

INDIR

Indexing Improving Information Retrieval

(on Jurisprudence Statements System indexed with Thesaurus of the
Federal Court of Accounts of Brazil - TCU)

Leonardo Augusto da Silva Pacheco e Marcus Vinícius Borela de Castro

Junho 2023

Abstract

The study aims to enhance the search capabilities of Information Retrieval systems. In addition to providing datasets JURIS-TCU-INDEX and JURIS-TCU for model training and search evaluation in Portuguese with high-quality contents, the study built an automatic indexer model named INDIR, trained using indexing data from JURIS-TCU-INDEX, which was used in three approaches to enhance searches in JURIS-TCU. The search pipeline on JURIS-TCU-INDEX based on INDIR outperformed the MT5-3B model in the evaluated metrics RANK1 and NDCG@12, demonstrating the effectiveness of a smaller yet well-trained model over larger ones. Using INDIR as a reranker and adding indexing data for query and document expansion improved search performance in the JURIS-TCU, benefiting searches in all analyzed metrics NDCG@10, PRECISION@50, and RECALL@100. The results also emphasize the importance of incorporating semantic similarity into search pipelines in the initial retrieval stage, alongside BM25, to address vocabulary mismatch cases while maintaining the effectiveness of BM25.

Code and data: <https://github.com/marcusborela/ind-ir>

1. Introdução

Baseado em MANNING (2019), *Information retrieval* (IR) pode ser compreendida como uma busca em documentos de natureza não estruturada (geralmente texto) por algo que satisfaça uma necessidade de informação. Um dos seus maiores desafios é o de incompatibilidade vocabular (*vocabulary mismatch*), identificado por Furnas et al. (1987), no qual os usuários empregam termos de consulta distintos daqueles utilizados nos documentos relevantes (por exemplo, "automóvel" ao invés de "carro").

Segundo LIN et al. (2022), existem três abordagens gerais para enfrentar esse desafio: enriquecer as representações de consulta para melhor corresponder às representações de documento, enriquecer as representações de documento para melhor corresponder às representações de consulta, e tentativas de ir além da correspondência exata de termos. Nessas tentativas, os autores afirmam que há uma crescente interseção entre as áreas de IR e *Natural Language Processing* (NLP). Saber se dois textos "significam a mesma coisa" (questão típica de NLP) está intimamente relacionado a saber se um texto é relevante para uma consulta (IR).

LIN (2022) sintetiza a busca como uma função de comparação que calcula as pontuações de pares de consulta e de documentos, ambos mapeados por codificadores para um espaço de representação, N_q e N_d , respectivamente. Segundo o autor, os bibliotecários, em suas atividades de indexação de textos, foram os que primeiro desempenharam o papel de codificação de documentos e de consultas, mapeando textos para um vetor *multi-hot* em que cada dimensão representa um descritor de conteúdo.

Atribuir um descritor (indexar) é declarar que ele possui alto grau de relevância para o conteúdo do documento, que o seu significado está fortemente associado a um conceito incorporado ao documento (Tinker, 1966, citado por Lancaster, 2004).

Segundo Brasil (2022), a indexação é uma operação que consiste em identificar os principais conceitos que caracterizam o conteúdo de um texto para obtenção de uma representação da informação relevante por meio de linguagem controlada e padronizada que

define os termos descritores. Ela compreende duas etapas distintas: a análise conceitual do assunto e a tradução dos conceitos em descritores existentes no vocabulário controlado.

Lancaster (2004) afirma que um mesmo texto pode ser indexado de formas distintas por indexadores diferentes, não só porque há uma dependência do vocabulário controlado utilizado (etapa de tradução) mas também porque a análise conceitual é moldada para ajustar os conceitos a uma clientela específica (usuários de um sistema de busca)¹.

Para o aperfeiçoamento da dupla IR e NLP, é necessário existirem *datasets* (conjuntos de dados) tanto para treinamentos de modelos quanto para avaliação de resultados quanto à relevância dos documentos retornados para uma busca na avaliação de novos métodos e modelos. É sabida a carência desses *datasets* na língua portuguesa.

São objetivos do presente trabalho:

1. construir um indexador automático para o sistema de enunciados de jurisprudências do TCU;
2. implementar e avaliar possíveis usos desse indexador na melhoria das buscas do sistema base. Daí o nome do projeto INDIR, por englobar a construção de um indexador (IND) e a sua aplicação em buscas (IR);
3. elaborar e publicar *dataset* JURIS-TCU-INDEX contemplando as indexações;
4. elaborar e publicar *dataset* JURIS-TCU com dados de buscas no sistema base de enunciados do TCU.

Os *datasets* publicados compõem um importante conjunto de informações para a Administração Pública, a base da Jurisprudência selecionada do Tribunal de Contas da União (TCU), que contém os enunciados de relevância jurisprudencial, conforme Brasil (2016), e o VCE (Vocabulário de Controle Externo), tesauro do TCU usado na indexação dos enunciados.

Na próxima seção, descreveremos os passos planejados para o alcance desses objetivos e na seção seguinte apresentaremos os experimentos realizados e seus resultados.

