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

Avaliação de Técnicas de Similaridade Textual na Uniformização de Jurisprudência

Thiago Alencar Gomes

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

Orientador
Prof. Dr. Marcelo Ladeira

Coorientador
Prof. Dr. Thiago de Paulo Faleiros

Brasília
2020

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

AT422a

Alencar Gomes, Thiago
Avaliação de Técnicas de Similaridade Textual na Uniformização de Jurisprudência / Thiago Alencar Gomes; orientador Marcelo Ladeira; co-orientador Thiago de Paulo Faleiros. -- Brasília, 2020.
71 p.

Dissertação (Mestrado - Mestrado Profissional em Computação Aplicada) -- Universidade de Brasília, 2020.

1. Mineração de Texto. 2. Processamento de Linguagem Natural. 3. Recuperação de Informação. 4. Pesquisa Jurídica. I. Ladeira, Marcelo, orient. II. de Paulo Faleiros, Thiago, co-orient. III. Título.

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

Avaliação de Técnicas de Similaridade Textual na Uniformização de Jurisprudência

Thiago Alencar Gomes

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

Prof. Dr. Marcelo Ladeira (Orientador)
CIC/UnB

Prof. Dr. Teófilo Emídio de Campos
CIC/UNB

Prof. Dr. Sandro José Rigo
Escola de Tecnologia/Unisinos

Prof. Dr. Marcelo Ladeira
Coordenador do Programa de Pós-graduação em Computação Aplicada

Brasília, 10 de Dezembro de 2020

Dedicatória

À minha mãe, minha esposa e a toda minha família que, com muito carinho e apoio, não mediram esforços para que eu chegasse até esta etapa de minha vida. Ao professor, orientador e coordenador do curso, pelo convívio, pelo apoio, pela compreensão e pela amizade.

Agradecimentos

Aos amigos e familiares , por todo o apoio e pela ajuda, que muito contribuíram para a realização deste trabalho. Aos professores, por todos os conselhos, pela ajuda e pela paciência com a qual guiaram o meu aprendizado. Ao meu orientador, que conduziu o trabalho com paciência e dedicação, sempre disponível a compartilhar todo o seu vasto conhecimento.

Resumo

A uniformização de jurisprudência é de extrema importância para a sociedade. Através dela é possível garantir maior celeridade processual e estabilidade jurídica à medida que novos casos são julgados com base na mesma tese jurídica de casos similares anteriores. Para garantir o efetivo acesso às teses, as bases textuais são indexadas em ferramentas de busca textual. Este estudo tem como contexto o Superior Tribunal de Justiça (STJ) que possui uma ferramenta legada de busca textual que fornece somente buscas baseadas em consultas booleanas com operadores lógicos e de proximidade complexos. Para facilitar a busca no corpus de decisões da Corte, a Secretária de Jurisprudência (SJR) fornece diversos produtos, como consultas pré-construídas para teses importantes e decisões agrupadas que possuem as mesmas teses. Assim, é possível acessar de forma otimizada a interpretação da legislação pelo STJ e acompanhar a sua evolução. O tempo dispendido na construção das consultas e no treinamento de servidores para utilização da ferramenta motiva este trabalho na avaliação da aplicação de outras técnicas de similaridade na recuperação de decisões. Como *baseline* utilizamos as consultas do sistema legado e comparamos com buscas a partir de textos, que descrevem as teses, escritos pelos servidores ou textos selecionados diretamente das decisões. Os resultados indicam que a utilização direta dos textos com modelos tradicionais (TF-IDF e BM25) pode substituir as consultas do sistema legado. Os modelos semânticos baseados em predição Word2Vec e BERT não apresentaram ganhos em relação aos modelos clássicos.

Palavras-chave: Mineração de Texto, Processamento de Linguagem Natural, Recuperação de Informação, Pesquisa Jurídica

Abstract

Jurisprudence is the set of all decisions of a judicial court and when they are organized efficiently they reflect the majority interpretation of the same court and thus consolidate an legal thesis used repeatedly. Hence, it is possible to guarantee faster judgments and legal stability as new cases are judged based on the same legal thesis as previous similar cases. This research investigates the use of text retrieval techniques on the Brazilian Superior Court of Justice decisions. The Court uses a legacy textual system that only provides complex Boolean queries. The training of new analysts on the tool takes between 2 and 3 months. This scenario motivates the research of other textual retrieval techniques that use text written in natural language as a queries. Through a historical base of legal theses descriptions written by the analysts, the decisions that those theses were extracted and the legacy system queries built to retrieve decisions with the same theses this work simulates two approaches. First, recovery of decisions after a user selects paragraphs with the legal opinion. Second, recovery of decisions after a user enters the legal opinion description in free text. The legacy system is used as baseline and compared with TF-IDF, BM25 retrieval models and prediction based semantic models Word2Vec and BERT. The results indicate that it is possible to replace the legacy system using classic and semantic textual retrieval using the decisions text as queries, with minimum intervention from the user.

Keywords: Text Mining, Natural Language Processing, Information Retrieval, Legal Search

Sumário

Lista de Figuras

Lista de Tabelas

Siglas

BERT *Bidirectional Encoder Representations from Transformers.*

BM25 *Best Match 25.*

BoW *Bag of Words.*

CBoW *Continuous Bag-Of-Words.*

CG *Cumulative Gain.*

DCG *Discounted Cumulative Gain.*

EPMI *Exponential Pointwise Mutual Information.*

ICAIL *International Conference on Artificial Intelligence and Law.*

IDCG *Ideal Discounted Cumulative Gain.*

JURIX *International Conference on Legal Knowledge and Information Systems.*

KLI *Kullback-Leibler Divergence for Informativeness.*

LMI *Local Mutual Information.*

MP Modelo Probabilístico.

MRR *Mean Reciprocal Rank.*

nDCG *Normalized Discounted Cumulative Gain.*

NPMI *Normalized Pointwise Mutual Information.*

PLM *Parsimonious Language Models.*

PMI *Pointwise Mutual Information.*

RI Recuperação da Informação.

RSLP Removedor de Sufixos da Lingua Portuguesa.

SJR Secretaria de Jurisprudência.

SRIJ Sistemas de Recuperação de Informação Jurídica.

STJ Superior Tribunal de Justiça.

STS *Short Sentence Similarity.*

TF-IDF *Term Frequency - Inverted Document Frequency.*

TREC *Text Retrieval Conference.*

USP Universidade de São Paulo.

VSM *Vector Space Model.*

WMD *Word Mover's Distance.*

Capítulo 1

Introdução

A aplicação de técnicas de inteligência artificial pode contribuir para automatizar e auxiliar a tarefa de uniformização da jurisprudência. Essa vertente é chamada de *legal analytics*. Ashley (2018) mostra em seu livro que diversas técnicas são utilizadas para esse propósito, tais como: mineração de textos, classificação de documentos, agrupamento de documentos, extração de conhecimento, recuperação de informação jurídica e reconhecimento de entidades jurídicas nomeadas. Opijnen e Santos (2017) explicitam que a abordagem mais utilizada pelos atores jurídicos são os Sistemas de Recuperação de Informação Jurídica (SRIJ).

Os SRIJ estão dentro do contexto de pesquisa das busca profissionais (tradução para o termo em inglês *professional search*) (Sanchez et al. 2020; Russell-Rose et al. 2018). Essa atividade consiste na intensa utilização, por profissionais especializados em domínios de conhecimento específico, de ferramentas de busca textual. Os autores destacam como desafios nessas atividades: a maioria dos sistemas somente fornece consultas booleanas; os operadores lógicos e de proximidade dessas consultas adicionam uma camada de complexidade alta, o que resulta em muito tempo para treinar novos especialistas; e a ordenação de relevância é realizada por ordem cronológica de documentos. *Pari passu* à utilização do sistema, os analistas criam e gerenciam extensos dicionários de sinônimos e tesauros. Como resultado, há um sentimento de confiança no sistema e, por outro lado, de resistência na adoção de outros sistemas.

Nesse contexto há um interesse contínuo de pesquisas na aplicação da metodologia de avaliação de recuperação de informação para mensurar a possibilidade de substituição de SRIJ legados com base em técnicas de recuperação textual mais recentes.

A área de Recuperação da Informação (RI) estuda diversas variáveis, como (Manning et al. 2009): representação vetorial dos documentos, otimização de compressão de índices, métodos estatísticos para avaliação de desempenho, métodos de *pooling*, avaliação de algoritmos em domínios de conhecimento diferentes, entre outros. Este trabalho tem

como escopo o estudo de diferentes técnicas de representação vetorial de documentos no domínio jurídico.

Existem inúmeras técnicas para representar documentos como vetores para sua recuperação através de medidas matemáticas. O *Vector Space Model* (VSM), geralmente abordado como um grande grupo das técnicas clássicas, divide o texto em termos (ou *tokens*) para construir um vocabulário comum entre todos os documentos. Cada termo pode ser ponderado ou valorado de diversas formas. Na fase de comparação de textos, os documentos candidatos são recuperados através do *match* exato de termos entre a consulta e documentos. Independente do valor de cada termo, essa abordagem não lida com características semânticas da linguagem natural (sinonímia, antônimas, homônimas, parônimas, polissemias, hiperônimos e hipônimos).

A evolução natural do *match* exato é a utilização de bases de conhecimento que mostre a proximidade entre termos com base em conceitos. Por exemplo, dicionários, tesauros e redes semânticas para que seja possível tratar sinônimos, antônimos, hiperônimos e hipônimos. A similaridade textual que utiliza essas bases é chamada de *Knowledge-Based*. Uma alternativa é a similaridade baseada na hipótese distribucional (Harris 1954) de termos em *corpus* de documentos (*Corpus-Based*). Nessa técnica, a partir de um *corpus* de documentos, os modelos supõem que palavras que ocorrem frequentemente no mesmo contexto possuem alguma relação semântica. Diversos modelos aplicam técnicas que variam de decomposição de matrizes, redes neurais rasas, até redes neurais profundas para extrair vetores que capturem essa propriedade contextual. Neste trabalho ao citar similaridade semântica estaremos nos referindo às técnicas baseadas em *corpus*.

Nesta pesquisa exploramos os modelos *Term Frequency - Inverted Document Frequency* (TF-IDF) (Sparck Jones 1988) e *Best Match 25* (BM25) (S. Robertson e Zaragoza 2009) do VSM. Para estudar a captura de semântica pelos modelos baseados em *corpus*, comparamos o Word2Vec (Mikolov, Chen et al. 2013) e *Bidirectional Encoder Representations from Transformers* (BERT) (Devlin et al. 2019). O primeiro lida com a proximidade semântica dos termos com estatística do *corpus* inteiro, o segundo gera vetores diferentes para o mesmo termo a depender do contexto que é utilizado.

1.1 Definição do Problema

No contexto do Superior Tribunal de Justiça, as decisões dos Ministros são armazenadas, organizadas e disponibilizadas pela Secretaria de Jurisprudência através de um processo de trabalho mapeado e bem definido. A correta execução do processo visa garantir a uniformização da jurisprudência com base em teses jurídicas. Logo, a similaridade e relevância dos documentos é definida pelas teses jurídicas que compõem a decisão.

Para dar apoio a esse tipo de pesquisa, o sistema atual somente fornece como meio de pesquisa a busca por consultas booleanas. O sistema legado atual possui as mesmas características apresentadas anteriormente por Sanchez et al. (2020) e Russell-Rose et al. (2018). Esse sistema suporta dois produtos da Secretaria: categorização de decisões (em principais e sucessivas) e catalogação de teses com consultas pré-programadas.

No primeiro produto, uma decisão principal possui diversos documentos sucessivos que contêm as mesmas teses jurídicas. Essa divisão é realizada com os objetivos de: reduzir informação redundante para comunidade jurídica; evidenciar a representatividade das decisões de cada ministro; e filtrar a quantidade de decisões que são processadas manualmente para evitar retrabalho, pois as decisões principais são fonte para extração de outras informações além dos dados textuais. O processo de trabalho é demonstrado pela Figura 1.1. Nessa figura, DJe é a sigla para Diário de Justiça Eletrônico.

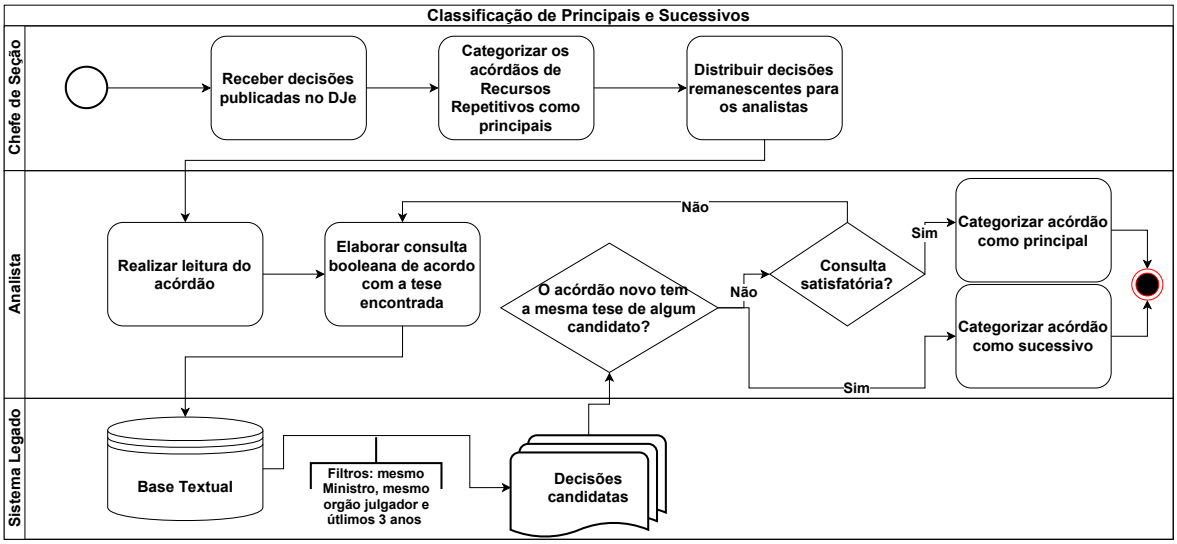


Figura 1.1: Fluxograma de trabalho da Seção de Classificação de Principais e Sucessivos.

O segundo produto tem por objetivo disponibilizar consultas prontas de teses jurídicas relevantes para o Tribunal. Essas consultas devem ser genéricas a fim de identificar na base do Tribunal o maior número de teses viáveis. O manual de trabalho destaca como de suma importância a elaboração de um bom critério de pesquisa, uma vez que a ausência de resgate de precedentes pertinentes leva à falsa compreensão do tema e do entendimento da Corte.

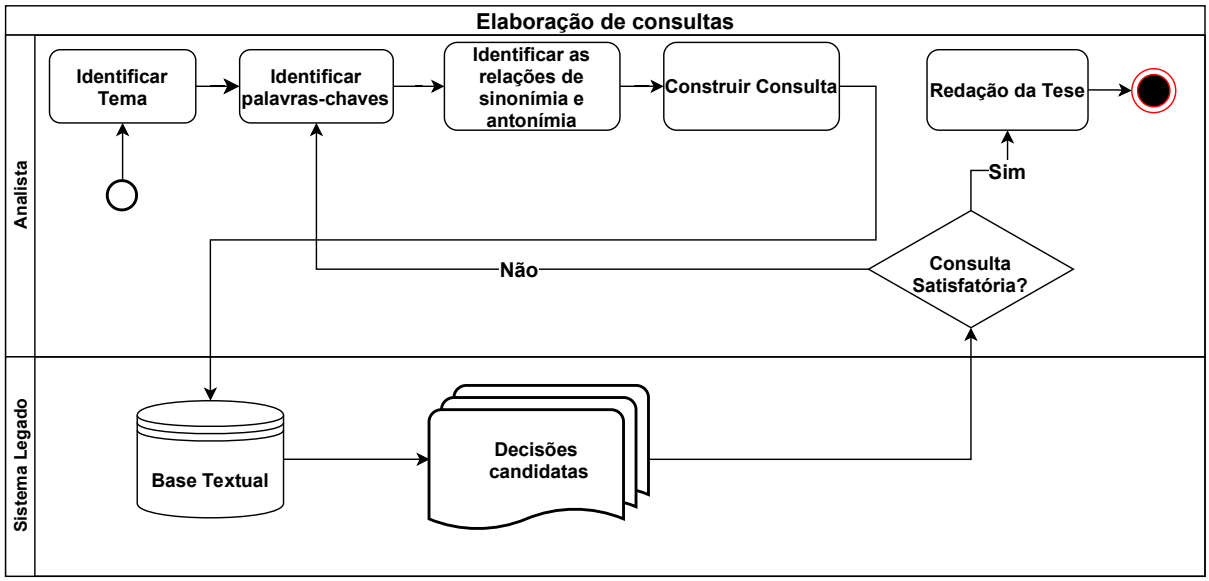


Figura 1.2: Fluxograma de trabalho da Seção de Jurisprudência em Teses.

A Figura 1.2 demonstra o fluxo de trabalho no manual de trabalho da equipe. Percebe-se que há um maior detalhamento nas etapas como a dedicação de uma etapa inteira para a identificação de sinônimos e antônimos.

Soma-se ao esforço necessário para construção da consulta o treinamento para servidores novos utilizarem o sistema de busca de forma eficaz dura entre 2 e 3 meses, e o volume crescente de decisões nos últimos anos. Até setembro de 2020, foram recebidos 262.404 processos e proferidas 383.651 decisões[1]. A Figura 1.3 demonstra o comparativo entre os anos. Dessa forma, o sistema de busca fornecido atualmente se apresenta como um gargalo e demanda muito retrabalho na realização do trabalho da Secretaria de Jurisprudência.

Nesse cenário e com base na revisão da sistemática da literatura, este trabalho possui as seguintes questões de pesquisa:

1. *Como comparar diferentes técnicas de similaridade textual no contexto da uniformização de jurisprudência do STJ?*

   A fase mais complexa da pesquisa em Recuperação da Informação (RI) é a modelagem do problema (Büttcher et al. 2016). Escolher os elementos pertinentes para o domínio legal e o *corpus* em questão é uma tarefa de extrema importância.

2. *Os modelos RI tradicionais podem ajudar os analistas a identificar documentos relevantes de forma menos custosa em comparação com o sistema legado?*

[1]Boletim Estatístico do Superior Tribunal de Justiça 2018. Disponível em:`http://www.stj.jus.br/webstj/Processo/Boletim/sumario.asp`.

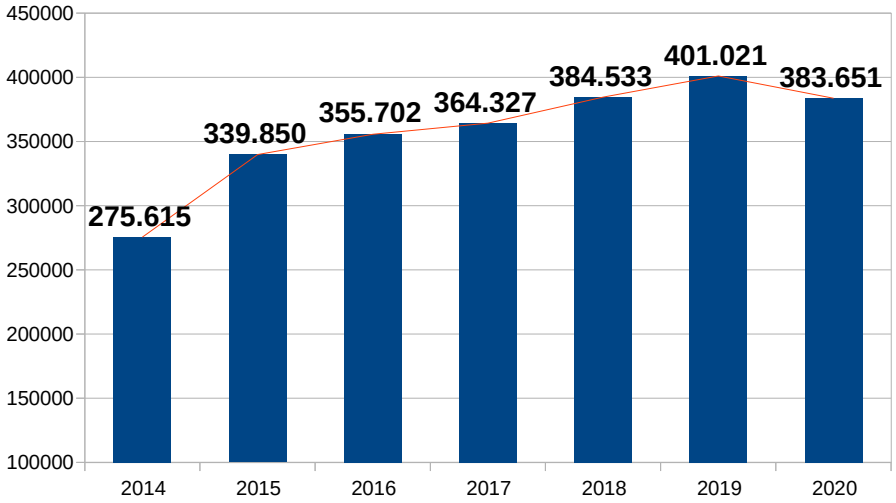


Figura 1.3: Evolução do quantitativo de decisões terminativas entre os meses de janeiro a setembro dos anos de 2014 a 2020 do STJ (Fonte: Superior Tribunal de Justiça 2020).

Apesar da confiança dos analistas no sistema legado com consultas Booleanas, há um elevado custo de tempo e conhecimento para que sejam elaboradas consultas para capturar relevância de tópicos. Enquanto os modelos tradicionais do *Vector Space Model* (VSM) foram modelados para capturar essa relevância a partir de consultas de texto puro (Zhai e Massung 2016).

3. *Podemos obter melhores resultados ao aplicar modelos semânticos?*

   Apesar dos métodos tradicionais facilitarem a consulta ao fornecerem uma boa ordenação de relevância sem operadores complexos, é possível que seu desempenho seja limitado pela ausência de similaridade semântica entre termos e que ainda seja necessária a adição manual de sinônimos por parte dos analistas. Por isso, exploramos a necessidade e impacto dos modelos semânticos na uniformização de jurisprudência do STJ.

4. *Como combinar modelos semânticos com os modelos tradicionais?*

   Os modelos semânticos tendem a facilitar a busca de informação ao aproximarem termos pela sua semântica, entretanto ao representar documentos com Word2Vec não há ponderação de relevância estatística de cada termo e os termos mais fre-

quentes acabam influenciando mais a direção do vetor no espaço vetorial. Diversos autores demonstram impacto positivo na combinação dos modelos tradicionais com semânticos (Galke et al. 2017; Kim et al. 2017; Zheng e Callan 2015).

1.2 Objetivos e contribuições

Esta pesquisa tem como objetivo geral: Avaliar o potencial da substituição do sistema legado (baseado em consultas booleanas) por novas técnicas de busca textual. Os objetivos específicos são:

1. Substituir a construção de consultas booleanas manuais por uma consulta em linguagem natural.
2. Avaliar a diferença de desempenho entre as técnicas semânticas e tradicionais.
3. Comparar o resultado das técnicas de similaridade aplicadas com o resultado das consultas de especialistas do sistema legado.

1.3 Justificativa

Alinhamento estratégico. O STJ é o tribunal superior que possui o dever constitucional, legal e regimental de dirimir divergências na aplicação de leis federais. Em seu Plano Estratégico 2015-2020[2], escolheu como sua missão: "Oferecer à sociedade prestação jurisdicional efetiva, assegurando uniformidade à interpretação da legislação federal". Como sua visão de futuro: "Tornar-se referência na uniformização da jurisprudência, contribuindo para a segurança jurídica da sociedade brasileira.", detalha que o Tribunal deve adotar práticas vanguardistas na construção e manutenção de uma jurisprudência coesa, a ponto de se tornar amplamente reconhecido no meio jurídico ao garantir previsibilidade na aplicação do Direito, devidamente justificada e motivada, com o fim de realizar justiça e contribuir para a estabilidade social.

Otimização de recursos humanos. Com o crescimento de aplicações de Inteligência Artificial no Judiciário Brasileiro, o foco da Administração do STJ está voltado para redução no tempo de rotinas de trabalho e aumento de produtividade. Teruel et al. (2018) demonstram que 35% do tempo dos profissionais jurídicos é dedicado à pesquisa textual. Investigações que avaliem métodos que reduzam esse custo são de suma importância para o STJ e toda comunidade jurídica.

[2]Aprovado pela Resolução STJ/GP n. 6 de 12 de maio de 2015.

Impacto social. Moreto (2012) explica que a uniformização de jurisprudência é um direito do jurisdicionado e um dever do Estado. A eficiência na sua execução traz celeridade processual e segurança jurídica na medida que, por meio de uma única tese, abrange partes ligadas por circunstâncias de fato em uma relação jurídica base ou origem comum e as tratam de forma isonômica e mitiga divergências.

1.4 Organização do Trabalho

Este trabalho está dividido da seguinte forma:

- Capítulo 2: discorre de forma sucinta sobre os componentes fundamentais de sistemas de RI, diferencia os modelos booleano do *Vector Space Model* (VSM), apresenta a teoria dos modelos TF-IDF, BM25, Word2Vec e BERT. Por fim, apresenta os trabalhos relacionados que avaliam recuperação de informação no domínio jurídico.
- Capítulo 3: descreve a metodologia TREC de forma detalhada e como é aplicada no escopo deste trabalho.
- Capítulo 4: apresenta o sistema construído, os dados utilizados e resultados alcançados.
- Capítulo 5: versa de forma sucinta sobre as respostas das questões de pesquisas e apresenta a pretensão de trabalhos futuros com base nos resultados alcançados.

