas abordagens léxicas, existem desafios mesmo quando o vocabulário utilizado é o mesmo. Nessas abordagens, em geral, a contribuição de uma palavra para o escore de relevância está vinculada à sua frequência no texto, e não a semântica. Uma palavra importante para um documento pode ter um peso menor no cálculo de relevância se aparecer poucas vezes. Esse problema pode ser

contornado com a expansão de documentos, situação em que palavras que já estão no texto são reforçadas (Nogueira et al, 2019), ou com a adoção de abordagens semânticas.

2.4. Modelos Neurais de Ranqueamento

Redes neurais podem ser utilizadas para codificar a *query* e o documento usando vetores numéricos n-dimensionais para extrair a semântica do texto (Bengio et al, 2003). Nessa abordagem, a função de ranqueamento pode ser qualquer medida de proximidade entre dois vetores, como a similaridade de cosseno. A representação vetorial da *query* ou do documento é chamada de *embeddings*, e os modelos para a sua extração são redes neurais treinadas especificamente para este fim (Bengio et al, 2003; Le e Mikolov, 2014).

Em tese, a representação semântica do texto resolve o problema de descasamento de vocabulário e a questão envolvendo a importância relativa das palavras no texto, visto ela codifica a semântica (sentido) do texto. Entretanto, na prática, há um problema de pesquisa em aberto para encontrar um bom modelo de geração de *embeddings* para representar bem qualquer coleção de textos – um modelo de *embeddings* treinados com textos genéricos pode não representar muito bem vocabulários de domínios específicos (Major, Surkis e Aphinyanaphongs, 2018; Chen et al, 2021). Assim, o BM25 continuou sendo um *baseline* robusto, de forma que, em muitas coleções de texto, a qualidade de uma pesquisa com BM25 supera a qualidade de alguns modelos vetoriais (Thakur et al, 2021). Essa realidade começou a ser alterada com a criação de uma nova arquitetura de rede neural – o *transformer*, que possibilitou evoluções significativas tanto na geração de *embeddings* quanto na função de ranqueamento.

2.4.1. Recuperação de Informações usando Transformers

Em 2017, pesquisadores do Google apresentaram uma nova arquitetura de redes neurais – o *transformer*. Inicialmente o modelo foi usado para aplicações de transdução de sequências, como a tradução de idiomas (Vaswani et al, 2017). O estado da arte até então utilizava arquiteturas de redes recorrentes, que processavam as sequências um passo por vez, dificultando a paralelização na etapa de treinamento. Já a arquitetura *transformer* processa todas as partes da sequência ao mesmo tempo, o que facilita a paralelização do treinamento (Vaswani et al, 2017; Tunstall et al, 2022). É um modelo que escala bem com grandes conjuntos de dados.

Essa característica possibilitou o treinamento de grandes modelos de linguagem (LLM – *large language models*), que conseguem representar estruturas básicas de um idioma. Apesar dos *transformers* terem sido usados inicialmente para tradução, essa arquitetura rapidamente deu origem a modelos que são utilizados em diversas outras áreas, inclusive em recuperação de informações.

2.4.1.1. BERT

Um dos primeiros modelos rapidamente empregado em sistemas de recuperação de informações foi o BERT (*Bidirectional Encoder Representations from Transformers*) (Devlin et al, 2019). É um modelo de linguagem bidirecional, ou seja, ele aprende a gerar ou processar textos em linguagem natural e, por ser bidirecional, cada palavra em uma sequência depende tanto das palavras à esquerda quanto à direita. O treinamento do BERT fornece um modelo genérico de linguagem que, posteriormente, pode ser especializado (*fine-tuning*) para tarefas específicas.

Até então, um modelo de linguagem era treinado especificamente para tarefas específicas. Para isso, era necessária uma grande quantidade de dados rotulados exclusivamente para aquela atividade. O treinamento do BERT, ao contrário, usa textos não rotulados escritos em linguagem natural, de forma que o modelo aprende qual é a estrutura básica do idioma presente no treinamento (Devlin et al, 2019).

Esse modelo pré-treinado serve como ponto de partida para a especialização do modelo em tarefas específicas, que usa dados rotulados específicos para a tarefa que se deseja resolver. Entretanto, a quantidade de dados rotulados necessários é consideravelmente menor do que a quantidade de dados não rotulados usado no treinamento. Além disso, isso também é válido para o custo computacional para o *fine-tuning* do modelo, que também é extremamente menor do que o custo para o treinamento.

Em sistemas de recuperação de informações, é possível usar o BERT e modelos semelhantes de algumas formas (Lin, Nogueira e Yates, 2021). Um de seus primeiros usos foi através de *fine-tuning* para que funcionasse como um classificador binário, indicando se um texto é ou não relevante para uma *query* (Nogueira e Cho, 2019). Após a especialização do modelo, na etapa de busca é possível estimar o grau de relevância da *query* com todos os documentos da base, ordenando-os (Lin, Nogueira e Yates, 2021).

Como o tempo de resposta desses modelos é consideravelmente maior do que o de modelos que usam a abordagem tradicional, dependendo do tamanho da coleção de documentos, a aplicação de classificadores binários em toda a base é inviável. Uma alternativa é usá-la como um segundo estágio de ranqueamento de documentos (re-ranqueamento). Ou seja, inicialmente é utilizado um estágio tradicional, como o BM25, que retorna alguns documentos (por exemplo, 500 ou 1.000) e, estes, são reordenados usando o classificador binário (Nogueira e Cho, 2019).

Também é possível usar modelos semelhantes para atacar o problema de descasamento de vocabulário expandindo *queries* (Zheng et al, 2020) e documentos (Nogueira et al, 2019) ou fornecendo representação semântica para textos (Formal et al, 2021; Santhanam et al, 2021).

2.4.1.2. Modelos Generativos (GPT e Llama)

Modelos generativos como o GPT (*Generative Pre-trained Transformer*), desenvolvido pela OpenAI, e o Llama, desenvolvido pela Meta, são modelos de linguagem

baseados em *transformers* utilizados em NLP. Diferente do BERT, que é bidirecional, o GPT e o Llama são unidirecionais, ou seja, utilizam apenas o contexto à esquerda de uma palavra para fazer predições. Estes modelos são treinados em uma grande quantidade de textos não rotulados usando o objetivo de prever a próxima palavra em uma sequência (Radford et al, 2019). Com esse propósito de prever a próxima palavra, esses novos modelos se mostraram versáteis em várias atividades que envolvem geração de texto, como correção de gramática, geração de código fonte, tradução de texto e conversas em linguagem natural.

No contexto de recuperação de informações, apresentam algumas limitações para serem usados diretamente como ferramentas de pesquisa, especialmente devido ao seu elevado custo computacional. Uma abordagem prática é utilizá-los em conjunto com sistemas tradicionais de recuperação de informações.

Uma possível aplicação é na reordenação de documentos. Por exemplo, o modelo pode ser usado em um segundo estágio de busca, onde o primeiro estágio consiste no uso de um método clássico (como o BM25) e, em seguida, um modelo generativo é usado para refinar a ordenação dos primeiros elementos retornados no estágio anterior, fornecendo um grau mais preciso de relevância (Sun et al, 2023).

Em relação ao problema de descasamento de vocabulário, podem ser usados na expansão de *queries* e documentos (Claveau, 2020). Com eles, não é necessário utilizar um dicionário fixo de palavras, pois os modelos podem ser usados para extrair a semântica das palavras no contexto em que elas se encontram, gerando novos sinônimos para o contexto correto. Além disso, na expansão de documentos, podem ser usados para reforçar partes importantes do texto ou para reescrever o texto em outras palavras.

3. Método

O Tribunal de Contas da União organiza e disponibiliza a sua jurisprudência em cinco base de dados. Dessas, destaca-se a base de jurisprudência selecionada, criada a partir de deliberações escolhidas por uma equipe do Tribunal a partir de critérios de relevância jurisprudencial e retratam entendimentos adotados pelo TCU sobre determinados assuntos (Brasil, 2024).