2. Metodologia

2.1. Extração e Tratamento de Dados

a. Vocabulário Controlado para indexações

Os termos do vocabulário controlado do TCU (VCE – Vocabulário de Controle Externo), suas definições, sinônimos, termos relacionados, traduções e outros metadados foram extraídos do Sistema VCE (BRASIL, 2019). Esses dados estão disponibilizados no arquivo doc.csv do *dataset* JURIS-TCU-INDEX, disponibilizado no repositório do projeto.

b. Jurisprudência Selecionada

Os enunciados da Jurisprudência foram extraídos de *view* de banco de dados do Sistema de Enunciados de Jurisprudência (e-Juris), sistema de uso interno da instituição. São disponibilizados publicamente pela Pesquisa Integrada do TCU à base de Jurisprudência Selecionada (link <https://pesquisa.apps.tcu.gov.br>). Esses dados formam o arquivo query.csv do JURIS-TCU-INDEX e o doc.csv do *dataset* JURIS-TCU, disponibilizados no repositório.

Juntamente com os enunciados, estão disponíveis sua indexação em termos do VCE (qrel.csv de JURIS-TCU-INDEX). Além de uma indexação livre (que nos dados atuais contempla até 9 termos), há três indexações obrigatórias e hierárquicas: área (entre 10 termos fixos do VCE), tema e subtema (quaisquer termos do VCE).

¹ Muitas instituições governamentais e privadas possuem tesouros. Esse trabalho tenta dar mais um bom uso para esses tesouros muitas vezes escondidos. Por isso, o nome INDIR também é uma referência ao personagem Indiana Jones da série de filmes criada por George Lucas, sempre na busca por relíquias e pelo seu uso para o bem e não para o mal, por nazistas. No caso deste projeto, para o bem das consultas!

c. Histórico de uso da Pesquisa à Jurisprudência

Os dados de acesso à base de Jurisprudência foram extraídos a partir de base de log da Pesquisa Integrada do TCU citada, uma base interna e cujos dados são sigilosos. Diversas consultas foram realizadas nesta base, sob os seguintes recortes:

- período de 12 meses, compreendendo acessos de junho de 2022 a maio de 2023;
- apenas consultas específicas realizadas na base de Jurisprudência Seleccionada;
- excluídas as consultas por todos os documentos da base (query="*");
- excluídas as consultas empregando operadores de proximidade.

Os dados recortados foram consolidados, agregados em consultas e documentos, anonimizados, e disponibilizados nos arquivos (LOG-JURIS-TCU), publicados no repositório:

- query.csv – buscas efetuadas, em ordem decrescente de execuções;
- doc-hits.csv – enunciados acessados, em ordem decrescente de número de acessos;
- query-doc-hits.csv – cruzamento dos enunciados com expressões de busca usadas.

2.2. Produção do *Dataset* de indexação

O *dataset* JURIS-TCU-INDEX serve como base de treinamento para a indexação automática dos enunciados. Foi construído de forma direta a partir dos dados extraídos e tratados dos sistemas de jurisprudência e VCE, e é composto pelos três arquivos a seguir:

- query.csv – enunciados da jurisprudência;
- doc.csv – documentos formados pela concatenação de informações de termos do VCE e seus metadados, como definição, sinônimos, termos relacionados e traduções;
- qrel.csv – indexações dos enunciados da jurisprudência por termos do VCE.

2.3. Produção do *Dataset* de avaliação

O *dataset* JURIS-TCU serve como base de avaliação para mecanismos de busca dos enunciados. É composto pelos três arquivos a seguir:

- query.csv – seleção de consultas à base de jurisprudência;
- doc.csv – os enunciados da jurisprudência;
- qrel.csv – escores de relevância com que determinado enunciado responde a uma consulta.

a. Seleção de consultas a partir do histórico de uso da Pesquisa

A partir do arquivo de log contendo buscas efetuadas, foram selecionadas as consultas com maior quantidade de execuções. Foram retiradas algumas delas, por exemplo, muito semelhantes a outras selecionadas, e ajustadas algumas expressões para que atendessem a uma busca mais genérica e não tanto enviesada para o sistema atual de busca, chegando a um grupo de 50 expressões de busca que denominamos grupo 1, correspondendo aos ids de 1 a 50.

b. Produção de consultas a partir dos enunciados

Os enunciados que formam o arquivo doc.csv são todos os presentes no sistema e-Juris, extraídos e tratados.

A partir do arquivo de log contendo enunciados acessados (doc-hits.csv), foram selecionados os 56 enunciados com maior quantidade de acessos. Destes, 6 pares foram agrupados, chegando a uma lista final de 50 enunciados mestres. Foi elaborada uma sequência de instruções ao LLM (ver prompt na Figura 1), para que produzisse 5 perguntas que pudessem ser respondidas por cada um destes enunciados. 5 perguntas produzidas passaram por um

*Elabore até 5 perguntas curtas e diretas que possam ser respondidas a partir do enunciado a seguir:
Em regra, o pregão é a modalidade de licitação adequada para a concessão remunerada de uso de bens públicos, com critério de julgamento pela maior oferta em lances sucessivos.”*

Figura 1 - Prompt usado para elaboração de perguntas baseado em enunciados.

processo de revisão e de ajuste manual gerando uma pergunta final, formando o grupo 3 com 50 perguntas completas, correspondendo aos ids de 101 a 150.

Essas 50 perguntas completas foram transformadas em expressões de busca, forma mais comumente empregada na Pesquisa Integrada do TCU, pela supressão de alguns termos e modificação de outros, chegando a um novo grupo de 50 expressões de busca (grupo 2, ids de 51 a 100). Os arquivos gerados estão registrados em pasta do projeto LLM-JURIS-TCU em *data/llm_juris_tcu*.

c. Produção de escores de relevância

Como base em cada uma das 150 consultas produzidas, foram realizadas diversas buscas sobre enunciados de jurisprudência, com resultados registrados em pasta do projeto *data/search/juris_tcu_basic*. Para cada consulta, foram selecionados 10 enunciados mais bem ranqueados (documentos positivos) e 5 enunciados selecionados aleatoriamente (documentos negativos).