Capítulo 2

Fundamentação Teórica

O problema focal de sistemas de Recuperação da Informação (RI) textual, nesta tese utilizamos o termo RI para abordar os sistemas de recuperação textual, é recuperar todos os documentos que contenham informação relevante para a necessidade de seus usuários, enquanto recuperam o mínimo possível de documentos não relevantes (Baeza-Yates e Ribeiro-Neto 2011). Tomando como exemplo uma necessidade de informação no domínio do STJ: *Achar todas as decisões colegiadas do STJ que definam os requisitos exigidos para demonstração do crime contra a honra.*

A descrição da necessidade de informação não necessariamente fornece os insumos necessários para consultar um sistema de RI. Geralmente, primeiro o usuário traduz a necessidade para uma consulta (*query*) ou um conjunto de consultas. A forma mais comum para representar consultas é um conjunto de termos soltos que possuem alguma relação com a necessidade de informação. A partir das consultas, o sistema de RI tem a tarefa de recuperar e apresentar para o usuário documentos que possuem a informação buscada.

Baeza-Yates e Ribeiro-Neto (2011) explicitam que a definição de relevância é fundamental para construção de RI efetivos. Por exemplo, esse conceito pode mudar por dimensões temporais, regionais e até depende em qual dispositivo a informação é exibida. Por isso, não existe sistema de RI perfeito. No domínio jurídico, Opijnen e Santos (2017) apresentam em seu trabalho uma taxonomia de conceitos de relevância.

De acordo com os autores, a modelagem de um sistema de Recuperação da Informação possui quatro componentes de alto nível:

- Q: um conjunto de necessidade de informação ou consultas.
- D: uma coleção ou *corpus* de documentos.
- F: um *framework* transforma as consultas e documentos textuais na mesma representação numérica.

- $R(q_i \in Q, d_j \in D)$: uma função que mede a similaridade das consultas em Q com os documentos em D. O valor fornecido por essa função é utilizado para ordenar os resultados.

A necessidade da informação pode ser transformada em consulta de diversas formas: um conjunto de termos digitados manualmente ou gerada através da seleção de trechos de texto pelo usuário. Definir o que compõe um documento (D) é outro ponto focal. Documentos podem ser todas as peças textuais de um processo ou parágrafos de cada peça. Nos modelos clássicos, o *framework* F é o componente responsável por preprocessar os documentos, criar um vocabulário, mapear o relacionamento dos documentos com os termos e ponderar os termos de acordo com o modelo escolhido. Nos modelos semânticos, F transforma os documentos em vetores densos.

2.1 Modelos Booleanos

Latha (2017) e Russell-Rose et al. (2018) afirmam que apesar de ser um dos primeiro modelos de recuperação textual, ainda é um dos mais utilizados. Ele é baseado na teoria dos conjuntos e na lógica booleana. Documentos são considerados relevantes para uma consulta quando os termos presentes nele fazem a fórmula da consulta verdadeira. Essa fórmula utiliza os operadores de lógica matemática de George Boole: conjuntivos (AND), disjuntivos (OR) ou negação (NOT). Em outras palavras, é um sistema que retorna documentos por *match* exato. Por exemplo, para uma consulta "crime AND hediondo AND (NOT qualificado)" seriam documentos relevantes os que contenham as palavras *crime* e *hediondo*, mas não contenham a palavra "qualificado".

A preferência dos usuários por esse modelo se baseia na fácil interpretação da relação das consultas com os resultados. Entretanto, são sistemas que não fornecem relevância gradual. Manning et al. (2009) e Baeza-Yates e Ribeiro-Neto (2011) destacam que alguns sistemas podem ser booleanos estendido. Nesses sistemas, há suporte para operadores de proximidade entre termos e operadores coringas. A Tabela 2.1 lista os operadores utilizados no sistema legado da Corte.

Tabela 2.1: Operadores utilizados pelo sistema legado do STJ. Tabela extraída do Manualde Pesquisa de Jurisprudência da Corte.

Operador	**Usado para**
E	Localizar palavras em qualquer lugar do documento, em qualquer ordem
OU	Localizar uma e/ou outra palavra, em qualquer ordem (Obs.:sempre entre parênteses)
NÃO	Excluir palavras ou argumentos de pesquisa inteiros
MESMO	Localizar palavras em um mesmo parágrafo, em qualquer ordem
COM	Localizar palavras em um mesmo parágrafo, em qualquer ordem
PROXn	Localizar palavras próximas, em qualquer ordem
ADJn	Localizar palavras adjacentes, na ordem direta
$	Localizar todas as variações a partir de um trecho de palavra
?	Localizar um caractere ligado ao trecho de palavra
( )	Agrupar itens de pesquisa / operador OU
“ ”	Localizar expressões exatas, operadores como palavras e números

2.2 Vector Space Model – VSM

O *Vector Space Model* é fundamentado na representação de textos como vetores de alta dimensionalidade em que cada componente do vetor representa um termo de um vocabulário (Büttcher et al. 2016). De maneira mais formal, a partir de um *corpus* de documentos é criado um vocabulário de termos W, cada termo $t_i \subset W$ recebe uma ponderação estatística. Manning et al. (2009) dividem essa ponderação em duas categorias: frequência do termo no documento (local) e na base de documentos (global).

A combinação mais utilizada dessas categorias é chamada de *Term Frequency - Inverted Document Frequency* (TF-IDF), que pondera com mais importância termos que aparecem frequentemente em um número pequeno de documentos, e pondera com menor importância termos que ocorrem com muita frequência em muitos documentos. Existem diversas variantes para realizar essa medição, mas a mais utilizada é representada na Equação 2.1, onde N representa o número de documentos no *corpus*. Nela, a ponderação local é representada por $tf(t_i, d_j)$, que representa a frequência do termo $t_i \subset W$ em um determinado documento $d_j \subset D$. Por sua vez, a ponderação global é representada pelo segundo operando da equação (idf), onde df_{t_i} representa a frequência do termo $t_i \subset W$ em todos os documentos do *corpus*. O objetivo do idf é determinar a especificidade de um termo, ou seja, o potencial de distinção que o termo fornece para os documentos em que aparece. Esse potencial está relacionado com o tópico ou tema contido no documento (Baeza-Yates e Ribeiro-Neto 2011).

$$TFIDF(t_i, d_j) = tf(t_i, d_j) \cdot \log(\frac{N}{df_{t_i}}) \tag{2.1}$$

No Modelo Probabilístico, a ordenação dos documentos é computado de acordo com a probabilidade de ser relevante. S. E. Robertson (1977) definiu o princípio da ordenação probabilístico como: se a resposta de um sistema de recuperação a cada solicitação é a ordenação de documentos em ordem decrescente de probabilidade, no qual a precisão das probabilidades estimadas depende de dados disponibilizados previamente ao sistema, a efetividade do sistema é diretamente proporcional a qualidade desses dados.

Com base nisso, diversos modelos foram definidos, sendo o mais conhecido Okapi BM25 (S. E. Robertson e Walker 1999). Ele utiliza ponderação estatística combinada com uma suposição probabilística. Introduz parâmetros na ponderação local dos termos (tf) para controlar a penalização no peso de um mesmo termo em documentos de tamanhos diferentes. Esta abordagem pressupõe que quanto maior o comprimento do documento, maior é a probabilidade de que os termos que correspondem à consulta estejam espalhados ao longo do documento e este documento seja menos específico do que outros documentos menores com os mesmos termos. A equação BM25 original e mais usada não altera o *idf* de um termo (Büttcher et al. 2016). A modificação ocorre na equação da frequência do termo (local).

$$bm25\_tf(t_i, d_j) = \frac{tf(t_i, d_j) \cdot (k+1)}{tf(t_i, d_j) + k \cdot (1 - b + b \cdot \frac{|d_j|}{\text{avgdl}})} \tag{2.2}$$

A Equação 2.2, introduz dois parâmetros para controlar a ponderação dos termos: k e b. O primeiro varia entre $[0, \infty]$ e funciona como limite superior para a ponderação local dos termos. O segundo varia entre $[0, 1]$ e controla o impacto que o tamanho do documento, em relação ao tamanho médio dos documentos no *corpus*, tem na relevância. Assim, captura a probabilidade dos termos de busca estarem mais dispersos ou menos dispersos de acordo com o princípio de ordenação probabilística. $|d_j|$ representa o tamanho de um documento calculado pela quantidade de termos. *avgdl* é a média de tamanho de documentos no *corpus*.

2.3 Modelos Semânticos

Gomaa, Fahmy et al. (2013) apresentam uma taxonomia para divisão de modelos semânticos textuais: baseados em conhecimento (*Knowledge-based*) e baseados no *corpus* (*Corpus-based*). O primeiro é a construção manual de dicionários, tesauros e bases léxicas (*lexicon* em inglês) que mapeiam o tipo de relação entre termos e medem sua distância dentro de cada tipo de relação. O exemplo mais conhecido é o WordNet (Miller 1998).

Por sua vez, os modelos baseados em *corpus*, partem da hipótese de distribuição linguística (Harris 1954): termos que ocorrem com muita frequência no mesmo contexto, tendem a ter uma relação semântica de similaridade. Lin (1998) utiliza como exemplo:

A bottle of tezgiiino is on the table.
Everyone likes tezgiiino.
Tezgiiino makes you drunk.
We make tezgiiino out of corn.

Apesar do termo Tezgiiino não existir, é possível inferir que é uma bebida alcoólica. Baroni et al. (2014) divide os modelos baseados em *corpus* em dois grandes grupos: baseado em contagem e baseado em predição. O objetivo dos dois grupos é o mesmo: produzir *word embeddings*. Neste trabalho utilizamos a definição de Almeida e Xexéo (2019) para *word embeddings*: vetores densos, distribuídos, de comprimento fixo, construídos usando estatísticas de co-ocorrência de palavras de acordo com a hipótese de distribuição linguística.

Os modelos de contagem iniciam sua abordagem com a construção de uma matriz de adjacência de termos, a partir dessa matriz é aplicada uma medida de associação que pondera a importância da coocorrência de dois termos em relação às outras coocorrências no *corpus*. Nesse grupo podemos citar modelos como o *Pointwise Mutual Information* (PMI) e seus derivados: *Exponential Pointwise Mutual Information* (EPMI), *Local Mutual Information* (LMI) e *Normalized Pointwise Mutual Information* (NPMI) (Aletras e Stevenson 2013).

Os modelos baseados em predição abordam o problema como uma tarefa de classificação: a partir de um *corpus*, geram exemplos de treinamento de forma não supervisionada em pares de termos e contextos. Na fase de treinamento, aplicam redes neurais para classificar um termo a partir de um contexto, ou um contexto a partir de um termo. Almeida e Xexéo (2019) apresentam uma revisão sistemática a aplicação de redes neurais para geração de *word embeddings*. Apesar do modelo de Mikolov, Chen et al. (2013) ter ganhado maior notoriedade nas aplicações práticas, as primeira publicações iniciaram em 2003 (Bengio et al. 2003).

O funcionamento genérico do modelo Wor2Vec segue os seguitnes passos:

1. Inicializa um conjunto de vetores para cada palavra no vocabulário.
2. Gera exemplos de treinamento a partir de janelas deslizantes sequenciais do *corpus*.
3. Multiplica os vetores do termo alvo com a matriz de contexto.
4. Minimiza o erro na predição dos termos próximos utilizando SGD (*Stochastic Gradient Descent*).

Dessa forma, quanto mais dois termos aparecem no mesmo contexto, maior será o reforço da proximidade deles. A Figura 2.1 mostra somente os exemplos positivos de treinamento. Como o tamanho de vocabulários tende a ser grandes, ajustar os pesos para todos os termos do vocabulário não é computacionalmente eficiente. Os autores utilizaram duas técnicas para abordar esse problema: *Hierarchical Softmax* e *Negative Sampling*. A premissa básica é escolher de forma inteligente somente alguns exemplos negativos em cada iteração.

O Word2Vec possui duas arquiteturas:

1. SkipGram.
2. *Continuous Bag-Of-Words* (CBoW)

Na primeira, dada uma palavra a rede tenta prever a probabilidade das palavras que aparecem no contexto. Através de uma janela deslizante, de tamanho definido previamente, pares de termos são extraídos. Por sua vez o CBoW simula uma tarefa de "preencha a palavra faltante", e para cada conjunto de palavras de contexto tenta prever a palavra central. A Figura 2.1 demonstra exemplos gerados a partir de uma sentença com janela de tamanho 2. As palavras em azul são as palavras de entrada da rede. Cada exemplo é um par de palavra e palavra de contexto. Quanto maior a janela, mais exemplos serão gerados.

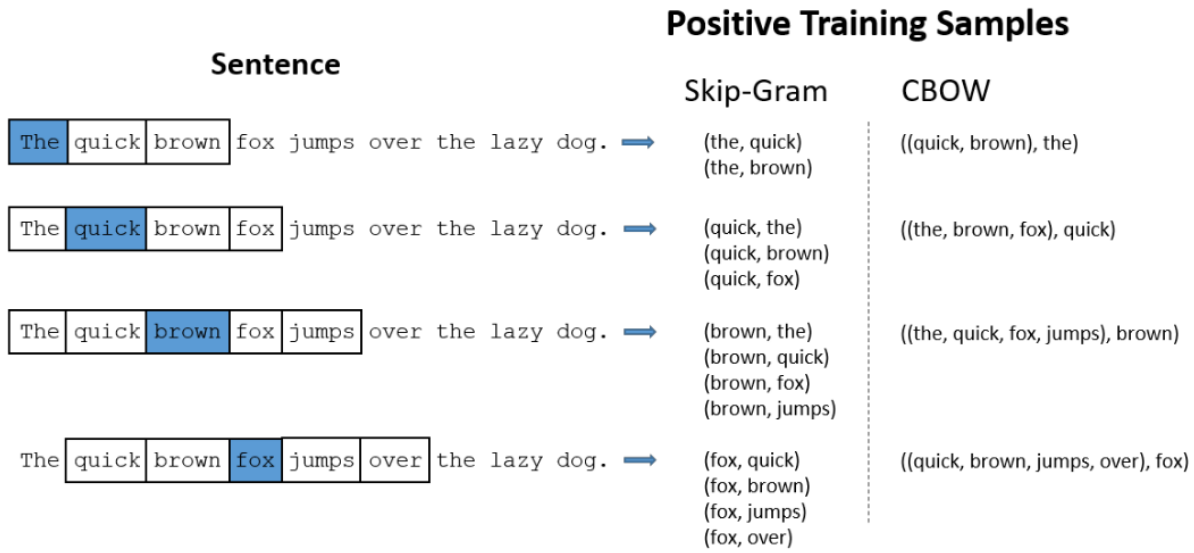


Figura 2.1: Exemplo da geração de exemplos de treinamento com uma janela deslizante com tamanho igual 2 (Fonte: McCormick 2019).

A Figura 2.2 demonstra a arquitetura da rede neural utilizada para o Skip-Gram. Dado um vocabulário de tamanho V e um valor para a dimensão N dos *embeddings*:

- Os vetores de entrada (x) e saída (y) são vetores de codificação binária (*one-hot*) dos termos w_i e w_c (termo alvo e termo contexto).
- Os vetores servem como indices para a matriz de embedding W de tamanho $V \times N$. Essa seleção é realizada através da multiplicação do vetor x com a matrix W.
- Então o vetor W da palavra w_c será utilizado como camada escondida.
- Em seguida a matriz de contexto W', de tamanho $N \times V$, é multiplicada pela camada escondida.
- Na última etapa, é utilizada a função de ativação Softmax para calcular a probabilidade do termo w_c estar no contexto de w_i.
- Por fim, o SGD é aplicado para ajuste dos pesos.
- A matriz W é utilizada como representação dos termos ao final do treinamento.

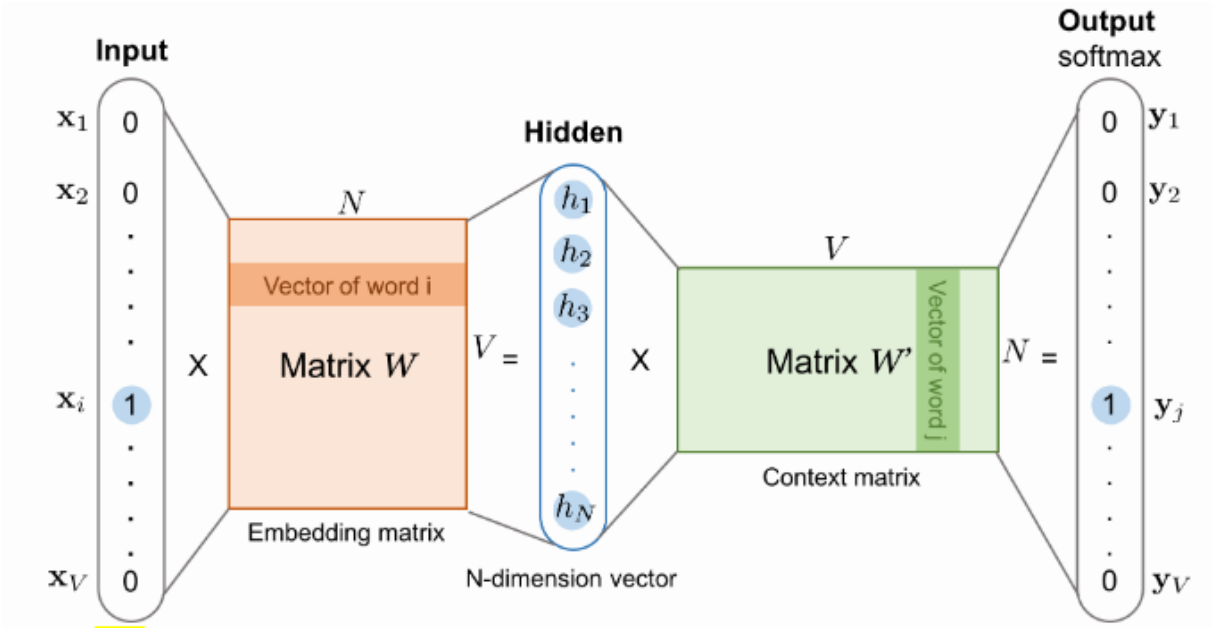


Figura 2.2: Arquitetura da rede neural Skip-Gram (Fonte: Weng 2017).

De maneira formal, um *corpus* é representado por uma sequência de termos ou palavras: $w_1, w_2, ..., w_T$. A partir de um valor c para a janela de contexto, o CBoW maximiza a função objetivo (Equação 2.3), enquanto o Skip-Gram maximiza a Equação 2.4. A diferença central é que vetor do conjunto de termos do contexto no CBoW é construído através da soma dos vetores de cada termo do contexto.

$$J(\theta) = \frac{1}{T}\sum_{t=1}^{T} \log p\left(w_t \mid \sum_{-c\leq j\leq c, j\neq 0} w_{t+j}\right) \tag{2.3}$$

$$J(\theta) = \frac{1}{T}\sum_{t=1}^{T} \sum_{-c\leq j\leq c, j\neq 0} \log p\left(w_{t+j} \mid w_t\right) \tag{2.4}$$

A probabilidade do termo do contexto w_c dado um termo alvo w_t é dado pela Equação 2.5. Onde u_{wt} é o vetor e *embedding* do termo alvo e u_{wc} o embedding do contexto. Entretanto, o cálculo do gradiente (denominador) tem custo proporcional ao tamanho do vocabulário V. Em um trabalho posterior, Mikolov, Sutskever et al. (2013) aplicaram duas soluções:

- Hierarquical Softmax (Morin e Bengio 2005): A saída da rede (após a ativação softmax) é serializada em uma árvore, onde cada folha é um termo e cada nó interno representa a probabilidade dos termos nas folhas. Essa técnica reduz a complexidade de $O(V)$ para $O(\log V)$.

- Negative Sampling: baseado no Noise Contrastive Estimation (NCE) (Gutmann e Hyvärinen 2010). Faz amostragem de exemplos negativos usando regressão logística. A probabilidade de escolha é retirada de uma distribuição uniforme de termos.

$$p\left(w_c \mid w_t\right) = \frac{\exp\left(v_{w_c}^T u_{w_t}\right)}{\sum_{w=1}^{W} \exp\left(v_w^T u_{w_t}\right)} \tag{2.5}$$

Diferente do VSM, no qual cada documento possui um vetor do tamanho do vocabulário V, os modelos semânticos baseados em redes neurais representam cada termo w com um vetor denso extraído de uma camada escondida. O Word2Vec utiliza o vetor da matriz W de tamanho pré-definido N. Para representar um documento é possível combinar os vetores de todos os termos do documento através da média ou soma dos vetores individuais, o que resulta em um vetor único de tamanho N.

Modelos preditivos como o Word2Vec não lidam com a semântica de forma completa. Por exemplo, termos antônimos como "bom" e "mau" possuem vetores próximo pois ocorrem no mesmo contexto frequentemente. Recentemente diversos modelos linguísticos (*language models*) que geram vetores diferentes a depender do contexto de cada termo estão sendo utilizados. Devlin et al. (2019) desenvolveram o *Bidirectional Encoder Representations from Transformers* (BERT), modelo que apresenta resultados de estado da arte em diversas tarefas de NLP (Z. Yang et al. 2019).

Ele utiliza mecanismos de atenção (*Transformers*) que aprendem relações contextuais entre termos. Esse mecanismo é composto de *Encoders* e *Decoders*. O BERT utiliza

somente os *Encoders*. Seu diferencial é que carrega a entrada de forma paralela, ao invés da forma tradicional sequencial.

O BERT é composto por Encoders empilhados e o modelo original possui dois tamanhos. O primeiro, Base, é composto por 12 camadas de Encoders. O segundo, Large, é composto por 24 camadas de Encoders. Entre cada Encoder há camadas de *feed-forward* com 768 e 1024 dimensões para os modelos Base e Large, respectivamente.

Na fase de treinamento ele é treinado em duas tarefas: *Masked Language Model* e *Next Sentence Prediction*. O primeiro tem o objetivo de prever tokens removidos de frases de forma aleatória e a segunda tem o objetivo de prever a frase seguinte a partir de uma anterior.

Como entrada o modelo recebe os tokens já separados pelo algoritmo Byte-Pair Encoding (BPE). Adicionalmente são adicionais dois tokens especiais: [CLS], no ínicio e é utilizado para consolidar os vetores de todos os tokens e segmentos; e o token [SEP], utilizado no fim da primeira e da segunda sentença na tarefa de *Next Sentence Prediction*. Para extrair *embeddings* de documentos é possível utilizar a combinação dos vetores produzidos por qualquer uma das 12 ou 24 camadas do modelo. Mais detalhes são explicados na Seção 4.3.3.

2.4 Função de similaridade

Diversas métricas são propostas para calcular a similaridade entre dois vetores. A similaridade do cosseno é a mais utilizada. Representada pela Equação 2.6, a similaridade do cosseno mede a distância entre dois vetores através do ângulo. Dessa forma, a magnitude do vetor não influencia o resultado. Em todos os modelos apresentados, os documentos e consultas são representados como vetores esparsos ou densos. Em ambos os casos, a partir de dois vetores de documentos A e B de tamanho n, a similaridade do cosseno é dada pelo produto escalar dos dois vetores, divididos pelo produto da normalização euclidiana de ambos os vetores.

$$sim(A, B) = \frac{\sum_{i=1}^{n} a_i \cdot b_i}{\sqrt{\sum_{i=1}^{n} a_i^2} \cdot \sqrt{\sum_{i=1}^{n} b_i^2}} \tag{2.6}$$

2.5 Métricas de avaliação

Para comparar os diversos sistemas, a escolha da métrica de desempenho é baseada na tarefa do usuário e na noção de relevância que orienta essa tarefa. Zhai e Massung (2016) explicam que há dois grandes grupos de métricas de avaliação de sistemas de recuperação da informação: as que avaliam a relevância dos documentos retornados e as que avaliam

a ordem dos documentos relevantes retornados. No primeiro grupo estão Precisão, Revocação e F-measure. Precisão, demonstrada pela Equação 2.7a, calcula a proporção de documentos relevantes D_r retornados em relação a todos documentos retornados. Revocação, demonstrada pela Equação 2.7b, calcula a proporção de documentos relevantes retornados em relação a todo conjunto de documentos relevantes existentes no *corpus* $D_r + D_{rnr}$, onde D_{rnr} representa os documentos relevantes não recuperados. Ambas as métricas variam entre 0 e 1. Naturalmente, os sistemas de recuperação de informação tendem a ter uma precisão menor quando há um aumento na quantidade de documentos retornados para se aumentar o revocação, para balancear essa decisão a média harmônica é realizada através da F-measure (Equação 2.7c). O valor de β controla a preferência entre precisão e revocação, mas na maioria das vezes utiliza-se o valor igual a 1. Nesse caso, a métrica dá peso igual para revocação e precisão e penaliza resultados onde somente um dos valores é alto.

$$Precisão = \frac{D_r}{D_r + D_{nr}} \tag{2.7a}$$

$$Revocação = \frac{D_r}{D_r + D_{rnr}} \tag{2.7b}$$

$$F_\beta = \frac{(\beta^2 + 1) \times P \times R}{\beta^2 \times P + R} \tag{2.7c}$$

Esse primeiro grupo é adequado quando a tarefa de recuperação de informação utiliza somente critérios de relevância binária: documento relevante ou não relevante. Entretanto, para cenários de recuperação de informação mais complexos, a tarefa do usuário é realizada em estágios como por exemplo a revisão bibliográfica em trabalhos de pesquisa acadêmica. Portanto, é necessário um critério de relevância com maior flexibilidade. Na tarefa utilizada como exemplo, o usuário realiza a pesquisa, utiliza os resumos dos artigos para decidir se irá realizar ou não a leitura completa do artigo, mas como é uma atividade muito custosa os critérios de relevância devem fornecer meios para que a segunda etapa (leitura do texto completo dos artigos selecionados) seja realizada em ordem de probabilidade dos artigos possuírem a informação buscada.

