Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Similaridade Semântica entre Acórdãos para Apoio na Formulação de Jurisprudência do TCU

Wagner Miranda Costa

Dissertação apresentada como requisito parcial para conclusão do
Mestrado Profissional em Computação Aplicada

Orientador
Prof. Dr. Glauco Vitor Pedrosa

Brasília
2023

Ficha catalográfica elaborada automaticamente,
com os dados fornecidos pelo(a) autor(a)

COSTA, WAGNER MIRANDA
C838s Similaridade Semântica entre Acórdãos para Apoio na Formulação de Jurisprudência do TCU / WAGNER MIRANDA COSTA; orientador GLAUCO VITOR PEDROSA. -- Brasília, 2023.
68 p.

Dissertação(Mestrado Profissional em Computação Aplicada) -- Universidade de Brasília, 2023.

1. Similaridade Semântica. 2. Word Embeddings. 3. Processamento de Linguagem Natural. 4. Representação Vetorial de Documentos. 5. Bag-of-Concepts. I. PEDROSA, GLAUCO VITOR, orient. II. Título.

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Similaridade Semântica entre Acórdãos para Apoio na Formulação de Jurisprudência do TCU

Wagner Miranda Costa

Dissertação apresentada como requisito parcial para conclusão do
Mestrado Profissional em Computação Aplicada

Prof. Dr. Glauco Vitor Pedrosa (Orientador)
PPCA/UnB

Prof. Dr. Thiago de Paulo Faleiros
CIC/UnB

Dr. Eduardo de Paula Costa
Corteva Agriscience

Prof. Dr. Gladston Luiz da Silva
Coordenador do Programa de Pós-graduação em Computação Aplicada

Brasília, 21 de dezembro de 2023

Dedicatória

À minha esposa, Del, por tudo, sempre.

Aos meus filhos, Fellipe e Cecília: o que tenho de mais valioso. Me fazem querer ser um pai melhor, todos os dias.

E aos meus pais, duas fortalezas.

Sei o tanto que torcem por mim. Amo todos vocês.

Agradecimentos

Agradeço ao apoio dos dirigentes e colegas do Tribunal de Contas da União, que viabilizaram e contribuíram para a minha participação nesse Mestrado. E à banca e ao meu orientador, prof. Dr. Glauco Pedrosa, pela precisa condução e pelo inestimável apoio durante todo o tempo de realização da pesquisa e da elaboração da dissertação.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES), por meio do Acesso ao Portal de Periódicos.

Resumo

Jurisprudência se refere ao conjunto de decisões reiteradas sobre determinado assunto, constituindo uma espécie de precedente judicial. No âmbito do Tribunal de Contas da União (TCU), órgão responsável por exercer o controle externo da Administração Pública Federal, a jurisprudência representa as interpretações consolidadas das normas aplicáveis à fiscalização financeira e operacional das contas públicas dos órgãos e entidades da União. Uma vez que a elaboração da jurisprudência é definida a partir de um agrupamento de acórdãos similares, é relevante desenvolver ferramentas automatizadas que auxiliem os especialistas responsáveis por esta atividade. Porém, essa é uma tarefa desafiadora para a área da computação, devido às especificidades do vocabulário presente nos textos dos acórdãos e ao volume massivo de dados a serem processados. Sendo assim, é necessário desenvolver abordagens escaláveis, eficazes e eficientes, e que possuam baixo custo computacional. Este trabalho apresenta o estudo e implementação de algumas abordagens para a representação desses documentos textuais, tanto em nível de palavra quanto em nível de conceito. Como contribuição, foi proposta uma nova abordagem denominada BoC-Th (*Bag of Concepts with Thesaurus*), que gera histogramas ponderados de conceitos definidos a partir da distância das palavras do documento ao seu respectivo termo similar dentro de um tesauro. Esta abordagem permite enfatizar palavras com maior significado no contexto, gerando, assim, vetores mais discriminativos. Realizaram-se avaliações experimentais comparando a abordagem proposta com as abordagens tradicionais para representação de documentos. O método proposto obteve resultados superiores entre as técnicas avaliadas para recuperação de documentos jurisprudenciais. O BoC-Th aumentou a precisão média em comparação às abordagens tradicionais, incluindo a versão original BoC (*Bag of Concepts*), ao mesmo tempo que foi mais rápido que as representações tradicionais BoW, BM25 e TF-IDF. A abordagem proposta contribuiu para enriquecer uma área com características peculiares, fornecendo um recurso para recuperação de informações textuais de forma mais precisa e rápida do que outras técnicas baseadas em processamento de linguagem natural.

Palavras-chave: Processamento de Linguagem Natural, Recuperação de Informação, Representação Vetorial de Documentos, Bag-of-Concepts, Word Embeddings

Abstract

Jurisprudence refers to the set of repeated decisions on a given subject, constituting a type of judicial precedent. Within the scope of the Federal Audit Court (TCU), the body responsible for exercising external control of the Federal Public Administration, jurisprudence represents the consolidated interpretations of the rules applicable to the financial and operational supervision of the public accounts of the Union's bodies and entities. Since the elaboration of jurisprudence is defined based on a grouping of similar rulings, it is important to develop automated tools that assist the specialists responsible for this activity. However, this is a challenging task for the area of computing, due to the specificities of the vocabulary present in the texts of the rulings and the massive volume of data to be processed. Therefore, it is necessary to develop scalable, effective and efficient approaches that have low computational cost. This work presents the study and implementation of some approaches for representing these textual documents, both at the word level and at the concept level. As a contribution, a new approach called BoC-Th (*Bag of Concepts with Thesaurus*) was proposed, which generates weighted histograms of concepts defined based on the distance of the words in the document to their respective similar term within a thesaurus. This approach allows us to emphasize words with greater meaning in the context, thus generating more discriminative vectors. Experimental evaluations were carried out comparing the proposed approach with traditional approaches for document representation. The proposed method obtained superior results among the techniques evaluated for recovering jurisprudential documents. BoC-Th increased average accuracy compared to traditional approaches, including the original BoC (*Bag of Concepts*), while also being faster than traditional BoW, BM25, and TF-IDF representations. The proposed approach contributed to enriching an area with peculiar characteristics, providing a resource for retrieving textual information more accurately and quickly than other techniques based on natural language processing.

Keywords: Natural Language Processing, Information Retrieval, Document Vector Representation, Bag-of-Concepts, Word Embeddings

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

BM25 Best Matching 25.

BoC Bag-of-Concepts.

BoC-Th Bag-of-Concepts with Thesaurus.

BoW Bag-of-Words.

CBOW Continuous Bag of Words.

CPC Código de Processo Civil.

IDF Inverse Document Frequency.

mAP mean Average Precision.

MPTCU Ministério Público junto ao Tribunal de Contas da União.

NLP Processamento de Linguagem Natural.

OCTC Ontology-Enabled Concept-Based Text Categorization.

SGBD Sistemas Gerenciadores de Bancos de Dados.

SRI Sistema de Recuperação de Informação.

STF Supremo Tribunal Federal.

TCU Tribunal de Contas da União.

TF-IDF Term Frequency–Inverse Document Frequency.

VCE Vocabulário de Controle Externo.

Capítulo 1

Introdução

Este capítulo apresenta a contextualização e motivação para o desenvolvimento deste trabalho, a abordagem proposta, as questões de investigação e a organização geral dos restantes capítulos desta dissertação.

1.1 Contextualização

De acordo com a Constituição Federal [4], em seu art. 71, o controle externo do Governo Federal é de responsabilidade do Congresso Nacional, e exercido com o auxílio do Tribunal de Contas da União (TCU). Dentre as competências do TCU, podem ser elencadas a responsabilidade pela fiscalização contábil, financeira, orçamentária, operacional e patrimonial dos órgãos e entidades públicas do país quanto à legalidade, legitimidade e economicidade. Em suma, o TCU é responsável por zelar pela boa gestão dos recursos públicos.

As ações de natureza técnica ou administrativa a cargo do TCU são estruturadas em processos, assim como em praticamente todo o serviço público. O processo é o mecanismo por meio do qual a competência institucional do Órgão é desempenhada, e seu resultado, ao fim de todo um trâmite processual, é expresso em um acórdão [5]. Em tribunais colegiados, como o TCU, o conjunto de decisões (acórdãos) sobre um mesmo tema, ainda que a respeito de casos concretos distintos, compõe uma jurisprudência [6].

Dentro do ordenamento jurídico ocidental, há dois sistemas vigentes: *common law* e *civil law*. O primeiro tem na figura dos juízes, bem como na dos advogados, um papel relevante da formulação do Direito, visto que sua aplicação não decorre essencialmente das leis, mas sim das decisões provenientes de situações concretas, como a jurisprudência. Já no *civil law*, a interpretação da legislação expressa rege a tomada de decisões, com a norma jurídica (as leis) possuindo grande valor no processo legal [7]. No Brasil, acolhe-se a tradição romano-germânica do *civil law*, vez que o sistema jurídico pauta-se na legislação

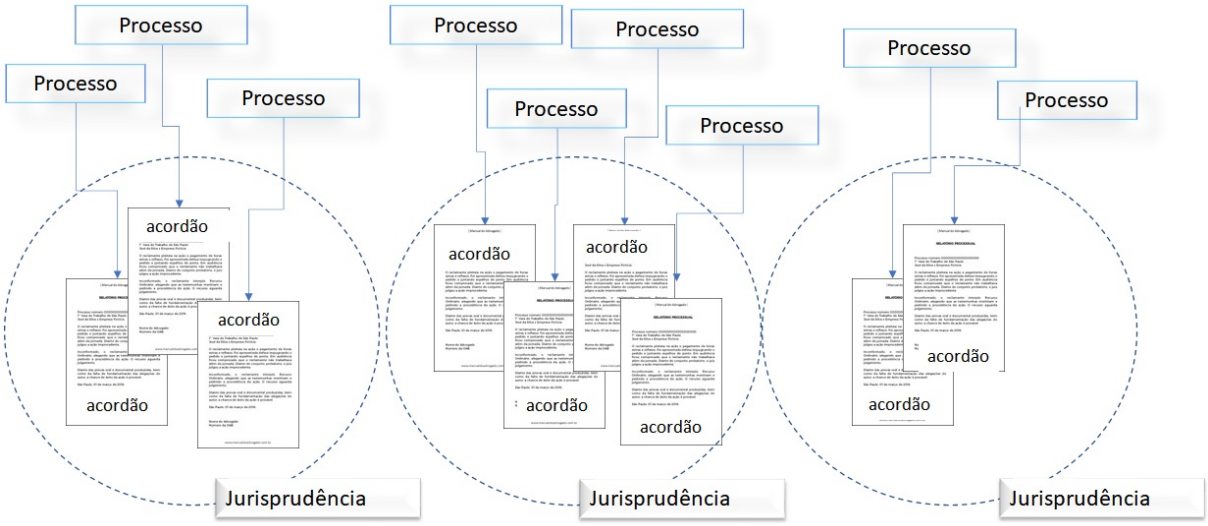


Figura 1.1: A decisão de um processo é apresentada em um instrumento jurídico denominado acórdão. O conjunto de acórdãos acerca de um determinado assunto constitui uma jurisprudência.

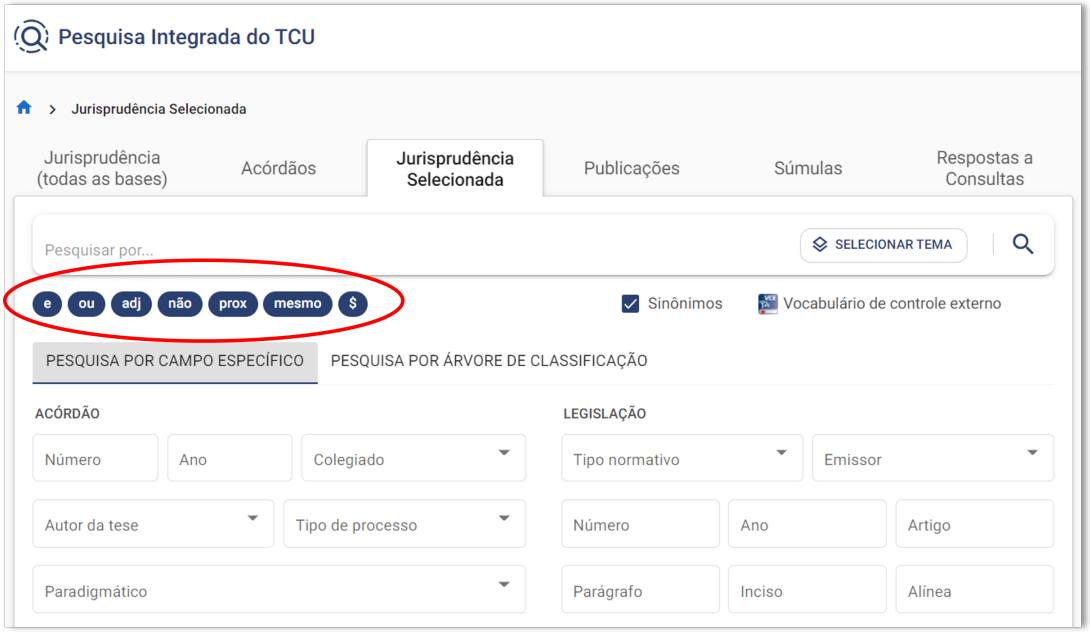
existente [8]. Contudo, há uma evidente aproximação entre ambos os princípios [9], com a adoção da jurisprudência como legitima fonte argumentativa. O novo Código de Processo Civil (CPC)[10] deixa isso claro, ao estabelecer que:

> Art. 489. São elementos essenciais da sentença:
> [...]
> § 1º Não se considera fundamentada qualquer decisão judicial, seja ela interlocutória, sentença ou acórdão, que:
> [...]
> VI -– deixar de seguir enunciado de súmula, jurisprudência ou precedente invocado pela parte, sem demonstrar a existência de distinção no caso em julgamento ou a superação do entendimento.

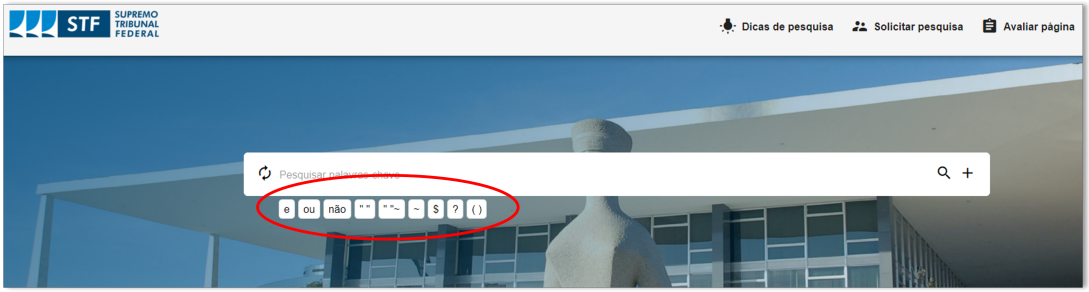
Desta forma, a jurisprudência se mostra essencial à estabilidade jurídica, considerando que esta é alcançada com a uniformização de entendimento e da aplicação de decisões similares. Por isso, é muito importante desenvolver soluções que auxiliem o operador de Direito (magistrados, advogados, assessores jurídicos, dentre outros) na elaboração de jurisprudência, a fim de mitigar as lacunas que possam existir na aplicação da lei, bem como na busca pela segurança jurídica.

1.2 Motivação e Desafios

Nos últimos anos, algumas ferramentas de pesquisa foram desenvolvidas com o objetivo de apoiar a atividade dos operadores do Direito quando da necessidade de se localizar



(a)



(b)

(c)

Figura 1.2: Ferramentas de pesquisa de jurisprudência de Tribunais Superiores, com destaque para os operadores de pesquisa: (a) Tribunal de Contas da União (TCU); (b) Supremo Tribunal Federal (STF); (c) Superior Tribunal de Justiça (STJ)
Fontes: (a) `www.tcu.gov.br` (b) `www.stf.jus.br` (c) `www.stj.jus.br`

a jurisprudência sobre temas específicos. No TCU, por exemplo, a Pesquisa Integrada[1] disponibiliza acesso à toda a base de dados indexada do Órgão. Ferramentas com esse propósito, que estão presentes nos sítios de órgãos jurídicos (ver Figura 1.2), utilizam operadores que localizam termos e expressões no conteúdo indexado. Esse tipo de busca pressupõe que o usuário que irá utilizar o sistema de busca tenha um certo nível de conhecimento sobre o texto a ser localizado, uma vez que a montagem da expressão de busca visa fornecer à ferramenta parâmetros que indiquem como termos ou partes de textos estão presentes nos documentos. Uma evolução desse tipo de busca é que o sistema consiga recuperar conteúdos com semânticas equivalentes ou aproximadas. A Tabela 1.1 relaciona os operadores mais comuns utilizados em ferramentas de pesquisa textual.

Operador	**Descrição**	**Exemplo**
AND	localiza documentos que possuam todos os termos informados na expressão	'edital AND roteador' retonará resultados que cotenham ambos os termos 'edital' e 'roteador'
OR	localiza documentos que possuam ao menos um dos termos informados na expressão	'autarquia OR estatal' retornará resultados que contenham a palavra 'autarquia' ou a palavra 'estatal'
NOT	despreza documentos que possuam o termo informado na expressão	'auditoria NOT conformidade' retornará resultados que contenham a palavra 'auditoria', mas eliminará aqueles que contenham a palavra 'conformidade'
$\sim$	localiza documentos em que termos informados estão a uma dada distância um do outro	'economia $\sim$4 produção' retornará resultados em que a palavra 'economia' está no máximo a 4 palavras de distância de 'produção'
*	substitui caracteres em uma expressão de busca	'susten*' retornará documentos que possuam as palavras 'sustentável', 'sustentação', 'sustentabilidade', etc.
""	localiza documentos que possuam o texto exato que está contido entre aspas	'"aprendizagem de máquina"' retornará resultados que contenham a exata expressão 'aprendizagem de máquina'
( )	agrupa termos de pesquisa para formação de expressões mais elaboradas	'(edital OR licitação) AND suspenso' retornará resultados que contenham as palavra 'suspenso' e 'edital' ou as palavras 'suspenso' e 'licitação'

Tabela 1.1: Exemplos de operadores utilizados em ferramentas de busca textual.