A Figura 3 mostra um exemplo de documento dessa base. Cada registro é estruturado em alguns campos que indicam, entre outras informações, o documento de origem de onde a jurisprudência foi extraída, sua data, quais as referências legais utilizadas, quem foi o relator, algumas palavras-chave que podem ser usadas para encontrá-lo, o excerto de onde a decisão foi extraída e, por fim, um enunciado que resume a decisão. Desses, os campos Enunciado e Excerto são os mais relevantes, pois descrevem o entendimento que pode ser extraído da decisão e o trecho da decisão que levou a esse entendimento.

A pesquisa de jurisprudência selecionada do TCU é léxica, baseada em lógica booleana e no BM25. A configuração atual usa o operador booleano AND entre todos os termos de pesquisa e faz o ranqueamento dos documentos usando o BM25 com

pesos diferentes para cada campo. Além disso, para mitigar o problema do descasamento de vocabulário, é utilizado o vocabulário de controle externo (VCE) do TCU (Brasil, 2019) para expansão de *queries*. Dessa forma, a pesquisa apenas retorna documentos que possuem todas as palavras (ou algum sinônimo catalogado no VCE) contidas na *query*.

Figura 3: Exemplo de documento da base de jurisprudência selecionada.

ACÓRDÃO:
Acórdão 945/2024-Segunda Câmara

DATA DA SESSÃO:
20/02/2024

RELATOR:
VITAL DO RÊGO

ÁREA:
Pessoal

TEMA:
Acumulação de cargo público

SUBTEMA:
Invalidez permanente

OUTROS INDEXADORES:
Remuneração, Acumulação, Ressarcimento ao erário, Proventos, Vedação

TIPO DO PROCESSO:
TOMADA DE CONTAS ESPECIAL

ENUNCIADO:
A invalidez permanente é incompatível com o exercício de qualquer cargo público, razão pela qual é indevida a acumulação de proventos de invalidez permanente com remuneração decorrente do exercício de outro cargo, cabendo restituição ao erário dos proventos recebidos durante a acumulação ilegal.

EXCERTO:
Voto:

Trata-se de tomada de contas especial instaurada pelo Fundo Nacional de Saúde - Ministério da Saúde (FNS/MS) em desfavor de [responsável], em razão do irregular acúmulo de aposentadoria, por invalidez, com o exercício do cargo de médica.

2. O recebimento indevido de recursos federais decorrente do acúmulo ilegal de aposentadoria com exercício de cargo de médica ocorreu no período de 31/12/2000 a 30/6/2017 e totalizou R$ 1.128.359,62, em valores nominais.

... ...

PUBLICADO:

- Boletim de Jurisprudência nº 482 de 11/03/2024
- Boletim de Pessoal nº 120 de 20/03/2024

ENUNCIADOS RELACIONADOS:

- A invalidez permanente é incompatível com o exercício de qualquer cargo público. Portanto, não é possível a acumulação de proventos de invalidez permanente com remuneração decorrente do exercício de outro cargo público.
- A invalidez permanente é incompatível com o exercício de qualquer cargo público. Portanto, não é possível a acumulação de proventos de invalidez permanente com remuneração decorrente do exercício de outro cargo público.

... ...

Fonte: Brasil. Tribunal de Contas da União (2024).

3.1. Modelagem dos Experimentos

Os experimentos deste trabalho foram modelados para usar a técnica de expansão de documentos com modelos de linguagem baseados em *transformers*. Essa abordagem foi adotada pois pouco altera a arquitetura da pesquisa atual, sendo exigido apenas um novo módulo para a expansão de textos que é inserido no momento da indexação. Assim, não modifica a etapa de consulta, não inserindo novo elemento susceptível à falha nesta fase. Isso é algo que deve ser considerado em sistemas em produção, pois modelos de linguagem dessa natureza exigem *hardware* especializado diferente dos utilizados por sistemas de busca tradicionais. Eventuais problemas na etapa de indexação podem ser corrigidos sem que impacte significativamente o usuário. Já na etapa da busca, eventuais problemas podem inviabilizar a pesquisa.

3.1.1. Descrição do Qrels da Base de Jurisprudência Selecionada

Para avaliar a pesquisa atual e verificar se as técnicas propostas melhorarão ou não a sua qualidade, é necessário utilizar um conjunto qrels para testes. Para isso, será considerado o conjunto disponibilizado por Pacheco e Borela (2023), que contém 150 *queries* de teste, cada uma relacionada a uma lista de documentos. As relações *query*-documento são anotadas com valores no conjunto $\{0, 1, 2, 3\}$ indicando se o documento é irrelevante, pouco relevante, relevante ou muito relevante para a *query*. As *queries* estão disponíveis no Anexo A.

As 150 *queries* padronizadas são divididas em 3 grupos distintos, cada um com 50 *queries*. O primeiro grupo representa termos de busca comuns extraídos do *log* da pesquisa de jurisprudência selecionada em produção, ou seja, são termos que foram muito pesquisados pelos usuários.

O segundo e o terceiro grupo são formados por *queries* sintéticas, geradas com o auxílio do GPT. Isso foi feito extraindo do log de pesquisa os 50 documentos mais acessados e, em seguida, enviando o campo Enunciado desses documentos ao GPT para que ele gerasse 5 perguntas curtas e diretas que pudessem ser respondidas por ele (prompt disponível na Figura 4 do Anexo B). As respostas foram revisadas e condensadas manualmente e originaram o terceiro conjunto de *queries*. Essas perguntas foram manualmente transformadas em expressões de busca típicas da pesquisa e correspondem ao segundo grupo de *queries*. Os conjuntos de *queries* possuem as seguintes características:

- Grupo 1 (1-50): *queries* reais usando termos de pesquisa. Possuem, em média, 3,5 palavras por *query*;
- Grupo 2 (51-100): *queries* sintéticas usando termos de pesquisa. Possuem, em média, 6,5 palavras por *query*;
- Grupo 3 (101-150): *queries* sintéticas em formato de perguntas. Possuem, em média 16,5 palavras por *query*.

A anotação de pares de relevância foi feita usando um processo automático com verificação e correção manual. A partir de uma arquitetura de busca de múltiplos estágios, selecionou-se 10 documentos bem ranqueados (exemplos positivos) e 5 documentos selecionados aleatoriamente (exemplos negativos). Para cada documento, solicitou-se ao GPT que produzisse um escore de relevância de 0 a 3 usando o prompt usado disponível na Figura 5 do Anexo B. O resultado obtido foi enviado para um especialista da área de jurisprudência do TCU e as correções indicadas foram incorporadas no qrels gerado (Pacheco e Borela, 2023).

3.2. Cenários de Teste

Com o qrels, é possível realizar o diagnóstico da pesquisa atual, estabelecendo uma linha de base para comparar com os métodos que serão testados. Além disso, também será utilizado como uma segunda linha de base o BM25 padrão com os

valores de $k_1 = 1{,}2$ e $b = 0{,}75$. Esses valores foram escolhidos pois funcionam relativamente bem para propósitos gerais[3]. Na indexação desse cenário base, foram usados apenas os campos Enunciado e Excerto, pois eles contêm as informações mais importantes da base de dados.

Para a avaliação, serão utilizadas as métricas $P@k$, $R@k$, $nDCG@k$ e $MRR@k$, com $k = \{5, 20\}$. A escolha desses valores para k se deve a forma como os usuários costumam navegar em páginas Web e em serviços de pesquisa. Ao investigar sistemas de busca Web, pesquisadores identificaram que a posição em que um elemento se encontra na página de resultados afeta consideravelmente a sua probabilidade de ser clicado. Em relação a um resultado na primeira posição, um resultado na segunda posição tem entre um a dois terços a menos de chance de ser clicado dependendo do tipo de pesquisa que está sendo realizada (Glick et al, 2014). Esse fato justifica o uso de $k = 5$, pois esse é o número usual de resultados visíveis na tela sem que o usuário tenha que rolar a página na interface da pesquisa integrada do TCU.