Para cada um dos 15 enunciados para cada consulta, foi elaborada uma instrução a LLM (prompt na Figura 2), para que produzisse um escore de relevância (de 0 a 3) para o enunciado em relação à consulta (no caso do grupo 3, relacionada como pergunta).

Você é um especialista na jurisprudência do Tribunal de Contas da União com o objetivo de avaliar se um enunciado de jurisprudência responde a uma pergunta.

Deve retornar um valor de escore de 0 a 3, sendo:

0 - irrelevante - o enunciado não responde a pergunta;

1 - relacionado - o enunciado apenas está no tópico da pergunta;

2 - relevante - o enunciado responde parcialmente a pergunta;

3 - altamente relevante - o enunciado responde a pergunta, tratando completamente de suas nuances.

Em seguida, explique a razão para a escolha do escore.

Por favor, responda no formato JSON, contendo as chaves Razão e Score;

o valor de Razão deve ser a motivação para a escolha do score;

o valor de Score deve ser o valor do score atribuído.

Pergunta: Qual é a modalidade de licitação adequada para a concessão remunerada de uso de bens públicos?

Enunciado de jurisprudência: Em regra, o pregão é a modalidade de licitação adequada para a concessão remunerada de uso de bens públicos, com critério de julgamento pela maior oferta em lances sucessivos.

Response:

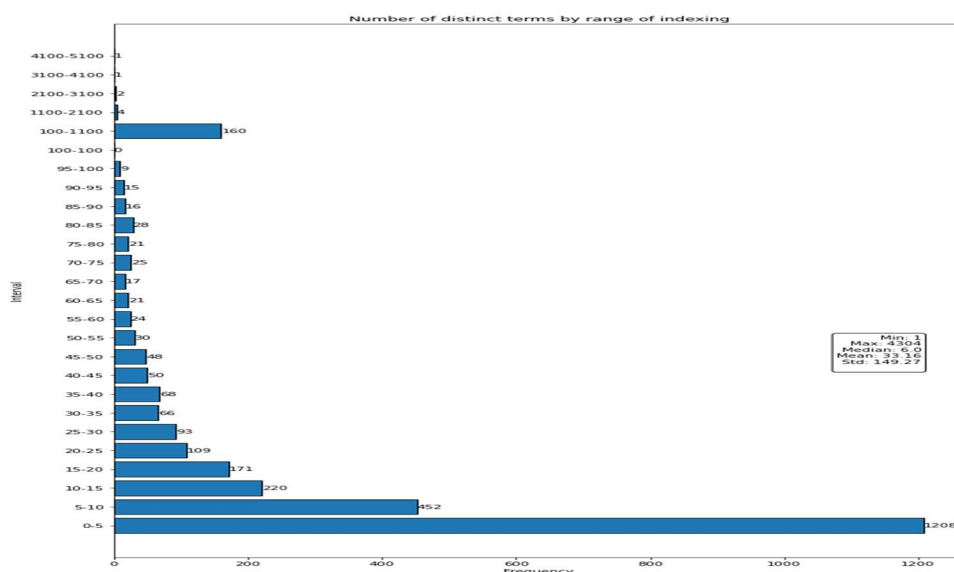
{'Razão': 'O enunciado responde diretamente à pergunta, indicando que a modalidade de licitação adequada para a concessão remunerada de uso de bens públicos é o pregão, com critério de julgamento pela maior oferta em lances sucessivos.', 'Score': 3}

Figura 2 - prompt usado para produção de escore de relevância.

2.4. Construção do indexador INDIR usando dados do JURIS-TCU-INDEX

Tratou-se da elaboração de um pipeline com estágio de rerranqueamento que a partir de enunciados (consultas no *dataset* JURIS-TCU-INDEX e documentos no *dataset* JURIS-TCU) retornasse os termos mais relevantes para sua indexação. Esse é um caso exemplar que integra, com visto na introdução, IR e NLP.

A impossibilidade de se usar um classificador (NLP) se deu pela grande quantidade de classes possíveis (13205, tendo sido usados até o momento 2859) e o baixo número de exemplos por classe para se treinar um modelo. Essa situação é evidenciada na Figura 3: uma grande parte dos termos não têm mais do que 5 indexações. A possibilidade de se usar classificação por grupos de termos também foi descartada pois o número de classes seria 11893, tendo cada agrupamento de 1 a 20 indexações, com média de 1.35 indexações (desvio padrão 1.13).



Na Figura 3 abaixo, pode-se observar as principais indexações existentes em enunciados por tema (top 50 termos). Ela evidencia o desbalanceamento dos dados por classe.