Atualmente, no âmbito do TCU, uma equipe de especialistas tem a incumbência de avaliar, dentre os acórdãos proferidos pela Corte, aqueles que possuem similaridade e relevância suficiente para que constem como enunciados de súmulas, que embora não determinem a convicção do Magistrado, certamente subsidiam sua interpretação sobre

[1] `https://pesquisa.apps.tcu.gov.br/#/pesquisa/jurisprudencia`

questões próximas. Um recurso computacional que possibilite a comparação semântica entre acórdãos e enunciados traria, portanto, uma desejável e valiosa instrumentalização a essa tarefa.

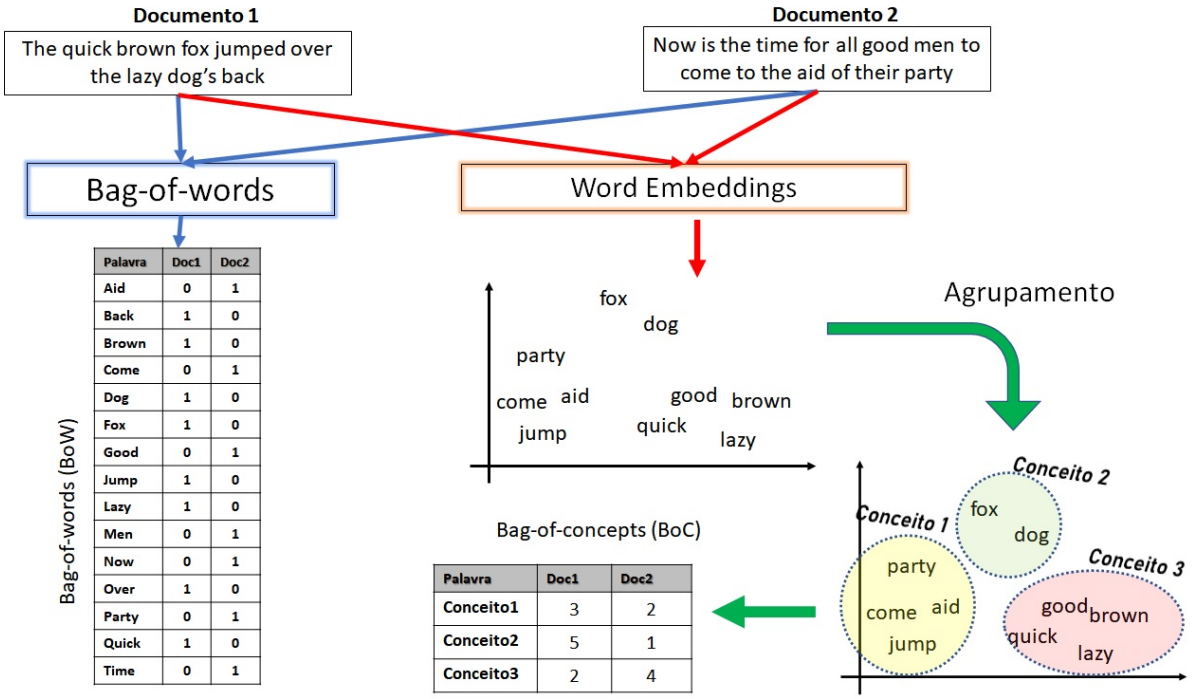
1.3 Problema de Pesquisa

No contexto da análise de similaridade entre jurisprudências e acórdãos do TCU, este estudo deverá lidar com dois principais desafios. Um deles é o contexto especializado de aplicação do trabalho, que é uma área cujos documentos possuem estruturas implícitas e uma linguagem peculiar com características únicas. O outro desafio é a busca por informações em uma base de dados com grande volume de documentos. Por isso, é importante desenvolver abordagens escaláveis que tenham execução com baixo custo computacional, a fim de retornar de maneira eficiente e eficaz documentos próximos à expectativa do usuário.

Do ponto de vista computacional, para a implementação de um Sistema de Recuperação de Informação (SRI), os documentos devem ser representados de forma que o computador possa interpretar e diferenciar cada documento dentro da coleção. Esta representação visa a retratar numericamente documentos de texto não estruturados para torná-los matematicamente computáveis [11]. Então, para um dado conjunto de documentos de texto $D = \{d_1, d_2, d_3, ..., d_n\}$, onde cada d_i representa um documento, o problema da representação textual é representar cada $d_i \in D$ como um ponto em um espaço dimensional S, onde a similaridade entre pares de pontos neste espaço dimensional S é definida através de uma função de distância.

A Figura 1.3 ilustra algumas abordagens de representação vetorial de documentos textuais. A linha azul mostra a modelagem Bag-of-Words (BoW) que é uma das técnicas mais simples e populares. Ela consiste em representar um documento por suas frequências de palavras e, por sua simplicidade, permite uma interpretabilidade intuitiva do vetor numérico gerado. No entanto, esse método sofre com a alta dimensionalidade dos seus vetores e desconsidera o impacto de palavras semanticamente similares. Para mitigar tal limitação, alguns modelos semanticamente mais enriquecidos foram propostos, tais como o modelo Word2Vec [12] - que é baseado em redes neurais profundas - cujo objetivo é representar palavras em vetores numéricos de tal forma que palavras semanticamente similares terão representações vetoriais semelhantes (linha vermelha). Uma maneira simples para representar um documento usando a abordagem Word2Vec é computar a média dos vetores das palavras que ocorrem no documento [13]. Porém, isso gera vetores de documentos menos representativos e falha em fornecer uma interpretação intuitiva por trás de seus vetores de documento gerados. A linha verde mostra a abordagem Bag-of-Concepts (BoC) [14], que

surgiu como um método alternativo de representação de documentos textuais que mitiga os pontos fracos das abordagens BoW e Word2Vec. A abordagem BoC cria conceitos através do agrupamento dos vetores-de-palavras gerados a partir do Word2Vec, e utiliza as frequências desses agrupamentos de conceitos para representar os documentos. Por meio de conceitos, a abordagem BoC incorpora o impacto de palavras semanticamente semelhantes, enriquecendo a representação semântica dos documentos textuais sem aumentar drasticamente a dimensionalidade do vetor-de-característica necessário para representá-lo.



Palavra	Doc1	Doc2
Aid	0	1
Back	1	0
Brown	1	0
Come	0	1
Dog	1	0
Fox	1	0
Good	0	1
Jump	1	0
Lazy	1	0
Men	0	1
Now	0	1
Over	1	0
Party	0	1
Quick	1	0
Time	0	1

Palavra	Doc1	Doc2
Conceito1	3	2
Conceito2	5	1
Conceito3	2	4

Figura 1.3: Visão geral de diferentes abordagens para representação vetorial de documentos. As linhas azuis indicam o modelo Bag-of-Words. As linhas vermelhas indicam técnicas de *word embedding* e as linhas verdes indicam a representação Bag-of-Concepts.

Em suma, o contexto deste trabalho foi avaliar técnicas de representação textual, tanto em nível de palavra quanto tem nível de conceito e analisar a possibilidade de se combiná-las a fim de aprimorar a busca por similaridade entre os acórdãos do TCU. Sendo assim, definiu-se a seguinte questão de pesquisa:

A partir de uma base massiva de enunciados de jurisprudência, como recuperar de forma eficaz e eficiente um conjunto de enunciados similares a um dado acórdão informado pelo usuário?

1.4 Objetivos

O objetivo geral deste trabalho foi propor e avaliar abordagens de representação textual que possibilitem recuperar acórdãos similares do TCU a fim de apoiar a formulação de jurisprudência.

Para atingir o objetivo geral, alguns objetivos específicos foram definidos:

1. Investigar abordagens não-supervisionadas para a representação vetorial de documentos textuais;
2. Construir uma base de dados de acórdãos e jurisprudências para realização de testes experimentais;
3. Validar experimentalmente, junto a uma equipe técnica especializada do TCU, o desempenho do sistema de recuperação proposto.

1.5 Organização da dissertação

O texto deste trabalho está organizado da seguinte forma:

- Capítulo 2 apresenta as fundamentações teóricas e os trabalhos correlatos que motivaram o desenvolvimento deste trabalho
- Capítulo 3 apresenta os materiais e métodos utilizados;
- Capítulo 4 apresenta os resultados obtidos em um estudo de caso envolvendo a recuperação de acórdãos e jurisprudências do TCU
- Capítulo 5 apresenta a conclusão e potenciais trabalhos futuros.

Capítulo 2

Fundamentação Teórica

Este capítulo apresenta uma revisão teórica e conceitual das técnicas utilizadas neste trabalho para representar palavras e documentos em um espaço vetorial, bem como os trabalhos correlatos que motivaram o desenvolvimento deste trabalho.

2.1 Recuperação de Informação

A recuperação da informação é realizada por meio do uso de um Sistema de Recuperação de Informação (SRI). Ao se utilizar um SRI, o usuário está interessado em recuperar "informação" sobre um determinado assunto. Essa característica é o que diferencia os SRI dos tradicionais Sistemas Gerenciadores de Bancos de Dados (SGBD). Conforme [15], as funções de um SRI são as seguintes: 1) representação das informações contidas nos documentos e expressas pelos processos de indexação e descrição dos documentos; 2) armazenamento e gestão física e ou lógica desses documentos e de suas representações; e 3) recuperação das informações ali contidas e dos próprios documentos armazenados no sistema.

O processo de recuperar informações consiste em identificar a partir de um conjunto de documentos (corpus) quais documentos são similares à necessidade de informação do usuário. A busca por expressões (palavras-chave) é uma maneira de expressar essa necessidade de informação do usuário. Os principais mecanismos de busca utilizados atualmente, como o Google, funcionam através de uma interface de consulta na qual o usuário informa palavras-chave a serem utilizadas como referência para a localização dos links, ordenados por relevância, para as páginas onde os termos foram encontrados. Dessa forma, cabe ao usuário analisar dentre esse enorme conjunto de respostas aquelas que melhor se ajustam à sua requisição. Uma tarefa que nem sempre é factível dado o volume de respostas retornado.

2.1.1 Recuperação de Documentos Textuais

A recuperação de documentos textuais é uma outra maneira de se buscar por "informação". Nessa abordagem o usuário informa ao SRI um documento e este, por sua vez, retorna um conjunto de documentos similares ao documento-consulta. A similaridade semântica entre documentos é um dos problemas cruciais na área do Processamento de Linguagem Natural [16]. Encontrar similaridade entre documentos é uma tarefa utilizada em vários domínios, tais como: recomendação de livros e artigos semelhantes ou identificação de documentos plagiados, por exemplo.

A Figura 2.1 descreve esquematicamente todo o processo envolvido na recuperação de documentos textuais. Esse processo inicia-se representando cada documento do corpus através de seu conteúdo. Durante essa representação vetorial são extraídos conceitos do documento através da análise de seu conteúdo e traduzidos em um vetor numérico. Esta representação identifica o documento e define seus pontos de acesso para a busca por similaridade, que consiste em aplicar uma função de distância sobre o vetor do documento de consulta fornecido pelo usuário com os vetores de cada documento do corpus. Essa função de distância retorna um valor de similaridade e, a partir desse valor, um ranking de documentos similares é fornecido ao usuário.

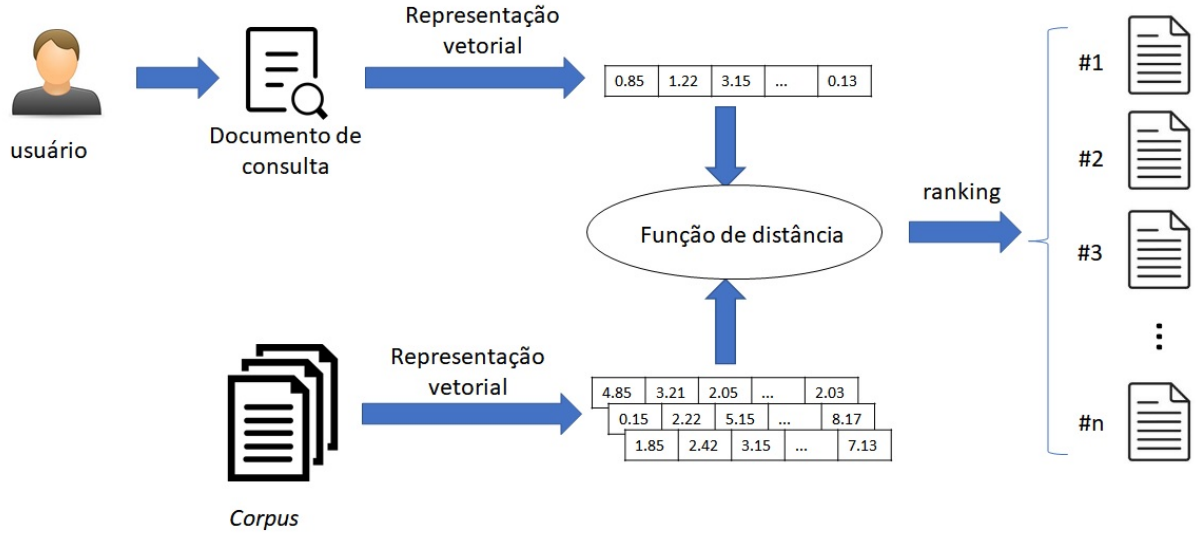


Figura 2.1: Arquitetura de um processo de recuperação de informação

2.1.2 Recuperação de Documentos Jurídicos

Pesquisas sobre busca e organização de documentos jurídicos são apresentadas nessa seção. O comportamento dos assessores de ministros do STF quanto à necessidade e forma de recuperação de dados foi investigado por [17], que constatou que a adequada infor-

mação jurídica é fundamental para o desempenho de suas atividades. Uma característica observada é que a busca da informação jurídica é baseada no ministro ao qual o assessor está vinculado, o que reforça a importância da existência de um suporte tecnológico para a localização de documentos e temas símeis.

Por sua vez, o trabalho de [18] estudou mecanismos de busca de jurisprudência utilizados por diferentes tribunais no Brasil. O trabalho permitiu confirmar que, de fato, as ferramentas de suporte à recuperação da informação colaboram fortemente nas tarefas de organização e refinamento da exploração dos dados – não-estruturados em grande parte. Magistrados ouvidos na pesquisa relataram que há uma intensa demanda por consultas em autos de decisões proferidas.

Uma arquitetura para agrupar documentos jurídicos a partir da sua vinculação jurisprudencial foi proposta por [19], em que foi apresentada uma estrutura de pré-processamento com *tokenização*, *lemmatização* e subsequente geração de vetores de atributos fundamentados por uma base léxica criada com a consolidação de diferentes dicionários (em português e latim), desambiguadores de termos probabilísticos e por regras, e a união dos tesauros da Justiça Federal (Tesauro da Justiça Federal - TLF) e o Vocabulário Controlado Brasileiro (VCB), mantido pelo Senado Federal. De forma sucinta, a solução proposta reune a jurisprudência em grupos obtidos em um processo de cluterização com refinamentos sucessivos limitados a um limite mínimo de vinte atributos por grupo, valor definido devido à baixa dimensionalidade dos vetores de atributos dos documentos. Com base nos centróides desses grupos, documentos jurídicos são classificados, tendo a jurisprudência pertencente ao respectivo grupo vinculada a si.

Com o objetivo de apoiar a etapa preliminar do processo de elaboração de leis da Câmara dos Deputados brasileira, o trabalho de [20] investigou variações do algoritmo de ranqueamento de relevância BM25, usual em recuperação de informações em representações Bag-of-Words de documentos. Como outros trabalhos semelhantes, esforça-se em aprimorar a fase de pré-processamento, com foco na redução de dimensionalidade por meio de *stemming* e *lemmatization* dos atributos.

Confirmando uma tendência na abordagem dos estudos sobre recuperação de jurisprudência, o trabalho de [21] avalia algoritmos de *stemmização* usados na montagem do corpus de bases de dados jurisprudenciais e conclui, contraintuitivamente, que rotinas menos agressivas – que resultam em menor redução de dimensionalidade de atributos – apresentam uma relação custo-benefício melhor, alcançando uma maior efetividade na recuperação de informações. Outra contribuição deste trabalho diz respeito ao rigor estatístico na aferição dos algoritmos, mediante cálculo de precisões do resultado das recuperações (*precision at document cutoff*, *R-Precision* e *mean Average Precision*) e seu uso na análise de significância estatística para responder a hipótese de pesquisa "a redução de

dimensionalidade degrada a recuperação de documentos jurisprudenciais?"para cada uma das técnicas avaliadas.

Investigando a aplicação de técnicas de comparação de similaridade entre teses jurídicas, [22] identificou que há ganhos marginais na utilização de modelos semânticos frente aos tradicionais modelos de representação baseados em frequência, mesmo quando combinados. Sugeriu, ainda, a utilização de resultados confirmados (relevantes) de consultas como base supervisionada de treinamento para ajuste dos modelos semânticos.

2.2 Representação Vetorial de Documentos

Para que um computador consiga estabelecer o grau de similaridade entre documentos é necessário, primeiramente, definir uma maneira de representar o texto contido nos documentos em uma estrutura quantificável (ou um objeto matemático – geralmente em forma vetorial), para que seja possível a realização de cálculos de similaridade com esse documento. Assim, converter um documento em um objeto matemático e definir uma medida de similaridade são basicamente os passos necessários para que máquinas realizem a análise de similaridade semântica entre documentos.

Nos modelos de espaço vetorial, cada documento de uma coleção (corpus) é representado como um vetor neste espaço. Essa representação permite comparar quaisquer dois documentos por similaridade (por exemplo, calculando a similaridade de cosseno entre dois vetores de documentos). Neste modelo, cada termo exclusivo em uma coleção de documentos é atribuído a uma dimensão em um espaço vetorial. Tanto documentos quanto consultas são expressos como vetores, em que as coordenadas refletem as frequências ou ponderações dos termos presentes nesses documentos ou consultas. As variantes do modelo de espaço vetorial que são abordadas nesta dissertação são as abordagens: Bag-of-Words (BoW), Term Frequency–Inverse Document Frequency (TF-IDF), Best Matching 25 (BM25), a técnica Word2vec e o modelo Bag-of-Concepts (BoC). A Figura 2.2 ilustra essas técnicas, que serão detalhadas a seguir.