A questão da interação do usuário com a página também justifica a escolha de $k = 20$, pois é essa a quantidade de resultados que a pesquisa exibe por página. Assim, admitindo que a pesquisa de jurisprudência selecionada tem a característica de ser usada em situações de trabalho, é esperada maior resiliência do usuário para pesquisar além dos primeiros documentos.

3.2.1. Expansão de Documentos

Neste artigo, serão testadas quatro abordagens de expansão de documentos: 1. usando o método docT5query; 2. por meio de extração de sinônimos; e 3. através da reescrita do enunciado; e 4. combinando docT5query com extração de sinônimos. Em todos os casos, a expansão é feita a partir do campo Enunciado do documento, pois é este campo que contém o sumário da decisão. Uma vez expandidos, os documentos contendo a expansão e os campos Enunciado e Excerto serão pesquisados com o BM25, com os mesmos parâmetros utilizados no cenário base ($k_1 = 1{,}2$ e $b = 0{,}75$).

A primeira abordagem a ser testada é com o método **docT5query** (Nogueira e Lin, 2019), que fez o *fine-tuning* do modelo de linguagem T5 (Raffel et al, 2019) para que ele sugira possíveis *queries* que possam ser utilizadas para pesquisar por um texto. Recentemente foi divulgado[4] um modelo em português, que será utilizado para gerar *queries* a partir do campo Enunciado. Serão consideradas duas configurações, com a expansão do documento com 1 e com 5 *queries.*

A segunda abordagem a ser testada utiliza a técnica de extração de sinônimos para os termos mais importantes do campo Enunciado. Para isso, serão utilizados modelos generativos de linguagem (GPT-3.5, GPT-4o e Llama 3 70B), conforme prompt da Figura 6 do Anexo B, que retorna as cinco palavras consideradas mais importantes

[3] https://www.elastic.co/pt/blog/practical-bm25-part-3-considerations-for-picking-b-and-k1-in-elasticsearch
[4] https://huggingface.co/doc2query/msmarco-portuguese-mt5-base-v1

para o texto e uma lista de sinônimos. O documento é expandido com o conteúdo retornado pelo prompt.

A terceira abordagem a ser testada utiliza reescrita do campo Enunciado usando modelos generativos de linguagem (GPT-3.55 e Llama 3 70 B), conforme prompt da Figura 7 do Anexo B, que reescreve o campo Enunciado usando outras palavras. Nesse caso, o documento foi expandido com a reescrita.

A quarta abordagem combina o **docT5query com a técnica de extração de sinônimos**, expandido o documento com o retorno das duas técnicas.

Assim, as 4 abordagens propostas foram testadas com diferentes combinações, produzindo resultados para 12 cenários de teste. Desses, 2 representam cenários base comparativos (*baselines*) e 10 são possibilidades de intervenção na pesquisa de jurisprudência selecionada usando técnicas de expansão de documento. Os cenários base são referentes à pesquisa atual e a uma pesquisa com BM25 com os documentos indexados usando apenas os campos Enunciado e Excerto. Nos 10 cenários envolvendo expansão de documentos, o método de busca também é o BM25 com a indexação envolvendo os campos Enunciado e Excerto e, além deles, os textos gerados pelas técnicas de expansão estudadas. O Quadro 1 descreve cada cenário.

Quadro 1 – Cenários de estudo

Cenário	docT5query	Sinônimos do Enunciado	Reescrita do Enunciado
Atual (*baseline*)	–	–	–
BM25 (*baseline*)	–	–	–
BM25 + docT5query(1)	1 *query*	–	–
BM25 + docT5query(5)	5 *queries*	–	–
BM25 + sinônimos(GPT-3.5)	–	GPT-3.5	–
BM25 + sinônimos(GPT-4o)	–	GPT-4o	–
BM25 + sinônimos(Llama)	–	Llama 3 70B	–
BM25 + reescrita(GPT3.5)	–	–	GPT-3.5
BM25 + reescrita(Llama)	–	–	Llama 3 70B
BM25 + docT5query(5) + sinônimos(GPT-3.5)	5 *queries*	GPT-3.5	–
BM25 + docT5query(5) + sinônimos(GPT-4o)	5 *queries*	GPT-4o	–
BM25 + docT5query(5) + sinônimos(Llama)	5 *queries*	Llama 3 70B	–
Em todos os casos envolvendo BM25, foram indexados apenas os campos Enunciado e Excerto e, se for o caso, o texto gerado pela técnica de expansão indicada.			

Fonte: Elaboração própria (2024).

[5] Devido aos resultados obtidos na abordagem anterior e considerando os custos envolvidos para executar o modelo GPT-4o, optou-se por não o testar nessa abordagem.

4. Resultados e Discussões

Esta seção analisa os resultados obtidos para os 3 conjuntos de *queries* disponíveis no qrels. É válido ressaltar as diferenças entre cada conjunto. O conjunto 1 (Quadro 2 do Anexo A) é formado por palavras-chave que efetivamente foram usadas por usuários reais da pesquisa. Os conjuntos 2 e 3 (Quadro 3 e Quadro 4 do Anexo A) foram gerados usando LLMs a partir dos enunciados mais acessados por usuários, sendo o conjunto 3 formado por *queries* na forma de perguntas e o conjunto 2 formado por *queries* no formato de palavras-chave.

Assim, o conjunto 1 representa a forma como os usuários pesquisam na base de jurisprudência selecionada, sendo, por isso, o mais relevante para esta análise. A importância do conjunto 3 reside no fato de que, com a proliferação de ferramentas de IA generativa, é possível que haja alguma alteração na forma de pesquisa dos usuários. Por fim, o conjunto 2 é um meio do caminho entre os conjuntos 1 e 3, ou seja, a pesquisa ainda é feita usando palavras-chave, mas a quantidade de termos utilizada é maior.

As Tabelas 1, 2 e 3 mostram as métricas $P@5$, $R@5$, $MRR@5$ e $nDCG@5$ para, respectivamente, os conjuntos de *queries* 1, 2 e 3. O maior valor para a métrica é indicado, nas tabelas, em negrito. Valores até um ponto percentual menor do que o máximo atingido são indicados em itálico.

Tabela 1 – Resultados para o conjunto de *queries* 1 ($k = 5$)

Cenário	$P@5$	$R@5$	$MRR@5$	$nDCG@5$
Atual (baseline)	0,2880	0,1155	0,3720	0,2653
BM25 (baseline)	0,2720	0,1106	0,5253	0,2824
BM25 + docT5query(1)	0,2920	0,1180	0,5187	0,3023
BM25 + docT5query(5)	0,3880	0,1568	0,6417	0,4044
BM25 + sinônimos(GPT-3.5)	0,3160	0,1276	0,5387	0,3224
BM25 + sinônimos(GPT-4o)	0,2920	0,1177	0,5357	0,3056
BM25 + sinônimos(Llama)	0,3120	0,1255	0,5380	0,3204
BM25 + reescrita(GPT3.5)	0,2800	0,1138	0,5040	0,2837
BM25 + reescrita(Llama)	0,2680	0,1090	0,4377	0,2643
BM25 + docT5query(5) + sinônimos(GPT-3.5)	**0,4320**	**0,1735**	0,6687	**0,4402**
BM25 + docT5query(5) + sinônimos(GPT-4o)	0,4120	*0,1655*	0,6537	0,4194
BM25 + docT5query(5) + sinônimos(Llama)	*0,4240*	*0,1703*	**0,6817**	*0,4343*

Fonte: Elaboração própria (2024).

Tabela 2 – Resultados para o conjunto de *queries* 2 ($k = 5$)

Cenário	$P@5$	$R@5$	$MRR@5$	$nDCG@5$
Atual (baseline)	0,4560	0,1892	0,8667	0,5639
BM25 (baseline)	0,5000	0,2077	0,8620	0,5713
BM25 + docT5query(1)	0,4960	0,2062	0,8920	0,5807
BM25 + docT5query(5)	0,5560	0,2315	0,8773	0,6235
BM25 + sinônimos(GPT-3.5)	0,5280	0,2196	0,8797	0,6015
BM25 + sinônimos(GPT-4o)	0,5280	0,2195	*0,9020*	0,6025
BM25 + sinônimos(Llama)	0,5280	0,2197	0,8690	0,5956
BM25 + reescrita(GPT3.5)	0,5120	0,2135	0,8867	0,5912
BM25 + reescrita(Llama)	0,5280	0,2193	*0,8940*	0,5992
BM25 + docT5query(5) + sinônimos(GPT-3.5)	*0,5680*	*0,2363*	0,8880	*0,6368*
BM25 + docT5query(5) + sinônimos(GPT-4o)	*0,5720*	*0,2383*	*0,8940*	*0,6404*
BM25 + docT5query(5) + sinônimos(Llama)	**0,5760**	**0,2403**	**0,9040**	**0,6431**

Fonte: Elaboração própria (2024).