Supondo que a relevância dos documentos pode ser tranformada em valores númericos, Järvelin e Kekäläinen (2000) e Burges et al. (2005) definiram um conjunto de métricas que comparam sistemas em relação a capacidade de ordernar documentos mais relevantes antes do documentos menos relevantes e, por fim, esses antes dos não relevantes.

Na tarefa exemplo da revisão bibliográfica, podemos definir os seguintes níveis de relevância:

Tabela 2.2: Ordenação do sistema hipotético A.

Ordenação	**Relevância**
1	2
2	1
3	2
4	0
5	1

Tabela 2.3: Ordenação do sistema hipotético B.

Ordenação	**Relevância**
1	1
2	0
3	2
4	1
5	2

- 2: resumos que descrevam o escopo, a metodologia, os modelos e resultados alcançados. A temática dos artigos tenha relevância com a pesquisa da tarefa da revisão bibliográfica em questão.
- 1: resumos que descrevam o escopo e a metodologia mas não os modelos e resultados. A temática dos artigos parece ter relevância, mas é necessário ler o texto inteiro para que se tenha certeza.
- 0: resumos que não tenham nenhuma relevância com a revisão bibliográfica.

Na segunda etapa da pesquisa, o usuário ordena os artigos para leitura com os de resumo rotulados com relevância 2 e caso esses não satisfaçam sua revisão, utilizaria os resumos menos relevantes. Considerando dois sistemas de ordenação hipotéticos A e B. As Tabelas 2.2 e 2.3 apresentam o resultado da ordenação de cada um respectivamente.

O trabalho dos autores citados começa por definir uma métrica simples: *Cumulative Gain* (CG). Representada pela Equação 2.8, é o somatório do valor da relevância (rel_i) dos documentos na ordenação sem levar em consideração sua posição. Ao somar os valores de relevância das Tabelas 2.2 e 2.3 o valor 6 para os dois sistemas, o que impossibilita a diferenciação da capacidade dos sistemas na tarefa do usuário.

A primeira etapa para conseguir mensurar o impacto da posição em que os documentos aparece na ordenação é aumentar a penalização (ou descontar) a medida que a posição na ordenação cresce. Esse desconto é o denominador da Equação 2.9 (*Discounted Cumulative Gain* (DCG) na posição n da ordenação). As Tabelas 2.4 e 2.5 mostram a memória de cálculo para nosso exemplo hipotético. Vemos que agora, é possível distinguir que o

Tabela 2.4: Métrica calculada para o sistema hipotético A.

Ordenação	**Relevância**	**Desconto**	**Relevância\Desconto**	**DCG**
1	2	1,00	3,0	3,0
2	1	1,58	0,63	3,63
3	2	2,00	1,5	5,13
4	0	2,32	0,0	5,13
5	1	2,58	0,39	5,51

Tabela 2.5: Métrica calculada para o sistema hipotético B.

Ordenação	**Relevância**	**Desconto**	**Relevância\Desconto**	**DCG**
1	1	1,00	1,0	1,0
2	0	1,58	0,0	1,0
3	2	2,00	1,5	2,5
4	1	2,32	0,43	2,93
5	2	2,58	1,16	4,09

sistema A é mais adequado e tem um desempenho melhor porque reduz o esforço do usuário.

$$\text{CG(n)} = \sum_{i=1}^{n} rel_i \tag{2.8}$$

$$\text{DCG(n)} = \sum_{i=1}^{n} \frac{2^{rel_i} - 1}{\log_2(i+1)} \tag{2.9}$$

Por fim, os autores apresentam o *Normalized Discounted Cumulative Gain* (nDCG): métrica ajustada em comparação com um sistema teórico com ordenação perfeita de todos os documentos mais relevantes ordenados antes dos menos relevantes (*Ideal Discounted Cumulative Gain* (IDCG)). A Tabela 2.6 apresenta o IDCG para nossa situação hipotética. Utilizando os valores apresentandos o sistema A e B teriam nDCG 0,94 e 0,68 respectivamente.

$$\text{nDCG(n)} = \frac{DCG(n)}{IDCG(n)} \tag{2.10}$$

Tabela 2.6: Cálculo do ranking ideal IDCG para a tarefa.

Ordenação	**Relevância**	**Desconto**	**Relevância\Desconto**	**IDCG**
1	2	1,00	3,0	3
2	2	1,58	1,89	4,89
3	1	2,00	0,5	5,39
4	1	2,32	0,43	5,82
5	0	2,58	0	5,82

Tabela 2.7: Exemplos de comparação de métrias entre dois sistemas hipotéticos (Zhai e Massung 2016).

Experimento I			**Experimento II**		
Consulta	**Sistema A**	**Sistema B**	**Consulta**	**Sistema A**	**Sistema B**
1	0,20	0,40	1	0,02	0,76
2	0,21	0,41	2	0,39	0,07
3	0,22	0,42	3	0,16	0,37
4	0,19	0,39	5	0,58	0,21
5	0,17	0,37	6	0,04	0,02
6	0,20	0,40	6	0,09	0,91
7	0,21	0,41	7	0,12	0,46
Média	0,20	0,40	Média	0,20	0,40

Tabela 2.8: Exemplo do teste de significância o experimento II (Zhai e Massung 2016).

Consulta	**Sistema A**	**Sistema B**	**Sign Test**	**Wilcoxon**
1	0,02	0,76	+	+0,74
2	0,39	0,07	−	-0,32
3	0,16	0,37	+	+0,21
4	0,58	0,21	−	-0,37
5	0,04	0,02	−	-0,02
6	0,09	0,91	+	+0,82
7	0,12	0,46	+	+0,34
Média	0,20	0,40	$p = 1,0$	$p = 0,9375$

Através de testes estatísticos de significância, podemos quantificar matematicamente se a métricas de desempenho de dois sistemas são de fato diferentes. Esse teste na RI indica quão provável uma diferença nas métricas se deve ao acaso.

Zhai e Massung (2016) mostram na Tabela 2.7 os valores da métrica de precisão em dois experimentos distintos. Os valores médios são iguais em ambos os experimentos e o Sistema B parece obter um desempenho muito superior. Ao observarmos os resultados detalhados por consulta no experimento I, é possível ver que em todas as consultas o sistema B obtém um desempenho melhor do que o sistema A. Já no segundo experimento, podemos notar uma variância maior por consulta.

Os testes de siginificância avaliam a variação nas métricas por consulta. Se houver uma grande variação, isso significa que os resultados podem variar de acordo com diferentes consultas, o que torna o resultado não confiável na comparação de resultados entre os sistemas.

O primeiro teste demonstrado na Tabela 2.8 é chamado de teste do sinal (*Sign Test*): quando o sistema B é melhor é utilizado o + e o sinal de − no caso inverso. Através desse método, temos 4 casos em que B é melhor e 3 casos em que A é melhor. Os resultados

parecem ser aleatórios nesse caso. Para conseguirmos diferenciar de fato, é necessário utilizar o p-valor: probabilidade de que os resultados são dados ao acaso. Nesse teste o p-valor é igual 1, o que demonstra que não há como afirmar uma diferença real entre os sistemas. O teste de Wilcoxon é um teste não paramétrico que considera a magnitude da diferença junto com o sinal, como demonstrado na coluna da tabela.

2.6 Trabalhos Relacionados

Para o levantamento dos artigos foi utilizada, nas base de dados Scopus (634 artigos encontrados) e Web of Science (221 artigos encontrados), a seguinte consulta:

> ((legal OR juridic) AND ((“document similarity”) OR (“information retrieval”) OR (“semantic similarity”) OR (“semantic search”) OR (“concept search”) OR (“e-discovery”) OR (“patent search”) OR (“embeddings”) OR (“word2vec”) OR (“deep learning”)))

Adicionalmente, foram levantados os periódicos e revistas especializados em data mining na área jurídica:

1. 136 artigos do *International Conference on Artificial Intelligence and Law* (ICAIL)
2. 234 artigos do *International Conference on Legal Knowledge and Information Systems* (JURIX)
3. 106 artigos da revista Artificial Intelligence and Law[1] da editora Springer

Para lista de artigos, foram selecionados primeiramente com base na relação do título do artigo ao trabalho aqui proposto; em seguida todos os artigos duplicados foram removidos e, por fim, foram lidos 57 artigos da lista final dos quais 4 foram escolhidos como trabalhos relacionados.

Diversos autores levantam desafios para que os sistemas de recuperação de informações jurídicas atendam de forma eficiente as demandas dos atores do direito (advogados, servidores e membros). Locke, Zuccon e Scells (2017) e Locke e Zuccon (2018) abordam a busca por decisões jurídicas relevantes para que advogados possam organizar e utilizar teses jurídicas de precedentes. Baseados estudos que estimam que 28% do tempo utilizado no trabalho jurídico é dedicado a busca de precedentes, trabalham com hipótese de que a eliminação da construção de consultas manuais pode reduzir o tempo para cada processo.

No primeiro trabalho, são avaliadas técnicas de geração automática de consultas através da seleção de termos com as ponderações: *Term Frequency - Inverted Document Frequency* (TF-IDF), *Kullback-Leibler Divergence for Informativeness* (KLI) e *Parsimonious*

[1] ISSN 1572-8382

Language Models (PLM). Utilizando as decisões da Suprema Corte dos Estados Unidos, foi gerada uma base com 63.916 documentos, 248 queries e 2.645 documentos anotados com relevância. Cada documento foi dividido em unidades menores representativas: sentenças e parágrafos. Com os documentos indexados na ferramenta ElasticSearch[2] e com a função score padrão BM25 (com os parâmetros $b = 0,75$ e $k1 = 1,2$) as consultas geradas automaticamente foram comparadas com as 248 elaboradas por advogados especialistas. Para medir a qualidade de sua abordagem os autores utilizaram as métricas de precisão na posição 1 e 5 (P@1 e P@5), a precisão média na posição 5 (AP@5) e a média de posicionamento recíproco – *Mean Reciprocal Rank* (MRR). Por fim, ao medir a significância estatísticas dos resultados, utilizando o teste t pareado, os autores chegaram a conclusão de que as consultas geradas automaticamente são, na média, tão efetivas quanto as manuais, mas são significativamente inferiores às melhores consultas dos especialistas, pois não consideram a semântica dos termos.

No segundo trabalho, os autores identificaram que sua base de dados, e outras utilizadas em trabalhos acadêmicos, não eram representativas em relação ao real problema. Assim, aumentaram sua base para 3.597.230 coletadas da mesma fonte. Desse *corpus*, foram selecionadas randomicamente 12 teses jurídicas. No total os especialistas jurídicos rotularam 2.572 decisões quanto a relevância. Com as mesmas técnicas do trabalho anterior, somente as métricas de avaliação foram alteradas. Acrescentou-se a precisão na posição 10 e 100, a precisão média e revocação na posição 10 e 100. Como conclusão os autores afirmam que a extração de termos através do TF-IDF tem melhor performance do que a extração por KLI e PLM.

Nejadgholi et al. (2017) apontam, no contexto jurídico, que apesar do aumento da capacidade de processamento de matrizes esparsas, por exemplo *Bag of Words* (BoW) com TF-IDF, a aplicação de semântica para mensuração de similaridade de textos é recente. Eles se baseam no trabalho de Grabmair et al. (2015), que aplicaram a busca semântica para recuperar decisões relacionadas a sequelas derivadas de vacinas. Nesse cenário, ponderam que os modelos que utilizam *embeddings* de palavras possuem alguma vantagens sobre os modelos tradicionais: o significado das palavras é definido pelo contexto, assim a proximidade semântica pode ser representada por vetores similares; e no modelo BoW não há como representar palavras fora do vocabulário de forma calculada.

Para estudar a viabilidade no contexto jurídico, os autores construíram um sistema de busca de casos sobre imigração no Canadá. Da mesma forma que outros autores citados anteriormente, primeiro utilizaram um protocolo de rotulação para dividir os textos em unidades semânticas menores: nesse caso dividiram os documentos em parágrafos que narravam fatos e que continham outro objetivo semântico o que resultou um total de

[2]https://www.elastic.co/

4.549.809 parágrafos. Treinaram um *Word Embeddings* utilizando o algoritmo FastText e a técnica SkipGram com 100 dimensões em uma base de 46.000 decisões judiciais da Suprema Corte Canadense. Em seguida selecionaram 15 parágrafos que continham fatos e rotularam parágrafos de outros casos com fatos semelhantes como relevantes. A partir os vetores de cada um dos 15 parágrafos utilizaram a similaridade do cosseno para recuperar parágrafos similares e atingiram precisão média de 78%. Os autores atribuíram o melhor resultado ao FastText pois dentre os parágrafos que continham a narração dos fatos, há muito erro ortográfico escrito por imigrantes e o FastText contorna esse problema aproximando vetores para palavras desconhecidas do vocabulário.

Sugathadasa et al. (2018) explicam em seu trabalho que há uma grande dependência dos especialistas do direito para recuperação de informação jurídica. Com base nisso, trabalham com a hipótese de que a construção de *Word Embeddings* para capturar a semântica dos termos especializados reduz a dependência de especialistas. Para validar essa hipótese, extraíram 2.500 documentos do site FindLaw[3] e para cada um desses documentos foram extraídos outros documentos que os citam. Com essa base *gold standard*, a partir de cada um dos 2.500 documentos, os autores utilizaram TF-IDF e Doc2Vec (uma variação do Word2Vec que leva em conta o contexto de cada documento) para comparar os documentos através da similaridade do cosseno. Mesmo para uma quantidade de documentos reduzidos, verificou-se que o modelo semântica obteve melhor revocação.

Por sua vez, Galke et al. (2017) focaram seu trabalho em testar o impacto dos modelos semânticos, especificamente *Word Embeddings*, na recuperação de informação em maior escala. Para isso escolheram bases *gold standard* abertas para a manter reprodutibilidade da pesquisa: NTCIR2, *Economics* e o *Reuters*. A primeira consiste em 134.978 documentos, dos quais foram extraídos 49 subgrupos de documentos com, em média, 43 documentos rotulados para cada grupo. O segundo consiste em 61.792 documentos, dos quais foram extraídos 4.518 subgrupos de documentos com, em média, 72 documentos rotulados por grupo. O último, consiste em 100.000 documentos de onde foram extraídos 102 subgrupos de documentos com, em média 3.143 documentos rotulados por grupo. Para servir de necessidade de informação, foram considerados como consultas simples os títulos de cada documento e como consultas completas a descrição, o resumo ou o texto completo. Para comparar e ordenar os documentos, os autores utilizaram TF-IDF, Word2Vec e Doc2Vec. A função de comparação utilizada foi a similaridade do cosseno e o *Word Mover's Distance* (WMD). Foram utilizadas as métricas *Mean Average Precision*, *Mean Reciprocal Rank* e *Normalized Discounted Cumulative Gain*. Os autores identificaram que, na média, a ponderação de termos por TF-IDF obteve melhores resultados.

[3] `https://www.findlaw.com/`

Apesar dos bons resultados que o BERT demonstra, adaptá-lo para tarefa de recuperação de informação é um desafio pelas seguinte características:

- Treinar um modelo do ínicio tem um custo muito elevado.
- Ele possui um limite de input de 512 tokens.
- Decidir qual é a melhor forma de realizar o pooling dos embeddings das 12 (ou 24 camadas) é uma questão a ser respondida em cada cenário (Devlin et al. 2019).

Para contornar esses pontos, pesquisadores geralmente:

- Utilizam o modelo somente na etapa de reordenação (Sanchez et al. 2020).
- Utilizam a sumarização de documentos (Anand Deshmukh e Sethi 2020).
- Transformam um documento em diversas observações de treinamando através do truncamento em frases de N tokens (MacAvaney et al. 2019).
- Realizam a ordenação dos documentos somente pela sentença mais próxima (W. Yang et al. 2019).

No domínio jurídico, Sanchez et al. (2020) apresentam uma avaliação de recuperação de informação mas com bases textuais de notícias. Em seu cenário, os especialistas do direito utilizam um sistema com consultas Booleanas para monitorar tópicos de interesse em sites de notícia. Utilizaram o BERT na fase de reordenação utilizando no resumo dos documentos. Comparado com outras técnicas de reordenação com *features* manuais, o modelo obteve *mean average precision* e revocação significativamente melhores.

É possível perceber uma variação de desempenho entre as diversas abordagens de recuperação textual e modelos, mesmo para um domínio específico. Os modelos léxicos tradicionais sempre apresentam resultados competitivos e os modelos semânticos parecem entregar somente ganhos marginais.A única constante entre os trabalhos são os textos escritos em língua inglesa. Esse cenário fortalece a importância da investigação em *corpus* diferentes, escritos em língua portuguesa e em um cenário prático. Diferente dos trabalhos descritos, não precisamos gerar consultas a partir de documentos longos. Utilizamos textos curtos que já descrevem exatamente a tese jurídica a ser pesquisada. Assumimos que os textos utilizados passam por uma criteriosa revisão ortográfica antes de sua publicação e por isso não utilizaremos modelos como o FastText. Por fim, para o modelo BERT não usamos a abordagem de predição de relevância ou reordenação, pela ausência de base histórica para ajuste. Geramos o vetor para cada texto a partir da extração dos *embeddings* diretamente das camadas.

Capítulo 3

Metodologia de Avaliação

Existem duas formas de modelar a avaliação sistemas de Recuperação da Informação (Büttcher et al. 2016; Sakai 2018): teste de laboratório e análise de logs de usuários. Devido a ausência de logs de usuário nos sistemas do STJ iremos focar na primeira forma de modelagem.

A base para os experimentos de laboratório é a metodologia Cranfield para avaliação de componentes de sistemas (Zhai e Massung 2016). Desenvolvida em 1960, é utilizada para avaliação empírica em diversas tarefas e não somente na Recuperação da Informação. A metodologia inicia pela construição de uma coleção de documentos de teste reutilizáveis e de métricas de mensuração de desempenho. Com esses componentes fixos, através da variação dos algoritmos e configurações de sistemas é possível estimar a diferença de performance. A coleção de documentos deve ser montada semelhante a uma coleção real do cenário de pesquisa. É necessário também definir um conjunto de consultas ou tópicos que simulam a necessidade de informação do usuário. Por fim, para cada necessidade de informação deve-se solicitar a rotulação da relevância pelos usuários que construíram e definiram as consultas ou tópicos.

A Figura 3.1 apresenta um cenário de uso da metodologia. O conjunto de consultas é representado por $Q_1, Q_2, \cdots, Q_{50}$. O *corpus* é representado pelos documentos $D_1, D_2, \cdots, D_{48}$. As rotulações ou julgamentos de relevância são representados no canto direito onde cada documento é marcado se é relevante (+) ou não relevante (−). Por exemplo, os documentos D_1 e D_2 são considerados relevantes em relação a consulta Q_1. D_3 não é considerado relevante para a mesma necessidade de informação.

Com o foco na consulta Q_1 podemos realizar uma comparação mais focada no usuário. Na figura, R_A e R_B representam os resultados do ranking do sistema A e B respectivamente. Levando em consideração os rótulos de relevância, decidir qual sistema tem um desempenho melhor deve ser uma comparação focada na tarefa do usuário. O sistema A pode ser considerado melhor porque retorna somente uma quantidade pequena de docu-

mentos, dos quais a maioria é relevante. Por outro lado, o sistema B pode ser considerado melhor porque retorna uma quantidade maior de documentos relevantes. Então é necessário definir na fase de metodologia e planejamento de estudo, qual será a tarefa do usuário de forma precisa.

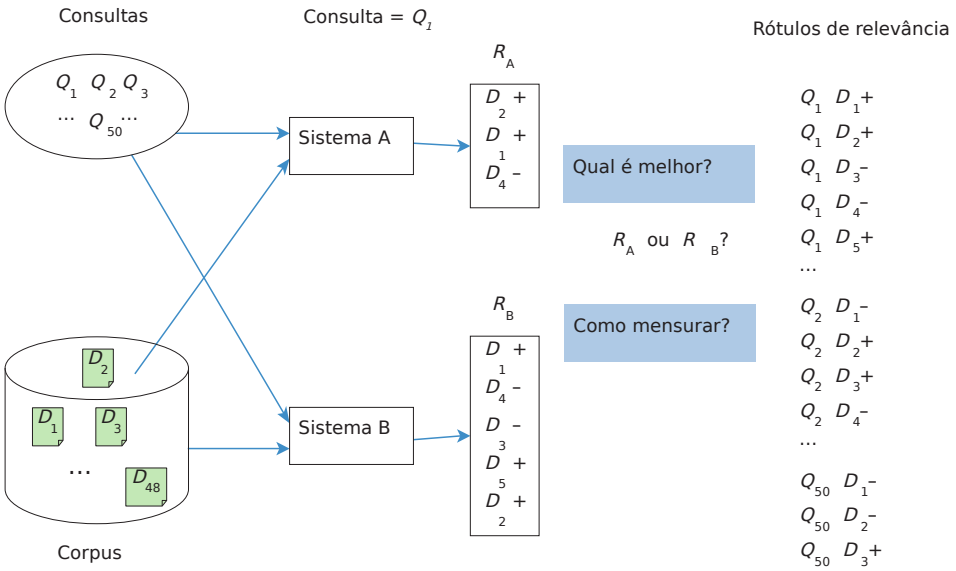


Figura 3.1: Representação da metodologia de avaliação Cranfield (Fonte: Zhai e Massung 2016).

Apesar de sua lógica simples e efetiva, a metodologia de Cranfield não é adequada para experimentos de laboratório para um grande volume de dados. Não é fáctivel a rotulação de bases grandes com milhares e até milhões de documentos. Para cenários mais atuais, a metodologia mais adequada é a utilizada na *Text Retrieval Conference* (TREC). Em 1990, a DARPA (*Defense Advanced Research Projects Agency* of US) financiou o NIST (*National Institute of Standards and Technology*) para construir uma grande coleção de testes a ser usada na avaliação de um projeto de pesquisa de texto, TIPSTER. Em 1991, o NIST propôs que esta coleção fosse disponibilizada para a comunidade de pesquisa em geral por meio de um programa chamado *Text Retrieval Conference*. O evento de avaliação anual começou em novembro de 1992. A partir desse ano, a conferência é realizada anualmente com o objetivo de:

- criar coleções de teste para um conjunto de tarefas de recuperação em domínios de conhecimento variados.

- promover os resultados das pesquisas de forma mais ampla possível.
- reunir pesquisadores para discussão e avaliação dos seus resultados.

Mais detalhes sobre a história da conferência são detalhadas no trabalho de Sanderson (2010). De forma objetiva, a metodologia TREC possui como componentes:

1. D: coleção de documentos ou *corpus* fixo que será compartilhado entre todos os pesquisadores.
2. Q: na conferência as consultas são agrupadas em Tópicos. Tópicos são descrições temáticas que detalham a necessidade de informação e o que seriam documentos relevantes para cada necessidade especifica.
3. J ou *qrels (Query Relevance Assessment)*: rótulos de relevância de documentos em relação as consultas e tópicos.
4. M: métrica de avaliação de desempenho.

Como as bases possuem um volume grande de documentos, o conjunto J não é definido previamente. Em uma etapa preliminar D e Q são disponibilizados para pesquisadores interessados em participar da competição. A primeira tarefa de cada equipe é gerar *runs*: resultado da aplicação de frameworks de modelagem (F) e funções de ranking ou similaridade (R) no *corpus* disponibilizado. Esse resultado é o *rank*, geralmente de 1.000, documentos com maior similaridade com a consulta ou tópico. Em posse das *runs*, a organização da conferenção utiliza a técnica de *pooling* para selecionar conjunto de documentos para que os rótulos de relevância sejam coletados.

A hipotése do *pooling* é se uma variedade grande de *runs* é utilizada para rotular documentos, o viés amostral tem menor influência na comparação entre sistemas (Lipani et al. 2016; Tonon et al. 2015). Diversas técnicas especificas de *pooling* são estudadas, mas a utilizada pela metodologia TREC é a *Depth@K*: a partir do conjunto de *runs* são extraídos os K melhores documentos do ranking de cada uma. A *pool* final é composta pela união da extração de cada *run*.