2.2.1 *Bag-of-Words (BoW)*

Uma abordagem comumente adotada e eficaz para a representação de documentos é o modelo *Bag-of-Words*. O modelo BoW atribui um vetor a um documento como $d = \{x_1, x_2, x_3, ..., x_l\}$, em que x_i denota o número normalizado de ocorrências do i-ésimo termo (palavra) no documento, e l é o tamanho da coleção de termos (palavras) dos documentos presentes na base de dados.

A abordagem BoW é um método simples, mas eficaz, para mapear um documento em um vetor de comprimento fixo. No entanto, a função de mapeamento no modelo BoW é

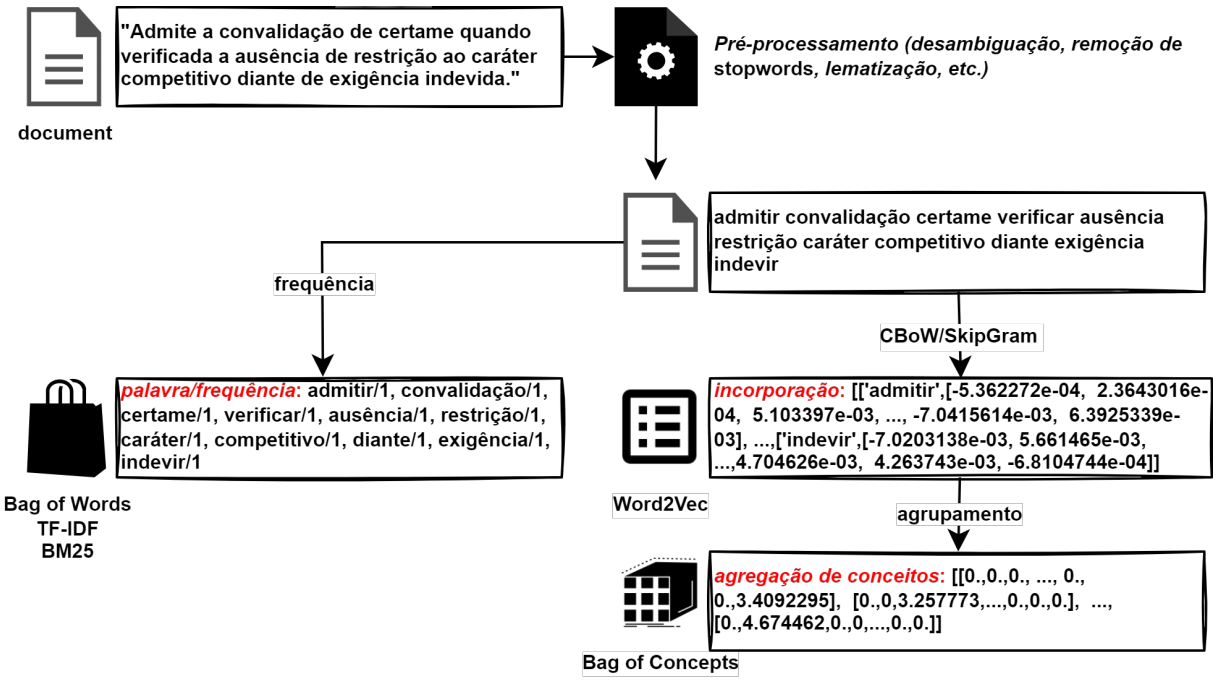


Figura 2.2: Diferentes abordagens de representação de documentos textuais: *Bag of Words*, *TF-IDF*, *BM25*, *Word2Vec* e *Bag of Concepts*.

hard (ou binária), visto que ela representa apenas a presença ou ausência de um termo base no documento. A função de mapeamento rígido tem várias limitações. Primeiramente, o vetor gerado para cada documento é extremamente esparso, pois um documento contém apenas uma porção muito pequena de todos os termos básicos do vocabulário. Em segundo lugar, as representações BoW podem não capturar efetivamente a semântica dos documentos, uma vez que documentos semanticamente semelhantes, mas com diferentes conjunto de palavras, serão mapeados para espaços vetoriais muito diferentes.

2.2.2 *Term Frequency–Inverse Document Frequency (TF-IDF)*

A técnica Term Frequency–Inverse Document Frequency (TF-IDF) tem origem no campo da recuperação de informações, tendo suas bases sido inicialmente propostas pelo alemão Hans Peter Luhn [23], na década de 1950, e pela britância Karen Spärck Jones [24], cientista da computação e pesquisadora em processamento de linguagem natural, na década de 1970. Essa técnica mede a importância relativa de uma palavra (termo) em um documento, considerando toda uma coleção de documentos (ou corpus). É dividida em duas partes:

1. Frequência do Termo (TF): corresponde à proporção do número de vezes que um termo aparece em um documento em relação ao total de termos neste documento.

2. Frequência Inversa do Documento (IDF): avalia o quão comum é um termo em toda a coleção de documentos. Confere um valor IDF maior a termos mais raros, ressaltando sua relevância.

Em resumo, TF mede a importância local de um termo em um documento específico, enquanto IDF mede a importância global do termo em relação a uma coleção de documentos. A multiplicação desses dois fatores resulta em um peso que destaca termos que são frequentes em um documento específico, porém raros na coleção global, destacando assim a relevância desses termos para o conteúdo do documento em questão.

A fórmula 2.1 calcula o peso de cada termo t em um documento d de uma coleção D, resultando em uma pontuação que reflete a relevância relativa deste termo naquele contexto específico. Termos que ocorrem frequentemente em um documento e são raros em outros recebem pesos mais elevados, enquanto termos comuns em toda a coleção são ponderados de forma menos significativa.

$$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \times \text{IDF}(t, D) \tag{2.1}$$

onde:

$$\text{TF}(t, d) = \frac{\text{Número de vezes que o termo } t \text{ ocorre em } d}{\text{Número total de termos em } d}$$

$$\text{IDF}(t, D) = \log\left(\frac{\text{Número total de documentos na coleção } D}{\text{Número de documentos que contêm o termo } t + 1}\right)$$

2.2.3 *Best Matching* (BM25)

A técnica Best Matching 25 (BM25) é uma variação aprimorada do modelo TF-IDF. Ela incorpora parâmetros ajustáveis que permitem uma melhor adaptação a diferentes tamanhos de documentos. Também busca atenuar riscos de saturação na pontuação causados pela influência de termos frequentes que são pouco discriminativos. Sua representação é dada pela equação 2.2:

$$\text{BM25}(q, d) = \sum_{i=1}^{n} \left(\frac{f(q_i, d) \cdot (k_1 + 1)}{f(q_i, d) + k_1 \cdot \left(1 - b + b \cdot \frac{\text{len}(d)}{\text{avg\_len}}\right)}\right) \cdot \log\left(\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5}\right) \tag{2.2}$$

em que:

q representa a consulta,

d representa o documento,

n é a frequência do termo na coleção,

$f(q_i, d)$ é a frequência do termo q_i no documento d,

N é o número total de documentos na coleção,

$\text{len}(d)$ é o comprimento do documento em termos de palavras,

avg_len é o comprimento médio dos documentos na coleção,

k_1, b são parâmetros ajustáveis.

A primeira parte da fórmula, relativa ao cálculo da frequência do termo (TF), pondera a contribuição da frequência na pontuação, com ajustes para evitar a saturação e para considerar o tamanho do documento. A segunda parte, referente à frequência inversa do documento (IDF), possui a incumbência de atribuir pesos aos termos com base em sua importância relativa (sua "raridade", o que os tornam mais informativos).

2.2.4 Vetores de Palavras (*Word Embeddings*)

A ideia central por trás das técnicas conhecidas como *word embeddings* (incorporação de palavras) é atribuir uma representação vetorial a cada palavra de modo que palavras que são semanticamente afins fiquem próximas umas das outras no espaço vetorial. Uma das técnicas de *word embeddings* mais populares é a técnica Word2Vec [1]. O Word2Vec é uma abordagem de Processamento de Linguagem Natural (NLP) fundamentada em uma teoria que propõe que cada palavra é distintamente caracterizada e, em certa medida, definida pelo contexto das outras palavras ao seu redor. Diferentemente das técnicas baseadas em frequência de termos, o Word2Vec mapeia palavras para vetores densos de números reais em um espaço multidimensional, onde a proximidade espacial expressa a similaridade semântica.

A Figura 2.3 mostra as duas arquiteturas usadas no Word2Vec: Continuous Bag of Words (CBOW) e Skip-gram. No modelo CBOW, o objetivo é prever uma palavra alvo a partir do contexto em que ela está inserida. Já no modelo Skip-gram, a ideia é prever o contexto (palavras vizinhas) dada uma palavra de entrada. Ambas as arquiteturas utilizam uma camada intermediária (*embedding layer*) para aprender as representações vetoriais.

Devido à sua arquitetura e ao treinamento não-supervisionado, o Word2Vec pode ser construído eficientemente em um corpus não anotado de grande escala. O Word2Vec

é capaz de codificar relações linguísticas significativas entre palavras em *embeddings* de palavras aprendidas.

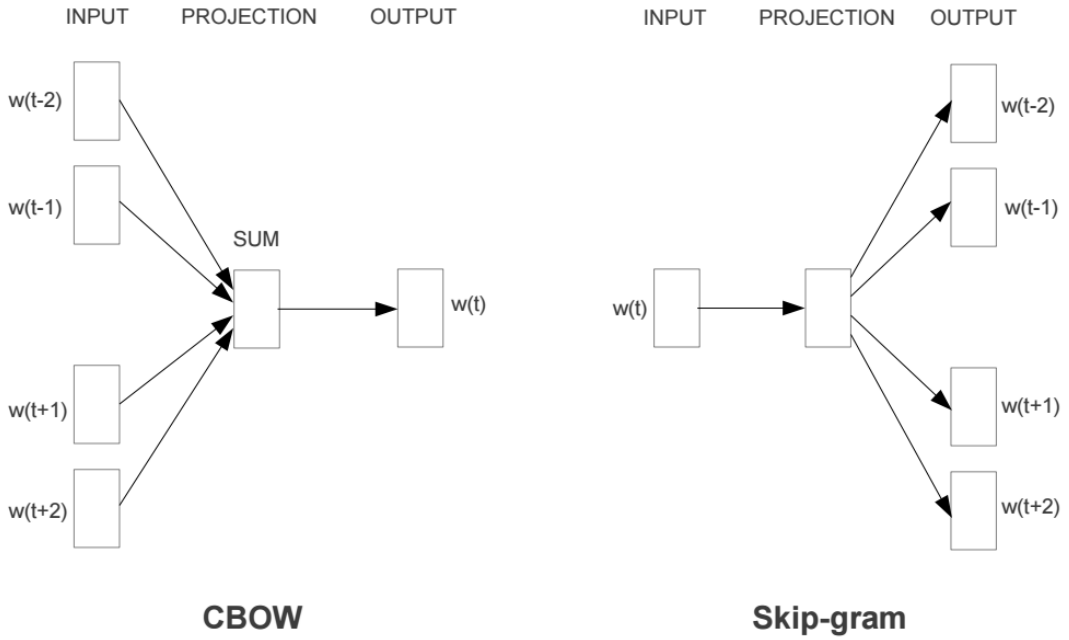


Figura 2.3: Arquiteturas do modelo Word2Vec. O CBOW é uma arquitetura de modelo de linguagem que se propõe a prever uma palavra de destino com base no contexto fornecido pelas palavras circundantes. O Skip-gram prevê as palavras circunvizinhas a partir de uma palavra de entrada. Fonte: [1]

Um grande mérito da abordagem *word embeddings* é que a semelhança semântica entre duas palavras pode ser convenientemente avaliada com base na medida de semelhança de cosseno entre suas correspondentes representações vetoriais.

Um exemplo da utilização de *embedding* em nível de documento para a avaliação de similaridade de textos jurídicos é apresentado em [25]. Usando o algoritmo *paragraph vector* [12], os autores consolidam os vetores de características do texto e estimam os demais parágrafos com o método *stochastic gradient descent*. Em seguida, obtém-se a orientação angular dos documentos, a partir do cálculo da similaridade de cosseno de seus vetores. Os documentos que possuírem orientações similares são considerados também semanticamente similares.

2.2.5 Bag-of-Concepts (BoC)

Uma outra maneira de representação vetorial de documentos é o Bag-of-Concepts (BoC). Nesta abordagem, as palavras de um documento são associadas à um conceito. Um conceito é um agrupamento de palavras semanticamente relacionadas. Esses conceitos são

obtidos a partir de um dicionário de conceito, e uma das formas de gerar esse dicionário é através do agrupamento das palavras obtidas por alguma técnica de *word embeddings*.

A Figura 2.4 mostra os passos da abordagem BoC, que está dividida em duas fases: codificação e sumarização. A fase de codificação atribui um conceito para cada palavra contida no documento, e a fase de sumarização contabiliza a frequência de cada conceito no documento.

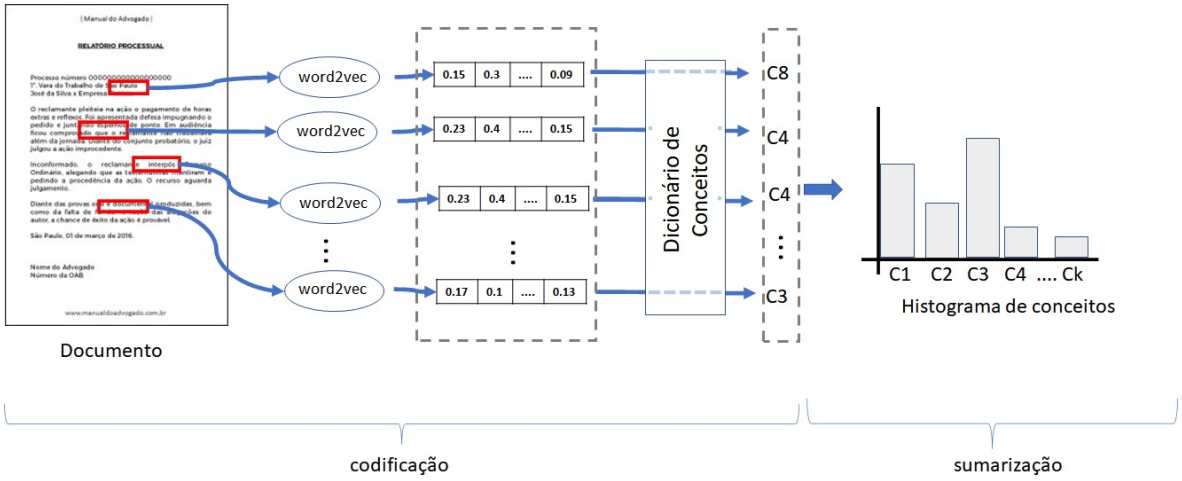


Figura 2.4: Etapas da abordagem BoC para representar um documento em um histograma de conceitos.

A abordagem BoC busca tratar algumas limitações apresentadas pelo BoW, como a ausência de representatividade semântica, além da alta dimensionalidade e esparsidade do vetores [14]. A definição de "conceito" como unidade de significado foi proposto por [26], que estudou a utilização de bases de conhecimento – no caso, a Wikipedia – na criação de Bag-of-Concepts. Como resultado, obteve-se um ganho de até 157% na modelagem de classificadores, em comparação com a modelagem Bag-of-Words.

Geração do Dicionário de Conceitos

Seja $W = \{w_1, w_2, ..., w_v\}$ o vocabulário que abrange todas as palavras existentes em um corpus, e v o tamanho desse vocabulário. Cada palavra $w_j \in W$ é representada por um vetor r-dimensional obtido por uma técnica de *word embeddings*, como Word2Vec, gerando uma matriz $S \in R^{v \times r}$. A abordagem BoC faz uso de um dicionário de conceitos $C = \{c_1, c_2, ..., c_k\}$, onde cada conceito $c_i \in R^r$ é o centroide de um grupo $C_i \subset S$, tal que:

$$c_i = \frac{1}{|C_i|} \sum_{w_j \in S} w_j, \quad \forall i \in \{1, 2, ..., k\} \tag{2.3}$$

em que, $|C_i|$ é a quantidade de palavras associadas ao grupo C_i e k é a quantidade de conceitos.

Pode-se utilizar diferentes algoritmos de agrupamento para a geração dos grupos de palavras. A abordagem tradicional, apresentada por [14], utiliza o algoritmo não-supervisionado *Spherical k-means*, que é indicado para o processamento de vetores esparsos e com grande dimensionalidade, que vem a ser o caso dos dados analisados neste trabalho.

Modelos Estendidos para o Bag-of-Concepts

A abordagem Bag-of-Concepts tem sido utilizada por diversos trabalhos com diferentes aplicações práticas, e por vezes têm sofrido modificações no intuito de se aprimorar a abordagem para uma efetiva identificação de similaridade entre documentos. A seguir, alguns desses trabalhos são apresentados.

O problema da sinonímia e da polissemia é considerado em [27], que ressalta a desvantagem na análise vetorial de textos curtos, uma vez que mesmo textos com significados similares podem não compartilhar das mesmas palavras.