Tabela 3 – Resultados para o conjunto de *queries* 3 ($k = 5$)

Cenário	$P@5$	$R@5$	$MRR@5$	$nDCG@5$
Atual (baseline)	0,0360	0,0164	0,1100	0,0510
BM25 (baseline)	0,5200	0,2340	0,9150	0,6030
BM25 + docT5query(1)	0,5200	0,2338	*0,9500*	0,6129
BM25 + docT5query(5)	0,5280	0,2374	0,9367	0,6202
BM25 + sinônimos(GPT-3.5)	0,5160	0,2317	0,9133	0,6035
BM25 + sinônimos(GPT-4o)	0,5280	0,2386	0,9340	0,6170
BM25 + sinônimos(Llama)	0,5120	0,2307	0,9217	0,6044
BM25 + reescrita(GPT3.5)	0,5440	*0,2446*	0,9400	0,6296
BM25 + reescrita(Llama)	**0,5560**	*0,2496*	**0,9550**	*0,6398*
BM25 + docT5query(5) + sinônimos(GPT-3.5)	0,5440	*0,2429*	0,9167	0,6251
BM25 + docT5query(5) + sinônimos(GPT-4o)	**0,5560**	**0,2506**	0,9400	**0,6402**
BM25 + docT5query(5) + sinônimos(Llama)	0,5440	*0,2444*	0,9267	0,6296

Fonte: Elaboração própria (2024).

Para o diagnóstico inicial da pesquisa, é importante compará-lo com o cenário BM25. Com o conjunto de *queries* 1, as métricas de precisão e revocação são melhores na configuração atual do que no BM25, com uma diferença de cerca de 5%. Entretanto, as métricas $MRR@5$e $nDCG@5$ são melhores no BM25 (cerca de 40% para MRR e 6% para $nDCG$). Embora a configuração atual traga um pouco mais de resultados relevantes que o BM25, este ordena melhor a lista retornada.

Na comparação com o conjunto de queries 2 e 3, a superioridade do BM25 em relação ao modelo atual é evidente. Isso ocorre pois a configuração atual da pesquisa exige que todos os termos pesquisados estejam presentes (usa por padrão o operador booleano AND entre as palavras da *query*). Dessa forma, aumentar o tamanho da *query* aumenta a probabilidade de um documento não ser encontrado. Já o BM25 padrão não exige a presença de todos os elementos da *query*.

Esse resultado é relevante pois mostra que o BM25, mesmo sem nenhum ajuste fino em seus parâmetros (b e k_1), ainda é um modelo de ranqueamento difícil de ser superado. Na pesquisa em tela, o BM25 produziu, em geral, resultados mais bem ordenados que a configuração atual. Esse resultado complementa o obtido por Gomes e Ladeira (2020), que avaliaram o sistema de pesquisa do Superior Tribunal de Justiça e concluíram que o BM25 supera o sistema legado de busca do órgão, que é baseado em uma pesquisa booleana. Na mesma linha, também complementa o resultado de Rosa, Rodrigues e Lotufo (2021), que indicaram que o BM25 é um forte *baseline* para sistemas de pesquisa com documentos de origem legal. Por isso, as comparações no restante desta seção serão feitas com o *baseline* BM25, a não ser que explicitamente dito o contrário.

A expansão de documentos usando docT5query foi testada com 1 e com 5 *queries*. Ambos os casos melhoraram os resultados da busca. Entretanto, nota-se uma superioridade no uso de 5 *queries* em todos os casos, com exceção da métrica $MRR@5$ para os conjuntos de *queries* 2 e 3. Com *5 queries*, a expansão superou os dois *baselines* nos 3 conjuntos de *queries.*

Observa-se, ainda, que a expansão com docT5query apresentou aumentos mais significativos no conjunto de *queries* 1 do que nos outros. Para o conjunto 1, a expansão com 5 *queries* elevou as métricas $P@5$, $R@5$ e $nDCG@5$ em aproximadamente 40% e o $MRR@5$ em cerca de 20%: a técnica ajudou a retornar mais documentos relevantes e a posicioná-los melhor no resultado da pesquisa. Para o conjunto de *queries* 2 e 3 os aumentos foram mais modestos, havendo melhorias de até 11% para o conjunto 2 e de até 3% para o conjunto 3, dependendo da métrica. Isso se justifica pelo fato de que os conjuntos 2 e 3 são de *queries* sintéticas, geradas a partir dos enunciados mais acessados. Esses conjuntos possuem, em média, mais palavras-chave diretamente retiradas do campo Enunciado do que o conjunto 1, representando com mais precisão o documento que se deseja encontrar.

Nogueira et al (2019), ao avaliarem a quantidade de *queries* indexadas em um *dataset* com trechos curtos em inglês, indicaram aumento aproximadamente linear na melhoria do MRR até 10 novas *queries* e, a partir daí, uma redução da métrica devido ao aumento do ruído inserido na base. Assim, considerando a base de jurisprudência selecionada e os resultados obtidos, a expansão com 5 *queries* é uma boa relação de compromisso, visto que melhorou consideravelmente todas as métricas do conjunto

de *queries* 1 e, nos conjuntos de *queries* 2 e 3, a piora em relação a utilizar apenas 1 *query*, quando houve, foi ínfima.

A expansão de documento usando extração de sinônimos do campo Enunciado foi testada com 3 modelos de linguagem distintos: GPT-3.5, GPT-4o e Llama 3 70B. Embora o GPT-4o seja mais robusto que o GPT-3.5, ambos tiveram resultados similares. Assim, devido aos custos envolvidos no uso do GPT-4o, não é interessante aplicá-lo neste problema. A extração de sinônimos com o Llama 3 70B, um modelo *open source*, trouxe resultados semelhantes aos da família GPT.

O enriquecimento de documentos usando sinônimos do enunciado promoveu avanços em relação aos cenários base. Entretanto, em geral, o incremento foi menos significativo do que o obtido com a expansão com docT5query com 5 *queries*.

A expansão de documentos com a reescrita do campo Enunciado foi testada com os modelos GPT-3.5 e Llama 3 70B[6]. No conjunto de *queries* 1, as métricas $R@5$ e $MRR@5$ apresentaram resultados piores do que quando usado apenas o BM25, sem expansão de documentos. Houve melhoria de menos de 0,5% no $nDCG@5$ e de menos de 3% na $P@5$. Para o conjunto de *queries* 2 e 3 houve alguma melhora nas métricas. No entanto, como esses conjuntos de *queries* também foram gerados por LLMs , esse resultado deve ser visto com cautela, pois os termos das *queries* gerados sinteticamente tendem a ser muito parecidos com os presentes no campo Enunciado, se distanciando de um cenário real.

Considerando os três primeiros métodos de expansão testados, o docT5query produziu os melhores resultados, seguido pela técnica de expansão com sinônimos. Por isso, foram feitos testes em um quarto método de expansão, combinando simultaneamente o docT5query com 5 *queries* e a geração de sinônimos usando GPT-3.5, GPT-4o e Llama 3 70B.

As melhores métricas foram obtidas com a aplicação conjunta dessas duas técnicas. Em relação ao conjunto de *queries* 1, a precisão e a revocação foram cerca de 50% maiores do que as obtidas em relação à configuração atual da pesquisa. O $MRR@5$ aumentou em quase 80% e, o $nDCG@5$, em aproximadamente 65%. Na comparação com o BM25, o aumento variou de 25% no $MRR@5$ a quase 60% para $P@5$. Também houve melhorias para os conjuntos de *queries* 2 e 3. No conjunto 2, as métricas, em sua maioria, melhoraram cerca de 13%. Já para o conjunto 3, cerca de 3%.