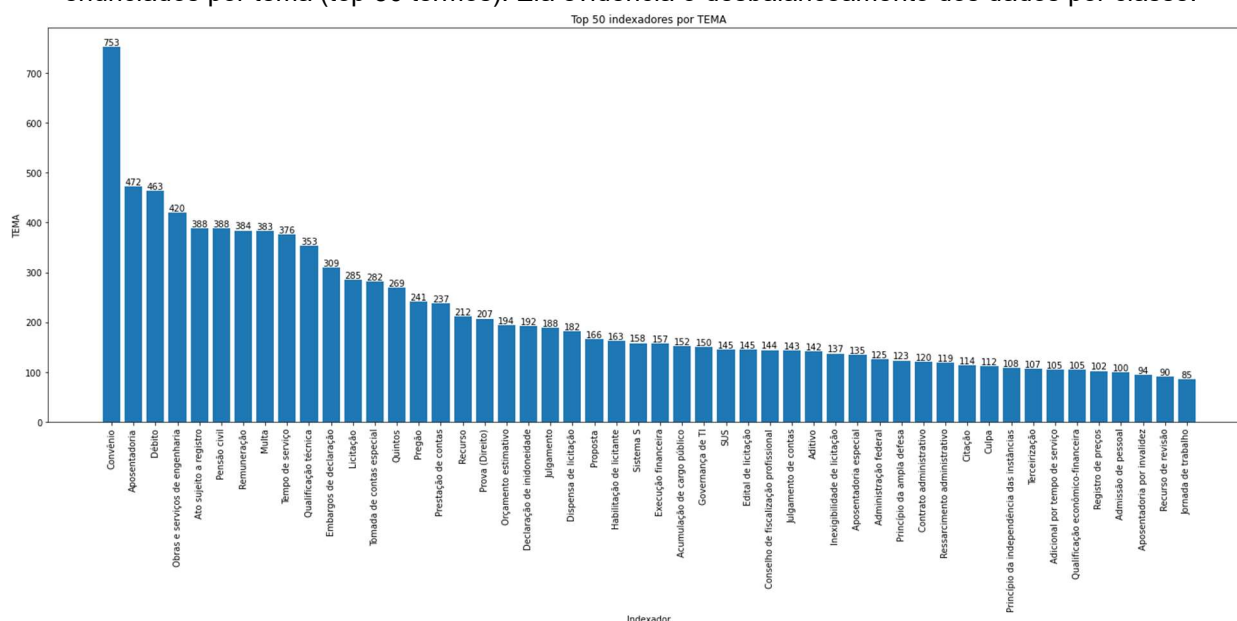


Figura 5 - Indexações por tema (top-50)

Na construção do pipeline de busca foram experimentadas versões com apenas um estágio (*retriever*) e com dois estágios (incluindo rerranqueamento). No primeiro estágio foram testados BM25, similaridade de sentença (STS) e uma combinação entre essas (JOIN). Como ranqueador foram experimentados três modelos disponíveis para a língua portuguesa e treinados dois desses modelos a partir de dados de treinamento gerados.

A qualidade do pipeline foi avaliada a partir de duas métricas: RANK1 e NDCG@12. Utilizou-se a métrica RANK1 pois ela evidencia a primeira posição de um documento relevante entre os retornados. Para um usuário que procura um termo para indexar um texto, essa métrica é uma boa indicativa da qualidade, pois pode evidenciar que o usuário não precisará navegar entre opções retornadas. A métrica NDCG@12 complementa a avaliação ao avaliar a proporção dos documentos relevantes que são retornados e sua ordem. O número doze equivale ao número máximo de termos usados, até o momento, na indexação de enunciados (*ground truth*).

2.5. Experimentação de uso do INDIR em buscas do JURIS-TCU

As versões de pipelines construídas na seção anterior que alcançaram melhor desempenho em JURIS-TCU-INDEX foram testadas para a melhoria da performance das buscas em JURIS-TCU. Foram experimentadas três abordagens de uso: expansão de consultas, expansão de documentos e como rerranqueador.

Como rerranqueador, além do próprio INDIR (PTT5-BASE-INDIR-400) foi empregado o modelo original, PTT5-BASE. Também foi experimentada uma inversão no prompt passado a cada um desses modelos, para se avaliar se o INDIR se beneficiaria de seu treinamento tendo os enunciados como consultas no JURIS-TCU-INDEX, já que no JURIS-TCU esses são documentos. Como primeiro estágio foram usados sempre BM25, STS e JOIN. Também foram experimentados pipelines sem rerranqueamento.

Considerando que Nogueira et al. (2019) melhoraram as buscas com a expansão de documentos com textos, perguntas geradas automaticamente, representativos do documento e que, como visto na introdução, também há a abordagem de se expandir as consultas, foram implementadas essas duas opções. Os documentos foram expandidos com os dados dos termos do VCE considerando-se duas origens: indexação feita pelo usuário (qrel em JURIS-TCU-INDEX) e indexação predita pelo próprio INDIR (PTT5-BASE-INDIR-400). Foram criadas versões diferentes de documentos conforme os metadados do VCE considerados na expansão: só termos, termos e seus sinônimos (quando existentes), termos e os termos com ele relacionados (quando existentes), e a união desses dois últimos, termos e seus sinônimos e termos

relacionados. As consultas foram expandidas em até dez termos sugeridos pelo INDIR. Também foi experimentada expansão das consultas com mais metadados dos termos: sinônimos ou termos relacionados (no máximo um grupo, de preferência os sinônimos).

Foram consideradas as métricas NDCG@10, PRECISION@50 e RECALL@100. Elas, em conjunto, cercam dimensões complementares de um sistema de busca. PRECISION@50 indica o grau de acerto do modelo entre os primeiros cinquenta documentos retornados, enquanto RECALL@100 indica o percentual entre os documentos relevantes efetivamente retornados entre os cem primeiros. NDCG@10 vai mais no detalhe, atentando para a ordem e para a posição dos relevantes entre os dez primeiros retornados. Para o cálculo das métricas, foram considerados relevantes os graus 2 e 3, sendo 0 e 1 não relevantes.