O trabalho de [28] utiliza-se de bases de conhecimento, tais como *Wikipedia* e *Probase*, para subsidiar a geração dos conceitos. Faz uso de *dataless classification*, o que permite efetuar classificação de dados não-anotados. Adicionalmente, a densificação dos vetores BoC é efetuada por meio de redes neurais, o que diminui sua esparsidade sem, no entanto, aumentar sua dimensão. Os autores indicam a obtenção de uma melhora de 1,6% no *score* de correlação e uma redução de 5% em erros de categorização, e avaliaram o modelo proposto sob três perspectivas:

1. relação semântica entre entidades, comparando-o aos modelos *Wikipedia Link-based Measure* (WLM), *Keyphrase Overlap RElatedness* (KORE), *Exclusivity-based Relatedness* (ExRel) e *Combined Information Content* (CombIC)

2. categorização conceitual, comparando-o aos modelos *Word embeddings trained on Wikipedia*, *Multiword Embeddings Trained on Wikipedia*, e *Entity-category embeddings*

3. classificação não-supervisionada de documentos, contrapondo-o a *Explicit Semantic Analysis* (ESA), *Word embeddings best match* e *Word Embeddings Hungarian algorithm*

Por sua vez, o trabalho de [29] não compara documentos individualmente mas busca encontrar padrões de similaridade entre conjuntos de documentos de mesma natureza, tais como entre textos completos e seus respectivos sumários, por meio da análise de distorção dos grafos resultantes do cálculo das distâncias de cosseno dos vetores de *embeddings*. Essa abordagem possibilita que textos de diferentes tamanhos e vetores de características

de diferentes dimensões possam ser comparados semanticamente, com resultados satisfatórios. Uma maior consistência na interpretabilidade semântica das *features* dos textos é encontrada em [30], que propõe o modelo Bag-of-Concepts-Clusters (BoCCl), que além de agrupar os conceitos semanticamente similares usa base de conhecimento probabilística (ProBase) e léxica (WordNet) para desambiguação de entidades. Algo semelhante é realizado por [31], porém na desambiguação dos próprios conceitos agrupados.

Um método de caracterização apoiado por ontologia específica de domínio, denominado OCTC, é apresentado por [32]. Neste trabalho, um documento sintaticamente marcado é convertido em uma representação conceitual conforme a frequência dos descritores dos conceitos existentes na ontologia. Essa frequência define o grau de relevância do conceito em determinado documento. O estudo afirma que a representação conceitual exige menos recursos computacionais em sua geração e que a acurácia do classificador baseado na abordagem OCTC é superior à de outras técnicas, como *Latent Semantic Analysis* (LSA) e Doc2Vec combinados com *Support Vector Machine* (SVM).

O trabalho de [33] atua no tratamento de homonímia por meio da substituição do *word embedding* dos homônimos pelo seu significado semântico, a partir do agrupamento de *Embeddings from Language Models* (ELMo). A localização de relações semânticas entre termos – hiperonímia e meronímia, por exemplo – também foi a abordagem de [34], embora no contexto de análise de sentimentos. Com a utilização de bases de dados semânticas para marcações de aspectos em *bag-of-concepts*, essa sistemática alcançou um incremento de até 70% na acurácia frente a realização da mesma tarefa por meio de algoritmos baseados em redes neurais e Naive Bayes.

2.3 Funções de Distância

Uma vez que documentos textuais estejam representados por vetores numéricos, é necessário utilizar-se de funções de distâncias para medir a similaridade entre esses vetores. Os valores da similaridade entre os vetores são utilizados no ordenamento dos documentos. Portanto, a função de distância utilizada pode influenciar se a recuperação de informação será boa ou ruim. Esta escolha impacta em todo o processo de recuperação de informação e, na literatura não há um consenso sobre uma medida similaridade que seja aplicável a todos os tipos de variáveis que podem existir numa base de casos. Geralmente, os pesquisadores empregam medidas variadas, como: Distância Euclidiana, Distância de Manhattan e Similaridade por Cossenos.

2.3.1 Distância Euclidiana

Dados dois pontos p e q presentes em um espaço vetorial n-dimensional, e correspondentes aos vetores $(p_1, p_2, ..., p_n)$ e $(q_1, q_2, ..., q_n)$, respectivamente, a Distância Euclidiana, que equivale ao comprimento do segmento de reta entre esses dois pontos, é obtida por[35]:

$$dist_{Euclidiana}(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} \tag{2.4}$$

No contexto do cálculo de similaridade semântica, quanto mais próximo de 0 for a Distância Euclidiana entre as representações vetoriais de dois documentos, maior a similaridade entre eles.

2.3.2 Distância Manhattan

A Distância de Manhattan, possui funcionamento similar à Distância Euclidiana, mas não é influenciado pela diferença de escala sobre o resultado já que não há elevação ao quadrado das características a serem medidas. Dados dois pontos p e q presentes em um espaço vetorial n-dimensional, e correspondentes aos vetores $(p_1, p_2, ..., p_n)$ e $(q_1, q_2, ..., q_n)$, respectivamente, a Distância Manhattan é defina como:

$$dist_{Manhattan}(p, q) = |(p_1 - q_1)| + |(p_2 - q_2)| + \cdots + |(p_n - q_n)| \tag{2.5}$$

2.3.3 Similaridade por Cossenos

Outra forma de se mensurar o grau de similaridade entre documentos é por meio do cálculo do cosseno do ângulo entre os vetores que os representam[36]. Dados dois vetores $(p_1, p_2, ..., p_n)$ e $(q_1, q_2, ..., q_n)$, esse calculo é dado pela equação 2.6. A Figura 2.5 demonstra graficamente essa forma de análise: a medição de similaridade entre documentos obtida por meio do cosseno da distância angular entre suas representações vetoriais possui valores que variam de -1 (totalmente dissimilares) a 1 (totalmente similares).

$$sim_{Cossenos}(p, q) = \frac{\sum_{i=1}^{n} p_i \cdot q_i}{\sqrt{\sum_{i=1}^{n} p_i^2}\sqrt{\sum_{i=1}^{n} q_i^2}} \tag{2.6}$$

2.4 Medidas de Avaliação de um SRI

Para avaliar o desempenho de um SRI é necessário medir o quão bem este sistema atende a necessidade de informação do usuário, ou seja, a qualidade do ranking gerado pelo sistema. O que se almeja ao se fazer uma busca em uma base documental é encontrar

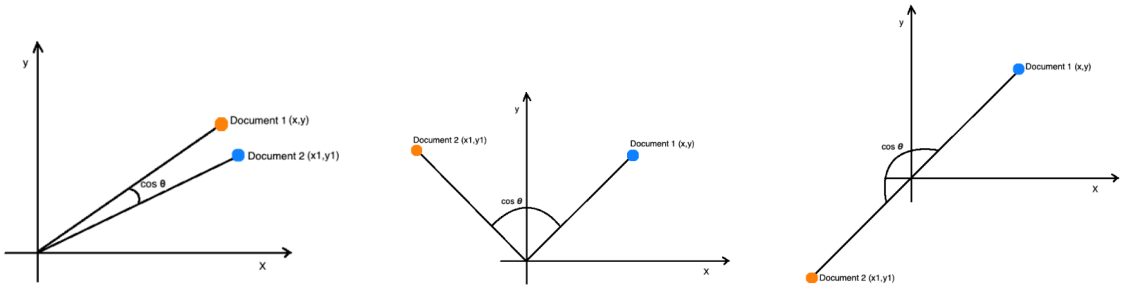


Figura 2.5: Exemplo gráfico da Similaridade por Cossenos: (a) cosseno entre documentos similares; (b) cosseno entre documentos com baixa similaridade; (c) cosseno entre documentos dissimilares. Fonte: [2]

documentos que sejam úteis (relevantes) para satisfazer a uma necessidade de informação, evitando recuperar itens não relevantes. Assim, após a execução de uma busca, pode-se dividir os documentos do corpus nos seguintes conjuntos:

- **R**: conjunto de documentos relevantes (apontados por especialistas);
- **A**: conjunto de documentos recuperados (documentos apresentados pelo sistema)

A Figura 2.4 mostra o diagrama da relação entre esses conjuntos. O usuário, ao utilizar um SRI, está interessado no conjunto dos documentos relevantes e recuperados.

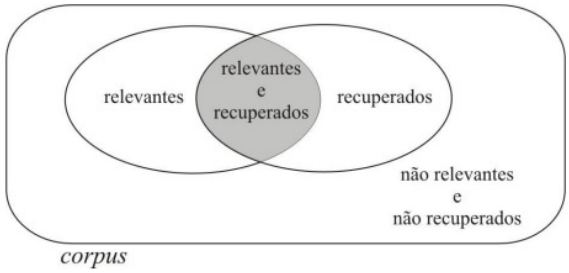


Figura 2.6: Conjuntos de elementos após se utilizar um Sistema de Recuperação de Informação

Segundo [37] as principais métricas para avaliação da qualidade do ranking gerado por um SRI são: **precisão** (*precision*) e **revocação** (*recall*). A equação 2.7 mostra o cálculo da Precisão, que é dada pela fração dos documentos recuperados que são relevantes. A equação 2.8 mostra o cálculo da Revocação, que é a fração dos documentos relevantes que são recuperados, ou seja, a revocação se refere a quão completos os resultados estão.

$$precisão = \frac{|R \cap A|}{|A|} \tag{2.7}$$

$$revocação = \frac{|R \cap A|}{|R|} \tag{2.8}$$

Em experimentos realizados na década de 1960, Cleverdon [37] observou o comportamento de mecanismos de recuperação de documentos indexados – chamados por ele de dispositivos (*devices*) – sob duas perspectivas: o incremento da probabilidade da recuperação de uma maior quantidade de documentos relevantes (*recall devices*), e a garantia de que documentos não-relevantes não são recuperados (*precision devices*). O autor constatou, décadas antes da popularização de mecanismos de busca como Google e Bing, que os sistemas de indexação e recuperação de informações são um "amálgama de dispositivos de revocação e precisão"e que eles [os dispositivos] "interagem entre si de maneira tão complexa"que seria necessário estabelecer novas métricas para a avaliação de diferentes SRI, que lidam com uma crescente base de documentos, sem necessariamente analisar a íntegra dessas grandes coleções. Isso foi alcançado com o trabalho realizado na TREC [38], que sugeriu novas métricas que permitissem essa mensuração sobre um conjunto limitado de documentos recuperados. Uma dessas métricas é utilizada neste estudo, a *Mean Average Precision* (mAP).

2.4.1 Relevância

Entretanto, a pedra de toque para a avaliação da performance de um SRI vem a ser a qualidade da indicação de relevância dos documentos recuperados. Embora não haja consenso, na literatura, sobre a definição de relevância, alguns autores buscaram aglutinar as variadas concepções. Cooper [39], ao mesmo tempo que sustenta a importância da relevância na teoria da recuperação de informação, ressente-se da então inexistência de uma única conceituação formal. Ele sugere uma definição em termos de implicação lógica, que mesmo matematicamente imprecisa, pode ser estendida para o campo de recuperação de informações expressas em linguagem natural. Quarenta e seis anos depois (1971 a 2017), Park, em estudo empírico da interpretação de relevância do ponto de vista do usuário da informação [40], frusta qualquer esperança de que o tempo e o avanço dos estudos nesse campo resultasse numa uniformização conceitual. Em abrangente revisão literária, enumera diferentes significados, onde é ressaltada a natureza idiossincrática da relevância, constatada pela estreita vinculação com a experiência e a expectativa do usuário interessado no resultado da pesquisa. Finaliza afirmando que a aferição de relevância é "um fenômeno complexo e que não pode ser representado como um estático e preciso relacionamento entre documentos e as questões de busca dos usuários".

No âmbito desta dissertação, a relevância será considerada em termos de similaridade semântica.

2.4.2 *Mean Average Precision* (mAP)

A *mean Average Precision* (mAP) sintetiza a relação precisão-revocação em um único valor entre 0 a 1 (quanto mais próximo de 1 mais acurado é o modelo). É obtida por meio da fórmula da Eq. 2.9 e que, por sua vez, equivale à média do *Average Precision* (AP) de cada uma de suas consultas calculada a partir da fórmula da Eq. 2.10, dadas por:

$$mAP = \frac{1}{N} \sum_{i=1}^{N} AP_i \tag{2.9}$$

$$AP = \frac{1}{|R|} \sum_{k=1}^{N} P(k)r(k) \tag{2.10}$$

em que N corresponde ao total de consultas do modelo, AP_i é a precisão média (*Average Precision*) da consulta i, $|R|$ é quantidade de documentos relevantes retornados pela pesquisa, N é a quantidade total de documentos retornados, $P(k)$ é a precisão na posição k (quantidade de documentos relevantes na posição k / quantidade de documentos retornados na posição k) e $r(k)$ indica a relevância do k documento (0 se não-relevante, 1 se relevante).

Para exemplificar, a Tabela 2.1 mostra o cálculo da *Average Precision* (AP) para uma consulta fictícia realizada por um SRI, que retornou um conjunto de cinco documentos $\{d_1, d_2, d_3, d_4, d_5\}$ com as respectivas indicações de relevância e ordenados por similaridade ao documento-consulta do usuário. Os documentos d_3 e d_5 não foram considerados relevantes pelo usuário.

Documentos Recuperados	d_4	d_5	d_3	d_1	d_2
k	1	2	3	4	5
relevante?	sim	sim	não	sim	não
Precisão em k	$1/1 = 1$	$2/2 = 1$	$2/3 = 0{,}67$	$3/4 = 0{,}75$	$3/5 = 0{,}6$
Relevância de k	1	1	0	1	0
$P(k)r(k)$	1	1	0	0,75	0
$AveragePrecision = \frac{1}{3}(1 + 1 + 0 + 0,75 + 0) \Rightarrow 0,92$					

Tabela 2.1: Simulação de cálculo do *Average Precision* sobre uma consulta fictícia. Fonte: Adaptado de [3]

2.4.3 *Recall@k*

A métrica *recall@k* é uma medida de desempenho que avalia a capacidade de um modelo em recuperar corretamente itens relevantes em uma lista classificada, levando em consideração apenas os primeiros "k"itens dessa lista. Ela mede a proporção de itens relevantes

que foram corretamente identificados pelo modelo entre os primeiros "k"itens apresentados (Eq. 2.11). Essa métrica é útil quando o foco está na identificação de todos os itens relevantes, ou em cenários nos quais a perda de itens relevantes é mais crítica do que a presença de itens irrelevantes na lista recuperada.

$$recall@k = \frac{\text{quantidade de itens relevantes recuperado até a posição } k}{\text{total de itens relevantes presentes na coleção}} \tag{2.11}$$

Documentos Recuperados	d_4	d_5	d_3	d_1	d_2
k	1	2	3	4	5
relevante?	sim	sim	não	sim	não
$recall@5 = \frac{3}{3} \Rightarrow 1$					

Tabela 2.2: Simulação de cálculo do *recall@k*, com $k = 5$, sobre uma consulta fictícia. Fonte: Autor, com adaptação de [3]

A Tabela 2.2 simula o cálculo do *recall@k* em um mesmo contexto de recuperação que o apresentado na Tabela 2.1. Uma informação adicional é de que o total de itens relevantes da coleção é igual a 3. Assim, observa-se que, no exemplo, o valor do indicador *recall@k* é mais expressivo que o *AveragePrecision*, pois todos os itens relevantes foram retornados na consulta.

Capítulo 3

Materiais e Métodos

Este capítulo apresenta o procedimento adotado para o desenvolvimento deste trabalho que, sistematicamente, compreende três etapas:

1. Entendimento do Negócio
2. Obtenção e Preparação dos Dados
3. Modelagem dos Dados

As duas primeiras etapas apresentam as informações, os dados e os pré-processamentos que foram utilizados para realizar os testes experimentais que serão descritos no próximo capítulo. A terceira etapa apresenta as técnicas implementadas e propõe uma nova abordagem de modelagem dos dados textuais que contempla uma combinação de técnicas da literatura.

3.1 Entendimento do negócio

O rito processual, no âmbito do Tribunal de Contas da União, tem origem com a autuação de um processo (Figura 3.1). Este instrumento organiza o conjunto de atos, comprovantes e manifestações oficiais acerca de uma determinada matéria.

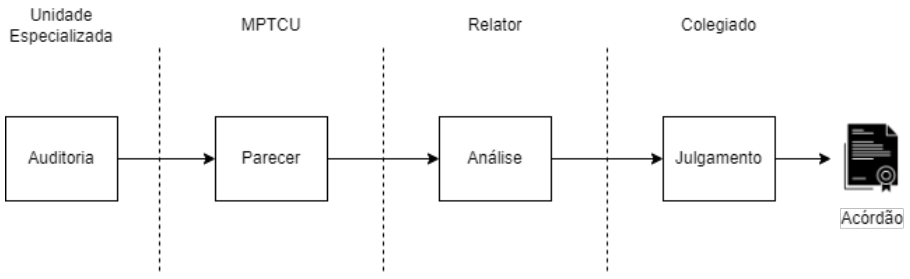


Figura 3.1: Fluxo de instrução processual e julgamento do TCU

Tipicamente, uma unidade especializada acerca do tema tratado no processo executa uma auditoria e apresenta seu resultado e uma proposta de encaminhamento ao ministro relator do processo. Algumas situações preveem a manifestação do Ministério Público junto ao Tribunal de Contas da União (MPTCU). Em seguida, o ministro relator procede à analise do relatório de auditoria e do eventual parecer do MPTCU – é discricionário acatar ou não as propostas da auditoria. Na sequência, o ministro relator submete uma minuta de decisão a um Colegiado do TCU. É este Colegiado, composto pelos Ministros do Tribunal, que apreciará o processo e decidirá sobre o assunto em tela. O instrumento que congrega a decisão adotada pela Corte é denominado **acórdão**.