Os experimentos também foram executados considerando os primeiros 20 resultados da pesquisa ($k = 20$). Assim como ocorreu com a análise com os 5 primeiros documentos ($k = 5$), as melhores métricas foram obtidas com a aplicação conjunta do docT5query e da inclusão de sinônimos para termos relevantes do campo Enunciado. O Anexo C mostra os resultados dos experimentos para $k = 20$. Tendo em vista os resultados obtidos para $k = 5$, para simplificar a visualização, o anexo suprimiu os resultados para os cenários docT5query com 1 *query* e a expansão de documentos envolvendo a reescrita do campo Enunciado.

[6] Devido aos resultados obtidos na abordagem anterior e considerando os custos envolvidos para executar o modelo GPT-4o, optou-se por não o testar nessa abordagem.

5. Conclusão

Este artigo analisou a aplicação de técnicas de expansão de documentos para melhorar a qualidade da pesquisa de jurisprudência selecionada do TCU. Essa abordagem adiciona novos textos aos documentos antes de sua indexação e é utilizada tanto para melhorar a ordenação dos resultados da pesquisa, quando termos importantes são reforçados, quanto para mitigar o problema de descasamento de vocabulário, quando palavras que ainda não estavam no documento são adicionadas a ele.

Como forma de estabelecer cenários base para comparação, foi realizada a avaliação da pesquisa atual e a aferição de sua eficácia em relação ao BM25 sem expansão de documentos. Os resultados dessa avaliação são significativos, pois mostraram que o BM25, mesmo sem expansão de documentos, retornou a lista de resultados com os documentos relevantes mais bem ordenados quando comparado com o cenário atual.

Essa questão é importante, pois a pesquisa atual de jurisprudência selecionada usa o operador booleano AND entre todos os termos de pesquisa e faz o ranqueamento dos resultados seguindo um método baseado em BM25 dando pesos diferentes para os diversos campos do documento. Ou seja, apesar de ser um método de pesquisa mais complexo, isso não se traduz em melhorias evidentes de qualidade.

Após a avaliação da pesquisa e o estabelecimento de cenários para comparação, foram consideradas quatro abordagens para expansão de documentos, com diferentes configurações. A primeira foi usando o método docT5query, que usa o modelo de linguagem T5 para sugerir possíveis *queries* que podem ser utilizadas para se pesquisar o documento. A segunda abordagem testada foi o uso de modelos generativos (GPT e Llama) para extração de sinônimos de palavras importantes do enunciado. A terceira abordagem foi a reescrita de todo o enunciado. A última abordagem combinou o método docT5query com a extração de sinônimos.

As técnicas de expansão de documentos produziram resultados melhores do que quando comparados com os cenários base. Dentre elas, os maiores incrementos nas métricas ocorreram com o uso do docT5query com a expansão usando 5 *queries*. A reescrita do campo Enunciado ou a extração de sinônimos de palavras relevantes desse campo também trouxeram resultados positivos, mas com amplitudes consideravelmente menores que as do método docT5query.

Importante ressaltar que, para esse problema, o uso de modelos generativos mais robustos e mais caros (GPT-4o) não necessariamente produziu resultados melhores quando comparados com modelos menores (Llama 3 70B e GPT-3.5). Assim, para a aplicação estudada, o uso do GPT-4o não se justifica. Além disso, na escolha entre o GPT-3.5 (modelo pago) e Llama 3 70 B (modelo *open source*), deve-se ponderar o custo para implementação, manutenção e uso, mas pode-se desconsiderar a eficácia dos modelos de linguagem, visto que produzem resultados semelhantes.

Também foi testada a expansão de documentos usando, simultaneamente, docT5query e geração de sinônimos, ocasião em que as melhores métricas foram obtidas.

Tendo em vista a discussão deste artigo, pode-se melhorar a qualidade da pesquisa de jurisprudência selecionada sem modificar sua arquitetura, mas alterando sua configuração atual para usar o BM25 padrão, onde o operador OR é utilizado entre os termos, e não o AND. Além disso, pode-se obter resultados consideravelmente melhores usando o método docT5query com 5 *queries* geradas a partir do campo Enunciado para enriquecer os documentos. O uso deste método enseja alteração apenas no processo de indexação, com a chamada ao modelo de linguagem antes de indexar o documento. É possível incrementar ainda mais a qualidade indexando também sinônimos extraídos do campo Enunciado a partir de modelos generativos como o GPT ou Llama.

Como trabalhos futuros, há outros dois grupos de técnicas que podem ser utilizados para atacar o problema de descasamento de vocabulário – a expansão de *queries* e a alteração do espaço vocabular das *queries* e documentos para o espaço semântico. Além disso, outras sugestões para trabalhos futuros é a criação de qrels para outras bases de dados do Tribunal, bem como testes dessas abordagens nas outras bases de dados.

Referências bibliográficas

BAEZA-YATES, Ricardo; RIBEIRO-NETO, Berthier. **Modern Information Retrieval**. 1. ed. Massachusetts: Addison-Wesley Longman Publishing Co., Inc., 1999.

BENGIO, Yoshua; DUCHARME, Réjean; VINCENT, Pascal; JAUVIN, Christian. **A neural probabilistic language model**. *Journal of Machine Learning Research*, v. 3, p. 1137–1155, 2003.

BRASIL. **Vocabulário de Controle Externo do Tribunal de Contas da União (VCE)**. 2019. Disponível em: https://portal.tcu.gov.br/vocabulario-de-controle-externo-do-tribunal-de-contas-da-uniao-vce.htm. Acesso em: 23 out. 2024.

BRASIL. **Acórdão nº 878/2022-Plenário do TCU**. 2022.

BRASIL. **Pesquisa Integrada do TCU**: manual completo da pesquisa de jurisprudência. 2024. Disponível em: https://pesquisa.apps.tcu.gov.br/. Acesso em 23 out. 2024.

BUENO, Mirelle; OLIVEIRA, Eduardo Seiti de; NOGUEIRA, Rodrigo; LOTUFO, Roberto A.; PEREIRA, Jayr Alencar. **Quati: A Brazilian Portuguese Information Retrieval Dataset from Native Speakers**. 2024. DOI: https://doi.org/10.48550/arXiv.2404.06976.

CHEN, Timothy L.; EMERLING, Max; CHAUDHARI, G. R.; CHILAKURU, Yeshwant R.; SEO, Youngho; VU, Thienkhai H.; SOHN, Jae H. **Domain specific word embeddings for natural language processing in radiology**. Journal of Biomedical Informatics, vol. 113, 2021. DOI: https://doi.org/10.1016/j.jbi.2020.103665.

CLAVEAU, Vincent. Query expansion with artificially generated texts. 2020. DOI: https://doi.org/10.48550/arXiv.2012.08787.

CROFT, Bruce; METZLER, Donald; STROHMAN, Trevor. **Search Engines: Information Retrieval in Practice**. 1. ed. Boston: Pearson, 2009.

DEVLIN, Jacob; CHANG, Ming-Wei; LEE, Kenton; TOUTANOVA, Kristina. **BERT: Pre-training of deep bidirectional transformers for language understanding**. In: PROCEEDINGS OF THE 2019 CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: HUMAN LANGUAGE TECHNOLOGIES, VOLUME 1 (LONG AND SHORT PAPERS). 2019. p. 4171–4186. Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/N19-1423.

FORMAL, Thomas; LASSANCE, Carlos; PIWOWARSKI, Benjamim; CLINCHANT, Stéphane. **SPLADE v2: Sparse lexical and expansion model for information retrieval**. 2021. DOI: https://doi.org/10.48550/arXiv.2109.10086.