Por fim, as *pools* de documentos são distribuídas para especialistas avaliarem a relevância e construírem o conjunto J ou *qrels* (*query relevance pairs*). Essa é a fase com maior custo no laboratório de avaliação de RI. Büttcher et al. (2016) usa como exemplo a conferência TREC-8. A competição tem 86.830 documentos no *corpus*, 50 tópicos e utilizou $K = 100$. Foram submetidas 71 *runs*, caso todas tivessem sem seu *rank* 100 primeiros documentos diferentes seriam necessárias $71 \cdot 50 \cdot 100 = 355.000$ avaliações de relevância. Entretando, em todas as conferências sempre há uma grande quantidade de

sobreposição de documentos no topo do *rank* das diversas *runs*. Esse processo é ilustrado na Figura 3.2.

No Capítulo 1 apresentamos dois processos de trabalho apoiados pelo sistema legado na busca por precedentes com a mesma tese jurídica: categorização de decisões entre principais e sucessivos; e construção da base de Jurisprudência em Teses[1].

Como resultado da primeira atividade, somente as decisões principais são indexadas na ferramenta legada para pesquisa. Apesar de uma base histórica de relacionamentos entre decisões principais e sucessivas, essa base apresenta os seguintes impedimentos para sua utilização na avaliação de acordo com a metodologia:

- Muitas decisões possuem mais de uma tese. Não há marcação para saber qual tese foi utilizada para criar uma relação entre uma decisão principal e uma sucessiva. Portanto, não há como criar o conjunto de tópicos da metodologia TREC.
- Não há log histórico sobre qual foi a consulta utilizada pelos analistas no momento da classificação.
- A associação entre decisões é realizada somente dentro das decisões de um mesmo magistrado, o que diminui a variabilidade estilística dos textos e afeta a correta comparação entre modelos.

Por sua vez, a segunda atividade apresenta todos os elementos adequados para a metodologia e serão utilizadas da seguinte forma:

- As teses cadastradas no sistemas são os Tópicos.
- As consultas do sistema legado para cada tese são abertas publicamente. A qualidade das consultas é garantida pois foram construídas e validadas pelo processo de trabalho até que os resultados fossem satisfatórios.
- Cada tese possui uma descrição em texto livre elaborado pelo analista que construiu as consultas.
- Cada tese possui as decisões que foram usadas para originar a consulta e a descrição.

Com esses elementos, nossos experimentos vão simular dois cenários: (1) o usuário digita uma consulta em texto livre que descreve a tese buscada; e (2) o usuário seleciona parágrafos de uma decisão que descrevem a tese que deseja buscar. Dessa forma, é possível demonstrar a redução de esforço que os sistemas testados podem oferecer.

Por limitações do sistema legado, somente as ementas são indexadas na sua base. Então para apoiar de forma completa o processo de trabalho a tarefa de recuperação tem 3 níveis de relevância, explicados na Tabela 3.1.

[1] `https://scon.stj.jus.br/SCON/jt/`

Tabela 3.1: Níveis de relevância para a tarefa de recuperação de informação.

Valor	**Nível**	**Descrição**
0	Não relevante	Decisões em que o analista pode afirmar pela ementa que a tese jurídica não está presente na ementa ou no inteiro teor.
1	Mesmo tema	Decisões em que o analista identifica o mesmo tema na ementa, mas sem a mesma tese. Como abordam o mesmo tema, há probabilidade que a tese possa ser encontrada no inteiro teor da decisão.
2	Mesma tese	Decisões em que a mesma tese buscada está no texto da ementa.

Dessa forma, nossa metodologia se consolida nas etapas:

1. Extração das ementas de decisões do sistema legado de consulta textual.
2. Extração das teses, que servirão como tópicos da metodologia TREC, do sistema de Jurisprudência em teses.
3. Utilização da métrica nDCG@25, 25 para simular o cenário de trabalho real, no qual os analistas avaliam 25 ementas a cada vez que elaboram uma consulta.
4. Geração das *runs* para cada algoritmo.
5. Realização do *pooling* dos 25 documentos com melhor posição nas *runs*.
6. Submissão da *pool* para um especialista do domínio jurídico para avaliação de relevância para geração do arquivo de *qrel*.
7. Verificação com teste pareado estatístico de Wilcoxon para dar suporte aos resultados.

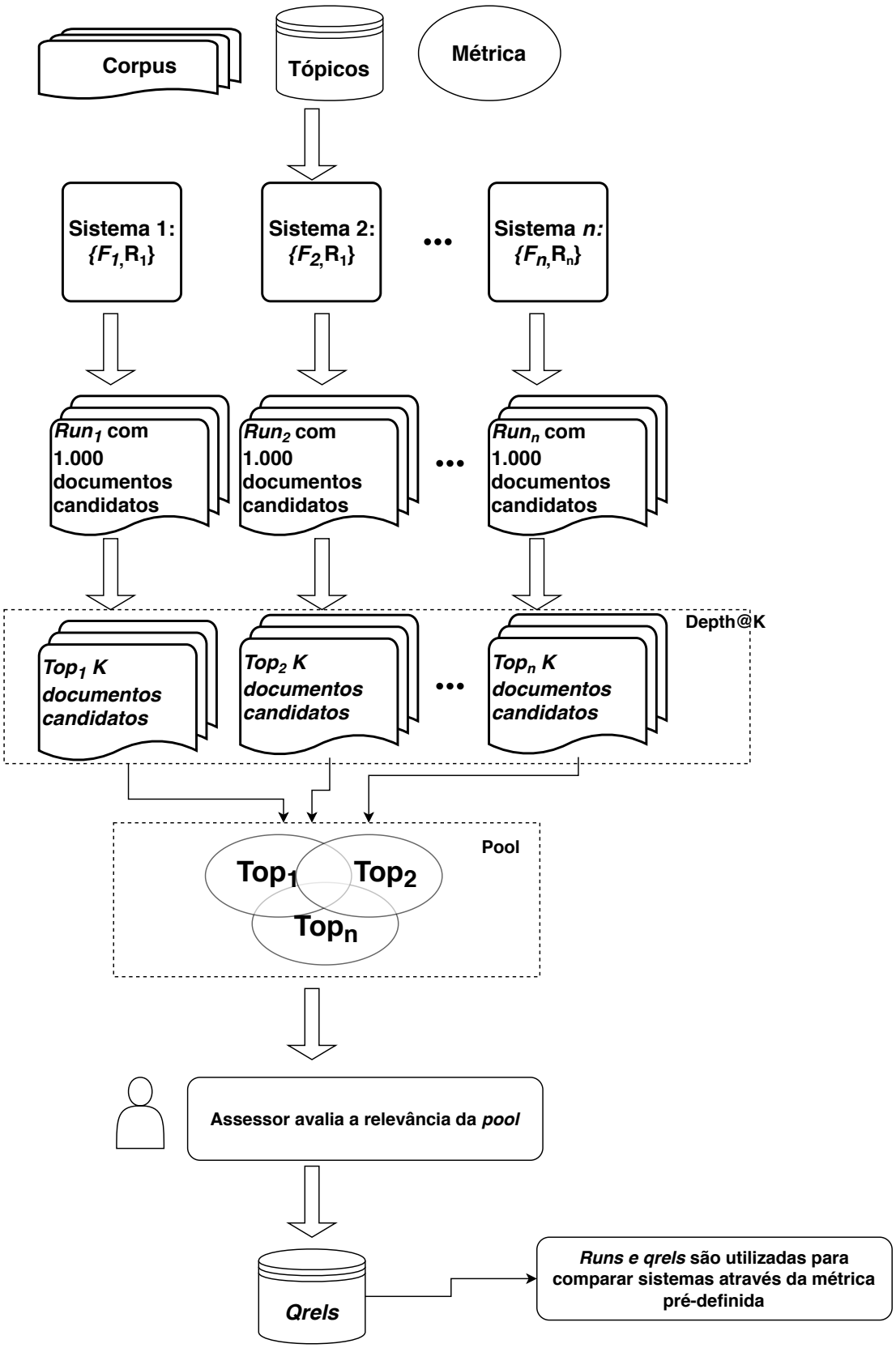


Figura 3.2: Fluxo da de trabalho da metodologia TREC.

Capítulo 4

Desenvolvimento da Pesquisa e Resultados

Neste Capítulo descrevemos a extração das teses jurídicas e os dados utilizados como consultas, em seguida o *corpus* textual utilizado como base de consulta. Assim como também explicamos os modelos utilizados, o seu treinamento e forma de vetorização dos textos utilizada para cada modelo. Por fim, descrevemos a execução do *Pooling* e a avaliação dos resultados para cada modelo.

4.1 Teses

As teses são agrupadas por ramo do direito e por temas no sistema. Para cada ramo, são selecionados temas e para cada tema são selecionadas várias teses para compor uma edição. A Figura 4.1 mostra as edições disponíveis para o Direito Administrativo. A edição 46 lista teses sobre o tema de desapropriamento e possui 14 teses. A Figura 4.2 mostra uma das teses catalogadas. Os acórdãos (decisões colegiadas do STJ) e decisões monocráticas (decisões emanadas por um único Ministro) listadas são a fonte de informação utilizadas pelo analista para construir a consulta e a descrição da tese. A descrição da tese é o texto destacado em azul. A Figura 4.3 ilustra um dos acórdãos utilizados e destaca o parágrafo que contém a mesma tese. Por fim a consulta pré-construída para recuperar decisões com a mesma tese:

> "desapropria$ com indireta com (interv$ ou participa$) com ("MP" ou parquet ou ministerial ou (ministério adj2 público)) não agrári$ não rescisória"

No total, foram extraídas 2005 teses do sistema[1]. Os seguintes filtros foram aplicados na limpeza dos dados extraídos:

[1]Extraídas no dia 23 de março de 2020.

Figura 4.1: Temas e edições das teses do ramo de Direito Administrativo.

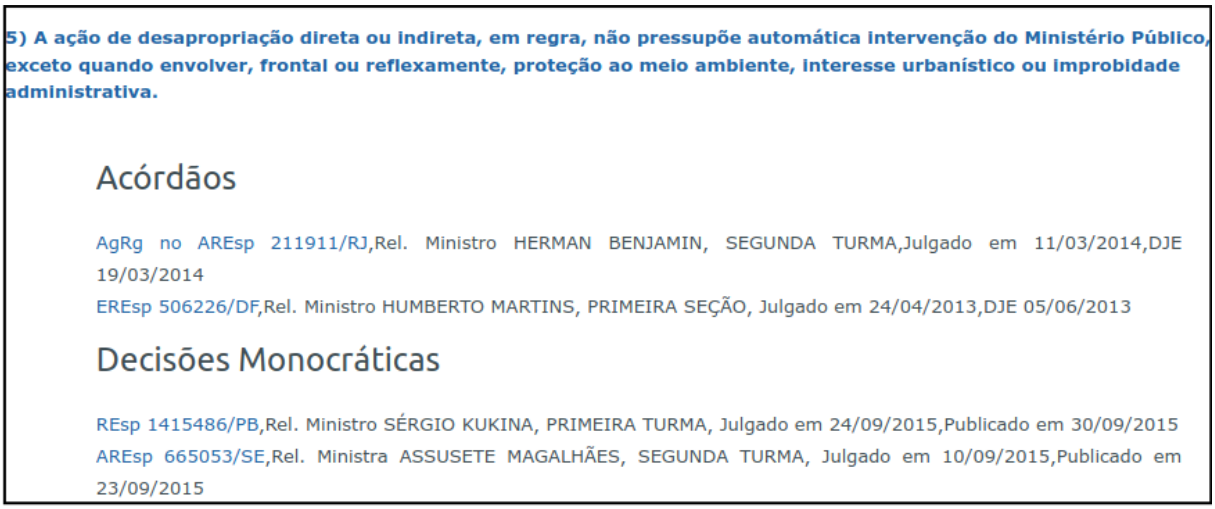
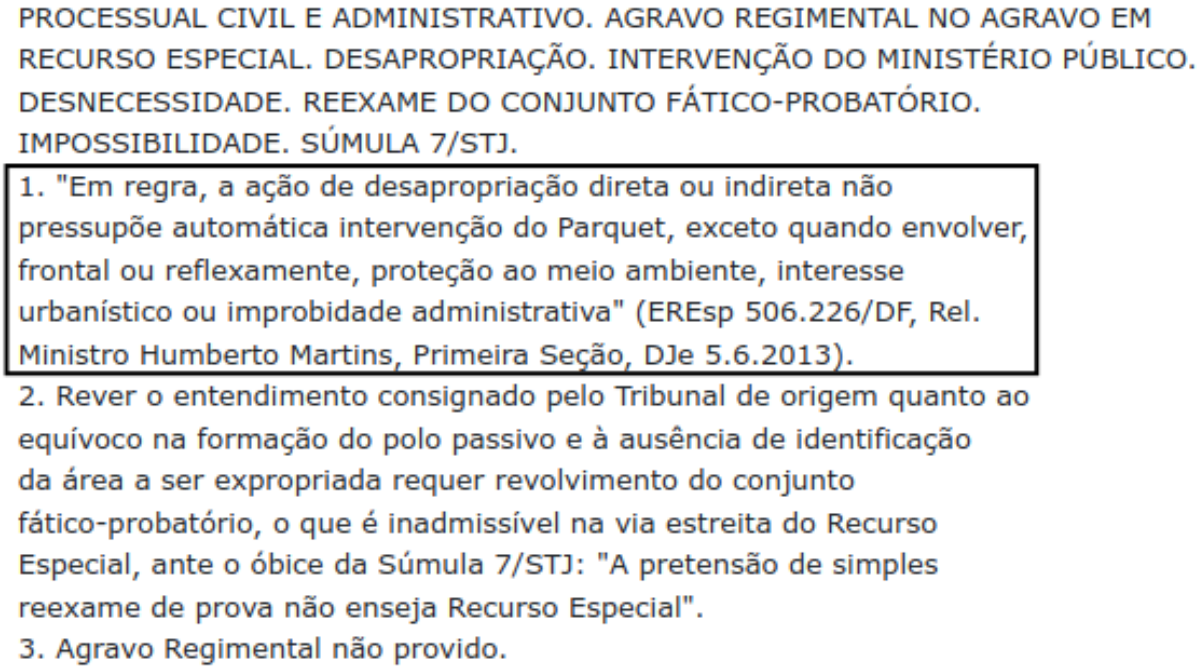


Figura 4.2: Exemplo de tese para a edição 46 do tema Desapropriação.


PROCESSUAL CIVIL E ADMINISTRATIVO. AGRAVO REGIMENTAL NO AGRAVO EM RECURSO ESPECIAL. DESAPROPRIAÇÃO. INTERVENÇÃO DO MINISTÉRIO PÚBLICO. DESNECESSIDADE. REEXAME DO CONJUNTO FÁTICO-PROBATÓRIO. IMPOSSIBILIDADE. SÚMULA 7/STJ.

1. "Em regra, a ação de desapropriação direta ou indireta não pressupõe automática intervenção do Parquet, exceto quando envolver, frontal ou reflexamente, proteção ao meio ambiente, interesse urbanístico ou improbidade administrativa" (EREsp 506.226/DF, Rel. Ministro Humberto Martins, Primeira Seção, DJe 5.6.2013).
2. Rever o entendimento consignado pelo Tribunal de origem quanto ao equívoco na formação do polo passivo e à ausência de identificação da área a ser expropriada requer revolvimento do conjunto fático-probatório, o que é inadmissível na via estreita do Recurso Especial, ante o óbice da Súmula 7/STJ: "A pretensão de simples reexame de prova não enseja Recurso Especial".
3. Agravo Regimental não provido.

Figura 4.3: Ementa do Acórdão (AgRg no AResp 211911/RJ) utilizado como fonte de informação para a tese sobre desapropriação.

1. 3 teses que não possuiam acórdãos como fonte de informação.
2. 67 teses removidas por não possuírem acórdãos no *corpus* utilizado.
3. 540 teses removidas que seus precedentes não estavam no *corpus* utilizado.
4. Das 1395 teses restantes, somente o acórdão mais recente dos listados foi utilizado.

4.2 *Corpus*

A nossa coleção D ou *corpus* de busca é composta por 160.714 ementas de decisões principais entre fevereiro de 2014 e abril 2019[2]. A estrutura de cada ementa é demonstrada na Figura 4.3. A estrutura básica é composta por:

1. Verbetação: primeiro parágrafo com termos todos em maísculo. Tem o objetivo de resumir em palavras chaves a decisão.
2. Parágrafos numerados: Todos os paragráfos, exceto o último, descrevem as teses da decisão. O último pargráfo é somente o resultado da decisão e não possui ne-

[2]Extração realizada em maio de 2019.

nhum valor semântico, esse resultado também é identificado por um indicador nos metadados.

4.3 Modelos

Nesta pesquisa foram utilizados os modelos TF-IDF e BM25 do VSM, os modelos Word2Vec pré-treinados em um *corpus* genérico e treinados em um *corpus* jurídico e o modelo BERT. Nesta Seção descrevemos os detalhes de treinamento e forma de utilização modelo.

4.3.1 TF-IDF e BM25

Os modelos *Term Frequency - Inverted Document Frequency* e *Best Match 25* são implementados de forma eficiente em diversos *frameworks* de indexação textual, como Lucene[3] (e as ferramentas que utilizam ele como núcleo: ElasticSearch[4] e Solr[5]), Indri[6] e Terrier[7].

O cálculo dos pesos é realizado em tempo de indexação, o que possibilita a comparação de documentos a partir de pesos gerados pelos próprios documentos. Nesse cenário, utilizamos o mesmo *corpus* de busca (160.714 ementas de decisões) para criar o vocabulário e calcular o peso dos termos.

Na etapa de pré-processamento para ambos modelos foram aplicadas as seguintes técnicas:

1. Remoção de frases entre parênteses.

2. Remoção de trechos utilizados somente para listar precedentes (*e.g.*, "1. Esta Corte possui entendimento de que o Sindicato possui legitimidade ativa para substituir os sucessores dos Servidores falecidos, independentemente do fato de o óbito ter ocorrido antes do ajuizamento da execução. **Precedentes**: REsp. 1.864.315/PE, Rel.Min. MAURO CAMPBELL MARQUES, DJe 25.6.2020 e AgInt no REsp. 1.578.639/RS, Rel. Min. NAPOLEÃO NUNES MAIA FILHO, DJe 19.11.2019.") ou reforçar que a tese da sentença anterior é oriunda de decisões anteriores (*e.g.*, "2. Acresça-se, outrossim, que é consolidado neste Superior Tribunal de Justiça a orientação segundo a qual o habeas *corpus*, porquanto vinculado à demonstração de plano de ilegalidade, não se presta à dilação probatória, exigindo prova pré-constituída das alegações, sendo ônus do impetrante trazê-la no momento da impetração, máxime quando se tratar de advogado constituído. **Precedentes**.").

[3] `https://lucene.apache.org`
[4] `https://www.elastic.co`
[5] `https://lucene.apache.org/solr`
[6] `http://www.lemurproject.org`
[7] `http://terrier.org`

3. Remoção de URLs (*Uniform Resource Locator*).

4. Transformação de termos de súmula em termos únicos. (*e.g.*, No trecho "2. A modificação do entendimento lançado no v. acórdão recorrido demandaria interpretação de cláusula contratual e revolvimento do suporte fático-probatório dos autos, providências inviáveis em sede de recurso especial, a teor do que dispõem as Súmulas 5 e 7 deste Pretório."o resultado da junção de termos seria "... que dispõem as sumula_5 sumula_7 deste Pretório ... ").

5. Remoção dos termos relacionados aos orgãos julgadores do Tribunal: turmas e seções.

6. Remoção de citações de Desembargadores convocados e Relatores.

7. Remoção de caracteres especiais de numeração ordinal.

8. Remoção de nome de Ministros.

9. Remoção de algarismos romanos.

10. Divisão do texto em termos (tokenização) por com expressão regular utilizando metacaracter de limites de palavras (\W+). São considerados caracters de palavras os alfanuméricos (a-z, A-Z, 0-9) e o caractere _ (sublinhado).

11. Transformação de todos os termos para minúsculo.

12. Remoção de símbolos de pontuação.

13. Remoção de acentos.

14. Remoção de termos números.

15. Remoção de *stopwords*.

16. Remoção de termos com menos de 2 caracteres.

17. Redução para o radical (*Stem*) com o algoritmo Removedor de Sufixos da Lingua Portuguesa (RSLP) (Orengo e Huyck 2001).

O pacote Gensim (Rehurek e Sojka 2010)[8] foi utilizado para o cálculo para o peso dos termos. Para o modelo TF-IDF os paramêtros padrões foram utilizados e executam a ponderação segundo a Equação 2.1 explicada no Capítulo 2. A mesma classe que calcula

[8]A versão utilizada foi a 3.8.0, em sua última versão publicada em 31/10/2020 diversas alterações foram realizadas nas classes de treinamento dos algoritmos.

a ponderação TF-IDF permite a alteração da função de ponderação local (tf) e da ponderação global (idf). Como explicado na fundamentação teórica, o BM25 somente altera a fórmula tf acrescentendo o parâmetro k (suavização da frequência do termo) e b (penalização para o tamanho do documento relativamente ao tamanho médio de documentos do *corpus*). Codificamos a função local com os valores recomendados pela literatura do BM25 com $b = 0,75$ e $k = 1,2$.

4.3.2 Word2Vec

Modelos semânticos preditivos possuem um custo computacional elevado e manter um treinamento contínuo a cada adição de documento no *corpus* é inviável. Spirling e Rodriguez (2019) explicam que comparar a utilização de modelos pré-treinados em *corpus* genéricos com a utilização de *corpus* treinados em textos do domínio da tarefa é uma questão essencial a ser tratada. Como modelo pré-treinado utilizamos os modelos Word2Vec da Universidade de São Paulo (USP) (Hartmann et al. 2017). O pré-processamento para esse modelo é descrito no trabalho dos autores e disponível publicamente no github[9]. Ao gerar os vetores por esses modelos foram seguidas as mesmas etapas dos autores para minizar a quantidade de termos fora do vocabulário. Utilizamos o SkipGram e CBoW de 300 dimensões. Esses serão referenciados como SKIP_NILC e CBOW_NILC.

Para o treino com documentos jurídicos foi disponibilizado pelo STJ um arquivo no formato XML (Extensible Markup Language) com todas as peças dos acórdãos entre os anos de 2009 e 2018. Um acórdão é constituído por três tipos de peças: (1) ementa, (2) Relatório e Voto e (3) Certidão de Julgamento. O primeiro já foi amplamente discutido anteriormente nesse trabalho, o segundo é onde o conteúdo completo da decisão se encontra e o último contém somente metadados mínimos para certificar a decisão.

Para treinamento realizamos o seguinte pré-processamento a partir dos arquivos disponibilizados (no total foram extraídas do arquivo 1.909.966 peças):

1. Seleção somente de peças do tipo (1) e (2).
2. Remoção de ementas do *corpus* de busca. Necessária para simular um modelo que não é constatemente treinado.
3. Utilização da biblioteca chardet[10] para detectar a codificação do texto. Foram encontradas 3 codificações distintas: WINDOWS-1252, ISO-8859-1 e UTF-8.
4. Extração do cabeçalho e da assinatura de cada documento.

[9] `https://github.com/nathanshartmann/portuguese_word_embeddings`.
[10] `https://github.com/chardet/chardet`.

5. Execução das etapas 12, 4, 9, 11, 17, 14, 16 e 13 descritas na Seção 4.3.1.

Para treinar os algoritmos foi utilizado o pacote Gensim. Dois modelos foram gerados, um SkipGram (SKIP_STJ) e outro CBoW (CBOW_STJ) com os seguintes paramêtros:

1. *alpha* = 0,025 – A taxa de aprendizagem (*learning rate*) inicial. Valor padrão da implementação original de Mikolov, Chen et al. (2013).
2. *window* = 5 – Quantidade de termos na janela deslizante.
3. *min_alpha* = 0,0001 – A taxa de aprendizagem cairá linearmente para min_alpha conforme o treinamento avança.
4. *hs* = 0 – Utilização do *Negative Sampling*.
5. *negative* = 5 – Quantas palavras utilizar como exemplos negativos.
6. *cbow_mean* = 1 – Realizar o pooling dos vetores dos termos no algoritmo CBoW utitlizando a média dos vetores de todos os termos de entrada.
7. *vector_size* = 300 – Tamanho do vetor de embedding.
8. *epochs* = 5 – Valor padrão da biblioteca.

Para gerar um vetor único para um documento utilizamos três abordagens:

1. Média dos vetores de todos os termos.
2. Média ponderada pelo *idf* de cada termo.
3. Média ponderada pelo BM25 de cada termo.