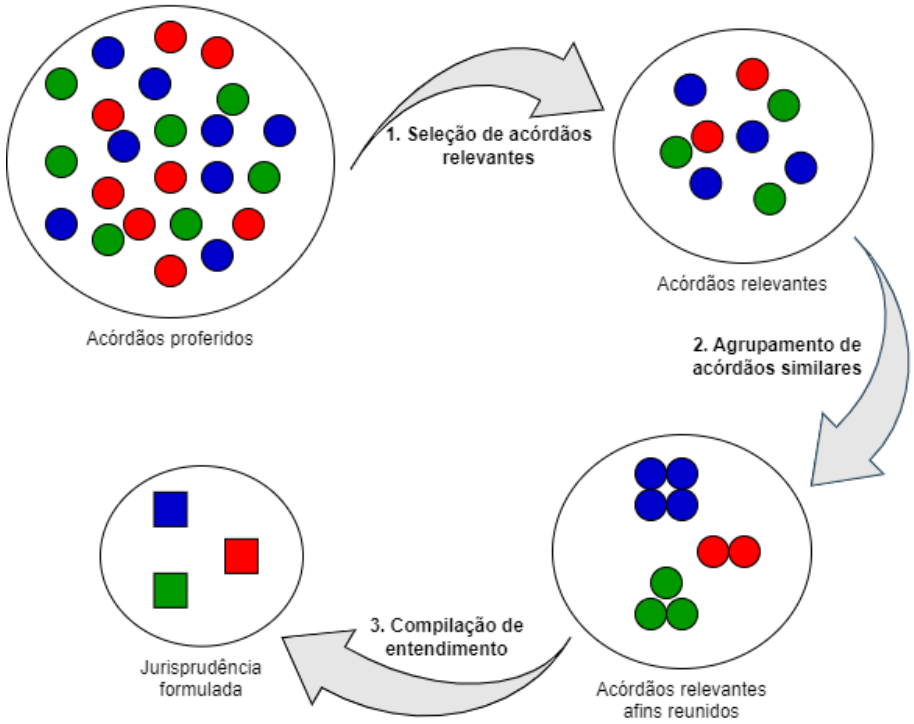


Figura 3.2: Fluxo de formulação de jurisprudência no TCU (cada cor ilustra uma dentre as áreas de atuação do Controle Externo): Na etapa (1) são identificados os acórdãos com conteúdo significativo para Jurisprudência. Em (2), os acórdãos selecionados são agrupados por afinidade e campos de atuação do TCU. Finalmente, na etapa (3), o entendimento da Corte sobre dada matéria é consolidado em enunciados de Jurisprudência. Estes enunciados são formalizados em um documento chamado súmula.

Define-se jurisprudência como um conjunto de decisões de um Tribunal a respeito de um tópico em particular, com o objetivo de pacificar seu entendimento sobre um assunto. No processo de elaboração da jurisprudência (Figura 3.2), especialistas do TCU utilizam heurísticas para reconhecer, dentre os acórdãos proferidos, aqueles que possuem maior

relevância jurisprudencial. Estes são então agrupados por similaridade e conforme as áreas de atuação do Controle Externo. É esta etapa a qual pretende-se instrumentalizar com a aplicação do resultado desta dissertação. Por fim, sumarizam o entendimento proveniente desses acórdãos em um novo documento, chamando **enunciado**. Eventualmente, sua consolidação é expressa em uma **súmula**, que é um instrumento instituído pelo Supremo Tribunal Federal (STF) em 1963 [41] que pode ser entendido como uma compilação formal da jurisprudência predominante da Corte.

3.2 Obtenção e Preparação dos Dados

3.2.1 Base de Dados

O TCU dispõe uma base de dados composta de um conjunto de 15.000 (quinze mil) enunciados de jurisprudências, publicamente disponível em seu Portal[1].

A jurisprudência é categorizada pelos especialistas por meio de três conceitos que compõem sua árvore de classificação: área, tema e subtema, conforme listado na Tabela 3.1. A característica *área* indica os campos de atuação do controle externo exercido pelo TCU. *Tema* e *subtema*, por sua vez, são refinamentos da classificação que permitem uma rotulação mais discriminativa. Ambos possuem uma relação transversal ao atributo *área*.

Característica	**Quantidade**
Área	10
Tema	356
Subtema	762

Tabela 3.1: Características da base de dados utilizada no experimento

A Figura 3.3 apresenta a quantidade de enunciados de jurisprudência por área de atuação do TCU. É possível observar que a área *Pessoal* representa o grupo com a maior quantidade de enunciados de jurisprudência. Isso é coerente com o fato de ser a área de atuação do TCU com a maior quantidade de acórdãos proferidos[2].

Área	**Quantidade de temas**
Competência do TCU	44
Contrato Administrativo	51
Convênio	32

[1] `https://pesquisa.apps.tcu.gov.br/#/pesquisa/jurisprudencia-selecionada`

[2] Relatório Anual de Atividades do TCU - ANO 2021 (`https://portal.tcu.gov.br/relatorio-anual-de-atividades-do-tcu.htm`)

Desestatização	13
Direito Processual	46
Finanças Públicas	51
Gestão Administrativa	51
Licitação	67
Pessoal	82
Responsabilidade	46

Tabela 3.2: Total de temas distintos por área de atuação.

A Tabela 3.2 lista a quantidade de diferentes temas vinculados aos enunciados de jurisprudência classificados em cada área de atuação. A Figura 3.4 apresenta essa proporção visualmente.

A Tabela 3.3 relaciona, dentre os 382 temas existentes, os 15 com maiores quantidades de enunciados classificados por estas características.

Tema	**Quantidade de enunciados**
Convênio	726
Débito	422
Aposentadoria	410
Obras e serviços de engenharia	394
Pensão civil	379
Ato sujeito a registro	365
Multa	362
Remuneração	353
Tempo de serviço	346
Qualificação técnica	334
Embargos de declaração	296
Licitação	271
Tomada de contas especial	263
Quintos	235
Prestação de contas	230

Tabela 3.3: 15 temas com maiores quantidades de enunciados.

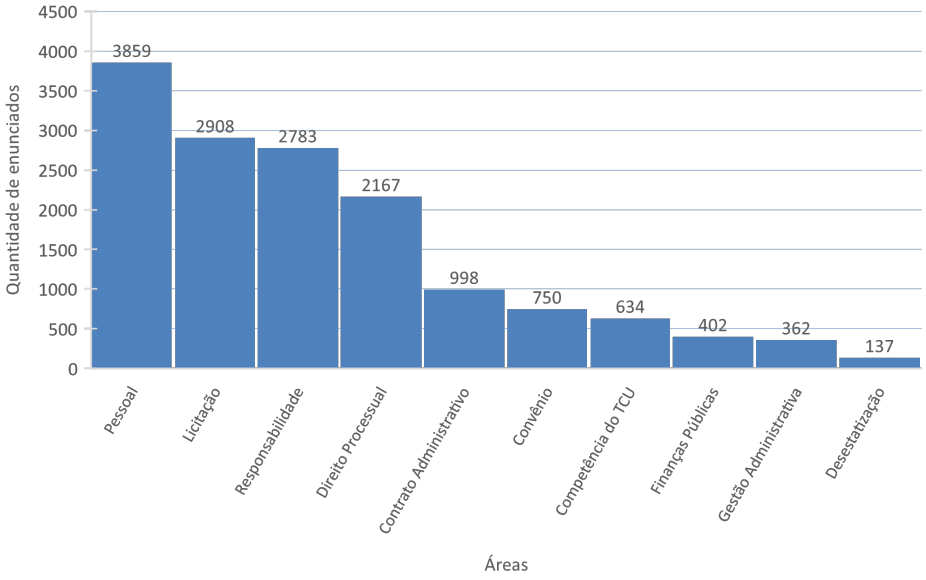


Figura 3.3: Histograma de enunciados de jurisprudência por área

3.2.2 Processamento dos Dados

A preparação adequada do texto desempenha um papel fundamental no desempenho eficaz de algoritmos de vetorização em Processamento de Linguagem Natural (NLP), garantindo que a informação textual seja devidamente estruturada e contextualizada.

A Figura 3.5 exibe as etapas de pré-processamento adotadas neste trabalho:

1. Conversão para minúsculas: todas as palavras são convertidas para minúsculas. Isso evita que o modelo trate palavras com maiúsculas e minúsculas como diferentes;
2. Remoção de números e caracteres especiais: eliminação de números, caracteres especiais, pontuações e símbolos que não contribuem para a compreensão do texto;
3. Remoção de *stopwords*: remoção de palavras comuns que não concorrem para o significado geral do texto, como artigos, preposições e pronomes;
4. *Tokenização* e lematização: *tokenização* é a segmentação do texto em unidades menores, chamadas de tokens. Em seguida, aplica-se a lematização, que é o processo linguístico de reduzir palavras flexionadas ou derivadas às suas formas base, conhecidas como "lemas", que é a forma que representa a palavra em sua classe gramatical. Sua utilidade é a de normalizar palavras e reduzir a variabilidade lexical, facilitando assim a análise de texto.

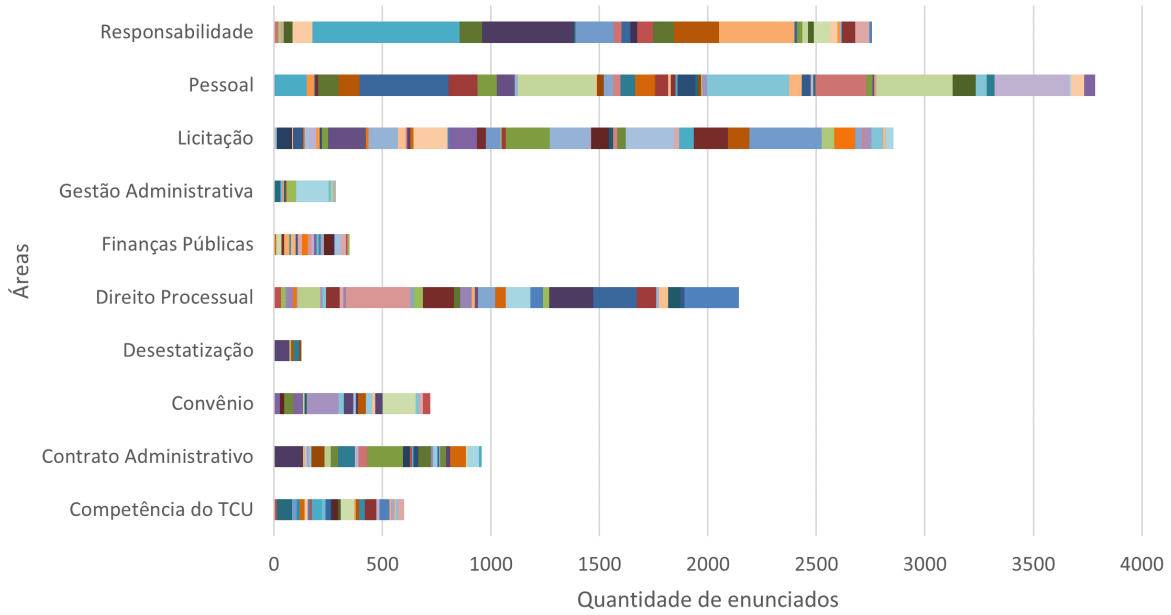


Figura 3.4: Distribuição de temas por área. Cada cor representa um tema.

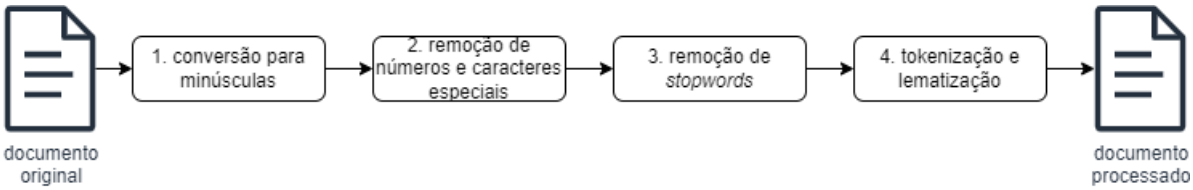


Figura 3.5: Etapas de pre-processamento realizadas nos textos

3.2.3 Tesauro

Tesauro (do grego: *thesaurus*; plural: *thesauri*) é um tipo de vocabulário controlado cujas relações entre seus termos são identificadas por meio de indicadores padronizados e que são empregados mutuamente [42]. De forma simplificada, pode ser definido como uma lista de termos com seus respectivos sinônimos – ou alguma outra relação semântica equivalente. Comumente, versa sobre um domínio do conhecimento, o que faz com que sua utilização seja bastante específica. No âmbito da área de atuação do TCU – controle externo – foi desenvolvido um tesauro denominado Vocabulário de Controle Externo (VCE)[3], criado com o objetivo de uniformizar a terminologia usada nas atividades institucionais do Tribunal, além de apoiar o tratamento da informação no Órgão. Está estruturado por funções de governo e possui descritores de Assunto, de Entidades e de Localidades. É composto por 5.123 tuplas de termos-sinonímias, conforme exemplificado na tabela 3.4.

[3] `https://portal.tcu.gov.br/vocabulario-de-controle-externo/`

Termo	Sinônimo
Tributo	Obrigação fiscal
Sentença	Decisão judicial
Renúncia de receita	Renúncia tributária
Governo eletrônico	e-GOV
Patrimônio mobiliário	Bens patrimoniais móveis
Incentivo fiscal	Gastos tributários
Whitelist	Remetentes confiáveis
Prejuízo	Dano contratual
Capacitação	Qualificação profissional
Racismo	Discriminação racial

Tabela 3.4: Exemplos de termos e respectivos sinônimos presentes no Vocabulário de Controle Externo - VCE (Tesauro do TCU).

A jurisprudência possui, em seus textos, vários dos termos contidos no VCE. O termo mais frequente é "hipótese", com 1.257 ocorrências. A tabela 3.5 exibe os cinco termos que aparecem com mais frequência nos enunciados.

Termo	Ocorrências
Hipótese	1.257
Exercício do cargo	835
Constituição Federal	800
Valor	793
Forma	538

Tabela 3.5: Cinco termos do VCE mais frequentes nos enunciados de Jurisprudência.

A Figura 3.6 mostra a frequência de utilização dos termos do VCE em toda a jurisprudência. Observa-se que a maioria dos enunciados possui até 2 termos do VCE em seus textos.

3.2.4 *Ground truth*

Conforme detalhado na seção 3.1, os especialistas em jurisprudência do TCU reúnem os acórdãos relevantes que possuem afinidade material para descreverem em um enunciado o que avaliam ser o juízo da Corte de Contas em relação a um determinado assunto. Ao fazerem isso, organizam a associação desses acórdãos entre si, bem como da vinculação destes ao enunciado resultante, e armazenam essas informações no banco de dados do sistema utilizado para esse fim.

A construção do *ground truth* deste trabalho é baseada tanto nessa heurística quanto nesses dados. Ele é constituído de 47 acórdãos, cada um associado a até 11 enunciados de jurisprudência (a média é de aproximadamente 4 enunciados por acórdão), totalizando

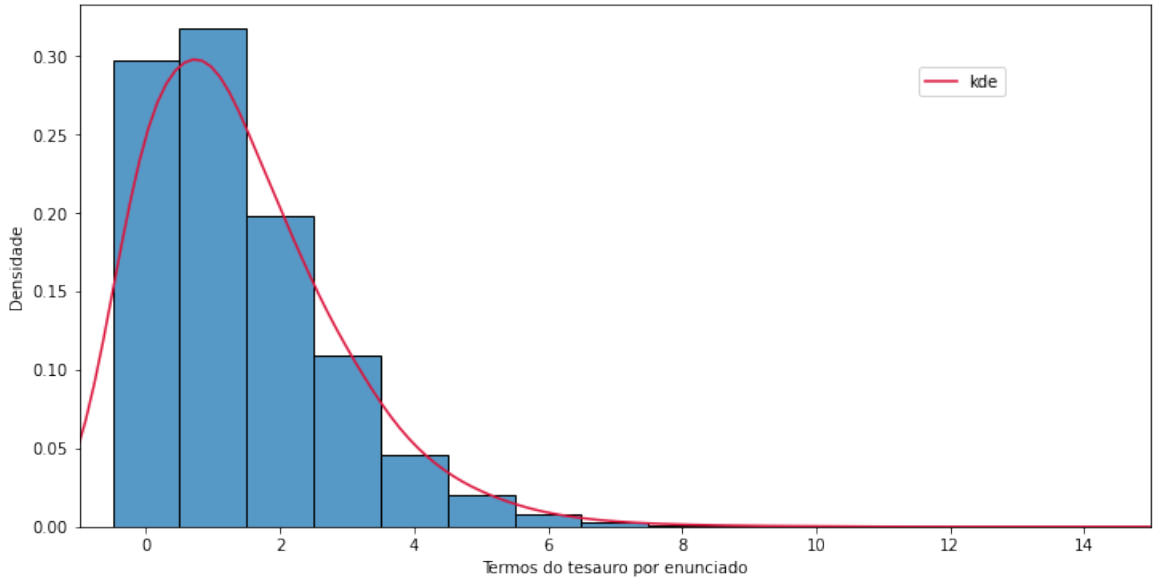


Figura 3.6: Distribuição de termos do VCE presentes nos enunciados de Jurisprudência.

172 documentos presentes no *ground truth*. A tabela 3.6 mostra a estrutura do *ground truth* e a tabela 3.7 o exemplifica com dados reais. Sucintamente, a validação do método proposto verifica se os enunciados retornados em uma consulta de similaridade a um dado acórdão pertencem à lista dos enunciados que lhe servem de referência. Em seguida, essa reposta é utilizada no ranqueamento dos documentos mais similares, essencial para o cálculo do *Mean Average Precision* e do *recall@k*, métricas selecionadas para comparação das técnicas examinadas nesta dissertação.

Acórdão	**Enunciados**				
acórdão 1	enunciado 1	enunciado 2	enunciado 3	...	enunciado n
acórdão 2	enunciado 4	enunciado 5	enunciado 6	...	enunciado m
acórdão 3	enunciado 7	enunciado 8	enunciado 9	...	enunciado p

Tabela 3.6: Estrutura do *ground-truth*.

Acórdão	Enunciados			
Auditoria na secretaria de políticas públicas de emprego do ministério do trabalho e emprego. Celebração de convênios cujos objetos não estão adequados à efetiva demanda de qualificação profissional. Ausência de critérios técnicos e objetivos para julgamento das propostas das entidades proponentes de convênios. Fraude em cotação de preços. Declaração de inidoneidade. Declaração de inabilitação de signatário de convênio em função de irregularidades graves. Multa. Determinações. Recomendações. (...)	Nos Planos Setoriais de Qualificação - Planseq originados de emendas parlamentares cabe ao Ministério do Trabalho e Emprego (MTE) a análise do plano de trabalho encaminhado, devendo o órgão legislativo ser cientificado caso constatada alguma inconsistência ou discrepância na proposta.	Não há possibilidade jurídica de se realizar substabelecimento de convênios, uma vez que esses ajustes devem ser celebrados nos termos dos objetivos institucionais das entidades convenentes, previstos em seus respectivos estatutos.		