GOMES, Thiago; LADEIRA, Marcelo. **A new conceptual framework for enhancing legal information retrieval at the Brazilian Superior Court of Justice**. In: PROCEEDINGS OF THE 12TH INTERNATIONAL CONFERENCE ON MANAGEMENT OF DIGITAL ECOSYSTEMS (MEDES). 2020. p. 26–29. DOI: https://doi.org/10.1145/3415958.3433087.

GLICK, Mark; RICHARDS, Greg; SAPOZHNIKOV, Margarita; SEABRIGHT, Paul. **How Does Ranking Affect User Choice in Online Search?**. *Review of Industrial Organization*, v. 45, p. 99–119, 2014. DOI: https://doi.org/10.1007/s11151-014-9435-y.

JUNIOR, Daniel da Silva; CORVAL, Paulo Roberto dos S.; PAES, Aline; OLIVEIRA, Daniel de. **Datasets for Portuguese Legal Semantic Textual Similarity: Comparing weak supervision and an annotation process approaches**. 2023. DOI: https://doi.org/10.48550/arXiv.2306.00007.

LE, Quoc V.; MIKOLOV, Tomas. **Distributed representations of sentences and documents**. 2014. DOI: https://doi.org/10.48550/arXiv.1405.4053.

LIN, Jimmy; NOGUEIRA, Rodrigo; YATES, Andrew. **Pretrained transformers for text ranking: BERT and beyond**. 2021. DOI: https://doi.org/10.48550/arXiv.2010.06467.

MAJOR, Vincent; SURKIS, Alisa; Aphinyanaphongs, Yindalon. **Utility of General and Specific Word Embeddings for Classifying Translational Stages of Research**. In: AMIA Annu Symp Proc. 2018. pp 1405–1414. 2018.

MANNING, Christopher D.; RAGHAVAN, Prabhakar; SCHÜTZE, Hinrich. **Introduction to Information Retrieval**. Cambridge: Cambridge University Press, 2008.

MITRA, Bhaskar; CRASWELL, Nick. **An Introduction to Neural Information Retrieval**. 2018. DOI: https://doi.org/10.1561/1500000061.

NOGUEIRA, Rodrigo; YANG, Wei; LIN, Jimmy; CHO, Kyunghyun. **Document expansion by query prediction**. 2019. DOI: https://doi.org/10.48550/arXiv.1904.08375.

NOGUEIRA, Rodrigo; CHO, Kyunghyn. **Passage re-ranking with BERT**. 2019. DOI: https://doi.org/10.48550/arXiv.1901.04085.

NOGUEIRA, Rodrigo; LIN, Jimmy. **From doc2query to docTTTTTquery**. 2019.

PACHECO, Leonardo; BORELA, Marcus. **Juris TCU**. Disponível em: https://github.com/marcusborela/ind-ir/tree/main/data/juris_tcu. 2023. Acesso em 23 out. 2024.

RADFORD, Alec; WU, Jeffrey; CHILD, Rewon; LUAN, David; AMODEI, Dario; SUTSKEVER, Ilya. **Language models are unsupervised multitask learners**. 2019.

RAFFEL, Colin; SHAZEER, Noam; ROBERTS, Adam; LEE, Katherine; NARANG, Sharan; MATENA, Michael; ZHOU, Yanqi; LI, Wei; LIU, Peter J. **Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer**. 2019. DOI: https://doi.org/10.48550/arXiv.1910.10683.

ROBERTSON, Stephen E.; WALKER, Steve; JONES, Susan; HANCOCK-BEAULIEU, Micheline; GATFORD, Mike. **Okapi at TREC-3**. In: PROCEDDINGS OF THE THIRD TEXT RETRIEVAL CONFERENCE (TREC-3). 1995.

ROSA, Guilherme Moraes; RODRIGUES, Ruan Chaves; LOTUFO, Roberto; NOGUEIRA, Rodrigo. **Yes, BM25 is a Strong Baseline for Legal Case Retrieval**. 2021. DOI: https://doi.org/10.48550/arXiv.2105.05686.

SALTON, G.; WONG, A.; YANG, C. S. **A vector space model for automatic indexing**. *Communications of the ACM*, v. 18, n. 11, p. 613-620, nov. 1975. DOI: https://doi.org/10.1145/361219.361220.

SANTHANAM, Keshav; KHATTAB, Omar; SAAD-FALCON, Jon; POTTS, Christopher; ZAHARIA, Matei. **ColBERTv2: Effective and efficient retrieval via lightweight late interaction**. 2021. DOI: https://doi.org/10.48550/arXiv.2112.01488.

SIQUEIRA, Felipe A.; VITÓRIO, Douglas; SOUZA, Ellen; SANTOS, José A. P.; ALBUQUERQUE, Hidelberg O.; DIAS, Márcio S.; SILVA, Nádia F. F.; DE CARVALHO, André C. P. L. F.; OLIVEIRA, Adriano L. I.; BASTOS-FILHO, Carmelo. **Ulysses Tesemõ: a new large corpus for Brazilian legal and governmental domain**. *Language Resources & Evaluation*, 2024. DOI: https://doi.org/10.1007/s10579-024-09762-8.

SUN, Weiwei; YAN, Lingyong; MA, Xinyu; WANG, Shuaiqiang; REN, Pengjie; CHEN, Zhumin; YIN, Dawei; REN, Zhaochun. **Is ChatGPT good at search? Investigating large language models as re-ranking agents**. 2023. DOI: https://doi.org/10.48550/arXiv.2304.09542.

THAKUR, Nandan; REIMERS, Nils; RÜCKLÉ, Andreas; SRIVASTAVA, Abhishek; GUREVYCH, Iryna. **BEIR: A heterogenous benchmark for zero-shot evaluation of information retrieval models**. 2021. DOI: https://doi.org/10.48550/arXiv.2104.08663.

TUNSTALL, Lewis; VON WERRA, Leandro; WOLF, Thomas. **Natural language processing with transformers: building language applications with Hugging Face**. 1. ed. Sebastopol, CA: O'Reilly Media, 2022. ISBN 9781098103245.

VASWANI, Ashish; SHAZEER, Noam; PARMAR, Niki; USZKOREIT, Jakob; JONES, Llion; GOMEZ, Aidan N.; KAISER, Łukasz; POLOSUKHIN, Illia. **Attention is all you need**. In: PROCEEDINGS OF THE 31ST INTERNATIONAL CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS (NIPS'17). DOI: https://doi.org/10.48550/arXiv.1706.03762.

WANG, Jiajia; HUANG, Jimmy X.; TU, Xinhui; WANG, Junmei; HUANG, Angela J.; LASKAR, Md Tahmid Rahman; BHUIYAN, Amran. **Utilizing BERT for Information Retrieval: Survey, Applications, Resources, and Challenges**. 2024. DOI: https://doi.org/10.48550/arXiv.2403.00784.

ZHENG, Zhi; HUI, Kai; HE, Ben; HAN, Xianpei; SUN, Le; YATES, Andrew. **BERT-QE: Contextualized query expansion for document re-ranking**. In: FINDINGS OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: EMNLP 2020. Online, 2020. p. 4718–4728. Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/2020.findings-emnlp.424.

Anexo A – *Queries* disponíveis no qrels

Quadro 2 – Conjunto de *queries* 1

ID	*Query*	ID	*Query*
1	técnica e preço	26	contrato com a administração pública
2	restos a pagar	27	inexigibilidade e singularidade
3	aditivo a contrato	28	mera participação e epp
4	adesão a ata de registro de preços	29	decisão judicial
5	sobrepreço e superfaturamento	30	fiscal de contrato
6	restrição a competitividade	31	medida cautelar
7	acréscimos e supressões	32	é possível a aplicação concomitante
8	obras e serviços de engenharia	33	reajuste de contrato
9	fiscalização de contratos	34	publicidade e concurso público
10	diárias e passagens	35	falecimento e multa
11	bens e serviços comuns	36	independência das instâncias
12	parcelas de maior relevância e valor significativo	37	planejamento estratégico
13	despesa sem cobertura contratual	38	sistema s e licitação
14	decreto-lei 4.657/1942	39	citação e validade
15	contas e materialidade	40	multa a particulares
16	inexequibilidade e comprovação	41	licitação e preço de mercado
17	impedimento de licitar e contratar	42	edital modificação
18	aditivo e obra	43	padronização marca
19	fraude a licitação	44	interesse recíproco

20	auditoria interna	45	pesquisa de preços
21	fracionamento de despesas	46	planilha de custos e formação de preços
22	novo e improrrogável prazo	47	publicidade e propaganda
23	inidoneidade de licitante	48	processo seletivo e sistema s
24	licitações e contratos	49	modalidade de licitação
25	exigência de atestado de capacidade	50	antecipação de pagamento

Fonte: Elaboração própria (2024).