4.3.3 BERT

O modelo BERT possui um elevado custo de treinamento (Sharir et al. 2020). Como resultado somente é viável a utilização de modelos pré-treinados. Neste trabalho abordamos 3 estratégias:

1. Explorar diferentes formas de extrair embeddings do modelo BERT treinado em *corpus* da língua portuguesa (Souza et al. 2020).
2. Realizar *fine-tuning* em uma tarefa de classificação da descrição de teses em ramo do direito.
3. Utilizar diretamente um modelo multi-línguas com *fine-tuning* para tarefa de Inferência com uma rede siamesa (Reimers e Gurevych 2020).

A primeira estratégia é baseada na discussão do paper original do modelo (Devlin et al. 2019). Nele os autores discutem que os vetores do modelo podem ser usados como *features* em outras tarefas e que a única forma de encontrar a melhor forma de extrair os vetores de cada camada depende de cada tarefa e cada conjunto de dados. A Figura 4.4 ilustra os resultados alcançados pelos autores. O modelo utilizado foi o menor com 12 camadas. As formas de *pooling* testadas pelos autores foram: primeira camada (ou camada de embedding) de entrada do modelo, última camada, soma ponderada das 12 camadas, penúltima camada, soma ponderada dos 4 últimos layers e a concatenação dos 4 últimos layers. A última abordagem obteve melhor resultado.

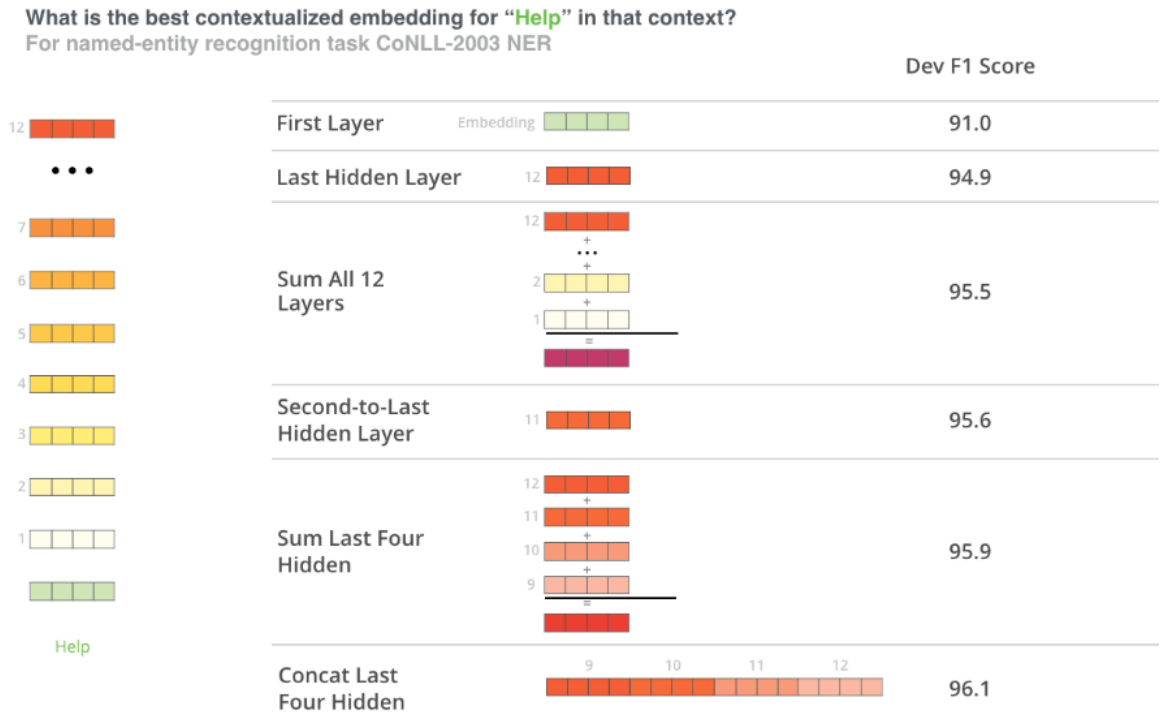


Figura 4.4: Avaliação do impacto da extração de *features* na tarefa de reconhecimento de entidade nos dados do CoNLL-2003 (Fonte: Alammar 2018).

Para a primeira estratégia utilizamos as biblioteca *transformers* (Wolf et al. 2020) com o BERT para português com 12 camadas e exploramos:

1. Extração de embeddings dos termos da primeira camada, com e sem o termo [CLS].
2. Extração de embeddings dos termos da última camada, com e sem o termo [CLS].
3. Extração de embeddings dos termos da penúltima camada, com e sem o termo [CLS].

4. Extração padrão do pacote *transformers*. Realiza a extração do embedding do termo [CLS] e aplica a função de ativação Tahn após uma camada linear totalmente conectada (*fully connected*).

O tamanho máximo para as sentenças após o pré-processamento foi fixado em 512 e os vetores resultantes tem dimensão de 768 posições. Não foram utilizadas as somas ponderadas por não ter o detalhamento de como é realizada a ponderação no artigo.

Para recuperação da informação, o ajuste do modelo para tarefas do domínio é realizado com bases históricas que possuem exemplos de treinamento de triplas: consulta, documento, grau de similaridade. Como a base para avaliação está sendo construída do zero, não há bases disponíveis para testar essa abordagem. Utilizando o BERT da língua portuguesa, realizamos um *fine-tuning* em uma tarefa de classificação com os dados de descrições das teses não utilizadas como consultas, sem a expansão de vocabulário. A tarefa é de classificação multiclasse dos trechos em seu respectivo ramo do direito. Foram utilizadas 1972 teses com divisão de 90% para treino e 10% para teste. A arquitetura padrão do pacote *transformers* foi utilizada, ela realiza o *pooling* padrão com uma camada de *Dropout* (com probabilidade 0,1) e uma camada totalmente conectada. Por padrão todas as 12 camadas do modelo Bert são ajustadas. Para ajuste dos pesos foi utilizado o AdamW com *learning rate* 0,00002 e *adam epsilon* 0,00000001. Com 4 epochs o modelo atingiu no conjunto de validação acurácia de 0,87. A partir desse modelo ajustados os vetores foram gerados utilizando o *pooling* padrão.

Por fim, na terceira estratégia utilizamos o pacote *sentence-transformers* (Reimers e Gurevych 2020) com o modelo BERT multi-línguas ajustado para a tarefa de *Short Sentence Similarity* (STS). Utilizamos o modelo *distiluse-base-multilingual-cased*.

4.4 *Pooling*

As Tabelas 4.1 a 4.5 mostram quais consultas foram usadas e qual granularidade das ementas foi utilizada por cada tipo de modelo. Primeiro, no sistema legado (Tabela 4.1) foi utilizada somente com a consulta booleana e só permite a comparação pelo texto integral da ementa. Em segundo (Tabela 4.2), os modelos tradicionais são utilizados com a combinação dos dois tipos de consulta com os dois níveis de granularidade, o que resulta em 8 modelos.

Os modelos Word2Vec (Tabela 4.3) adicionam mais uma dimensão. Ao gerar os vetores para um texto, foi utilizada a média simples dos vetores do texto, a média ponderada pelo *idf* e pelo BM25. Ao ponderar os vetores de cada termo é possível aproximar mais o vetor do documento para os termos de maior relevância. A combinação dos elementos da tabela resulta em 48 modelos.

Tabela 4.1: Tabela de pooling para o sistema legado.

Modelo	**Consulta**	**Granularidade *Corpus***
Legado	Consulta Booleana	Toda a ementa

Tabela 4.2: Tabela de pooling para modelos tradicionais.

Modelo	**Consulta**	**Granularidade *Corpus***
TF-IDF	Descrição da Tese	Toda a ementa
BM25	Parágrafo Selecionado	Por Parágrafo

Em seguida, Tabela 4.4 as 6 possibilidades de extração dos vetores do modelo BERT. Para a granularidade do *corpus*, somente é utilizada a recuperação por parágrafo pela restrição de tamanho de 512 tokens. A combinação resulta em 12 modelos. Por fim, a Tabela 4.5 mostra que o BERT multi-línguas e ajustado adicionam mais 4 modelos. No total são 73 modelos diferentes comparados e utilizados para formação do *pooling* de documentos.

Na recuperação por parágrafo, cada documento recebe o rank igual ao rank do seu parágrafo de melhor rank.

Para suporte a gestão da atividade de rotulação de relevância foi construído um sistema web. Através do sistema o especialista do domínio consegue visualizar todas as teses (Figura 4.5), selecionar os parágrafos de uma ementa que contém a tese (Figura 4.6), acessar painel de gerenciamento de *pools* (Figuras 4.7 a 4.8) e rotular a relevância dos documentos na *pool* de cada tese (Figura 4.9). Nessa última, o especialista não tem visualização de quais modelos contribuíram para adicionar aquela ementa na *pool* e nem do *rank* de cada documento. Esse é um requisito da metodologia TREC.

Para validação da metodologia e resposta das questões de pesquisa foram selecionadas 11 teses aleatoriamente, uma para cada ramo do direito. Através da similaridade do cosseno, cada um dos 73 sistemas gerou uma *run* com 1000 melhores documentos. Para selecionar o *pool* de documentos para avaliação foi utilizado o *Depth@k* com $k = 25$. Esse valor é derivado da quantidade de ementas que os analistas analisam antes de encerrar uma consulta ou refazê-la.

Tabela 4.3: Tabela de pooling para modelos Word2Vec.

Modelo	**Ponderação**	**Consulta**	**Granularidade *Corpus***
SKIP_NILC	TF-IDF	Descrição da Tese	Toda a ementa
CBOW_NILC	BM25	Parágrafo Selecionado	Por parágrafo
SKIP_STJ	Média		
CBOW_STJ			

Tabela 4.4: Tabela de pooling para modelos BERT português.

Modelo	**Método de Extração**	**Consulta**	**Granularidade *Corpus***
Bert português	Primeira Camada com CLS	Descrição da Tese	Por Parágrafo
	Primeira Camada sem CLS	Parágrafo Selecionado	
	Penúltima Camada com CLS		
	Penúltima Camada sem CLS		
	Última Camada com CLS		
	Última Camada sem CLS		

Tabela 4.5: Tabela de pooling para modelos BERT ajustado e multi-línguas.

Modelo	**Método de Extração**	**Consulta**	**Granularidade *Corpus***
Bert multi-línguas	CLS com ativação Tanh	Descrição da Tese	Por Parágrafo
Bert ajustado		Parágrafo Selecionado	

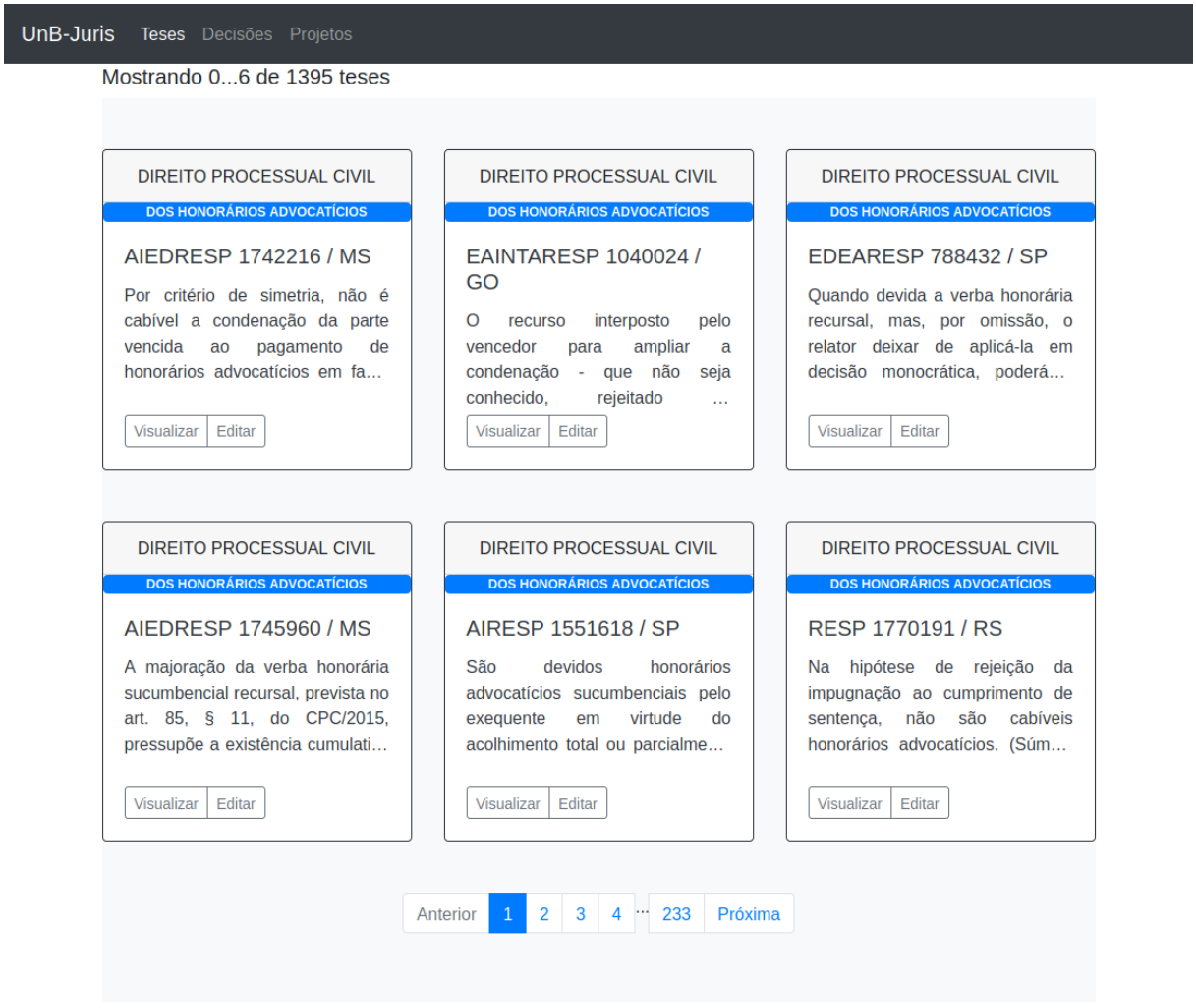


Figura 4.5: Tela do sistema para acesso as teses.

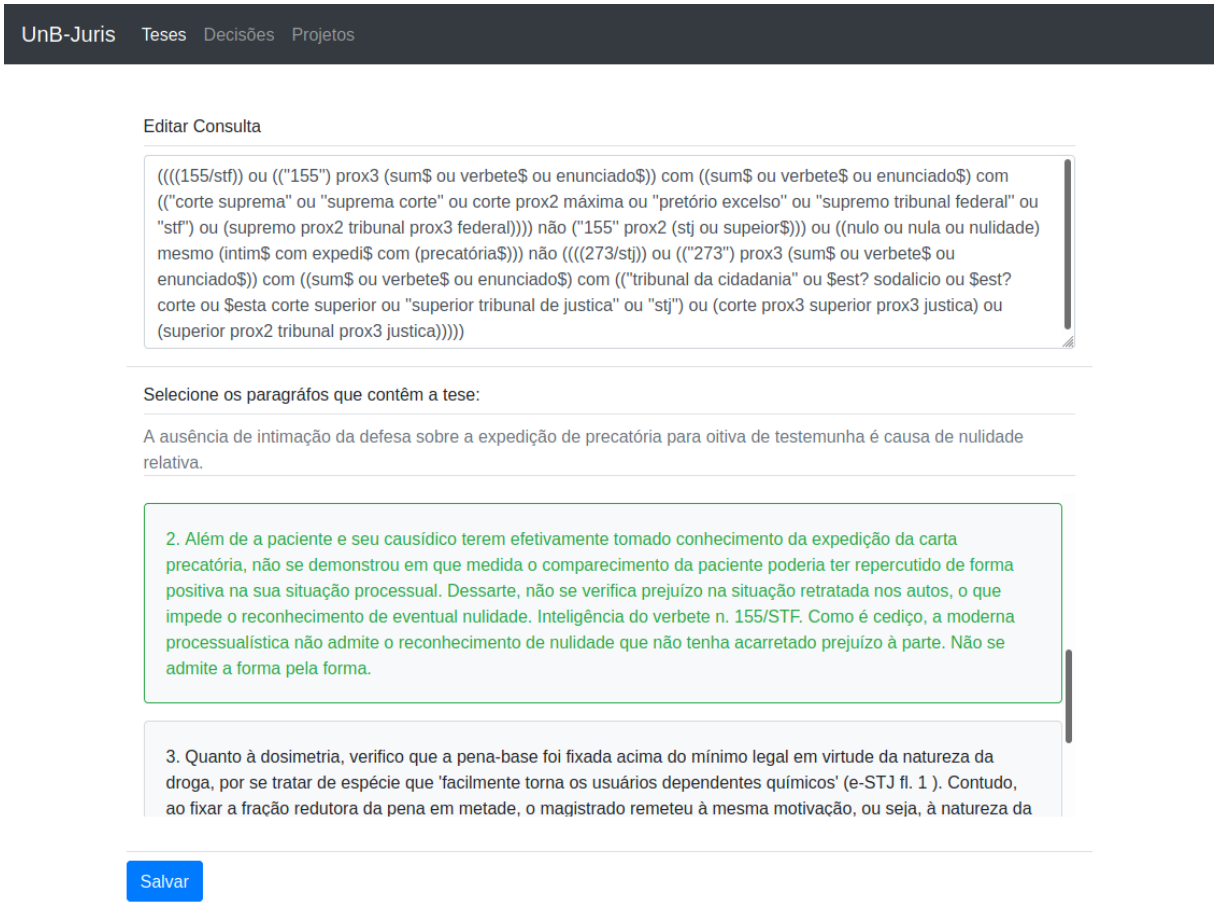


Figura 4.6: Tela do sistema de edição de teses. Junto com a consulta do sistema legado, os parágrafos da ementa são exibidos e o especialista pode selecionar o(s) que possue(m) a tese.

Como o sistema legado rodou na base de produção, foram eliminadas das *runs* respectivas as decisões que não estavam contidas dentro do nosso *corpus*. A Tabela 4.6 mostra que houve uma variação no quantitativo de decisões na pool em cada ramo do direito. Quanto menor o número, mais os *ranks* dos diversos sistemas foram parecidos. As tabelas do Apêndice A mostram a descrição das teses escritas pelos analistas, a consulta construída e os parágrafos da ementa mais recente utilizada para extração da consulta. Percebe-se que há uma variância estilística alta e que em determinados casos, os parágrafos selecionados não estão em sequência, como no caso da tese de 1374.

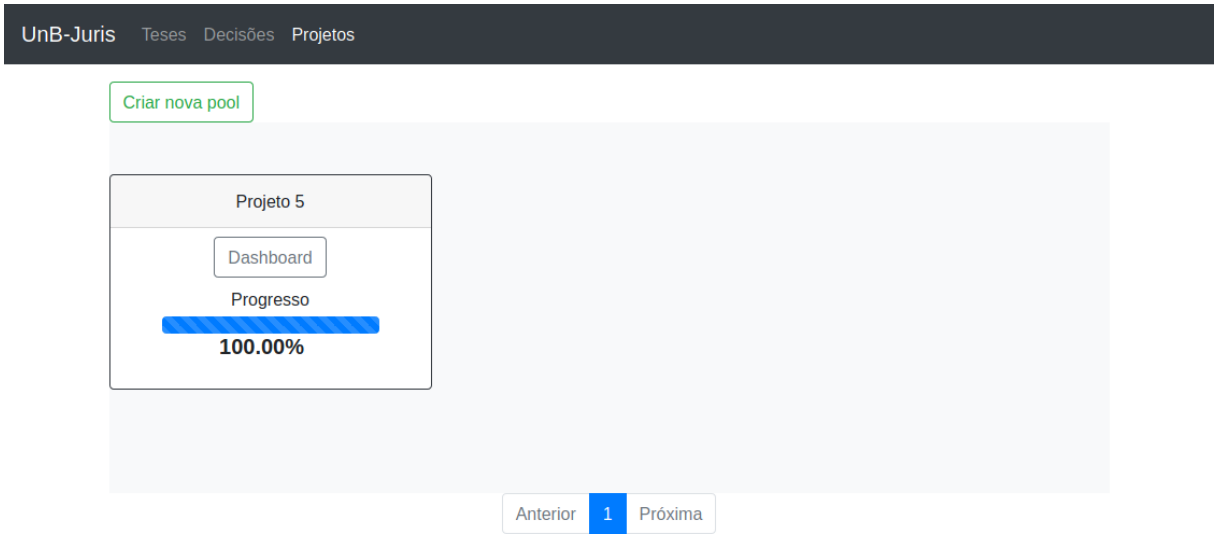


Figura 4.7: Painel de gestão de *pools* para rotulação de relevância.

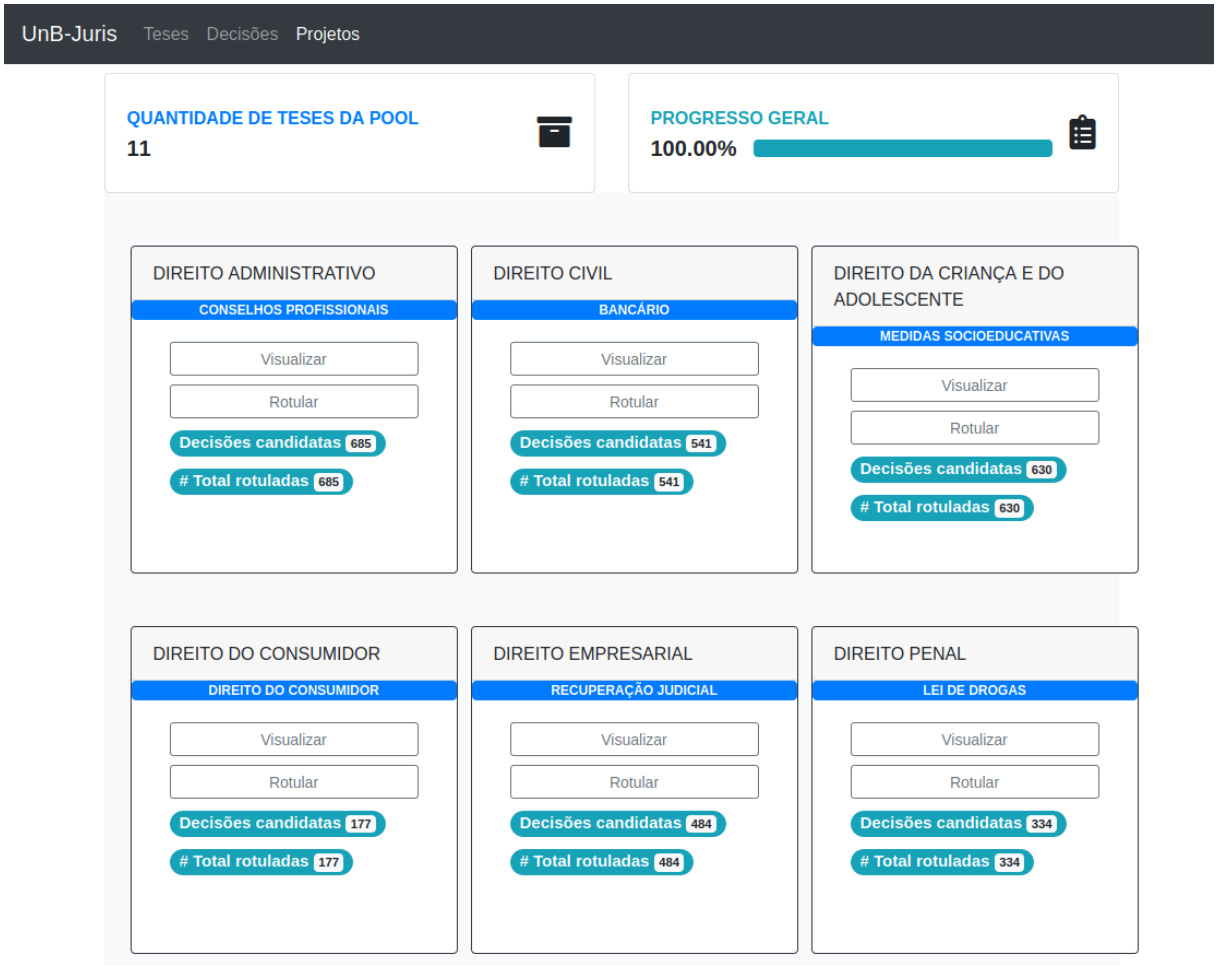


Figura 4.8: Painel de gestão de uma *pool* específica. Mostra o progresso geral, o quantitativo de documento por tese e o progresso por tese.

Figura 4.9: Tela de rotulação de relevância.

Tabela 4.6: Quantidade de documentos por tese selecionada.