Continuação da página anterior

Acórdão	Enunciados			
Representação. Contrato de prestação de serviços de informática. Opção indevida por concorrência, em detrimento do pregão. Impertinência dos atributos técnicos pontuáveis. Irregularidades na planilha de formação de preços das licitantes. Remuneração desvinculada de resultados, ausência de critérios de aceitabilidade e mensuração de serviços. Procedência. Determinações. Recomendação. Verificadas falhas na contratação de prestação de serviços decorrentes de prática equivocadas disseminadas na administração (...)	Não deve ser incluída a parcela 'reserva técnica', bem como os tributos IRPJ e CSL, nos orçamentos básicos e nos formulários para proposta de preços de contratação de serviços terceirizados de tecnologia da informação.	Nas licitações e contratações de serviços de tecnologia da informação, a Administração deve estabelecer previamente em plano de trabalho a justificativa da necessidade dos serviços, em harmonia com as ações previstas no Planejamento (...)	...	Em licitações e contratações de serviços de tecnologia da informação (TI) , a Administração deve definir metodologia de avaliação de qualidade dos serviços a serem prestados, abrangendo a definição de variáveis objetivas, a exemplo do grau de conformidade com as especificações inicialmente estabelecidas e do número de falhas detectadas no produto obtido, entre outras(...)

Continuação da página anterior

Acórdão	Enunciados			
Prestação de contas. Exercício de 2001. Terceirização para desempenho de tarefas do quadro permanente. Terceirização no hospital universitário para prestação de serviços do quadro permanente. Não-prestação de contas por parte da fundação de apoio. Contratação de servidores da universidade pela fundação de apoio sem papel nas fundações de apoio universitárias. Comprovação de compatibilidade de horário. Concessão e autorização de uso de instalações da universidade sem licitação. Alteração de contrato (...)	A participação de servidores de instituições federais de ensino superior e de pesquisa científica e tecnológica em atividades de interesse de Fundação de Apoio deverá se dar sem prejuízo de suas atribuições funcionais, vedada a participação durante a jornada de trabalho (...)	A realização do concurso vestibular das universidades públicas insere-se entre as atividades típicas que ensejam a participação das fundações de apoio.		

Tabela 3.7: Exemplos reais de documentos do *ground truth*.

3.2.5 Amostragem da Base de Dados

Realizar testes em amostras oferece benefícios em muitos contextos de pesquisa e análise de dados. A utilização de amostras proporciona uma abordagem mais eficiente, economizando tempo e recursos. Analisar apenas uma parte representativa do conjunto de dados completo é particularmente útil em grandes conjuntos de dados.

Além disso, experimentações realizadas em amostras ajudam a mitigar o risco de viés e erros sistemáticos. Ao selecionar aleatoriamente uma amostra representativa, os resultados dos testes têm maior probabilidade de refletir com precisão as características do conjunto de dados completo, reduzindo a possibilidade de generalizações inadequadas. A validade interna dos experimentos é fortalecida quando as amostras são cuidadosamente controladas para evitar fatores que possam distorcer os resultados de maneira sistemática. Isso é crucial para garantir que as conclusões obtidas a partir dos testes sejam mais confiáveis e aplicáveis a toda uma população.

A escolha da técnica de amostragem depende das características específicas da população, dos objetivos da pesquisa e dos recursos disponíveis. Levy & Lemeshow ressaltam que métodos de amostragem mais sofisticados são mais apropriados em situações em que uma amostra aleatória simples não seja suficiente para representar com precisão toda a diversidade da população [43]. Caso existam variações na população, com subgrupos significativos, ou heterogeneidade nas características relevantes ao estudo, amostragem estratificadas ou por conglomerados são mais adequadas. Contudo, para a realização das avaliações deste trabalho, optou-se por selecionar uma amostra aleatória simples da base de dados de 15.000 enunciados, na medida em que se confirmou a preservação de significância de características selecionadas na amostra obtida.

O tamanho de uma amostra aleatória simples é obtido de acordo com a equação 3.1, conforme descrito em [44].

$$\text{tamanho da amostra} = \frac{Z^2_{\alpha/2} \cdot p \cdot (1-p)}{E^2} \tag{3.1}$$

em que:

$Z^2_{\alpha/2}$ é o valor crítico do grau de confiança,

p é a proporção populacional de indivíduos pertencentes à categoria interessada, e

E representa a margem de erro.

Considerando um grau de confiança de 95% (valor de $Z^2_{\alpha/2}$ é 1,96), a proporção populacional da categoria interessada de 50% (por ser desconhecida), e uma margem de erro de 5%, obtemos um tamanho de amostra de 385 elementos (Eq. 3.2).

$$\text{tamanho da amostra} = \frac{1,96^2 \cdot 0,5 \cdot (1-0,5)}{0,05^2} \approx 385 \tag{3.2}$$

A tabela 3.8 apresenta a comparação entre algumas características da base de dados original e da amostra. Importante observar que os atributos "quantidade média de palavras

por documento"e "tamanho médio das palavras no corpus"conservam, na amostra, valores bastante similares dos presentes na base de dados original, com respectivos desvios-padrão também próximos. O tamanho do corpus da amostra é de 557 documentos, pois soma-se ao tamanho calculado para a amostra os 172 documentos que compõem o *ground truth*.

Característica	**Base original**	**Amostra**
tamanho do corpus (em #docs)	15.000	557
tamanho do vocabulário (#palavras distintas)	9.497	
#palavras	376.873	14.523
min/max #palavras por doc	4 / 346	5 / 108
qtde média palavras por doc (± dp)	25,124 (± 12,283)	26,073 (± 12,317)
tamanho médio palavra no corpus (± dp)	7,714 (± 0,665)	7,707 (± 0,648)

Tabela 3.8: Comparação de características entre a base de dados original e a amostra utilizada neste estudo.

Para verificar a normalidade da distribuição das características "tamanho médio de palavras"e "quantidade média de palavras por documento", aplicou-se o teste de *Shapiro-Wilk*. O *p-valor* obtido foi menor que o nível de significância de 5%, o que implica na rejeição da hipótese nula e sugere que os dados não seguem uma distribuição normal. Por este motivo, para avaliar se ambas as coleções de dados (base original e amostra) possuem a mesma distribuição, usou-se o teste não-paramétrico de *Mann-Whitney U*. Ele é uma alternativa ao teste *t de Student* quando a suposição de normalidade não é atendida e os dados são independentes. Sua hipótese nula afirma que não há diferença significativa entre as duas amostras. Como o resultado desse teste, para as duas características, foi maior que o nível de significância de 0,05, a hipótese nula não foi rejeitada, confirmando que os grupos são estatisticamente iguais. Essa análise é comprovada por meio da visualização dos diagramas de caixa (Figura 3.7), onde constata-se que as medianas estão simetricamente centradas, do mesmo modo que há equivalência nas amplitudes dos interquartis.

3.3 Modelagem dos Dados

Uma vez que os documentos textuais (acórdãos e enunciados) não estão em formato adequado para a recuperação de informação, faz-se necessário a aplicação de métodos para representá-los vetorialmente.

Existem diferentes técnicas para representação vetorial de documentos textuais, tal como apresentado na Seção 2.2 desta dissertação. Para fins comparativos, neste trabalho essas técnicas foram agrupadas em duas abordagens diferentes de representação:

- Representação em Nível de Palavra

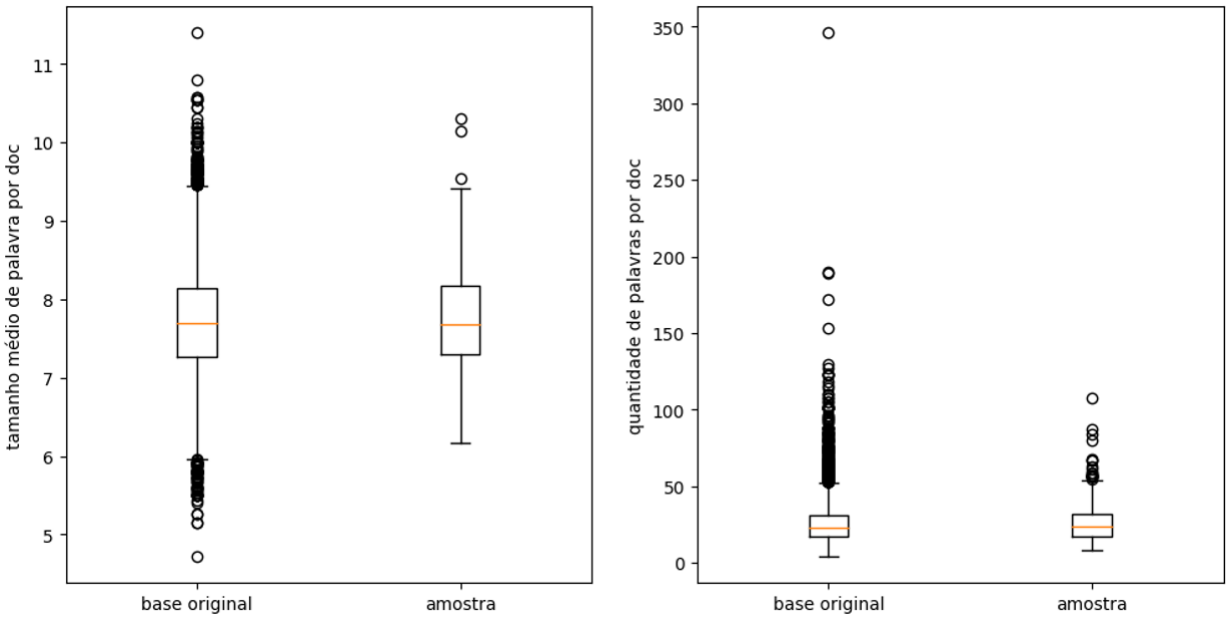


Figura 3.7: Diagramas de caixa com as características "tamanho médio de palavras por documento"e "quantidade de documentos por documento"da base original e da amostra.

BoW (seção 2.2.1)

TF-IDF (seção 2.2.2)

BM25 (seção 2.2.3)

- Representação em Nível de Conceito

  BoC (seção 2.2.5)

Além dessas técnicas tradicionais da literatura, neste trabalho foi proposta e avaliada uma nova abordagem que contempla, simultaneamente, estes dois níveis de representação. A técnica proposta denominada de Bag-of-Concepts with Thesaurus (BoC-Th) é baseada nas técnicas TF-IDF (nível de palavra) e BoC (nível de conceito).

A proposta da técnica BoC-Th é ponderar cada palavra do documento pela distância ao termo mais próximo do tesauro: quanto mais distante, menos peso essa palavra terá na representação do documento. A motivação para essa proposta é enfatizar as palavras/termos que são diretamente relacionadas ao linguajar peculiar dos documentos jurídicos e, dessa forma, os documentos textuais serão representados de uma maneira mais discriminativa.

A seguir será apresentada em detalhes a abordagem proposta BoC-Th.

3.3.1 Bag-of-Concepts with Thesaurus (BoC-Th)

A Figura 3.8 ilustra as etapas da abordagem proposta, que é uma combinação das técnicas em nível de palavra (IDF) e em nível de conceito (BoC). A linha azul corresponde às etapas da técnica BoC e a linha verde corresponde ao cálculo do IDF de cada palavra. O diferencial inovador da técnica proposta BoC-Th está representado pela linha em vermelho, que consiste em ponderar os conceitos de cada palavra pela distância do seu termo mais similar à um tesauro e também pelo seu valor IDF.

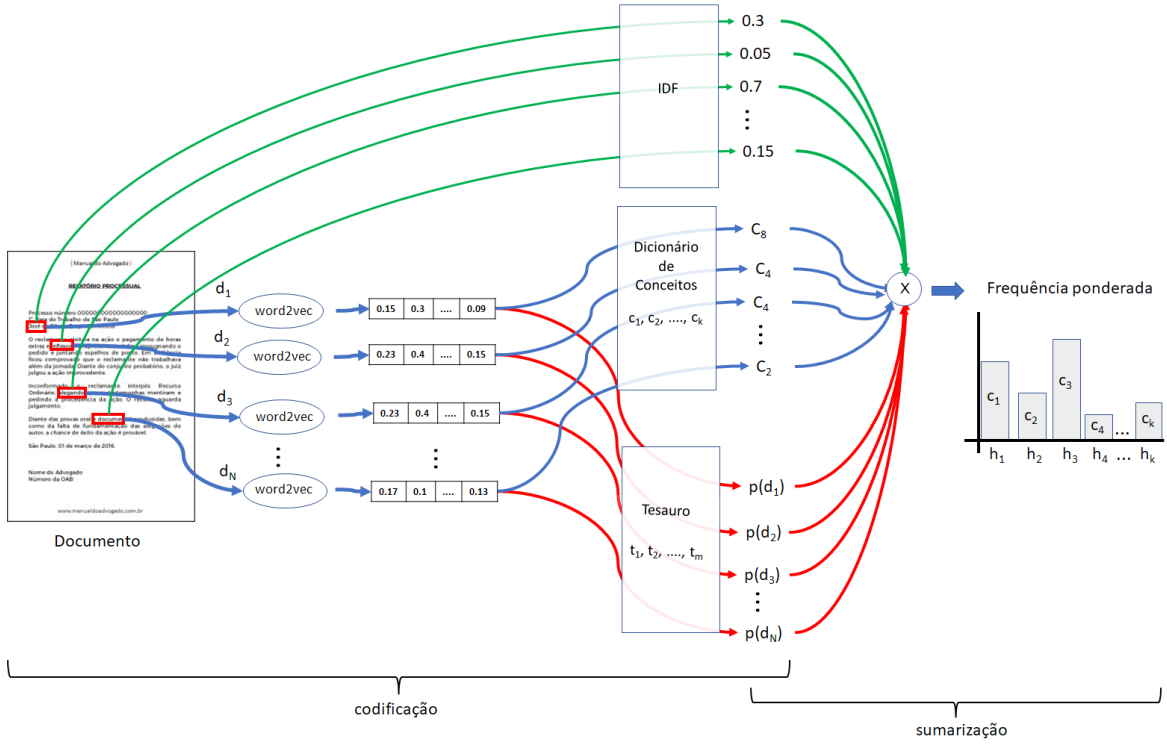


Figura 3.8: Fluxograma da técnica proposta BoC-Th que combina representação em nível de conceitos com representação em nível de palavras. A linha verde associa cada palavra do documento ao seu valor IDF, a linha azul atribui um conceito para cada palavra do documento, a linha em vermelho contabiliza a distância de cada palavra ao seu termo mais similar em um tesauro. O resultado final é um histograma de conceitos ponderados pela distância e pelo valor IDF de cada palavra do documento.

Definindo formalmente a técnica BoC-Th, considere $D = \{d_1, d_2, ..., d_N\}$ um documento com N palavras e $C = \{c_1, c_2, ..., c_k\}$ um dicionário com k conceitos. Cada palavra $d_i \in D$ está associada ao conceito $c_j \in C$ através da seguinte função:

$$\phi(d_i) = \{\alpha_1^i, \alpha_2^i, ..., \alpha_k^i\} \tag{3.3}$$

em que α_j^i é um valor binário que associa d_i ao conceito c_j. A técnica BoC-Th considera que cada palavra $d_i \in D$ será atribuída à apenas um conceito, ou seja:

$$\alpha_j^i = \begin{cases} 1, & \text{se } j = \arg\min_{j \in \{1,\ldots,k\}} ||d_i - c_j||_2^2 \\ 0, & \text{caso contrário.} \end{cases} \tag{3.4}$$

Considerando $T = \{t_1, t_2, ..., t_m\}$ um tesauro com m palavras, o diferencial inovador da técnica BoC-Th consiste em ponderar os conceitos na representação final do documento vinculado a distância de cada palavra $d_i \in D$ à sua palavra mais próxima $t_r \in T$. Em outras palavras, o vetor numérico final que irá representar o documento D será dado por um histograma ponderado de conceitos $H = [h_1, h_2, ...h_k]$, onde k é a quantidade de conceitos existentes no dicionário de conceitos e cada $h_j \in H$ é definido como:

$$h_j = \sum_{i=1}^{N} p(d_i).\alpha_j^i.\text{IDF}(d_i), \qquad \forall j \in \{1, 2, ..., k\} \tag{3.5}$$

em que,

$$p(d_i) = 1 - \arg\min_{t_r \in T} ||d_i - t_r||_2^2 \tag{3.6}$$

A função de ponderação $p : R^d \rightarrow R$ retorna a distância da palavra d_i ao seu termo correspondente mais similar no tesauro T. Se $d_i \in T$, então $p(d_i) = 1$ e, portanto, o conceito associado à palavra d_i será contabilizado integralmente no histograma. Caso contrário, quanto mais distante for a palavra d_i das palavras do tesauro, menor será a contribuição do conceito da palavra d_i no histograma final do documento. A importância relativa IDF da palavra d_i estabelece a ponderação final.

3.4 Execução da avaliação

3.4.1 Cálculo da Similaridade Semântica

Similaridade semântica é a medida estabelecida nesta pesquisa para se avaliar o grau de proximidade entre dois documentos, de forma a possibilitar a aferição da eficácia das técnicas de vetorização estudadas. Seu valor é obtido a partir do cálculo do cosseno dos ângulos entre as representações vetoriais dos respectivos documentos, conforme apresentado na seção 2.3.3. O Algoritmo 1 descreve a lógica de cálculo dos valores de similaridade.

Algoritmo 1: Cálculo de similaridade semântica entre acórdãos e enunciados de jurisprudência

Entrada: *groundtruth*: base de referência para avaliação da similaridade; *jurisprudência*: corpus da amostra da base de jurisprudência
Saída : valores de similaridade do par *ground truth* × jurisprudência de cada uma das técnicas

```
Consolidação dos vetores BoC-Th com valores IDF de cada termo, com prevalência da dimensão dos vetores de conceitos;
foreach acórdão do groundtruth do
    Recupera vetores do acórdão em nível de palavra;
    Recupera vetores do acórdão em nível de conceito;
    foreach enunciado da jurisprudência do
        Recupera vetores do enunciado em nível de palavra;
        Recupera vetores do enunciado em nível de conceito;
        Calcula a similaridade cosseno entre os vetores do acórdão e os vetores do enunciado, nas respectivas técnicas BoW, TF-IDF, BM25, BoC, BoC-Th (sem IDF), e BoC-Th;
    end
end
```

3.4.2 Configuração da Implementação

O código-fonte[4] para a geração dos vetores de características e a execução dos cálculos dos indicadores de similaridades foi implementado na linguagem de programação Python 3.8, com o suporte das bibliotecas scikit-learn, Gensim, Pandas, Matplotlib e NumPy. Este código foi executado em ambiente Google Colab, utilizando uma máquina Tesla T4 GPU com 40 núcleos 1.59 GHz e 16 GB de memória RAM.