Quadro 3 – Conjunto de *queries* 2

ID	*Query*
51	concessão remunerada de uso de bens públicos modalidade
52	citação válida falecimento
53	garantia contratual patrimônio líquido mínimo
54	garantia de participação patrimônio líquido mínimo
55	despesas sem cobertura contratual multa
56	uso de áreas comerciais em aeroportos pregão
57	cessão das áreas comerciais de centrais públicas de abastecimento de gêneros alimentícios licitação
58	capacidade bens pertinentes e compatíveis com o objeto da licitação comprovação
59	extrapolação dos limites para alterações consensuais qualitativas de contratos de obras e serviços
60	responsabilidade do gestor sucessor omissão do antecessor
61	projeto de parceria público-privada modalidade
62	objeto do certame contrato social licitante
63	responsabilidade gestão dos recursos Fundo Municipal de Saúde
64	pesquisa de mercado orçamento fornecedores
65	pesquisa de preços obrigatória em licitações
66	vantajosidade adesão à ata
67	compensação entre acréscimos e supressões contratos administrativos
68	vedado restabelecimento item suprimido
69	marca aquisição cartuchos
70	SUS recursos União competência do TCU
71	preços referenciais sistemas oficiais estimativa de custos
72	ponderação técnica e preço
73	pontuação desarrazoada licitações técnica e preço
74	prorrogação de contrato administrativo término prazo de vigência
75	atestado de capacidade técnica serviços advocatícios conselho de fiscalização profissional
76	reposição de importâncias indevidamente percebidas boa-fé servidores erro escusável por órgão/entidade
77	aquisição direta risco a saúde
78	pregão divulgação dos preços estimados no edital
79	instrutor treinamento inexigibilidade de licitação
80	reajuste de preços e repactuação
81	recomposição reajuste
82	retenção de pagamentos perda da regularidade fiscal
83	irregulares despesas desnecessárias e anteriores à celebração do contrato
84	vínculo trabalhista ou societário pré-existente competitividade das licitantes
85	técnica e preço profissionais quadro permanente
86	erro material desclassificação da proposta
87	vedação à inclusão documento habilitação diligência
88	exigência comprovação de quantitativos mínimos características semelhantes
89	extensão sanção de impedimento de licitar e contratar
90	habilitação técnico-operacional obras e serviços de engenharia
91	controle adesão
92	alvará de funcionamento habilitação jurídica

93	acumulação de cargo secretário municipal professor
94	prescrição da pretensão punitiva julgamento das contas
95	Banco de Preços em Saúde referência medicamentos
96	alteração contratual serviços já previstos no edital
97	licitação princípio da legalidade estrita afastado
98	renovação de contrato serviços de natureza continuada pesquisa de preços
99	contratações de software cartas de exclusividade
100	assistência médica a servidores

Fonte: Elaboração própria (2024).

Quadro 4– Conjunto de *queries* 3

ID	*Query*
101	Qual é a modalidade de licitação adequada para a concessão remunerada de uso de bens públicos?
102	A citação é considerada válida após o falecimento do responsável se a defesa já foi apresentada?
103	A prestação de garantia contratual é permitida juntamente com a exigência de patrimônio líquido mínimo?
104	A exigência de garantia de participação é permitida juntamente com a exigência de patrimônio líquido mínimo?
105	Quais são as consequências para os responsáveis que realizam despesas sem cobertura contratual?
106	O pregão é adequado para concessões de uso de áreas comerciais em aeroportos?
107	Quais normas devem ser observadas na cessão das áreas comerciais de centrais públicas de abastecimento de gêneros alimentícios?
108	É necessário exigir ao licitante comprovar que já forneceu bens pertinentes e compatíveis com o objeto da licitação?
109	É permitido a extrapolar os limites estabelecidos no art. 65, §§ 1º e 2º, da Lei 8.666/1993 para alterações consensuais qualitativas de contratos de obras e serviços?
110	Qual a responsabilidade do gestor sucessor quando o gestor antecessor omitiu o dever de prestar contas?
111	Qual é a modalidade de licitação utilizada para contratar serviços técnicos necessários à estruturação de projeto de parceria público-privada relativo a infraestrutura de rede de iluminação pública?
112	O que acontece se não houver compatibilidade entre o objeto do certame e as atividades previstas no contrato social da empresa licitante?
113	Quem é responsável pela gestão dos recursos do Fundo Municipal de Saúde?
114	É obrigatória pesquisa de mercado na elaboração do orçamento-base da licitação?
115	A pesquisa de preços correntes no mercado é obrigatória em licitações?
116	O que fazer para verificar a vantajosidade da adesão à ata?
117	É permitida a compensação entre acréscimos e supressões nos contratos administrativos?
118	Quais são as condições para que o restabelecimento total ou parcial de quantitativo de item anteriormente suprimido por aditivo contratual não configure compensação vedada pela jurisprudência do TCU?
119	A Administração pode indicar preferência por marcas em licitações para aquisição de cartuchos de tinta?
120	O TCU é responsável pela fiscalização das ações e serviços de saúde pagos com recursos repassados pela União no âmbito do Sistema Único de Saúde?
121	O que deve ser feito se não for possível obter preços referenciais nos sistemas oficiais para estimativa de custos em processos licitatórios?
122	É necessária ponderação entre a pontuação técnica e a de preço em uma licitação do tipo técnica e preço?
123	Como a pontuação desarrazoada pode limitar a competitividade nas licitações do tipo técnica e preço?
124	É possível prorrogar um contrato administrativo fora do término do prazo de vigência?

125	Qual é o problema na exigência de atestado de capacidade técnica para contratação de serviços advocatícios por conselho de fiscalização profissional?
126	É necessária reposição de importâncias indevidamente percebidas, de boa-fé, por servidores, em virtude de erro escusável de interpretação de lei por parte do órgão/entidade?
127	É possível aquisição direta quando a falta de produto ou serviço pode colocar em risco a saúde das pessoas?
128	É obrigatória a divulgação dos preços estimados no edital nos pregões para aquisição de medicamentos?
129	A contratação de professores, conferencistas ou instrutores para ministrar cursos de treinamento ou aperfeiçoamento de pessoal se enquadra na inexigibilidade de licitação?
130	Qual é a diferença entre reajuste de preços e repactuação?
131	É possível a aplicação da recomposição mesmo após a aplicação do reajuste previsto no contrato?
132	A retenção de pagamentos é permitida em caso de perda da regularidade fiscal?
133	Por que despesas desnecessárias e anteriores à celebração do contrato são consideradas irregulares?
134	Como o vínculo trabalhista ou societário pré-existente afeta a competitividade das licitantes no certame licitatório?
135	É permitido atribuir pontuação a uma empresa licitante nas licitações de técnica e preço com base na posse de determinados tipos de profissionais em seu quadro permanente?
136	Um erro material deve levar à desclassificação antecipada da proposta?
137	Há vedação à inclusão de novo documento destinado a atestar condição de habilitação preexistente em sede de diligência?
138	A exigência de comprovação de quantitativos mínimos em obras ou serviços com características semelhantes é legal?
139	Qual a extensão da sanção de impedimento de licitar e contratar?
140	Quais documentos são exigidos para habilitação técnico-operacional em certames de obras e serviços de engenharia?
141	Qual a finalidade do controle das autorizações de adesão?
142	Em que situações a apresentação do alvará de funcionamento é permitida na habilitação jurídica?
143	É permitida a acumulação do cargo de secretário municipal com o cargo de professor?
144	A prescrição da pretensão punitiva do TCU afeta o julgamento das contas?
145	O Banco de Preços em Saúde é uma referência de preços válida para a aquisição de medicamentos?
146	É possível a inclusão de serviços já previstos no edital em uma alteração contratual?
147	O princípio da legalidade estrita pode ser afastado em favor de outros princípios no procedimento licitatório?
148	Quais fontes devem ser prioritárias na pesquisa de preços para a renovação do contrato de serviços de natureza continuada?
149	Devem ser aceitas nas contratações de software as cartas de exclusividade emitidas pelos próprios fabricantes?
150	Qual é a regra geral para a contratação de entidade para prestação de serviços de assistência médica a servidores?