ID	Ramo do Direito	Tema	#Decisões na *pool*
1374	DIREITO DO CONSUMIDOR	JUROS	177
1270	DIREITO PREVIDENCIÁRIO	PREVIDÊNCIA COMPLEMENTAR	303
560	DIREITO PENAL	LEI DE DROGAS	334
955	DIREITO TRIBUTÁRIO	IMPOSTOS MUNICIPAIS	335
1423	DIREITO REGISTRAL	REGISTROS PÚBLICOS	397
635	DIREITO EMPRESARIAL	RECUPERAÇÃO JUDICIAL	484
1474	DIREITO CIVIL	BANCÁRIO	541
1238	DIREITO PROCESSUAL PENAL	NULIDADES NO PROCESSO PENAL	590
914	DIREITO DA CRIANÇA E DO ADOLESCENTE	MEDIDAS SOCIOEDUCATIVAS	630
120	DIREITO ADMINISTRATIVO	CONSELHOS PROFISSIONAIS	685
1995	DIREITO PROCESSUAL CIVIL	DOS HONORÁRIOS ADVOCATÍCIOS	776
Total			5252

4.5 Avaliação

Para gerar a métrica nDCG foi utilizada a biblioteca oficial das conferências TREC[11]. Para computar os valores a biblioteca recebe dois arquivos: *qrels* e *runs*. O primeiro é um arquivo único e contém todas os pares de avaliação de relavância entre a tese e os documentos das *pools* avaliadas. O segundo é um conjunto de arquivos, um para cada sistema ou framework de modelagem e possui para cada tópico o *rank* de documentos.

O primeiro modelo a ser analisado é o sistema legado com as consultas elaboradas pelos analistas da jurisprudência (Tabelas A.1 a A.11). Utilizamos a métrica nDCG@25 pelo mesma motivação da escolha de k no método de *pooling*, simular o uso do sistema por usuários que analisam até 25 ementas após realizar uma pesquisa. As consultas elaboradas pelo analista são muito detalhadas e específicas, os seus valores de ganhos cumulativos são baixo por reduzirem a quantidade de decisões relevantes retornadas. A Tabela 4.7 mostra que através das consultas especializadas, o sistema alcança um nDCG@25 médio de 0,6523 com um intervalo de confiança de 0,14, em um nível de confiança de 95%.

Para capturar melhor o coportamento dos outros modelos escolhemos variar as fontes das consultas (descrição da tese ou parágrafos da ementa selecionados) e variar a granularidade de recuperação para o *ranking*: comparar o vetor da consulta com o vetor da ementa inteira ou com o vetor de cada parágrafo da ementa e considerar somente o parágrafo mais similar. O primeiro iremos nomear de similaridade comum (sim_comum) e o segundo iremos nomear de *dismax*, em referência ao nome desse tipo de abordagem em ferramentas de recuperação de informação. Essa abordagem nos fornece 4 combinações: (1) descrição da tese e sim_comum, (2) descrição da tese e *dismax*, (3) parágrafo selecionado e sim_comum e (4) parágrafo selecionado e *dismax*.

Para avaliação dos modelos tradicionais (TF-IDF e BM25) são obtidos 4 resultados para cada um. A Tabela 4.8 mostra em negrito estão destacados a melhor perfomance por abordagem (linhas). O valor em itálico mostra a média geral de todas os resultados. A maior proporção de melhores resultados está na abordagem da consulta gerada a partir dos parágrafos selecionados. Mas através da média da última linha, percebemos que não há tanta diferença entre as abordagens.

Pela diferença absoluta, o modelo TF-IDF apresenta uma melhora de 13.3% em relação ao *baseline*. Ao utilizar o teste pareado de Wilcoxon temos as seguintes hipóteses:

- H_0: A diferença mediana é zero.
- H_A: A diferença mediana não é zero com $\alpha = 0{,}05$

[11] Disponível em `https://github.com/usnistgov/trec_eval`

Tabela 4.7: Valores ordenados do nDCG@25 para o sistema legado por tese com a média.

Tópico	nDCG@25
1374	0,9806
560	0,9270
1474	0,7797
914	0,7566
1238	0,7503
1423	0,7147
120	0,6754
635	0,5884
1270	0,4269
955	0,3756
1995	0,2006
Média	0,6523 ± 0,14

Ao executar o teste obtemos um p-valor = 0,18352. Um p-valor superior a 0,05 não é estatisticamente significativo e indica forte evidência para a hipótese nula. Então não há suporte para afirmar que realmente o modelo TF-IDF obteve melhor desempenho.

O BM25 obteve resultados muito próximos do TF-IDF. Apresenta uma melhora de 14.9% em relação a média do *baseline*. Ao executarmos o teste obtemos um p-valor = 0,13104, o que demonstra o mesmo resultado na comparação entre o sistema legado e o TF-IDF. Portanto, apesar na melhora absoluta de valores da métrica, os dados não são suficientes para confirmar essa diferença. A Figura 4.10 ilustra os resultados por abordagem de consulta e similaridade. É possível perceber que não há diferença entre os modelos tradicionais.

Quanto aos modelos Word2Vec, a combinação de 2 arquiteturas (CBoW e Skip-Gram), com 2 *corpus* de treinamento (USP/NILC e STJ), com 3 estratégias para gerar o vetor

Tabela 4.8: nDCG@25 por tópico para o TF-IDF.

Modelo	Tópico	Descrição da tese		Parágrafo		Média
		sim_com (1)	dismax (2)	sim_com (3)	dis_max (4)	
TF-IDF	1474	0,4888	**0,6921**	0,1416	0,3657	0,42205
TF-IDF	1270	0,4959	0,2984	**0,9578**	0,8293	0,6453
TF-IDF	1995	0,5681	0,5668	**0,7576**	0,6330	0,6314
TF-IDF	914	0,6585	0,7106	**1**	**1**	0,8422
TF-IDF	955	0,7643	0,7147	0,7820	**0,8108**	0,7679
TF-IDF	1423	0,7764	0,7680	**0,7952**	0,7687	0,7770
TF-IDF	1238	**0,7803**	0,7222	0,5123	0,3114	0,5815
TF-IDF	120	**0,8293**	0,8260	**0,8293**	**0,8293**	0,8285
TF-IDF	560	0,8815	0,9452	0,8296	**0,9442**	0,9001
TF-IDF	635	0,8916	**0,9231**	0,6419	0,4790	0,7339
TF-IDF	1374	**1**	**1**	**1**	**1**	1
Média		0,7395 ± 0,0991	0,7425 ± 0,115	**0,7498** ± 0,147	0,7247 ± 0,15	*0,7391* ± 0,0954

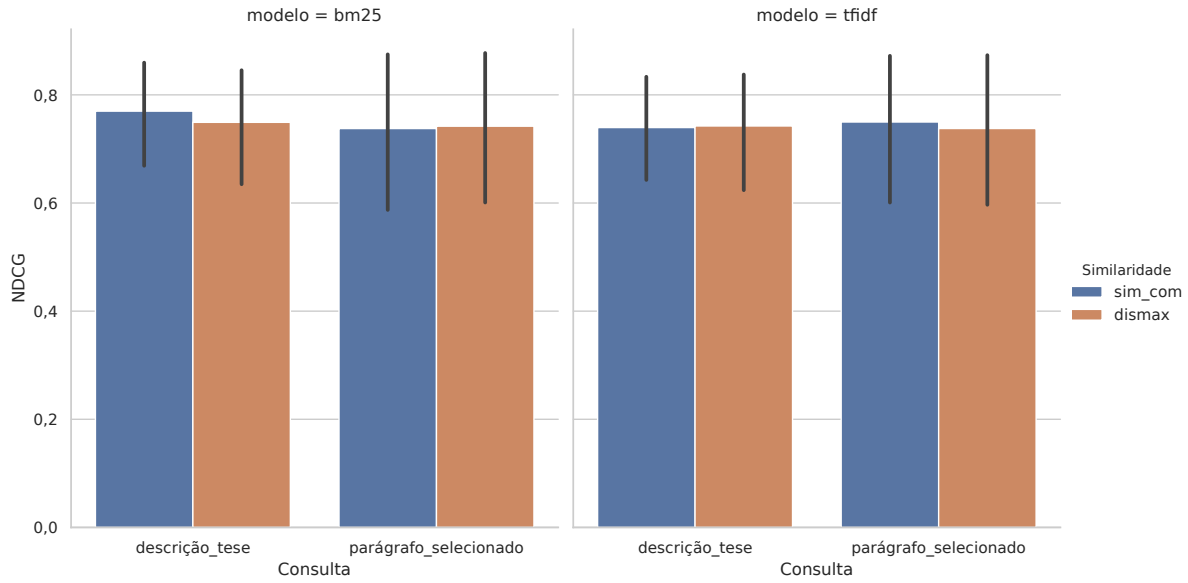


Figura 4.10: Gráfico de barra com intervalo de confiança da média do nDCG@25 para os modelos TF-IDF e BM25.

médio dos textos (média, média ponderada pelo *idf* e média ponderada pelo BM25) e com 2 estratégias de similaridade, resultou em 48 modelos testados. No total são 528 métricas se apresentarmos os resultados por tópico como nas tabelas dos modelos tradicionais. Por essa razão, a Tabela 4.10 apresenta os resultados nDCG@25 médios dos modelos.

Percebe-se pelos valores em negrito, os quais indicam qual foi a melhor desepenho dentre as técnicas de similaridade do mesmo modelo, que a utilização da descrição da tese e a similaridade por parágrafo está em maior quantidade. Uma indicação que o modelo semântico Word2Vec pode ser mais indicado para textos curtos. Os valores médios

Tabela 4.9: nDCG@25 por tópico para o BM25.

Modelo	Tópico	Descrição da tese		Parágrafo		Média
		sim_com (1)	dismax (2)	sim_com (3)	dis_max (4)	
BM25	1995	0,4352	0,5780	0,5824	**0,6336**	0,5573
BM25	1270	0,5778	0,2961	0,9294	**0,9735**	0,6942
BM25	1474	0,6644	**0,7796**	0,2541	0,3187	0,5042
BM25	1423	0,6888	**0,7672**	0,7230	0,7511	0,7325
BM25	955	0,7442	0,7378	**0,8791**	0,8128	0,7935
BM25	1238	**0,7486**	0,6695	0,4507	0,3843	0,5633
BM25	120	**0,8293**	0,8334	0,8334	0,8334	0,8324
BM25	635	**0,8984**	0,8436	0,4958	0,4859	0,6809
BM25	914	0,9086	0,8595	**1**	**1**	0,9420
BM25	560	**0,9735**	0,8767	0,9670	0,9693	0,9466
BM25	1374	**1**	**1**	**1**	**1**	1
Média		**0,7699** ± 0,103	0,7492 ± 0,111	0,7377 ± 0,151	0,7420 ± 0,149	*0,7497* ± 0,1

sublinhados indicam os melhores modelos na comparação da utilização de ponderação ou não. Todos os melhores desempenhos são verificados com a utilização da ponderação. O valor em itálico da coluna de média mostra que o modelo SkipGram treinado no *corpus* jurídico possui o melhor desempenho entre todos os modelos Word2Vec.

Utilizamos o Word2Vec com melhor desempenho para realizar o teste estatístico e comparar com os modelos tradicionais. Ao realizar o teste para comparar o BM25 e o SkipGram_STJ ponderado com BM25, obtemos o p-valor = 0,50926. Para o TF-IDF temos p-valor = 0,64552. Então não há evidências suficientes para rejeitar a hipótese nula de que a mediana da diferença de desempenho é igual a zero.

Para avaliar se a utilização de um modelo treinado em um *corpus* do domínio jurídico e um pre-treinando em *corpus* de domínio aberto, realizamos o teste para os melhores modelos de cada abordage:cbow_nilc_2 com cbow_stj_2 e skip_nilc_3 com skip_stj_3. Em ambos os casos obtemos o p-valor = 0,3843.

Em seguida avaliamos se os ganhos advindos da ponderação dos vetores são estatisticamente significantes. Para o o CBoW_NILC a ponderação resoltou em um incremento de 6,6%. Para o SkipGram_NICL resultou em 3% de melhora. Para o CBoW_STJ e SkipGram_STJ representou um ganho de 20% e 8%, respectivamente. As Figuras 4.11 a 4.14 ilustram de forma mais clara a diferença e utiliza o intervalo de confiança para mostrar a variância em cada abordagem. Apesar dos ganhos é possível visualizar que os intervalos de confiança se sobrepõem.

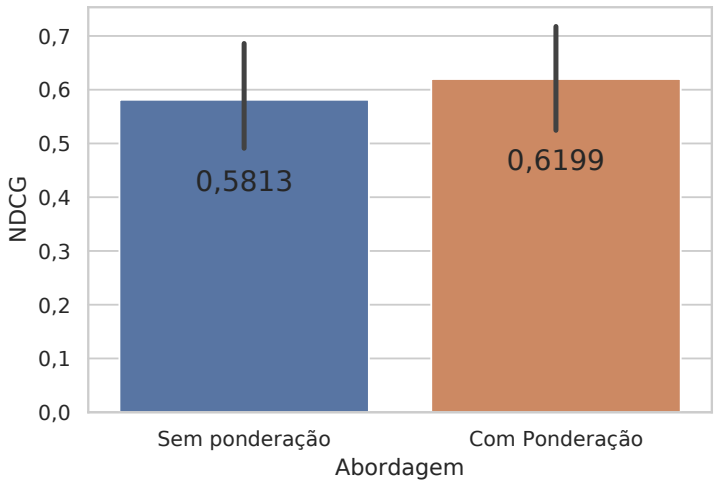


Figura 4.11: Comparação entre CBoW_NILC com e sem ponderação.

A Tabela 4.11 apresenta os resultados para os modelos BERT. É possível visualizar que não há quase nenhum impacto, negativo ou positivo, na utilização do vetor do termo

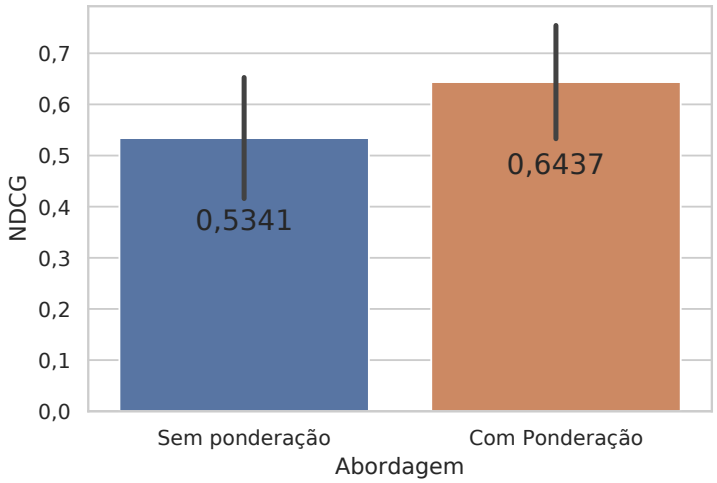


Figura 4.12: Comparação entre CBoW_STJ com e sem ponderação.

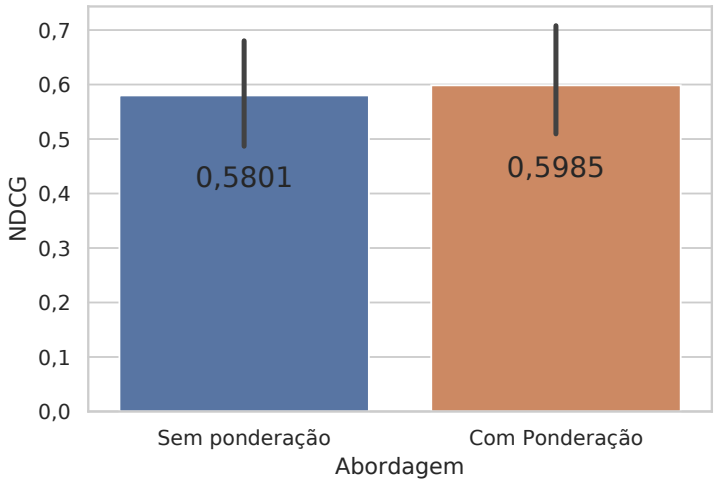


Figura 4.13: Comparação entre SkipGram_NILC com e sem ponderação.

[CLS]. Para alguns casos como o u_cls e u_scls o resultado é o mesmo. O modelo ajustado obteve pior desempenho. E o modelo multi-línguas, obteve o melhor resultado médio.

Ao realizar o teste de Wilcoxon entre o modelo multi-línguas e o skip_stj_3 (Word2Vec com arquitetura SkipGram treinando no *corpus* jurídico com melhor desempenho em valores absolutos) obtemos p-valor = 0,0164. Então há evidências suficientes para rejetiar a H_0 e sustentar a hipótese altenartiva de que a diferença entre o desempenho do Word2Vec e BERT é significativa.

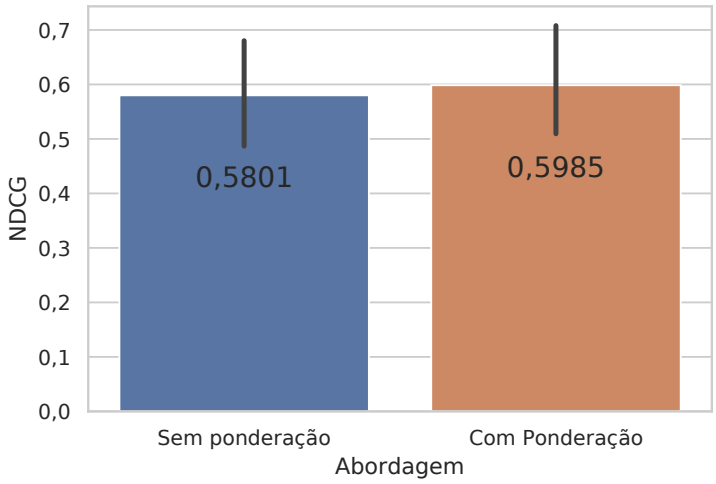


Figura 4.14: Comparação entre SkipGram_STJ com e sem ponderação.

Tabela 4.10: nDCG@25 por tópico para os modelos Word2Vec.

Identificador	Arquitetura	*Corpus*	Ponderação	Descrição da tese		Parágrafo		Média
				sim_com (1)	dismax (2)	sim_com(3)	dis_max(4)	
cbow_nilc_1	CBOW	NILC	média	0,5398	**0,6607**	0,5601	0,5649	0,5814 ± 0,0529
cbow_nilc_2	CBOW	NILC	idf	0,6006	**0,6559**	0,6069	0,6163	0,6199 ± 0,0243
cbow_nilc_3	CBOW	NILC	bm25	0,5694	**0,6585**	0,6110	0,6047	0,6109 ± 0,0359
skip_nilc_1	SkipGram	NILC	média	0,5686	**0,6874**	0,5178	0,5466	0,5801 ± 0,073
skip_nilc_2	SkipGram	NILC	idf	0,5610	**0,6380**	0,5592	0,6098	0,5920 ± 0,0378
skip_nilc_3	SkipGram	NILC	bm25	0,5602	**0,6624**	0,5556	0,6159	0,5985 ± 0,0496
cbow_stj_1	CBOW	STJ	média	0,4507	**0,6343**	0,4558	0,5956	0,5341 ± 0,0928
cbow_stj_2	CBOW	STJ	idf	0,6182	0,6377	0,6538	**0,6652**	0,6437 ± 0,02
cbow_stj_3	CBOW	STJ	bm25	0,6187	0,6337	**0,6563**	0,6350	0,6359 ± 0,0152
skip_stj_1	SkipGram	STJ	média	0,6462	**0,7652**	0,6120	0,6474	0,6677 ± 0,0657
skip_stj_2	SkipGram	STJ	idf	0,6950	0,7497	0,7009	**0,7514**	0,7243 ± 0,0299
skip_stj_3	SkipGram	STJ	bm25	0,7075	0,7415	0,7073	**0,7436**	*0,7250* ± 0,0199

Tabela 4.11: nDCG@25 médios dos modelos BERT.

Identificador	Detalhes	Descrição da tese dismax (1)	Parágrafo dismax (2)	Média
u_cls	Média dos vetores da última camada	0,5082	0,4241	0,4661
u_scls	Média dos vetores da última camada sem token [CLS]	0,5028	0,4184	0,4606
puc_cls	Média dos vetores da penúltima camada	0,4172	0,4046	0,4109
puc_scls	Média dos vetores da penúltima camada sem token [CLS]	0,4194	0,4045	0,41195
pc_cls	Média dos vetores da primeira camada	0,3201	0,3338	0,32695
pc_scls	Média dos vetores da primeira camada sem token [CLS]	0,3202	0,3340	0,3271
ajustado	Vetor do token [CLS]	0,3413	0,3837	0,3625
nli	Vetor do token [CLS] do modelo multi-línguas	0,4847	0,4896	**0,4871**

Capítulo 5

Conclusão

Este trabalho apresentou uma pesquisa dentro do domínio de Recuperação da Informação (RI) textual e jurídico, Sistemas de Recuperação de Informação Jurídica (SRIJ). Através da análise do trabalho do Superior Tribunal de Justiça (STJ) em disponibilizar a uniformização de jurisprudência de forma eficiente para a sociedade, foi escolhido o processo de trabalho do tribunal para o produto Jurisprudência em Teses. Dentro desse contexto, foram selecionadas 11 teses jurídicas (de diferentes ramos do direito) e as respectivas consultas do sistema legado da Corte. Com um *corpus* de busca das ementas disponibilizadas entre os anos de 2014 e 2019 foram exploradas as questões de pesquisas. A primeira, relativa a qual metodologia seria mais adequada para comparação de técnicas de similaridade textuais diferentes, foi respondia pela revisão da literatura e escolha da metodolgia *Text Retrieval Conference* (TREC).

A segunda, relativa à comparação de modelos tradicionais de RI com o sistema legado, apresentou resultados positivos. O modelo TF-IDF obteve uma melhora média de 13.3% em comparação com o sistema legado. O modelo BM25 obteve uma melhora média de 14.9%.

Para responder a terceira pergunta, relativa a possibilidade de melhoria dos resultados anteriores com modelos semânticos, foram testados os modelos Word2Vec e *Bidirectional Encoder Representations from Transformers* (BERT). Não houve diferença significativa entre modelos tradicionais e semânticos Word2Vec. Dentre os modelos semânticos, o Word2Vec obteve resultados melhores estatisticamente significantes em comparação aos modelos BERT.

Por fim, a última pergunta objetivava responder como combinar modelos semânticos e tradicionais. Cada vetor dos termos foram ponderados com seu *idf* e $bm25$ ao consolidar os embeddings de documentos. Houve uma pequena melhora na comparação com o embedding pela média simples, mas sem diferenças significativas.

O baixo desempenho do BERT comparado aos outros modelos pode ter sido resultado da falta de uma base histórica de *Short Sentence Similarity* (STS) do domínio jurídico para realizar o ajuste do modelo. Quanto ao modelo ajustado treinado, seu baixo desempenho pode ser resultante da tarefa de classificação utilizada não ser de similaridade entre textos.

Como resultado da pesquisa, também foi publicado um artigo (Apêndice B) na conferência MEDES'20 (The 12th International Conference on Management of Digital EcoSystem) entitulado "*A new conceptual framework for enhancing legal information retrieval at the Brazilian Superior Court of Justice*".

Apesar dos testes estatísticos não evidenciarem diferença significativa entre os modelos tradicionais e o sistema legado, a pesquisa expôs que é possível facilitar a pesquisa de teses jurídicas com os modelos tradicionais, visto que a digitação da tese jurídica ou a seleção de um parágrafo é uma atividade com menor complexidade do que a construção da consulta booleana.

Através dos resultados, visualizamos que a possibilidade de aumentar a quantidade de teses e de rotulações de relevância pode trazer resultados mais significativos. Essa base pode ser utilizada para ajuste dos modelos contextuais. Apesar da exploração de diversas configurações de cada modelo, ainda existem muitas abordagens a serem testadas no futuro. Uma delas seria a utilização do resultado das consultas prontas do sistema legado, presumindo que a maioria delas é relevante para a tese buscada, e utilizar como base de treinamento fracamente supervisionada para ajuste dos modelos contextuais ou modelos de aprendizado de *rank*.

Bibliografia

[1] Kevin D. Ashley. *Artificial intelligence and legal analytics: new tools for law practice in the digital age*. Cambridge University Press, 2018 (ver p. 1).

[2] Marc van Opijnen e Cristiana Santos. "On the concept of relevance in legal information retrieval". Em: *Artificial Intelligence and Law* 25.1 (mar. de 2017), pp. 65–87. ISSN: 1572-8382. DOI: 10.1007/s10506-017-9195-8. URL: https://doi.org/10.1007/s10506-017-9195-8 (ver pp. 1, 8).

[3] Luis Sanchez et al. "Easing Legal News Monitoring with Learning to Rank and BERT". Em: *Advances in Information Retrieval*. Ed. por Joemon M. Jose et al. Cham: Springer International Publishing, 2020, pp. 336–343. ISBN: 978-3-030-45442-5 (ver pp. 1, 3, 24).