3. Experimentos

3.1. Produção e publicação do *dataset* de Indexação – JURIS-TCU-INDEX

Foi produzido o *dataset* JURIS-TCU-INDEX, formado por 16057 consultas (query.csv), que são os enunciados de jurisprudência, 13205 documentos (doc.csv), que são os termos do VCE e 94809 registros de relevância (qrel.csv), que são as indexações de enunciados por termos do VCE. As figuras da seção 2.4 permitem uma visão quantitativa de seus dados. A Figura 6 evidencia que o tamanho dos textos dos documentos (doc.csv) é relativamente baixo assim como as consultas (Figura 7). Isso favorece seu uso por modelos com limitação no número de tokens no texto de entrada (caso dos rankers utilizados no trabalho, que possuem limite de 512). O *dataset* encontra-se publicado no github do projeto (data\juris-tcu-index).

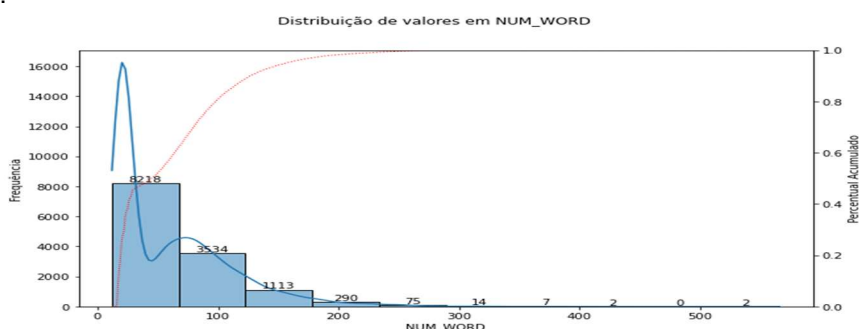


Figura 6 - Histograma de número de palavras em textos do VCE (documentos do JURIS-TCU-INDEX)

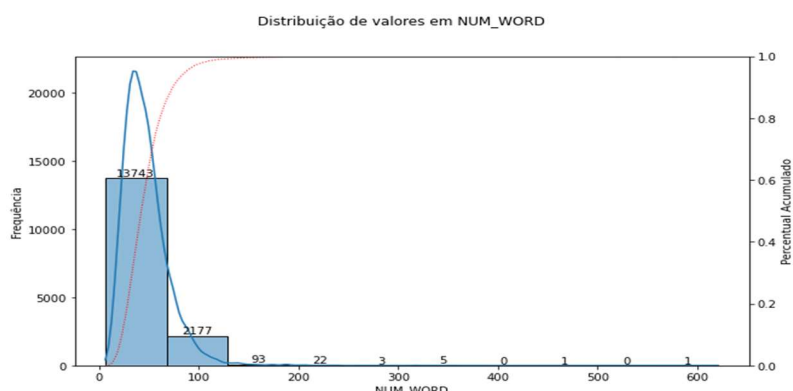


Figura 7 - Histograma de número de palavras em enunciados (consultas do JURIS-TCU-INDEX, documentos no JURIS-TCU)

3.2. Produção e publicação do *dataset* de Avaliação – JURIS-TCU

Foi produzido o *dataset* JURIS-TCU, formado por 150 consultas (query.csv), organizadas nos 3 grupos citados anteriormente, 16057 enunciados (doc.csv) e 2250 avaliações

de relevância, 15 por consulta (qrel.csv). O *dataset* encontra-se publicado no github do projeto (*data\juris-tcu*).

O LLM empregado para geração das perguntas e geração do escore das avaliações foi o GPT 4.0, produzido a partir de serviço disponibilizado em nuvem privada Azure.

Para realização de buscas, foram utilizados cinco pipelines diferentes:

1. BM25: busca esparsa BM25;
2. BM25 + Reranker: busca BM25 retornando 300 registros, seguida de rerranqueamento;
3. STS: busca densa por similaridade de sentença;
4. STS + Reranker: busca STS (topk também de 300), seguida de rerranqueamento;
5. (BM25 | STS) + Reranker: estágio 1 com JOIN (união) dos top-300 de cada BM25 e de STS, seguida de rerranqueamento.

Para todos os pipelines com rerranqueamento, foi empregado o modelo T5-like unicamp-dl/mt5-3B-mmarco-en-pt não ajustado para os dados do domínio. Nas buscas por similaridade, os embeddings foram gerados pelo modelo Bert-like rufimelo/Legal-BERTimbau-sts-large-ma-v3. Esse modelo foi usado em todas as abordagens com STS neste trabalho.

A Figura 8 evidencia que pipelines que empregam BM25 e que empregam rerranqueador tendem a mostrar resultados mais similares entre todos os grupos de consultas, considerando-se o percentual de enunciados coincidentes retornados. Observa-se também maior diferença de resultados para consultas do grupo 1, em especial quanto ao uso do rerranqueador após BM25, mas também as comparações com o pipeline STS+Reranker.