Fonte: Elaboração própria (2024).

Anexo B – Prompts

Figura 4: Prompt para gerar perguntas respondidas pelo campo Enunciado.

```
Elabore até 5 perguntas curtas e diretas que possam ser respondidas a partir do enunciado a seguir:
{ENUNCIADO}
```

Fonte: Elaboração própria (2024).

Figura 5: Prompt para extrair um escore de relevância.

```
[System]
Você é um especialista na jurisprudência do Tribunal de Contas da União com o objetivo de  avaliar se um enunciado de jurisprudência responde a uma pergunta.
Deve retornar um valor de escore de 0 a 3, sendo:
0 - irrelevante - o enunciado não responde a pergunta;
1 - relacionado - o enunciado apenas está no tópico da pergunta;
2 - relevante - o enunciado responde parcialmente a pergunta;
3 - altamente relevante - o enunciado responde a pergunta, tratando completamente de  suas nuances.
Em seguida, explique a razão para a escolha do escore.
Por favor, responda no formato JSON, contendo as chaves Razão e Score;
o valor de Razão deve ser a motivação para a escolha do score;
o valor de Score deve ser o valor do score atribuído.

[User]
Pergunta: {QUERY}
Enunciado de jurisprudência: {ENUNCIADO}
```

Fonte: Elaboração própria (2024).

Figura 6: Prompt para extrair sinônimos de palavras relevantes do campo Enunciado.

```
[System]
Você é um especialista em sistemas de busca que usam o algoritmo BM25 e está trabalhando na indexação de uma base de dados de jurisprudência do Tribunal de Contas da União. Essa base está sofrendo com o problema de descasamento de vocabulário, ou seja, o usuário usa termos de pesquisa que não estão no enunciado da jurisprudência. Trata-se de um problema comum, pois o usuário não sabe como o enunciado está escrito.

Para mitigar esse problema, você lerá um enunciado e escolherá as cinco palavras mais relevantes desse enunciado. Em seguida, escolherá dois ou três sinônimos para cada palavra.

O formato de sua resposta deve ser:
- [Palavra 1]: [Sinônimos]
- [Palavra 2]: [Sinônimos]
- [Palavra 3]: [Sinônimos]
- [Palavra 4]: [Sinônimos]
- [Palavra 5]: [Sinônimos]

Tudo o que você responder será indexado. Por isso, forneça apenas a lista das cinco palavras mais relevantes e seus sinônimos. Não inclua nenhuma instrução ou explicação sobre sua resposta.

[User]
{ENUNCIADO}
```

Fonte: Elaboração própria (2024).

Figura 7: Prompt para reescrever o campo Enunciado.

[System]
Você é um especialista em sistemas de busca que usam o algoritmo BM25 e está trabalhando na indexação de uma base de dados de jurisprudência do Tribunal de Contas da União. Essa base está sofrendo com o problema de descasamento de vocabulário, ou seja, o usuário usa termos de pesquisa que não estão no enunciado da jurisprudência. Trata-se de um problema comum, pois o usuário não sabe como o enunciado está escrito.

Para mitigar esse problema, além do enunciado original, será indexada uma versão reescrita do enunciado usando sinônimos. Dessa forma, espera-se que o usuário da pesquisa tenha maior probabilidade de encontrar o que procura.

Sua tarefa é reescrever o enunciado. Procure sinônimos mais comuns para as palavras usadas no enunciado original. O público alvo da pesquisa é o cidadão em geral, com variados graus de instrução. Por isso, o enunciado deve ser reescrito de forma simplificada.

Tudo o que você responder será indexado. Por isso, não forneça nada além da reescrita do enunciado.

[User]
{ENUNCIADO}

Fonte: Elaboração própria (2024).

Anexo C – Resultados dos experimentos para os 20 primeiros resultados da pesquisa

Tabela 4 – Resultados para o conjunto de *queries* 1 ($k = 20$)

Cenário	$P@20$	$R@20$	$MRR@20$	$nDCG@20$
Atual (baseline)	0,2180	0,3616	0,4043	0,3415
BM25 (baseline)	0,2090	0,3473	0,5446	0,3525
BM25 + docT5query(5)	0,2530	0,4142	0,6586	0,4466
BM25 + sinônimos(GPT-3.5)	0,2200	0,3648	0,5563	0,3786
BM25 + sinônimos(GPT-4o)	0,2150	0,3563	0,5586	0,3702
BM25 + sinônimos(Llama)	0,2210	0,3663	0,5508	0,3781
BM25 + docT5query(5) + sinônimos(GPT-3.5)	*0,2580*	0,4226	0,6765	**0,4629**
BM25 + docT5query(5) + sinônimos(GPT-4o)	**0,2640**	**0,4327**	0,6602	*0,4606*
BM25 + docT5query(5) + sinônimos(Llama)	*0,2600*	*0,4257*	**0,6922**	*0,4616*

Fonte: Elaboração própria (2024).

Tabela 5 – Resultados para o conjunto de *queries* 2 ($k = 20$)

Cenário	$P@20$	$R@20$	$MRR@20$	$nDCG@20$
Atual (baseline)	0,1570	0,2611	0,8667	0,4457
BM25 (baseline)	0,2710	0,4538	0,8665	0,5618
BM25 + docT5query(5)	*0,2860*	0,4783	0,8816	0,5951
BM25 + sinônimos(GPT-3.5)	*0,2830*	0,4723	0,8845	0,5884
BM25 + sinônimos(GPT-4o)	0,2700	0,4501	*0,9049*	0,5772
BM25 + sinônimos(Llama)	0,2750	0,4594	0,8735	0,5773
BM25 + docT5query(5) + sinônimos(GPT-3.5)	**0,2920**	**0,4889**	0,8925	**0,6087**
BM25 + docT5query(5) + sinônimos(GPT-4o)	*0,2840*	0,4740	*0,8965*	*0,6018*
BM25 + docT5query(5) + sinônimos(Llama)	*0,2870*	0,4781	**0,9085**	*0,6055*

Fonte: Elaboração própria (2024).

Tabela 6 – Resultados para o conjunto de *queries* 3 ($k = 20$)

Cenário	$P@20$	$R@20$	$MRR@20$	$nDCG@20$
Atual (baseline)	0,0090	0,0164	0,1100	0,0364
BM25 (baseline)	0,2690	0,4762	0,9175	0,5867
BM25 + docT5query(5)	*0,2770*	0,4891	**0,9425**	*0,6124*
BM25 + sinônimos(GPT-3.5)	*0,2780*	0,4909	*0,9389*	*0,6097*
BM25 + sinônimos(GPT-4o)	*0,2730*	0,4840	0,9153	0,5945
BM25 + sinônimos(Llama)	*0,2730*	0,4840	*0,9340*	0,6004
BM25 + docT5query(5) + sinônimos(GPT-3.5)	*0,2740*	0,4851	0,9237	0,5988
BM25 + docT5query(5) + sinônimos(GPT-4o)	*0,2810*	*0,4964*	0,9187	*0,6085*
BM25 + docT5query(5) + sinônimos(Llama)	**0,2830**	**0,5012**	*0,9400*	**0,6182**

Fonte: Elaboração própria (2024).

Instituto Serzedello Corrêa

Missão

Aprimorar a Administração Pública em benefício da sociedade por meio do controle externo

Visão

Ser referência na promoção de uma Administração Pública efetiva, ética, ágil e responsável

TCU TRIBUNAL DE CONTAS DA UNIÃO