[4] Tony Russell-Rose, Jon Chamberlain e Leif Azzopardi. "Information retrieval in the workplace: A comparison of professional search practices". Em: *Information Processing & Management* 54.6 (2018), pp. 1042–1057 (ver pp. 1, 3, 9).

[5] Christopher D. Manning, Prabhakar Raghavan e Hinrich Schütze. *Introduction to information retrieval*. Cambridge University Press, 2009 (ver pp. 1, 9, 10).

[6] Zellig S Harris. "Distributional structure". Em: *Word* 10.2-3 (1954), pp. 146–162 (ver pp. 2, 12).

[7] Karen Sparck Jones. "A Statistical Interpretation of Term Specificity and Its Application in Retrieval". Em: *Document Retrieval Systems*. GBR: Taylor Graham Publishing, 1988, pp. 132–142. ISBN: 0947568212 (ver p. 2).

[8] Stephen Robertson e Hugo Zaragoza. "The Probabilistic Relevance Framework: BM25 and Beyond". Em: *Found. Trends Inf. Retr.* 3.4 (abr. de 2009), pp. 333–389. ISSN: 1554-0669. DOI: 10.1561/1500000019. URL: https://doi.org/10.1561/1500000019 (ver p. 2).

[9] Tomas Mikolov, Kai Chen et al. *Efficient Estimation of Word Representations in Vector Space*. 2013. URL: http://arxiv.org/abs/1301.3781 (ver pp. 2, 12, 37).

[10] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. arXiv: 1810.04805 [cs.CL] (ver pp. 2, 15, 24, 38).

[11] Superior Superior Tribunal de Justiça. *Boletim Estatístico do Superior Tribunal de Justiça*. Nov. de 2020. URL: https://www.stj.jus.br/webstj/Processo/Boletim/?vPortalAreaPai=183 (ver p. 5).

[12] S. Büttcher, C.L.A. Clarke e G.V. Cormack. *Information Retrieval: Implementing and Evaluating Search Engines*. INFORMATION RETRIEVAL. MIT Press, 2016. ISBN: 9780262528870. URL: https://books.google.com.br/books?id=2c3RCwAAQBAJ (ver pp. 4, 10, 11, 25, 27).

[13] ChengXiang Zhai e Sean Massung. *Text data management and analysis: a practical introduction to information retrieval and text mining*. Morgan & Claypool, 2016 (ver pp. 5, 16, 20, 25, 26).

[14] Lukas Galke, Ahmed Saleh e Ansgar Scherp. "Evaluating the Impact of Word Embeddings on Similarity Scoring in Practical Information Retrieval". Em: *INFORMATIK 2017* (2017) (ver pp. 6, 23).

[15] Sun Kim et al. "Bridging the gap: Incorporating a semantic similarity measure for effectively mapping PubMed queries to documents". Em: *Journal of Biomedical Informatics* 75 (2017), pp. 122–127. ISSN: 1532-0464. DOI: https://doi.org/10.1016/j.jbi.2017.09.014. URL: http://www.sciencedirect.com/science/article/pii/S1532046417302186 (ver p. 6).

[16] Guoqing Zheng e Jamie Callan. "Learning to Reweight Terms with Distributed Representations". Em: *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '15. Santiago, Chile: Association for Computing Machinery, 2015, pp. 575–584. ISBN: 9781450336215. DOI: 10.1145/2766462.2767700. URL: https://doi.org/10.1145/2766462.2767700 (ver p. 6).

[17] Milagro Teruel et al. "Legal text processing within the MIREL project". Inglês. Em: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Ed. por Georg Rehm, Víctor Rodríguez-Doncel e Julián Moreno-Schneider. Miyazaki, Japan: European Language Resources Association (ELRA), mai. de 2018. ISBN: 979-10-95546-18-4 (ver p. 6).

[18] Mariana Capela Lombardi Moreto. "O precedente judicial no sistema processual brasileiro". Tese de dout. Universidade de São Paulo, 2012. DOI: 10.11606/t.2.2012.tde-15052013-162737 (ver p. 7).

[19] Ricardo A. Baeza-Yates e Berthier A. Ribeiro-Neto. *Modern Information Retrieval - the concepts and technology behind search, Second edition*. 2011 (ver pp. 8–10).

[20] K Latha. *Experiment and Evaluation in Information Retrieval Models*. CRC Press, 2017 (ver p. 9).

[21] Stephen E Robertson. "The probability ranking principle in IR". Em: *Journal of documentation* 33.4 (1977), pp. 294–304 (ver p. 11).

[22] Stephen E Robertson e Steve Walker. "Okapi/keenbow at trec-8". Em: *TREC*. Vol. 8. Citeseer. 1999, pp. 151–162 (ver p. 11).

[23] Wael H Gomaa, Aly A Fahmy et al. "A survey of text similarity approaches". Em: *International Journal of Computer Applications* 68.13 (2013), pp. 13–18 (ver p. 11).

[24] George A Miller. *WordNet: An electronic lexical database*. MIT press, 1998 (ver p. 11).

[25] Dekang Lin. "Automatic retrieval and clustering of similar words". Em: *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 2*. 1998, pp. 768–774 (ver p. 12).

[26] Marco Baroni, Georgiana Dinu e Germán Kruszewski. "Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors". Em: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Baltimore, Maryland: Association for Computational Linguistics, jun. de 2014, pp. 238–247. DOI: 10.3115/v1/P14-1023. URL: https://www.aclweb.org/anthology/P14-1023 (ver p. 12).

[27] Felipe Almeida e Geraldo Xexéo. "Word Embeddings: A Survey". Em: *CoRR* abs/1901.09069 (2019). arXiv: 1901.09069. URL: http://arxiv.org/abs/1901.09069 (ver p. 12).

[28] Nikolaos Aletras e Mark Stevenson. "Evaluating topic coherence using distributional semantics". Em: *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Long Papers*. 2013, pp. 13–22 (ver p. 12).

[29] Yoshua Bengio et al. "A neural probabilistic language model". Em: *Journal of machine learning research* 3.Feb (2003), pp. 1137–1155 (ver p. 12).

[30] Chris McCormick. *The Inner Workings of Word2Vec*. 2019 (ver p. 13).

[31] Lilian Weng. *Learning Word Embedding*. Out. de 2017. URL: https://lilianweng.github.io/lil-log/2017/10/15/learning-word-embedding.html (ver p. 14).

[32] Tomas Mikolov, Ilya Sutskever et al. "Distributed representations of words and phrases and their compositionality". Em: *Advances in neural information processing systems*. 2013, pp. 3111–3119 (ver p. 15).

[33] Frederic Morin e Yoshua Bengio. “Hierarchical probabilistic neural network language model.” Em: *Aistats*. Vol. 5. Citeseer. 2005, pp. 246–252 (ver p. 15).

[34] Michael Gutmann e Aapo Hyvärinen. “Noise-contrastive estimation: A new estimation principle for unnormalized statistical models”. Em: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. 2010, pp. 297–304 (ver p. 15).

[35] Zhilin Yang et al. “Xlnet: Generalized autoregressive pretraining for language understanding”. Em: *Advances in neural information processing systems*. 2019, pp. 5753–5763 (ver p. 15).

[36] Kalervo Järvelin e Jaana Kekäläinen. “IR evaluation methods for retrieving highly relevant documents”. Em: *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2000)*. Ed. por Emmanuel Yannakoudakis et al. Athens, Greece: ACM, 2000, pp. 41–48. ISBN: 1-58113-226-3. DOI: 10.1145/345508.345545. URL: http://doi.acm.org/10.1145/345508.345545 (ver p. 17).

[37] Chris Burges et al. “Learning to Rank Using Gradient Descent”. Em: *Proceedings of the 22nd International Conference on Machine Learning*. ICML ’05. Bonn, Germany: Association for Computing Machinery, 2005, pp. 89–96. ISBN: 1595931805. DOI: 10.1145/1102351.1102363. URL: https://doi.org/10.1145/1102351.1102363 (ver p. 17).

[38] Daniel Locke, Guido Zuccon e Harrisen Scells. “Automatic Query Generation from Legal Texts for Case Law Retrieval”. Em: *Information Retrieval Technology Lecture Notes in Computer Science* (2017), pp. 181–193. DOI: 10.1007/978-3-319-70145-5_14 (ver p. 21).

[39] Daniel Locke e Guido Zuccon. “A Test Collection for Evaluating Legal Case Law Search”. Em: *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval - SIGIR 18* (2018). DOI: 10.1145/3209978.3210161 (ver p. 21).

[40] Isar Nejadgholi, Renaud Bougueng e Samuel Witherspoon. “A Semi-Supervised Training Method for Semantic Search of Legal Facts in Canadian Immigration Cases”. Em: *Legal Knowledge and Information Systems* (2017), p. 125 (ver p. 22).

[41] Matthias Grabmair et al. “Introducing LUIMA”. Em: *Proceedings of the 15th International Conference on Artificial Intelligence and Law - ICAIL 15* (2015). DOI: 10.1145/2746090.2746096 (ver p. 22).

[42] Keet Sugathadasa et al. “Legal document retrieval using document vector embeddings and deep learning”. Em: *Science and Information Conference*. Springer. 2018, pp. 160–175 (ver p. 23).

[43] Anup Anand Deshmukh e Udhav Sethi. “IR-BERT: Leveraging BERT for Semantic Search in Background Linking for News Articles”. Em: *arXiv e-prints* (2020), arXiv–2007 (ver p. 24).

[44] Sean MacAvaney et al. “CEDR: Contextualized embeddings for document ranking”. Em: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2019, pp. 1101–1104 (ver p. 24).

[45] Wei Yang, Haotian Zhang e Jimmy Lin. “Simple applications of BERT for ad hoc document retrieval”. Em: *arXiv preprint arXiv:1903.10972* (2019) (ver p. 24).

[46] T. Sakai. *Laboratory Experiments in Information Retrieval: Sample Sizes, Effect Sizes, and Statistical Power*. The Information Retrieval Series. Springer Singapore, 2018 (ver p. 25).

[47] Mark Sanderson. *Test collection based evaluation of information retrieval systems*. Now Publishers Inc, 2010 (ver p. 27).

[48] Aldo Lipani et al. “The Impact of Fixed-Cost Pooling Strategies on Test Collection Bias”. Em: set. de 2016. DOI: 10.1145/2970398.2970429 (ver p. 27).

[49] Alberto Tonon, Gianluca Demartini e Philippe Cudre-Mauroux. “Pooling-based continuous evaluation of information retrieval systems”. Em: *Information Retrieval Journal* 18 (out. de 2015). DOI: 10.1007/s10791-015-9266-y (ver p. 27).

[50] Viviane Moreira Orengo e Christian Huyck. “A stemming algorithm for the portuguese language”. Em: *Proceedings Eighth Symposium on String Processing and Information Retrieval*. IEEE. 2001, pp. 186–193 (ver p. 35).

[51] Radim Rehurek e Petr Sojka. “Software Framework for Topic Modelling with Large Corpora”. Em: *IN PROCEEDINGS OF THE LREC 2010 WORKSHOP ON NEW CHALLENGES FOR NLP FRAMEWORKS*. 2010, pp. 45–50 (ver p. 35).

[52] Arthur Spirling e P Rodriguez. “Word embeddings: What works, what doesn’t, and how to tell the difference for applied research”. Em: (2019) (ver p. 36).

[53] Nathan Hartmann et al. “Portuguese word embeddings: Evaluating on word analogies and natural language tasks”. Em: *arXiv preprint arXiv:1708.06025* (2017) (ver p. 36).

[54] Or Sharir, Barak Peleg e Yoav Shoham. “The Cost of Training NLP Models: A Concise Overview”. Em: *arXiv preprint arXiv:2004.08900* (2020) (ver p. 37).

[55] Fábio Souza, Rodrigo Nogueira e Roberto Lotufo. “BERTimbau: pretrained BERT models for Brazilian Portuguese”. Em: *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*. 2020 (ver p. 37).

[56] Nils Reimers e Iryna Gurevych. “Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation”. Em: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, nov. de 2020. URL: https://arxiv.org/abs/2004.09813 (ver pp. 37, 39).

[57] Jay Alammar. *The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning)*. Dez. de 2018. URL: http://jalammar.github.io/illustrated-bert (ver p. 38).

[58] Thomas Wolf et al. “Transformers: State-of-the-Art Natural Language Processing”. Em: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, out. de 2020, pp. 38–45. URL: https://www.aclweb.org/anthology/2020.emnlp-demos.6 (ver p. 38).

Apêndice A

Consultas Utilizadas Para Cada Tese

As Tabelas A.1 a A.11 apresentam os dados de cada tese selecionada na amostragem aleatória. São 11 teses, uma por cada ramo do direito, extraídas do produto Jurisprudência em Teses do Superior Tribunal de Justiça. O campo "Descrição da Tese"é o texto livre escrito pelo analista para resumir a tese. O campo "Parágrafo selecionado"contém os parágrafos extraídos da ementa mais recente para simular a recuperação de outras decisões com a mesma tese após o usuário selecionar um parágrafo de texto livre. Por fim, o campo "Consulta Booleana"contém a consulta elaborada pelos analistas no sistema legado.

Tabela A.1: Dados da Tese 1374.

Ramo do Direito	DIREITO DO CONSUMIDOR
Tema	JUROS
Descrição da Tese	É admitida a revisão das taxas de juros remuneratórios em situações excepcionais, desde que caracterizada a relação de consumo e que a abusividade (capaz de colocar o consumidor em desvantagem exagerada - art. 51, §1 º, do CDC) fique cabalmente demonstrada, ante as peculiaridades do julgamento em concreto
Parágrafo Selecionado	1. A Segunda Seção do Superior Tribunal de Justiça, no julgamento do REsp nº 1. 61.53 /RS, Relatora a Ministra Nancy Andrighi, submetido ao regime dos recursos repetitivos, firmou posicionamento no sentido de que: 'a) As instituições financeiras não se sujeitam à limitação dos juros remuneratórios estipulada na Lei de Usura (Decreto 22.626/33), Súmula 596/STF; b) A estipulação de juros remuneratórios superiores a 12% ao ano, por si só, não indica abusividade; c) São inaplicáveis aos juros remuneratórios dos contratos de mútuo bancário as disposições do art. 591 c/c o art. 4 6 do CC/02; d) É admitida a revisão das taxas de juros remuneratórios em situações excepcionais, desde que caracterizada a relação de consumo e que a abusividade (capaz de colocar o consumidor em desvantagem exagerada art. 51, §1º, do CDC) fique cabalmente demonstrada, ante às peculiaridades do julgamento em concreto'. 3. É válida a cláusula contratual que prevê a cobrança da comissão de permanência, calculada pela taxa média de mercado apurada pelo Banco Central do Brasil, de acordo com a espécie da operação, tendo como limite máximo o percentual contratado (Súmula nº 294/STJ). 5. A mora restou configurada, pois não houve o reconhecimento da abusividade nos encargos exigidos no período da normalidade contratual (juros remuneratórios e capitalização).
Consulta Booleana	(revisão com taxa adj4 (remuneratór$ ou juros) com (excepcion$ ou especia$) com (consum$ ou abusi$ ou desvantag$ ou ((lei com ("8078"$ ou "8.078"$ ou "008078")) ou (CDC ou Código adj2 defesa adj2 consumidor)) com art$ prox10 ("51" ou "00051" ou "51") com ("par. 1" ou "par. 1." ou "§1" ou "§1º" ou (par$ "00001"))))

Tabela A.2: Dados da Tese 1270.

Ramo do Direito	DIREITO PREVIDENCIÁRIO
Tema	PREVIDÊNCIA COMPLEMENTAR
Descrição da Tese	As contribuições devolvidas pelas entidades de previdência privada ao associado devem ser atualizadas monetariamente pelo IPC - Índice de preços ao consumidor.
Parágrafo Selecionado	1. É devida a restituição da reserva de poupança a ex-participantes de plano de previdência privada, devendo ser corrigida conforme índices que melhor reflitam a real inflação da época, mesmo que o estatuto da entidade estabeleça critério diverso, devendo ser incluídos os expurgos inflacionários (Súmula n. 289/STJ). 2. A atualização monetária das contribuições restituídas pela instituição de previdência privada deve ser calculada pelo IPC, por ser o índice que melhor representa a perda do poder aquisitivo da moeda (Recurso Especial repetitivo n. 1.183.474/DF).
Consulta Booleana	previdencia prox3 (privada ou $plement$) com (devol$ ou restituição) mesmo IPC

Tabela A.3: Dados da Tese 560.

Ramo do Direito	DIREITO PENAL
Tema	LEI DE DROGAS
Descrição da Tese	O crime de associação para o tráfico de entorpecentes (art. 35 da Lei n. 11. 343/2006) não figura no rol taxativo de crimes hediondos ou de delitos a eles equiparados.
Parágrafo Selecionado	1. De acordo com a Jurisprudência desta Corte Superior, ante a ausência de previsão no rol do art. 2º da Lei 8.072/9 , o crime de associação para o tráfico previsto no art. 35 da Lei 11.343/06 não é crime hediondo ou equiparado.
Consulta Booleana	((associação prox3 trafico) ou ”35” prox10 (ldr$ ou lei adj3 (drog$ ou tráfico ou entorpecentes) ou ”11.343”$ ou ”11343”$)) com (ser ou considera$ ou equipar$ ou caracteri$ ou integr$ ou rol) prox15 (hediond$ ou equipara$) não (”ECA”)

Tabela A.4: Dados da Tese 955.

Ramo do Direito	DIREITO TRIBUTÁRIO
Tema	IMPOSTOS MUNICIPAIS
Descrição da Tese	Cabe ao contribuinte comprovar a ausência de notificação do lançamento tributário pelo não recebimento do carnê de cobrança do IPTU.
Parágrafo Selecionado	2. Há nesta Corte jurisprudência consolidada no sentido de que a notificação do lançamento do IPTU e das taxas municipais ocorre com o envio da correspondente guia de recolhimento do tributo para o endereço do imóvel ou do contribuinte, com as informações que lhe permitam, caso não concorde com a cobrança, impugná-la administrativa ou judicialmente.
Consulta Booleana	(IPTU$ mesmo (onus ou prova$ ou proba$ ou demonstr$) com (entreg$ ou receb$ ou envi$) com (carnê ou guia))

Tabela A.5: Dados da Tese 1423.

Ramo do Direito	DIREITO REGISTRAL
Tema	REGISTROS PÚBLICOS
Descrição da Tese	A vaga de garagem que possui matrícula própria no registro de imóveis não constitui bem de família para efeito de penhora.
Parágrafo Selecionado	3. O entendimento adotado pela Corte de origem está em consonância com o posicionamento deste Tribunal firmado na Súmula 449/STJ, verbis: 'A vaga de garagem que possui matrícula própria no registro de imóveis não constitui bem de família para efeito de penhora'.
Consulta Booleana	(((((((”000449” mesmo stj) e sum).ref.) ou (449/stj)) ou ((”449”) prox4 (sum$ ou verbete$ ou enunciado$)) com ((sum$ ou verbete$ ou enunciado$) com ((”tribunal da cidadania” ou $est? sodalicio ou $est? corte ou $esta corte superior ou ”superior tribunal de justica” ou ”stj”) ou (corte prox3 superior prox3 justica) ou (superior prox2 tribunal prox3 justica)))) não (”449” prox2 (stf ou suprem$)))) ou ((box$ ou vaga) com (estacionamento ou garagem) com (bem adj2 família ou $penhora$))

Tabela A.6: Dados da Tese 635.

Ramo do Direito	DIREITO EMPRESARIAL
Tema	RECUPERAÇÃO JUDICIAL
Descrição da Tese	A recuperação judicial é norteada pelos princípios da preservação da empresa, da função social e do estímulo à atividade econômica, a teor do art. 47 da Lei n. 11.101/2005.
Parágrafo Selecionado	3. A Lei n. 11.101/2005 visa à preservação da empresa, à função social e ao estímulo à atividade econômica, a teor de seu art. 47.
Consulta Booleana	recuperação adj2 judicial e ((preservação adj4 (empresa ou empresarial) com função adj2 social com estímulo) ou art$ adj2 47 ou superação adj2 crise)

Tabela A.7: Dados da Tese 1474.

Ramo do Direito	DIREITO CIVIL
Tema	BANCÁRIO
Descrição da Tese	O Código de Defesa do Consumidor é aplicável às instituições financeiras.
Parágrafo Selecionado	1. Aplicação do Código de Defesa do Consumidor às instituições bancárias e, assim, a possibilidade de inversão do ônus da prova (art. 6º, VIII, do CDC).
Consulta Booleana	(((((((”000297” mesmo stj) e sum).ref.) ou (297/stj)) ou ((”297”) prox4 (sum$ ou verbete$ ou enunciado$)) com ((sum$ ou verbete$ ou enunciado$) com ((”tribunal da cidadania” ou $est? sodalicio ou $est? corte ou $esta corte superior ou ”superior tribunal de justica” ou ”stj”) ou (corte prox3 superior prox3 justica) ou (superior prox2 tribunal prox3 justica)))) não (”297” prox2 (stf ou suprem$))) ou (aplica$ ou incide$) com (CDC ou consum$) com (instituiç$ adj2 (financeir$ ou bancari$) ou banco$)) não ((indenização prox4 (Fundo adj2 Garantidor adj2 Crédito)) ou (compensação adj2 crédito) ou (exibição) ou (pecunia) ou (procon) ou (anual) ou (cooperativ$) ou (inovação adj2 recursal) ou (imovel))

Tabela A.8: Dados da Tese 1238.

Ramo do Direito	DIREITO PROCESSUAL PENAL
Tema	NULIDADES NO PROCESSO PENAL
Descrição da Tese	A ausência de intimação da defesa sobre a expedição de precatória para oitiva de testemunha é causa de nulidade relativa.
Parágrafo Selecionado	2. Além de a paciente e seu causídico terem efetivamente tomado conhecimento da expedição da carta precatória, não se demonstrou em que medida o comparecimento da paciente poderia ter repercutido de forma positiva na sua situação processual. Dessarte, não se verifica prejuízo na situação retratada nos autos, o que impede o reconhecimento de eventual nulidade. Inteligência do verbete n. 155/STF. Como é cediço, a moderna processualística não admite o reconhecimento de nulidade que não tenha acarretado prejuízo à parte. Não se admite a forma pela forma.
Consulta Booleana	((((155/stf)) ou ((”155”) prox3 (sum$ ou verbete$ ou enunciado$)) com ((sum$ ou verbete$ ou enunciado$) com ((”corte suprema” ou ”suprema corte” ou corte prox2 máxima ou ”pretório excelso” ou ”supremo tribunal federal” ou ”stf”) ou (supremo prox2 tribunal prox3 federal)))) não (”155” prox2 (stj ou supeior$))) ou ((nulo ou nula ou nulidade) mesmo (intim$ com expedi$ com (precatória$))) não ((((273/stj)) ou ((”273”) prox3 (sum$ ou verbete$ ou enunciado$)) com ((sum$ ou verbete$ ou enunciado$) com ((”tribunal da cidadania” ou $est? sodalicio ou $est? corte ou $esta corte superior ou ”superior tribunal de justica” ou ”stj”) ou (corte prox3 superior prox3 justica) ou (superior prox2 tribunal prox3 justica)))))

Tabela A.9: Dados da Tese 914.

Ramo do Direito	DIREITO DA CRIANÇA E DO ADOLESCENTE
Tema	MEDIDAS SOCIOEDUCATIVAS
Descrição da Tese	É possível a incidência do princípio da insignificância nos procedimentos que apuram a prática de ato infracional.
Parágrafo Selecionado	3. In casu, se a Corte estadual deixou de analisar a possibilidade de efetiva aplicação do princípio da insignificância por entendê-la incabível no âmbito do Estatuto da Criança e do Adolescente. A pretensão de reconhecer a incidência do indiferente penal nesta via implicaria, em princípio, indevida supressão de instância, uma vez que a questão não foi objeto de exame no acórdão impetrado, que se limitou a enfrentar a eleição do tratamento mais adequado ao caso. 4. Considerando que o Superior Tribunal de Justiça já firmou entendimento no sentido da possibilidade de aplicação do princípio da bagatela às condutas regidas pelo Estatuto da Criança e do Adolescente (HC 276.358/SP, Rel. Ministro Nefi Cordeiro, Sexta Turma, DJe 22/09/2014), faz-se necessária a análise acerca de sua efetiva aplicação no presente caso. 5. Na aplicação do princípio da insignificância, devem ser utilizados os seguintes parâmetros: a) conduta minimamente ofensiva; b) ausência de periculosidade do agente; c) reduzido grau de reprovabilidade do comportamento; e d) lesão jurídica inexpressiva, os quais devem estar presentes, concomitantemente, para a incidência do referido instituto.
Consulta Booleana	(princípio adj2 (insignificância ou bagatela) ou bagatela$ ou mínima adj2 lesao) mesmo (ato adj2 infracional ou menorista ou ECA ou Estatuto adj2 criança ou ("8.069"$ ou "8069"$ ou "008069"$))

Tabela A.10: Dados da Tese 120.