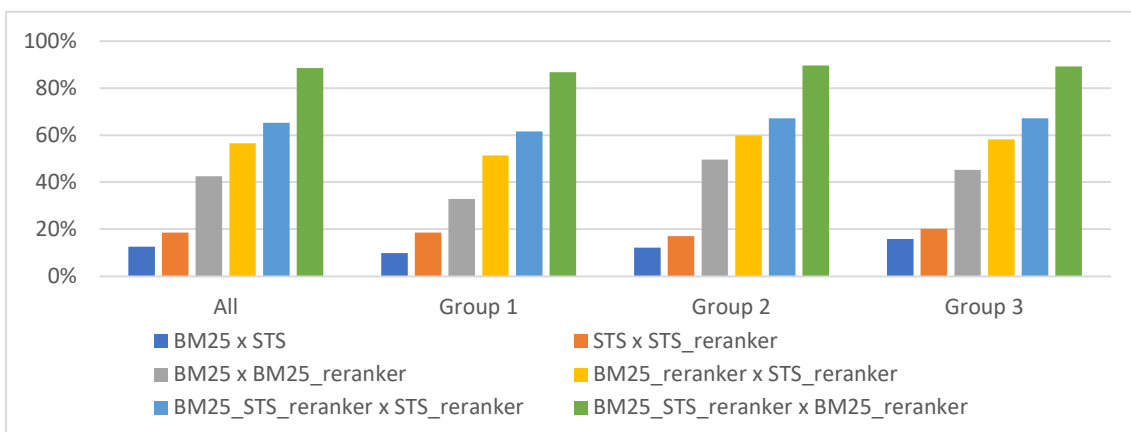


Figura 8 - Similaridade de resultados entre pipelines de busca

Como visto na seção 2.3, as consultas do grupo 3 foram geradas a partir de enunciados mais acessados da base. Na próxima figura, pode-se observar se esses enunciados originais são reencontrados após cada pipeline de busca, e em qual posição no ranking. Observa-se que o percentual de enunciados perdidos em pipelines baseados em busca densa é bem maior, e que, nesse caso, o rerranqueador melhora bastante a colocação desses enunciados. Já nos pipelines que possuem busca esparsa BM25, há um percentual muito maior de enunciados

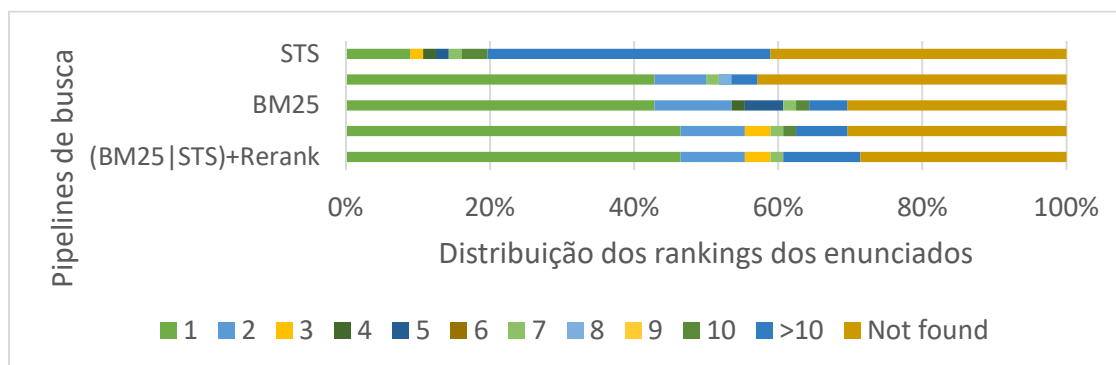


Figura 9 - Enunciados reencontrados - rankings

reencontrados e em primeira posição no ranking, mesmo sem o rerranqueador, que tem um impacto relativamente menor.

Os 10 documentos candidatos a positivos foram extraídos dos top-10 do pipeline 5, (BM25 | STS) + Reranker. Os 5 documentos candidatos a negativos foram selecionados a partir do pipeline 1, BM25, aleatoriamente dos top-1000, excluídos os 10 documentos candidatos a positivos.

Na figura 8 pode-se observar a distribuição de escores nos documentos, por método de seleção. Os documentos candidatos a positivos receberam em média avaliações mais relevantes, e os candidatos a negativos o contrário. No consolidado, há uma distribuição relativamente próxima entre todos os escores possíveis. Na figura 9, pode-se observar a variação de escores médios entre os grupos de consulta produzidos: o grupo 1 possui maiores médias de escores, o grupo 2 o contrário.

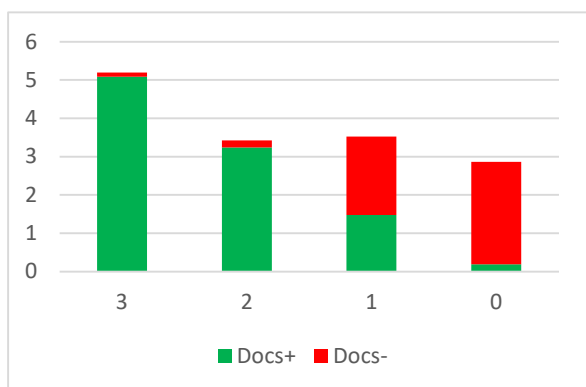


Figura 8 - Média de avaliações por escore

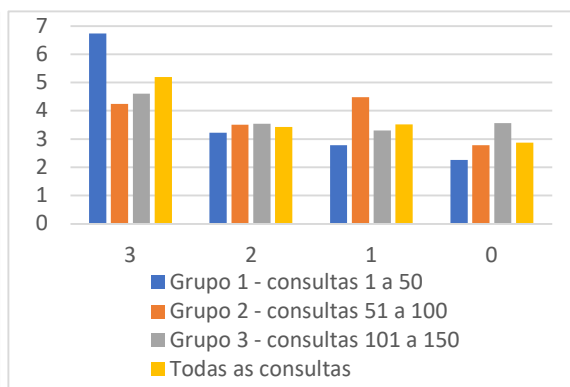


Figura 9 - Média de escores por grupo de consultas

Outra comparação realizada pode ser visualizada na figura 10. Para cada consulta do *dataset*, dentre os top-20 enunciados retornados pela busca atualmente em produção (que emprega BM25 e operador “e” implícito), a média de retornados é identificada na coluna TCU_SEARCH. Observa-se que a pesquisa não consegue trazer os 20 documentos para todas as consultas, a média é próxima a 11. Dos 15 enunciados avaliados no *dataset* (TCU_JURIS), menos de 4 coincidem com a busca atual. As demais colunas identificam os pipelines de busca empregados, que partem de uma quantidade de documentos de 300 a 1000. Claramente, os pipelines que empregam BM25 são os que conseguem cobrir a maior parte dos enunciados trazidos na busca atual.