3.4.3 Parâmetros das técnicas avaliadas

Word2Vec

Os vetores de palavras foram gerados por meio da biblioteca Gensim. Os seguintes parâmetros foram utilizados:

- *size*: dimensão dos vetores de palavra resultantes. Foram experimentadas geração de vetores de palavras com 300, 400 e 500 dimensões.
- *window*: Tamanho da janela de contexto. Foi mantido o valor padrão de 5, o que significa que são consideradas as cinco palavras à esquerda e as cinco à direita de uma palavra-alvo.

[4] `https://github.com/wagnermcosta/bocth.git`

- *min_count*: limite mínimo de quantidade que uma palavra aparece no corpus para poder ser considerada na vetorização. Em situações como a deste estudo, em que uma área específica do conhecimento está sendo trabalhada, palavras raras podem ter um impacto significativo na interpretação do contexto. Incluir todas as palavras pode melhorar a capacidade do modelo de entender sutilezas no uso de linguagem. Por esse motivo, foi utilizado o valor de 1.
- sg (skip-gram): Define o algoritmo a ser utilizado; se sg=0, o CBOW é usado, se sg=1, é utilizado o Skip-Gram. O algoritmo utilizado neste estudo foi o Skip-Gram (valor = 1).
- *epochs*: número de iterações completas sobre o conjunto de dados durante o treinamento. Para se evitar um sobreajuste do modelo ("overfitting"), que pode ocorrer com um número alto de iterações, manteve-se o valor padrão de 5.

BM25

No caso da técnica BM25, o parâmetro b, que controla como o comprimento de um documento afeta a pontuação de relevância, manteve o valor igual a 0,75. O parâmetro $k1$, que controla as características de saturação da frequência do termo na pontuação de um determinado documento, conservou o valor de 1,2.

Dicionário de conceitos

Para as técnicas BoC e BoC-Th, é preciso indicar o tamanho do dicionário de conceitos. Foram experimentados modelos de dicionários com 300, 400 e 500 conceitos. Os algoritmos de agrupamento utilizados para a geração desses dicionários foram o *k-Means* e o *Spherical k-Means*. A diferença básica entre eles reside no fato de que o *Spherical k-Means* é mais especializado para conjuntos de dados que estão presente em uma distribuição multidimensional, o que vem a ser o caso dos vetores de palavras deste estudo. Para ambos os algoritmos, os valores-padrão foram utilizados, exceto o valor que define o número de *clusters* (grupos), que corresponde ao tamanho do dicionário de conceitos desejado.

Definição do valor de k

A seleção do valor específico de k na métrica *recall@k*, no contexto da recuperação de documentos para a formulação de jurisprudência, é uma decisão estratégica que pode ter um impacto significativo na realização da atividade de análise jurídica. A escolha de $k = 100$ destaca a importância de recuperar uma quantidade considerável de documentos relevantes, pois isso influencia diretamente a análise subsequente realizada pelos especialistas. A

atual prática dos especialistas – que é manual – contempla a seleção empírica de poucas dezenas de acórdãos e enunciados, previamente elegíveis perante a fixação de um tema ou assunto relevante. Dessa maneira, a recuperação dos 100 documentos mais relevantes para a consulta jurisprudencial em questão é orientada pelo entendimento de que essa quantidade de documentos é suficiente para abranger uma representação significativa dos enunciados e decisões relevantes para a formulação de jurisprudência.

3.4.4 Metodologia

A avaliação de eficácia das técnicas foi divida em onze etapas. A Figura 3.9 ilustra esse fluxo, que consiste em:

1. Obtenção dos dados: etapa em que foram feitas as montagem do corpus e do *ground truth* e realizada a análise exploratória das bases geradas (descritas nas seções 3.2.1 e 3.2.4);
2. Preparação dos dados (descrita na seção 3.2.2);
3. Seleção da amostra (descrita na seção 3.2.5);
4. Definição dos parâmetros (descrita na seção 3.4.3);
5. Vetorização da amostra e do *ground truth* em nível de palavra, nas abordagens BoW, TF-IDF e BM25 (descritas nas seções 2.2.1, 2.2.2 e 2.2.3);
6. Vetorização da amostra e do *ground truth* em nível de conceito, nas abordagens BoC, Boc-Th (sem IDF) e Boc-Th (descritas nas seções 2.2.5 e 3.3.1);
7. Cálculo de similaridade semântica para cada uma das técnicas (descrito na seção 3.4.1);
8. Cálculo do indicador mean Average Precision (mAP) (descrito na seção 2.4.2);
9. Cálculo do indicador *recall@k*, com $k = 100$ (descrito na seção 2.4.3);
10. Consolidação dos resultados: cálculo das médias e respectivos desvios-padrão de cada série de indicador mAP e *recall@k*;
11. Validação estatística: verificação da significância estatística entre os grupos de valores de similaridade.

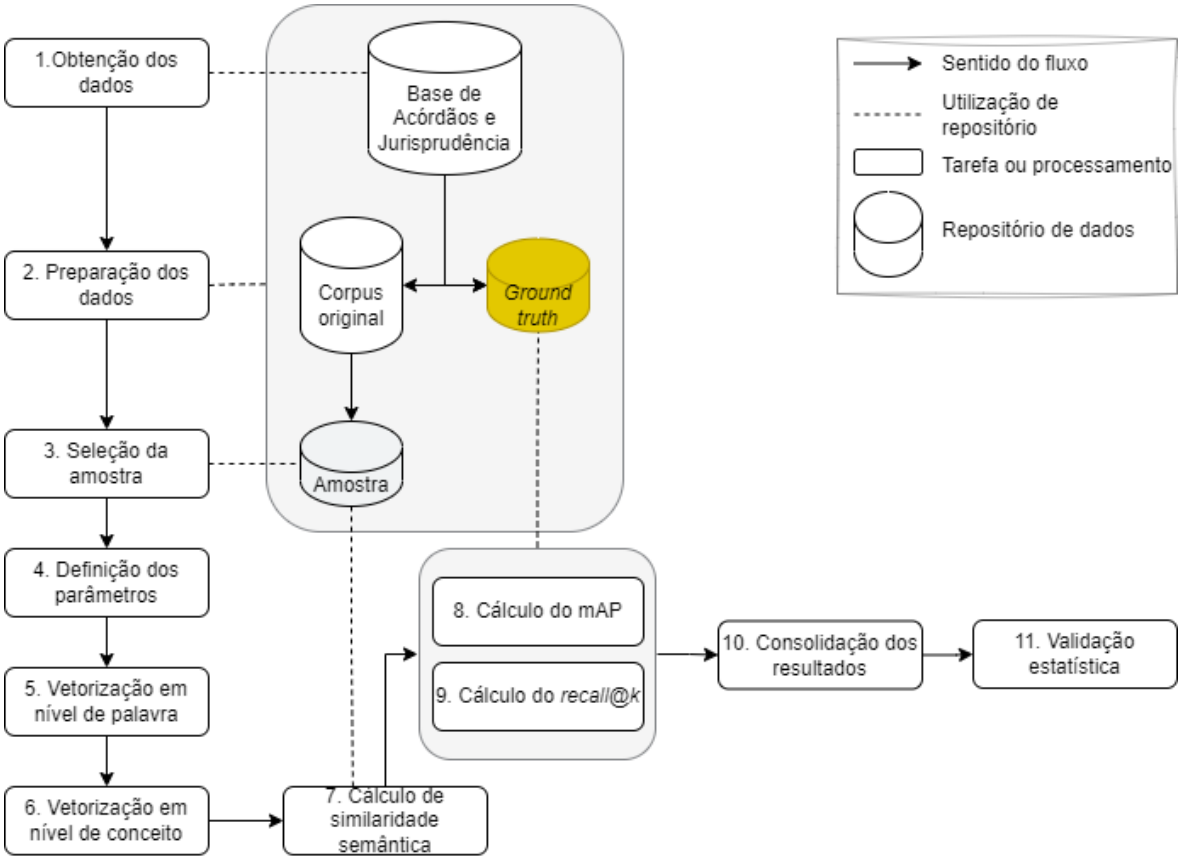


Figura 3.9: Fluxo de avaliação das técnicas de vetorização.

Capítulo 4

Resultados e Discussões

Este capítulo apresenta os testes experimentais que foram realizados a fim de avaliar a eficiência e eficácia de diferentes abordagens de representação textual na tarefa de recuperar enunciados similares à um dado acórdão fornecido pelo usuário como consulta.

4.1 Técnicas Avaliadas

As seguintes técnicas foram avaliadas, aplicando-se a metodologia apresentada na seção 3.4.4:

- BoW (descrita na seção 2.2.1)
- TF-IDF (descrita na seção 2.2.2)
- BM25 (descrita na seção 2.2.3)
- BoC (descrita na seção 2.2.5)
- BoC-Th (descrita na seção 3.3.1)

Adicionalmente, verificou-se o comportamento da técnica proposta BoC-Th desconsiderando sua vinculação com a abordagem IDF, a qual foi denominada "BoC-TH sem IDF".

O desempenho de cada uma das técnicas foi avaliado sobre um estudo de caso, envolvendo a busca e análise de similaridade de acórdãos e jurisprudências do Tribunal de Contas da União (TCU). A construção do *ground truth* utilizado nos testes experimentais está detalhado na seção 3.2.4.

4.2 Resultados Obtidos

Os resultados desse trabalho foram avaliados sob duas perspectivas:

- eficiência: se refere ao tempo gasto pela técnica para executar a busca;
- eficácia: corresponde à capacidade da técnica em recuperar os objetos mais relevantes nas primeiras posições do ranking ou uma maior quantidade de documentos relevantes na lista de sugestão de tamanho estabelecido.

Para se medir a eficácia foram utilizadas duas métricas: mean Average Precision (mAP) (seção 2.4.2) e *recall@k* (seção 2.4.3). A eficiência foi obtida a partir da medição do tempo necessário para o cálculo de similaridade de 1.000 documentos. A Tabela 4.1 exibe os valores obtidos nos testes realizados:

Técnica	**Configuração**	**mAP (± dp)**	**recall@100 (± dp)**	**Desempenho (s)**
BoW	não se aplica	0,088 (± 0,197)	0,217 (± 0,325)	0,0285
TF-IDF	não se aplica	0,095 (± 0,205)	0,256 (± 0,340)	0,0267
BM25	não se aplica	0,087 (± 0,189)	0,267 (± 0,358)	0,0223
BoC	D300C400-SKM	0,097 (± 0,211)	0,309 (± 0,359)	0,0157
Boc-Th (sem IDF)	D400C400-KM	0,098 (± 0,173)	0,284 (± 0,362)	**0,0137**
BoC-Th	D400C400-SKM	**0,102 (± 0,187)**	**0,320 (± 0,367)**	0,0156

Tabela 4.1: Resultado dos experimentos.
Configuração: D=dimensão do vetor de palavras, C=quantidade de conceitos, KM=k-Means, SKM=Spherical k-Means
Desempenho: tempo, em segundos, para cálculo da similaridade de 1.000 documentos.

4.2.1 Resultados de eficiência

Uma das principais preocupações dos sistemas de recuperação de informação é o tempo de execução para a comparação do documento de consulta (*query*) com os demais documentos existentes na base de dados. Por esse motivo, a avaliação da eficiência das técnicas é indispensável.

A formulação de jurisprudência é um processo complexo e intensivo que demanda análise cuidadosa de decisões passadas, interpretação de leis e normas, assim como a consolidação de entendimentos jurídicos. Tradicionalmente, essa tarefa é desempenhada de forma manual por especialistas jurídicos e por magistrados. Dado que o TCU lida com um volume significativo de processos e documentos legais diariamente, a tarefa de revisar, analisar e interpretar cada decisão para formular sua jurisprudência é demorada e propensa a erros. Identificar decisões semelhantes ou precedentes relevantes exige uma busca minuciosa através de vastos bancos de dados, muitas vezes sem o suporte adequado de ferramentas especializadas.

Embora etapas dessa atividade possam ser apoiadas por meio de soluções de tecnologia, a consolidação de entendimentos e a formação de jurisprudência a partir de diversas

decisões não pode prescindir da participação humana, visto que a análise deve considerar nuances, contextos e interpretações que demanda sua sensibilidade. Avalia-se que a automação na localização e análise de decisões pode resultar em significativa redução de tempo, já que ferramentas especializadas processam grandes volumes de dados em um curto período.

As técnicas em nível de palavra possuem vetores com grande dimensionalidade – equivalente ao tamanho do vocabulário, o que, no caso deste estudo, corresponde a 9.497 palavras (ver Tabela 3.8). Ademais, representações vetoriais no estilo Bag-of-Words são naturalmente esparsas. Por esta razão, necessitam de maior tempo de processamento para a execução das operações de álgebra linear necessárias aos cálculos de similaridade, frente às representações vetoriais densas utilizadas em modelos de incorporação de palavras, como o Word2Vec, base vetorial para as técnicas em nível de conceito.

A completa digitalização do processo de trabalho que suporta as atividades de formulação de jurisprudência implica na exigência de que a leitura de documentos seja feita, primordialmente, a partir da tela de um computador. Nessas condições, um leitor médio lê, com compreensão apropriada, uma média de 244 palavras por minuto, de acordo com [45], que estudou os efeitos da velocidade e compreensão de leitura realizado sob o suporte de um monitor de vídeo. Tendo em consideração que o tamanho médio de um documento do corpus de jurisprudência é de 25 palavras (ver Tabela 3.8), um especialista do TCU seria capaz, então, de realizar a leitura de cerca de 9 documentos de jurisprudência por minuto. Essa leitura compreenderia a interpretação e a análise de eventual similaridade com o documento de referência. Conforme apresentado na tabela 4.1, o tempo calculado para a análise de similaridade entre documentos, obtido com o uso da técnica proposta BoC-Th, é de 1.000 documentos em 15 centésimos de segundo. Portanto, incorporar esse recurso em uma solução de tecnologia que apoie o trabalho dos especialistas trará um ganho de desempenho muito expressivo em relação ao tempo dispendido no atual formato manual da execução da tarefa.

Todos os valores de desempenho foram computados extraindo-se a média aritmética do tempo gasto para a recuperação dos vetores que representam cada acórdão do *ground truth* e cada enunciado da jurisprudência e subsequente cálculo da similaridade por cosseno entre eles. Em cada configuração avaliada foram realizadas cinco execuções completas de cálculo de similaridade entre os documentos. Ou seja, um total de 47 acórdãos presentes no *ground truth* $\times$ 557 enunciados presentes na base de dados de amostra $\times$ 5 execuções = 130.895 cálculos efetuados para cada uma das seis técnicas analisadas.

4.2.2 Resultados de eficácia

Em sistemas de recuperação de informação que lidam com documentos jurídicos, é essencial garantir que nenhum documento relevante seja negligenciado em uma lista de documentos recuperados, mesmo que essa lista tenha uma quantidade pré-definida de itens a serem retornados. Diferentemente de métricas que possuem seu resultado vinculado à posição que um documento se encontra em uma lista de documentos retornados, a *recall@k* retorna um *score* maior quanto mais elevada for a porcentagem de documentos relevantes incluída na seleção final, mesmo que não ocupem necessariamente as primeiras posições.

Isso é importante, pois em contextos jurídicos a exaustividade na identificação de documentos pertinentes muitas vezes supera a importância de se classificar os documentos mais relevantes no topo da lista. A jurisprudência requer uma compreensão abrangente e representativa de casos similares, e uma técnica com alta revocação (*recall*, descrito na seção 2.4) contribui diretamente para esse objetivo, garantindo uma cobertura mais completa, e minimizando o risco da omissão de documentos relevantes.

Conforme os dados exibidos na Tabela 4.1, a técnica BoC-Th possui os melhores resultados frente às demais avaliadas. Tanto na métrica mAP quanto na *recall@k* (k=100), o desempenho é superior. BoC-Th atinge um mAP de 0,102 (± 0,187) e recall@100 de 0,320 (± 0,367). Sob a ótica da métrica mAP, a técnica Boc-Th obteve resultado 4% melhor quando comparada à melhor técnica em nível de conceito, e 7,3% superior à melhor técnica de vetorização em nível de palavra. Quando se observa pelo prisma da métrica *recall@k*, essa diferença é ainda mais destacada: 19,8%, quando confrontada com a melhor técnica em nível de palavra. A melhoria no desempenho da BoC-Th se deve à incorporação do vocabulário controlado de controle externo (seção 3.2.3), que aprimora a representação semântica e contribui para uma melhor precisão e revocação na recuperação.

O Gráfico 4.1 exibe os valores do indicador *recall@k* para as técnicas de vetorização estudadas, quando o valor k varia. Sua análise revela padrões consistentes de desempenho para os diferentes valores de k avaliados. A métrica BoC-Th demonstra amplamente o melhor desempenho dentre as técnicas analisadas. Isso confirma que a incorporação de um tesauro ao modelo tornam mais discriminativas as representações vetoriais e, consequentemente, propiciam uma localização mais apurada de documentos relevantes para os especialistas em jurisprudência.