Ramo do Direito	DIREITO ADMINISTRATIVO
Tema	CONSELHOS PROFISSIONAIS
Descrição da Tese	O exame de suficiência instituído pela Lei n. 12.249/2010, que alterou o art. 12, § 2º, do Decreto-Lei n. 9.295/1946, será exigido de contadores e de técnicos em contabilidade que completarem o curso após a vigência daquela lei.
Parágrafo Selecionado	II - No caso, verifico que o acórdão recorrido adotou entendimento pacífico nesta Corte, segundo o qual, o exame de suficiência será exigido daqueles que ainda não haviam completado o curso técnico ou superior em Contabilidade sob a égide da legislação pretérita.
Consulta Booleana	((((exam$ ou prova) adj3 sufic$ ou "12249"$ ou "12.249"$) e tec$ adj4 contab$ e tec$ adj4 contab$ com (exam$ ou prova) adj3 sufic$)) ou (exam$ ou prova) adj3 sufic$ com direito adj adquirido com contab$

Tabela A.11: Dados da Tese 1995.

Ramo do Direito	DIREITO PROCESSUAL CIVIL
Tema	DOS HONORÁRIOS ADVOCATÍCIOS
Descrição da Tese	Não é possível a compensação de honorários advocatícios quando a sua fixação ocorrer na vigência do CPC/2015 - art. 85, § 14.
Parágrafo Selecionado	VI - É necessário ressaltar também que a jurisprudência do STJ, firmada na vigência do CPC/73, orientava-se pelo entendimento de que não é possível a compensação dos honorários advocatícios fixados em processos distintos. Nesse sentido: AgInt no AgInt no REsp n. 1.609.915/RS, Rel. Ministro Mauro Campbell Marques, Segunda Turma, julgado em 15/12/2016, DJe 19/12/2016; AgRg no REsp n. 1.563.629/RS, Rel. Ministro Mauro Campbell Marques, Segunda Turma, julgado em 24/11/2015, DJe 2/12/2015; REsp n. 1.527.590/RS, Rel. Ministro Herman Benjamin, Segunda Turma, julgado em 26/5/2015, DJe 5/8/2015). VII - Não bastassem esses fundamentos, atente-se que a decisão contra a qual o município interpôs o agravo de instrumento, ao que tudo indica, foi proferida já na vigência do Código de Processo Civil de 2015, que, no seu art. 85, § 14, veda a compensação de honorários advocatícios.
Consulta Booleana	((honorar$ ou sucumbe$ ou (art$ adj2 ”85”)) mesmo compensa$ com (”NÃO” adj2 possivel ou impossib$ ou veda$) com (fixa$ ou arbitr$ ou proferid$) com (ambito ou vig$ ou durant$) e (process$ adj2 civil adj2 2015 ou NCPC ou CPC?2015 ou CPC? 15 ou ”13105”$ ou ”13.105”$ ou ”013105”) ou (honorar$ ou sucumbe$ ou (art$ adj2 ”85”)) com compensa$ com (”NÃO” adj2 possivel ou impossib$ ou veda$)) e art$ adj2 (”85” ou ”00085”) prox2 (par$ ou ”§”) prox2 (”14” ou ”00014”) prox4 (process$ adj2 civil adj2 2015 ou NCPC ou CPC?2015 ou CPC?15 ou ”13105”$ ou ”13. 105”$ ou ”013105”)

Apêndice B

Artigo: *A new conceptual framework for enhancing legal information retrieval at the Brazilian Superior Court of Justice*

A new conceptual framework for enhancing legal information retrieval at the Brazilian Superior Court of Justice

Thiago Gomes
Marcelo Ladeira
alencargomesthiago@gmail.com
mladeira@unb.br
University of Brasilia
Brasilia, Federal District, Brazil

ABSTRACT

Effective retrieval of jurisprudence (case-law) is imperative to achieve consistency and predictability for any legal system. In this work, we propose and proceed to an empirical evaluation of a framework for jurisprudence retrieval of the Brazilian Superior Court of Justice in order to ease the task of retrieval of other decisions with the same legal opinion. The experimental results shown that our approach based on text similarity performs better than the legacy system of the Court based on Boolean queries. The building of complex Boolean queries is very specialized and we aim to offer a tool able to use free text as queries without any operator. With the legacy system as baseline, we compare the TF-IDF traditional retrieval model, the BM25 probabilistic model and the Word2Vec model. Our results indicate that the Word2Vec Skip-Gram model, trained on a specialized legal corpus and BM25 yield similar performance and surpasses the legacy system. Combining BM25 model with embedding models improved the performance up to 19%.

CCS CONCEPTS

• **Applied computing** → Law; • **Information systems** → *Retrieval models and ranking*.

KEYWORDS

Natural Language Processing; Information Retrieval; Word Embedding; E-Discovery

ACM Reference Format:
Thiago Gomes and Marcelo Ladeira. 2020. A new conceptual framework for enhancing legal information retrieval at the Brazilian Superior Court of Justice. In Proceedings of the *12th International Conference on Management of Digital EcoSystems (MEDES '20), November 2–4, 2020, Virtual Event, United Arab Emirates.* ACM, Abu Dhabi, UAE, 4 pages. https://doi.org/10.1145/3415958.3433087

MEDES '20, November 2–4, 2020, Virtual Event, United Arab Emirates

ACM ISBN 978-1-4503-8115-4/20/11…$15.00
https://doi.org/10.1145/3415958.3433087

1 INTRODUCTION

Judicial jurisprudence (precedents) are past court decisions that state a legal opinion or thesis. In Common or Civil Law judicial systems, jurisprudence aims to deliver consistency and predictability for future decisions for cases with similar facts [24]. The Brazilian Superior Court of Justice (STJ) is responsible for the standardization of all matter of law, i.e. settle the legal opinion on cases with the same matters, concerning the Brazilian Non-constitutional Federal Legislation. Until November 2020, the court issued more than 257 thousands decisions[1]. The Court has an internal department to organize this decisions and facilitate the retrieval of jurisprudence by the internal and external users, in order to enhance the information retrieval of those decisions.

It is estimated that 35% of lawyer's labor time is consumed on retrieving relevant cases and prior related legal opinion [21]. Enhancing this task in legal information retrieval is an important subject and still an open issue with several facets [24]. When a case is provided with complete jurisprudence it is resolved more definitely at an early stage in the judicial process. Moreover, there is the technology facet. Most of the Courts information retrieval systems rely on legacy closed systems and only provide Boolean searches [1–3].

The complexity of legal domain information states as a challenge in defining the best technique to measure the similarity between court decisions [15]. Furthermore, there is a lack of annotated data for building and evaluating text mining models in the legal domain [13], especially in Portuguese language [4, 5]. The complexity and barriers of textual Boolean search system are further detailed in [19, 20].

These similar issues were explored with the decisions of the Supreme Court of India and United States. In the first [11, 12, 15], the authors explored textual similarity and citation network similarity. Their works indicate good results using paragraphs as query source with TF-IDF or Word2Vec. In the later [13, 14], the authors compared query built by lawyers specialized in textual retrieval with automatic retrieval based on the source document text. The results indicate BM25 as the best retrieval model, but the automatic approach does not surpass the most complex specialists queries.

In this scenario our contributions are threefold: we propose a conceptual framework for continuous evaluation of retrieval models against the legacy Boolean system in the Brazilian court, that can be applied to all Brazilian courts; we discuss the results of applying this

[1]Statistical Bulletin of the Superior Court of Justice. Available at https://www.stj.jus.br/webstj/Processo/Boletim/?vPortalAreaPai=183&vPortalArea=584

framework with several retrieval models and demonstrate possible solutions for the text retrieval in legal domain; and we contribute to the textual retrieval research area by showing the impact of the combination of word embedding and traditional models.

2 METHODOLOGY

To conduct a effective information retrieval research there are four main components that must be defined:

- Q: a set of queries that represents user needs of information.
- D: a collection or corpus of documents that may contain the information.
- M: a model that puts documents and queries in the same numerical representation.
- $R(q_i \in Q, d_j \in D)$: is a ranking function that defines the ordering of relevance of all documents in D relative to a query in Q.

The queries were extracted based on 1395 legal theses contained in the "Jurisprudence in Theses"[2] system of the court. This system organizes those theses in groups by law branches and themes. Each thesis is composed by:

- A law branch: represents the area of the law in Brazilian legislation. In total there are 11 (e.g., Criminal Law or Tax Law).
- A description: when an analyst selects a pertinent thesis, it describes in his own words (e.g, "To characterize the crime of calumny, it is essential that the agent who attributes someone something defined as a crime has knowledge of imputation falsity.").
- Docket: the source decision that the analyst knows that contains the selected legal opinion.
- Legacy system query (LSQ): in daily work, the analyst build several Boolean queries in the legacy system query language, until some quality criteria are achieved. The final query is published as an off-the-shelf search point to speed up future searches for the same thesis by internal and external users.

From this fields we use as queries: the description, the paragraphs of the docket that contains the legal opinion and the LSQ. Our corpus (D) is composed of 160,714 decisions dockets from 2016 to 2019 from the STJ. Dockets are a summary of a legal case and are structured as follow:

(1) Catchphrase: as in [22] the first paragraph is a sequence of keywords that aims to digest the decision.
(2) Enumerated Paragraphs: a sequence of free text paragraphs that describes the legal case and the judge's legal opinion. The final paragraph states the bureaucratic orientation of the decision and has no significant content.

We consider two approaches for ranking dockets. First, we compare the query vector against the vector of the whole text of the documents in D. Second, we compare the query against each paragraph of the documents.

As ranking function we employ the cosine similarity (CS) [6]. The dot product of the two vectors is divided by the Euclidean normalization of the vectors. Hence, the length of the documents is normalized and the similarity is taken by the angle between the two vectors.

The following models (M) were evaluated (All models were built with Gensim [17] python package):

- Traditional retrieval models Term Frequency-Inverse Document Frequency(TF-IDF) [10] and Best Match 25(BM25) [18]. In these models, documents vectors are sparse and the components are terms weights that represents they importance relative to the corpus. In the first, greater weight is given to rare terms that are more likely to differentiate documents topics. The main difference of BM25 is the probabilistic assumption that the bigger the document is, compared to the average document size of the corpus, it is more likely that terms that represent the user need are dispersed and the topic of the document is different from the user need. This models were built with the D corpus.
- Word2Vec [16] document embeddings is based on the distributional hypothesis [7], that states that similar words occur in a similar context. It is implemented as a shallow neural network with the objective to maximize the log of the probability of terms that occurs near each other. Each term in this model is a dense vector, hence to represent a document it is common to average or sum all terms vectors contained in a document to represent it. We explore 4 Word2Vec models: Skip-Gram (SKIP_NILC) and Continuous Bag of Words (CBOW_NILC) trained on a large multi domain corpus [8]; Skip-Gram (SKIP_STJ) and Continuous Bag of Words (CBOW_STJ) trained on 1,909,966 legal cases.
- Weighted Word2Vec. When taking the average of all vectors terms in a document the document vector is skewed towards the more frequent terms. This bias can be minimized by using the weighted average of the vectors with the traditional models. We employ the weighting with BM25 model.

To evaluate different algorithms in a IR task, it is necessary to add two extra components [23]: a set of relevance judgments or assessments J (that represents for each query in Q if documents in D satisfies the user need) and a set of runs R (that represents a ranking of the documents recovered by each model). A single run in our research is a combination of a query (thesis description or selected paragraphs), a model in M to represent queries and documents, the cosine similarity function and a method that defines the granularity of comparison between documents and queries (take the cosine similarity by whole document or by paragraphs, as in [12]). The J and R collections are usually called Gold Standard Corpus. We apply our conceptual framework in this stage. as shown in Fig. 1.

In step 1, all the combination of queries and source documents are employed. From the ranking of all models, we take the union of the top 25 best ranked documents in step 2 (Depth@k [23]). At step 3, the legal specialist evaluates the relevance of the ranked documents per legal opinion. The relevance assessments used in this paper are three: "Not relevant" for dockets that do not contain the legal opinion; "Relevant" for documents that discuss the same subject in the docket and are candidates for further analysis; and "Same Thesis" if the docket text has exactly the same legal opinion.

Finally, we compare retrieval models with the metric Normalized Discounted Cumulative Gain (NDCG) [9]. This metric takes into

[2]Court System where pre-built queries are available to speed legal opinion retrieval. Available at https://scon.stj.jus.br/SCON/

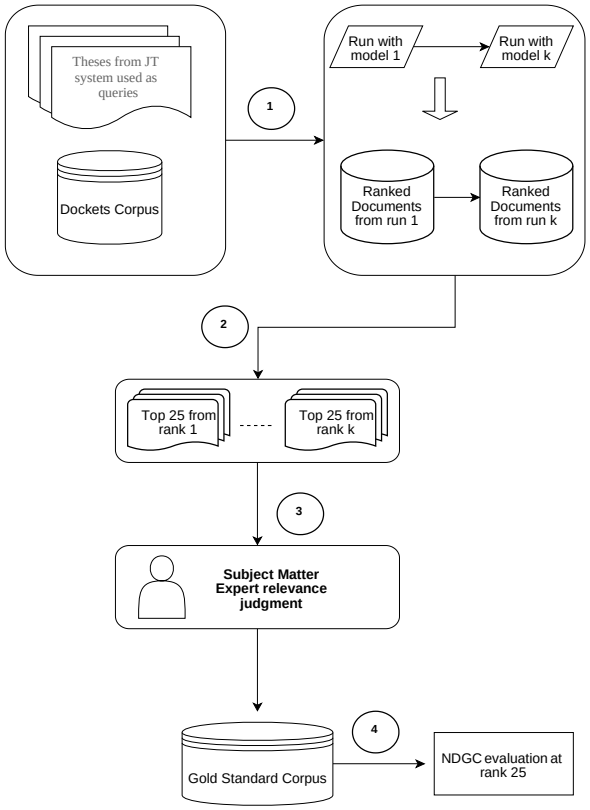


Figure 1: Conceptual Framework for Pooling and evaluation.

account if the system ranks the most relevant documents, in our case the documents with the same thesis, before the other marginally relevant documents. This way is possible to simulate whether our proposed approach will ease the search effort compared to the legacy system.

3 RESULTS AND DISCUSSION

For our initial experiments, we random selected 11 theses, one for each law branch. Our baseline model, complex Boolean queries in the legacy system, achieved a NDCG@25 (we use this cut off of 25 because the analysts only read 25 dockets before rebuild the query or end the search) of 0.654 ± 0.139.

Table 1 represents the mean NDCG@25 metric for all models across the 11 theses, except for runs with weighted word embedding models. Comparing the word embedding models with the lexical models, TF-IDF and BM25, we observe that lexical models yield better result. Only the SKIP_STJ embedding model gives similar results when using the thesis description as query source and ranking documents per paragraphs.

This results indicate that in the STJ dockets there are some pattern on the writing style of the text and traditional models are a promising replacement for the legacy system. In addition, we observe that the Skip-Gram model trained on a large legal domain has comparable results with the traditional models when the query is the free text thesis description and the document granularity is per paragraph. This shows that embedding models, for our domain, are more appropriate for small text segments.

Fig. 2 shows that for the embedding models, ranking dockets per paragraphs and search by thesis description has a higher impact. We observe, in Fig. 3, that this behavior does not occur when using selected paragraphs as source for queries.

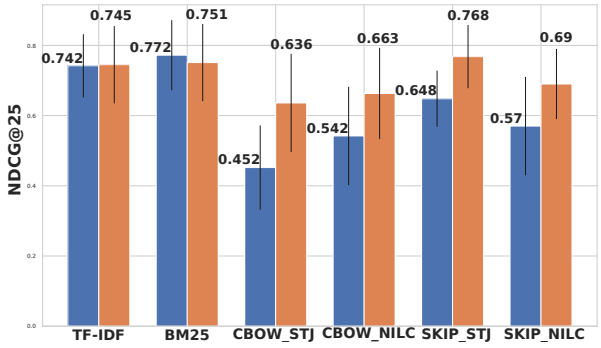


Figure 2: Average NDCG@25 for queries source as thesis description grouped by model and search granularity with confidence interval. Blue bars represent search ranked by the whole document text and orange ones represent results ranked per paragraph.

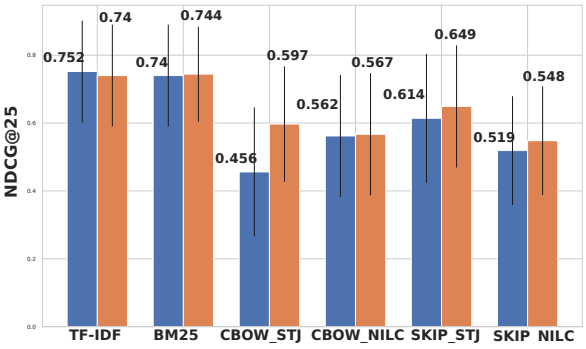


Figure 3: Average NDCG@25 for queries source as selected paragraphs grouped by model and search granularity with confidence interval. Blue bars represent search ranked by the whole document text and orange ones represent results ranked per paragraph.

We choose only to weight with the BM25 model, since there is no significant difference between the two lexical models and BM25 has a marginally better score than TF-IDF. Table 2 shows, in the generic models, an improvement of 5% and 3%, for CBOW and Skip-Gram respectively. For the models trained on specialized corpus, there

Table 1: Mean NDCG@25 across all 11 theses, with a confidence interval, of each model without any embedding weighting. Best performance are in bold.

Model	Query Source: Thesis description		Query Source: Selected Paragraphs		Mean
	All Document text	Per Paragraph	All Document text	Per Paragraph	
TF-IDF	0.742 ± 0.09	0.745 ± 0.11	**0.752** ± 0.15	0.740 ± 0.15	0.745 ± 0.06
BM25	**0.772** ± 0.1	0.751 ± 0.11	0.740 ± 0.15	**0.744** ± 0.14	**0.752** ± 0.06
CBOW_STJ	0.452 ± 0.12	0.636 ± 0.14	0.456 ± 0.19	0.597 ± 0.17	0.535 ± 0.08
CBOW_NILC	0.542 ± 0.14	0.663 ± 0.13	0.562 ± 0.18	0.567 ± 0.18	0.583 ± 0.08
SKIP_STJ	0.648 ± 0.08	**0.768** ± 0.09	0.614 ± 0.19	0.649 ± 0.18	0.670 ± 0.07
SKIP_NILC	0.570 ± 0.14	0.690 ± 0.1	0.519 ± 0.16	0.548 ± 0.16	0.582 ± 0.07

are gains of 19% and 8% in the same way. This results indicates promising use of both approaches in the STJ legal text retrieval.

Table 2: NDCG@25 gain by BM25 in Weighted Words Centroid.

Model	Raw Average	Weighted with BM25	Gain
CBOW_NILC	0.583 ± 0.08	0.638 ± 0.08	5%
SKIP_NILC	0.582 ± 0.07	0.600 ± 0.07	3%
CBOW_STJ	0.535 ± 0.08	0.638 ± 0.08	19%
SKIP_STJ	0.670 ± 0.07	0.727 ± 0.06	8%

4 CONCLUSION

This works explored information retrieval in the legal domain in a real-world scenario. We applied our conceptual framework for evaluation to build the golden standard corpus and evaluate the performance of several methods against a complex legacy system. The results demonstrate that this research can be continuous and be used to evaluate newer models, like BERT.

With the results it is possible to assume that the legacy system can be substituted by BM25 based systems. The user will not have to type complex queries and will be possible to retrieve jurisprudence only by selection paragraphs of the dockets or describing the thesis as free text.

The use of word embedding is still a open question and we could not conclude that, in our corpus, they represent advantage compared to lexical models. Further, investigation with more relevance assessments and other models will be carried out.

ACKNOWLEDGMENTS

To the Department of Computer Science of University of Brasilia (UnB) that supported this entire research.

REFERENCES

[1] Bruna Armonas Colombo, Pedro Buck, and Vinicius Miana Bezerra. 2017. Challenges When Using Jurimetrics in Brazil—A Survey of Courts. *Future Internet* 9, 4 (2017), 68.

[2] Rhuan Barros, André Peres, Fabiana Lorenzi, Leandro Krug Wives, and Etiene Hubert da Silva Jaccottet. 2018. Case Law Analysis with Machine Learning in Brazilian Court. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*. Springer, 857–868.

[3] Yen-Liang Chen, Yi-Hung Liu, and Wu-Liang Ho. 2013. A text mining approach to assist the general public in the retrieval of legal documents. *Journal of the American Society for Information Science and Technology* 64, 2 (2013), 280–290.

[4] Pedro Henrique Luz de Araujo, Teófilo E de Campos, Renato RR de Oliveira, Matheus Stauffer, Samuel Couto, and Paulo Bermejo. 2018. LeNER-Br: A Dataset for Named Entity Recognition in Brazilian Legal Text. In *International Conference on Computational Processing of the Portuguese Language*. Springer, 313–323.

[5] Diego de Vargas Feijó and Viviane Pereira Moreira. 2018. RulingBR: A Summarization Dataset for Legal Texts. In *International Conference on Computational Processing of the Portuguese Language*. Springer, 255–264.

[6] Wael H Gomaa, Aly A Fahmy, et al. 2013. A survey of text similarity approaches. *International Journal of Computer Applications* 68, 13 (2013), 13–18.

[7] Zellig S Harris. 1954. Distributional structure. *Word* 10, 2-3 (1954), 146–162.

[8] Nathan Hartmann, Erick Fonseca, Christopher Shulby, Marcos Treviso, Jessica Rodrigues, and Sandra Aluisio. 2017. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. *arXiv preprint arXiv:1708.06025* (2017).

[9] Kalervo Järvelin and Jaana Kekäläinen. 2017. IR evaluation methods for retrieving highly relevant documents. In *ACM SIGIR Forum*, Vol. 51. ACM New York, NY, USA, 243–250.

[10] Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation* (1972).

[11] Sushanta Kumar. 2014. *Similarity Analysis of Legal Judgments and applying 'Paragraph-link'to Find Similar Legal Judgments*. Ph.D. Dissertation. PhD thesis, International Institute of Information Technology Hyderabad.

[12] Sushanta Kumar, P Krishna Reddy, V Balakista Reddy, and Aditya Singh. 2011. Similarity analysis of legal judgments. In *Proceedings of the Fourth Annual ACM Bangalore Conference*. ACM, 17.

[13] Daniel Locke and Guido Zuccon. 2018. A test collection for evaluating legal case law search. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 1261–1264.

[14] Daniel Locke, Guido Zuccon, and Harrisen Scells. 2017. Automatic Query Generation from Legal Texts for Case Law Retrieval. In *Asia Information Retrieval Symposium*. Springer, 181–193.

[15] Arpan Mandal, Raktim Chaki, Sarbajit Saha, Kripabandhu Ghosh, Arindam Pal, and Saptarshi Ghosh. 2017. Measuring similarity among legal court case documents. In *Proceedings of the 10th Annual ACM India Compute Conference*. ACM, 1–9.

[16] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[17] Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA, Valletta, Malta, 45–50. http://is.muni.cz/publication/884893/en.

[18] Stephen Robertson and Hugo Zaragoza. 2009. *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc.

[19] Tony Russell-Rose, Jon Chamberlain, and Leif Azzopardi. 2018. Information retrieval in the workplace: A comparison of professional search practices. *Information Processing & Management* 54, 6 (2018), 1042–1057.

[20] Luis Sanchez, Jiyin He, Jarana Manotumruksa, Dyaa Albakour, Miguel Martinez, and Aldo Lipani. 2020. Easing Legal News Monitoring with Learning to Rank and BERT. In *Advances in Information Retrieval*, Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins (Eds.). Springer International Publishing, Cham, 336–343.

[21] Milagro Teruel, Cristian Cardellino, Fernando Cardellino, Laura Alonso Alemany, and Serena Villata. 2018. Legal text processing within the MIREL project. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (Miyazaki, Japan, 7-12), Georg Rehm, Víctor Rodríguez-Doncel, and Julián Moreno-Schneider (Eds.). European Language Resources Association (ELRA), Paris, France.

[22] Edwin Thuma and Nkwebi Peace Motlogelwa. 2017. On the importance of Legal Catchphrases in Precedence Retrieval.. In *FIRE (Working Notes)*. 92–94.

[23] Alberto Tonon, Gianluca Demartini, and Philippe Cudré-Mauroux. 2015. Pooling-based continuous evaluation of information retrieval systems. *Information Retrieval Journal* 18, 5 (2015), 445–472.

[24] Marc Van Opijnen and Cristiana Santos. 2017. On the concept of relevance in legal information retrieval. *Artificial Intelligence and Law* 25, 1 (2017), 65–87.