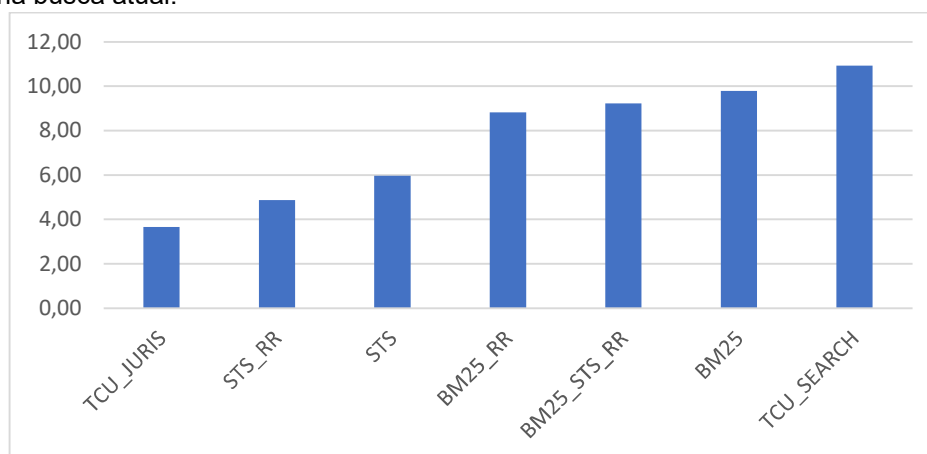


Figura 10 - Documentos retornados entre os top-20 da Pesquisa do TCU

3.3. Construção do Indexador INDIR

Os dados dos textos dos termos VCE foram carregados em um índice local Elastic Search. Foram experimentados diversos pipelines de busca. O desempenho foi avaliado usando-se um subconjunto de avaliação de 100 queries obtidas com o comando sample do pandas

(semente 123), salvo em `data\train_juris_tcu_index\juris_tcu_index_validation_query_id.csv`. Na entrada para o rerranqueador limitou-se o tamanho máximo da consulta em 350 tokens, de forma a sobrar, nos casos extremos, 162 tokens para os documentos, geralmente menores (ver Figuras 6 e 7 acima). O código foi implementado em python usando-se a biblioteca `haystack`².

As configurações experimentadas de *retriever* envolveram BM25, STS e JOIN, uma combinação entre essas. Além de versões de mono-estágio (só o *retriever*), também foram experimentados pipelines com dois estágios, usando-se inicialmente como ranqueadores dois modelos disponíveis para a língua portuguesa no diretório *unicamp-dl* do Hugging Face: *mMiniLM-L6-v2-pt-v2 (XLMRobertaForSequenceClassification)* e *ptt5-base-pt-msmarco-100k-v2 (T5ForConditionalGeneration)*. Em todas as configurações, utilizou-se o limite de 300 registros (top-k). Como não lograram um bom desempenho, experimentou-se então um modelo maior do mesmo repositório *mt5-3B-mmarco-en-pt (T5ForConditionalGeneration)* que alcançou resultados melhores, com diferença na casa de 10 pontos em média no NDCG@12 sobre o melhor resultado até então.

Porém, como os resultados ainda não estavam satisfatórios (NDCG@12 de 37.6% e RANK1 de 4.5 no critério “total”, união das 4 categorias³ área, tema, subtema e extra) optou-se por se treinar um modelo específico para a missão, denominado INDIR.

Por falta de disponibilidade de recursos (computacionais e tempo), não treinamos o modelo maior de 3B de parâmetros, mas os 2 outros ranqueadores acima apresentados que de forma resumida chamamos de MINILM e PPT5-BASE. A esperança era que uma boa base de treinamento levasse a um resultado frutífero. O primeiro a ser treinado foi o menor, MINILM, até para uma prototipação e validação dos dados de treinamento.

A geração dos dados de treinamento foi uma experiência à parte. Foram inicialmente quatro gerações que não levaram a modelos com melhores desempenhos. Variou-se a quantidade limite de registros por classe (termo VCE): sem limite e com limite em 50 ou 100 exemplos positivos (de relevantes). Os exemplos negativos (não relevantes) eram obtidos considerando-se os primeiros retornados que já não fossem positivos pelo melhor pipeline de busca até então, BM25 e o rerranqueador *mt5-3B*. Ou seja, os negativos não eram tão negativos assim. Eram “quase positivos”. Os resultados demonstravam que somente no critério “área”, que só tem 10 classes e bem distintas, os modelos treinados conseguiram bons resultados (NDCG@12 acima de 50%). Então veio o momento de eureka, mas de algo já sabido na literatura, mas em que se estava falhando: o modelo estava acertando na área pois os exemplos negativos eram bem diferentes dos positivos, como “pessoal” e “licitação”. E errava nos outros casos pois os exemplos negativos eram próximos dos positivos, tipo “Contrato” e “Contratação” (exemplos hipotéticos). E, por consequência, os modelos não aprendiam como diferenciar entre casos tão próximos. Gerou-se então uma nova versão de dados de treinamento com 405198 registros, obtendo-se para cada positivo cinco negativos de uma amostragem entre os primeiros mil retornados em uma busca BM25. Houve um esforço de se evitar um grande desbalanceamento entre as classes limitando-se a 50 positivos por classe, e 200 no total das 4 classes⁴. Desses registros, foram retirados 2460 registros de indexação das 100 queries de avaliação. E, durante o treinamento, foram separados 2014 registros para uma base de