Validação Estatística

Aplicaram-se testes de significância estatística para verificar se há diferenças entre os valores das métricas usadas na mensuração das técnicas de vetorização. Inicialmente,

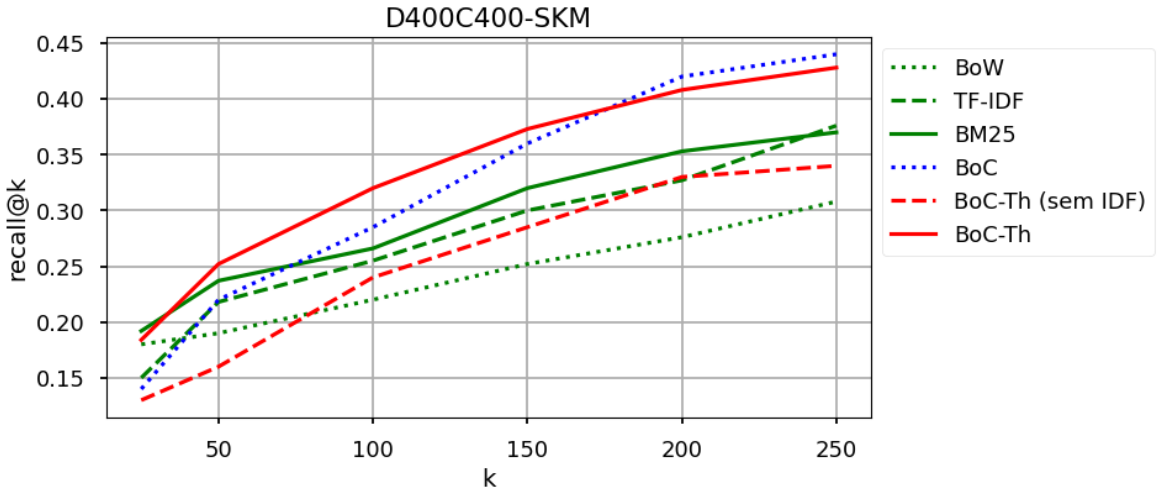


Figura 4.1: Variação da métrica *recall@k* para as técnicas estudadas.

aplicou-se o teste de *Shapiro-Wilk* para verificação de normalidade das distribuições. Os *p*-valores obtidos foram menores que o nível de significância de 5%, o que implica na rejeição da hipótese nula e sugere que os dados não seguem uma distribuição normal. Assim, decidiu-se por aplicar o teste *t de Student* e o teste não-paramétrico de *Wilcoxon*, para se ampliar a cobertura de verificação. Ambos são indicados para dados relacionados, com a diferença de que o *t de Student* é aplicado quando há distribuição normal das amostras enquanto o *Wilcoxon* é utilizado quando os dados não seguem uma distribuição normal. Para os dois testes, a hipótese nula assinala que não há diferença significativa entre as amostras, ao contrário da hipótese alternativa.

Teste	**Métrica**	**p-valor**
T-Test Pareado	mAP BoC-Th x TF-IDF	0,6447
T-Test Pareado	recall@k BoC-Th x TF-IDF	0,1273
T-Test Pareado	recall@k BoC-Th x BM25	0,2069
T-Test Pareado	recall@k BoC-Th x BoW	0,0310
Wilcoxon	mAP BoC-Th x TF-IDF	0,5255
Wilcoxon	recall@k BoC-Th x TF-IDF	0,1613
Wilcoxon	recall@k BoC-Th x BM25	0,2083
Wilcoxon	recall@k BoC-Th x BoW	0,0604

Tabela 4.2: Resultados dos Testes de Significância Estatística.

Analisando-se os resultados apresentados na Tabela 4.2, os p-valores acima de 0,05 nos *t-tests* entre BoC-Th e TF-IDF (com os dados das métricas mAP e *recall@k*) e BoC-Th e BM25 (*recall@k*) não fornecem evidências suficientes para rejeitar a hipótese nula de igualdade das médias. Entretanto, o *t-test* entre BoC-Th e BoW para r*ecall@k* revelou uma diferença significativa (p-valor de 0,0310), corroborada pelo teste de Wilcoxon (p-valor

de 0,0604). Esses resultados destacam uma vantagem estatisticamente significativa para BoC-Th em comparação com uma representação vetorial em nível de palavras, quando observada a métrica *recall@k*.

Capítulo 5

Conclusão

Este trabalho apresentou um estudo de técnicas para representação vetorial de documentos jurídicos produzidos no âmbito da atuação do Tribunal de Contas da União (TCU). A proposta foi analisar métodos computacionais que pudessem auxiliar os magistrados e os auditores na busca por similaridade entre os acórdãos do TCU e, assim, aprimorar uma instrumentalização necessária e útil para a instituição, de forma que seja possível aplicar esses modelos nas atividades de formulação de sua jurisprudência.

Neste sentido, foram estudadas e implementadas algumas técnicas tradicionais de representação textual da literatura e também foi proposta e avaliada uma nova abordagem chamada BoC-Th, que se destaca por adotar um enfoque em nível de conceito aplicada ao contexto específico do controle externo exercido pelo Tribunal de Contas da União. A técnica BoC-Th incorpora a utilização de um tesauro, o que enriquece sua representação semântica. Complementarmente, essa técnica pondera as palavras do corpus original com base em seus valores IDF, gerando um histograma de conceitos mais discriminativos na busca por similaridade.

Testes experimentais foram realizados com dados reais da instituição, comparando diferentes abordagens de representação textual. Os resultados mostraram que a técnica BoC-Th obteve um desempenho superior em comparação às demais técnicas analisadas: BoW, TF-IDF, BM25 e BoC. Sua aplicação prática consiste na localização de enunciados de jurisprudência semelhantes a acórdãos já proferidos, fornecendo um suporte valioso para os especialistas na tarefa de formulação da jurisprudência do TCU. A abordagem BoC-Th, ao capturar de maneira mais precisa as relações semânticas entre acórdãos e decisões precedentes, viabiliza uma busca por similaridade mais otimizada, agilizando o processo de pesquisa e análise jurisprudencial e contribuindo para uma formulação mais eficiente e fundamentada da jurisprudência do TCU, o que concorre para a consistência e a efetividade de suas decisões.

Observou-se, contudo, que o resultado é sensível à qualidade do vetor-de-palavras

gerado. Apesar de que o uso de um tesauro especializado amplie a robustez a ruídos dos dados de treinamento, ambiguidades podem escapar ao modelo, e relações semânticas entre as palavras podem não ser capturadas na geração dos vetores.

Por fim, e embora a abordagem proposta tenha se mostrado efetiva, ainda é necessário validar a técnica proposta em outras situações, e aprimorá-la para uso em diferentes cenários, por exemplo:

- aperfeiçoar o tratamento de polissemias, fenômeno que afeta a acurácia das comparações de similaridade, tal como apontado por [32]. A presença de múltiplos significados para uma palavra pode levar a ambiguidades semânticas. Quando uma palavra polissêmica é utilizada, o contexto específico em que ela aparece torna-se crucial para determinar o significado correto, podendo resultar em comparações semânticas imprecisas;
- analisar o quanto o modelo de geração de vetores-de-palavra interfere na geração de conceitos semanticamente discriminativos. Alguns modelos podem ser mais sensíveis a relações semânticas específicas, especialmente hiperonímia, hiponímia, holonímia e meronímia[1]. Isso é particularmente relevante em aplicações que exigem a compreensão de associações semânticas mais sofisticadas, como analogias;
- estudar a viabilidade de se estender a técnica proposta com a adoção de modelos baseados em uma arquitetura *Transformer*, como SBERT (*Sentence Bidirectional Encoder Representations from Transformers*)[46].

Em decorrência da pesquisa desenvolvida neste trabalho, dois artigos foram submetidos e aceitos em conferências Qualis da área da computação:

- Costa, Wagner e Pedrosa, Glauco. *Legal Information Retrieval Based on a Concept-Frequency Representation and Thesaurus.* In: 25th International Conference on Enterprise Information Systems (ICEIS), 2023, Prague. p. 303.
- Costa, Wagner e Pedrosa, Glauco. *A Textual Representation Based on Bag-of-Concepts and Thesaurus for Legal Information Retrieval.* In: X Symposium on Knowledge Discovery, Mining and Learning (KDMiLe), 2022, Campinas. p. 114.

[1]Hiperonímia/hiponímia: relação entre termos gerais (hiperônimos) e seus termos específicos (hipônimos). Ex: fruta (hiperônimo) e maçã (hipônimo).
Holonímia: relação em que uma palavra refere-se ao todo do qual outra é parte. Ex: "árvore"é holônimo de "folha".
Meronímia: relação contrária à hiponímia, em que uma palavra refere-se a uma parte de outra. Ex: "folha"é merônimo de "árvore".

Referências

[1] Mikolov, Tomas, Kai Chen, Greg Corrado e Jeffrey Dean: *Efficient estimation of word representations in vector space*. arXiv preprint arXiv:1301.3781, 2013. xi, 14, 15

[2] Januzaj, Ylber e Artan Luma: *Cosine similarity – a computing approach to match similarity between higher education programs and job market demands based on maximum number of common words*. International journal of emerging technologies in learning, 17(12):258–268, 2022. xi, 20

[3] Muneem, Rana Abdul: *What is the mean average precision in information retrieval?*, 2023. `https://www.educative.io/answers/what-is-the-mean-average-precision-in-information-retrieval`, acesso em 2023-01-11. xiii, 22, 23

[4] Brasil, Governo Federal do: *Constituição da República Federativa do Brasil*, 1988. `https://www.senado.leg.br/atividade/const/con1988/con1988_14.12.2017/CON1988.pdf`, acesso em 2022-04-03. 1

[5] Filho, Manoel Goncalves Ferreira: *Direitos humanos fundamentais*. Saraiva Educação S.A., 2016, ISBN 9788502208520. 1

[6] Tucci, José Rogério Cruz e: *Notas sobre os conceitos de jurisprudência, precedente judicial e súmula*, 2015. `https://www.conjur.com.br/2015-jul-07/paradoxo-corte-anotacoes-conceitos-jurisprudencia-precedente-judicial`, acesso em 2022-08-01. 1

[7] Bevilacqua, Helga: *Civil law e common law: a diferença entre os sistemas jurídicos*, 2021. `https://blog.sajadv.com.br/civil-law-e-common-law-a-diferenca-entre-os-sistemas-juridicos/`, acesso em 2022-08-01. 1

[8] Campos, Fernando Teófilo: *Sistemas de common law e de civil law: conceitos, diferenças e aplicações*, 2018. `https://jus.com.br/artigos/62799/sistemas-de-common-law-e-de-civil-law-conceitos-diferencas-e-aplicacoes`, acesso em 2022-08-06. 2

[9] Jales, Tulio de Medeiros: *False distances and real differences between common law and civil law/falsos distanciamentos e reais diferencas entre common law e civil law*. Revista eletrônica de direito processual, 18(1):377, 2017, ISSN 1982-7636. 2

[10] Brasil, Governo Federal do: *Lei nº 13.105, Código de Processo Civil*. 2

[11] Yan, Jun: *Text Representation*, páginas 3069–3072. Springer US, Boston, MA, 2009, ISBN 978-0-387-39940-9. https://doi.org/10.1007/978-0-387-39940-9_420. 5

[12] Le, Quoc e Tomas Mikolov: *Distributed representations of sentences and documents*. Em Xing, Eric P. e Tony Jebara (editores): *Proceedings of the 31st International Conference on Machine Learning*, volume 32 de *Proceedings of Machine Learning Research*, páginas 1188–1196, Bejing, China, 22–24 Jun 2014. PMLR. https://proceedings.mlr.press/v32/le14.html. 5, 15

[13] Xing, Chao, Dong Wang, Xuewei Zhang e Chao Liu: *Document classification with distributions of word vectors*. Em *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*. IEEE, dezembro 2014. https://doi.org/10.1109/apsipa.2014.7041633. 5

[14] Kim, Han Kyul, Hyunjoong Kim e Sungzoon Cho: *Bag-of-concepts: Comprehending document representation through clustering words in distributed representation*. Neurocomputing (Amsterdam), 266:336–352, 2017, ISSN 0925-2312. 5, 16, 17

[15] Araujo, Vera Maria Araujo Pigozzi de: *Sistemas de recuperação da informação: uma discussão a partir de parâmetros enunciativos*. Volume 24, páginas 137–143, 2012. 8

[16] Silva, Leandro, Sarajane Peres e Clodis Boscarioli: *Introdução a Mineração de Dados com aplicações em R*. dezembro 2017, ISBN 9788535284478. 9

[17] *Análise da busca, uso e avaliação dos serviços da biblioteca pelos assessores de ministros do supremo tribunal federal em relação as suas necessidades de informação jurídica*. Revista Ibero-americana de Ciência da Informação, 8(2):283–284, 2015, ISSN 1983-5213. 9

[18] Calheiros, Tânia Da Costa e Silvana Drumond Monteiro: *Mecanismos de busca de jurisprudência: um instrumento para a organização do conhecimento e recuperação da informação no ambiente jurídico virtual*. Em Questão, 23(2):146–166, 2017, ISSN 1807-8893. 10

[19] *Agrupamento e categorização de documentos jurídicos*, 2011. 10

[20] Souza, Ellen Polliana, Douglas Vitório, Gyovana Moriyama, Luiz Santos, Lucas Martins, Mariana Souza, Márcio Fonseca, Nádia Félix, André Carvalho, Hidelberg Albuquerque e Adriano Oliveira: *An Information Retrieval Pipeline for Legislative Documents from the Brazilian Chamber of Deputies*, páginas 119–126. dezembro 2021, ISBN 9781643682525. 10

[21] Oliveira, Robert A. N. de e Methanias C. Junior: *Experimental analysis of stemming on jurisprudential documents retrieval*. Information, 9(2), 2018, ISSN 2078-2489. https://www.mdpi.com/2078-2489/9/2/28. 10

[22] Gomes, Thiago Alencar: *Avaliação de técnicas de similaridade textual na uniformização de jurisprudência*. Universidade de Brasília - UnB, 2021. 11

[23] Luhn, H. P.: *A statistical approach to mechanized encoding and searching of literary information.* IBM Journal of Research and Development, 1(4):309–317, 1957. 12

[24] Sparck Jones, Karen: *A statistical interpretation of term specificity and its application in retrieval.* Journal of documentation, 28(1):11–21, 1972. 12

[25] Renjit, Sara e Sumam Mary Idicula: *Cusat nlp@ aila-fire2019: Similarity in legal texts using document level embeddings.* Em *FIRE (Working Notes)*, páginas 25–30, 2019. 15

[26] Mourino Garcia, Marcos Antonio, Roberto Perez Rodriguez e Luis E Anido Rifon: *Biomedical literature classification using encyclopedic knowledge: a wikipedia-based bag-of-concepts approach.* PeerJ (San Francisco, CA), 3:e1279–e1279, 2015, ISSN 2167-8359. 16

[27] Wang, Fang, Zhongyuan Wang, Zhoujun Li e Ji Rong Wen: *Concept-based short text classification and ranking.* Em *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, página 1069–1078, New York, NY, USA, 2014. Association for Computing Machinery, ISBN 9781450325981. https://doi.org/10.1145/2661829.2662067. 17

[28] Shalaby, Walid e Wlodek Zadrozny: *Learning concept embeddings for dataless classification via efficient bag-of-concepts densification.* 61(2):1047–1070, 2019, ISSN 0219-1377. 17

[29] Hematialam, Hossein, Luciana Garbayo, Seethalakshmi Gopalakrishnan e Wlodek W. Zadrozny: *A method for computing conceptual distances between medical recommendations: Experiments in modeling medical disagreement.* Applied Sciences, 11(5), 2021, ISSN 2076-3417. https://www.mdpi.com/2076-3417/11/5/2045. 17

[30] Li, Pengfei, Kezhi Mao, Yuecong Xu, Qi Li e Jiaheng Zhang: *Bag-of-concepts representation for document classification based on automatic knowledge acquisition from probabilistic knowledge base.* Knowledge-Based Systems, 193:105436, 2020, ISSN 0950-7051. https://www.sciencedirect.com/science/article/pii/S0950705119306604. 18

[31] Rajabi, Zeinab, Mohammad Reza Valavi e Maryam Hourali: *A context-based disambiguation model for sentiment concepts using a bag-of-concepts approach.* Cognitive computation, 12(6):1299–1312, 2020, ISSN 1866-9956. 18

[32] Lee, Yen Hsien, Paul Jen Hwa Hu, Wan Jung Tsao e Liang Li: *Use of a domain-specific ontology to support automated document categorization at the concept level: Method development and evaluation.* 174:114681, 2021, ISSN 0957-4174. 18, 51

[33] Lee, Younghoon: *Systematic homonym detection and replacement based on contextual word embedding.* 53(1):17–36, 2020, ISSN 1370-4621. 18

[34] Mehanna, Yassin S e Massudi Bin Mahmuddin: *A semantic conceptualization using tagged bag-of-concepts for sentiment analysis.* IEEE access, 9:118736–118756, 2021, ISSN 2169-3536. 18

[35] Tabak, John: *Geometry: the language of space and form*. Infobase Publishing, 2014. 19

[36] Xia, Peipei, Li Zhang e Fanzhang Li: *Learning similarity with cosine similarity ensemble*. Information Sciences, 307:39–52, 2015. 19

[37] Cleverdon, Cyril: *The cranfield tests on index language devices*. Em *Aslib proceedings*. MCB UP Ltd, 1967. 20, 21

[38] Voorhees, Ellen M, Donna K Harman *et al.*: *TREC: Experiment and evaluation in information retrieval*, volume 63. Citeseer, 2005. 21

[39] Cooper, William S: *A definition of relevance for information retrieval*. Information storage and retrieval, 7(1):19–37, 1971. 21

[40] Park, Taemin Kim: *The nature of relevance in information retrieval: An empirical study*. The library quarterly, 63(3):318–351, 1993. 21

[41] Almeida, Fernando Dias Menezes de: *Memória jurisprudencial: Ministro Victor Nunes*, páginas 31–36. Supremo Tribunal Federal, 2006. 26

[42] NISO, National Information Standards Organization: *Guidelines for the construction, format, and management of monolingual controlled vocabularies*. NISO Press, 2005. 29

[43] Levy, Paul S e Stanley Lemeshow: *Sampling of populations: methods and applications*. John Wiley & Sons, 2013. 35

[44] Berenson, Mark L, DM Levine e David Sephan: *Estatística: teoria e aplicações usando microsoft excel em português*, 2005. 35

[45] Dyson, Mary e Mark Haselgrove: *The effects of reading speed and reading patterns on the understanding of text read from screen*. Journal of research in reading, 23(2):210–223, 2000, ISSN 0141-0423. 46

[46] Reimers, Nils e Iryna Gurevych: *Sentence-bert: Sentence embeddings using siamese bert-networks*. arXiv preprint arXiv:1908.10084, 2019. 51