² Ela foi estendida localmente com um novo módulo (`monot5_limit.py`) para uso com modelos SEQ2SEQ (arquitetura T5) e foi implementado um tratamento de limite parametrizado de tamanho das queries para o módulo de classificação previamente existente (`sentence_transformers_limit.py`)

³ As buscas, por opção de projeto, tiveram um filtro que restringia a termos do VCE que já foram usados na indexação da classe em questão (área, tema, subtema e extra). Posteriormente ao término do projeto percebeu-se a retirada dessa restrição não alterou o desempenho. Por exemplo, para a classe área, que na prática é um problema de classificação, o filtro limitava a busca entre os únicos dez termos já usados. Adicionalmente, experimentou-se o critério “total_gte_5” para uma busca mais restritiva (*gte*, “greater than or equal”), restringindo a termos já usados pelo menos cinco vezes. Também houve um filtro adicional restringindo a termos do VCE que não fossem “localização”, “verbo” e “organização”, uma vez que esses foram usados apenas 894 vezes das 94809 indexações realizadas (0.94%) envolvem 53,5% dos descritores do VCE. Também não foram analisadas as buscas que devido ao limite de 300 (top-k no retriever) não trouxeram documentos relevantes. Ilustrando, no critério total, houve um caso. No tema, 2 (STS) e 4 (BM25).

⁴ Ver detalhes no código em `code\generate_train_data_juris_tcu_index.ipynb`. Dados em `data\train_juris_tcu_index\train_data_juris_tcu_index.7z`.

validação. O treinamento dos modelos foi realizado em uma máquina local com GPU 3090 de 24gb com rastro no Neptune.ai⁵.

A Figura 9 permite uma melhor visualização das diferenças alcançadas entre os modelos INDIR treinados e os seus concorrentes na métrica Rank1 nos dados de avaliação⁶. Os modelos treinados possuem sufixo INDIR_XX onde XX indica o percentual de uma época treinado. Treinou-se por 4 épocas o MINILM e o PTT5-BASE, por isso MINILM_INDIR_400 e PTT5_INDIR_400. Também foi experimentado o PTT5_INDIR_106 treinado em 106% de uma época. Percebe-se que o aumento do número de época de 106% para 400% não levou a uma grande diferença entre os modelos PTT5-INDIR. Inclusive, na métrica NDCG@12 no critério tema, o PTT5-INDIR-106 superou o PTT5-INDIR-400 (caso único). Pode ser um indicativo de que treinar o PTT5-INDIR-400 por mais épocas pode não levar a melhores resultados.

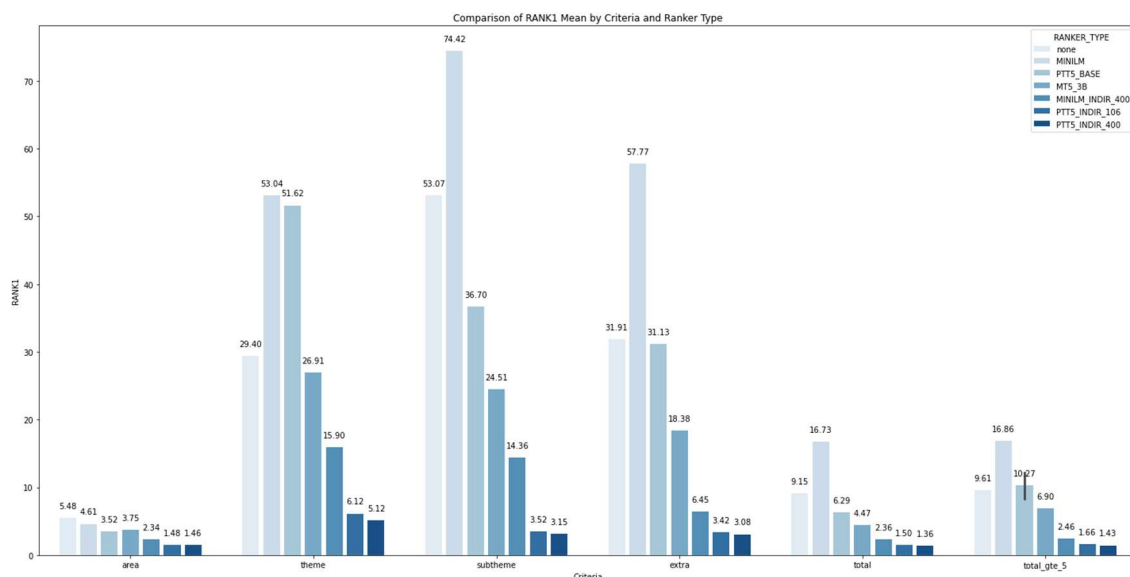


Figura 10 - Comparativo da métrica RANK1 em JURIS-TCU-INDEX

Os resultados com dados de avaliação nas métricas consideradas estão detalhados na Tabela 1. Como pode ser observado, os resultados alcançados tanto pelo MINILM-INDIR e principalmente pelos PTT5-INDIR foram bem superiores aos dos outros modelos (e da versão sem ranqueador, com “----” na tabela). O MINILM-INDIR-400 venceu o MT5-3B em todos os critérios na métrica Rank1 e em 4 dos 7 critérios no NDCG@12. E os PTT5-INDIR superaram o modelo maior MT5-3B em todos os 7 critérios, com uma diferença em média de 17.19 (INDIR-106) e 18.31 (INDIR-400). Comprovou-se assim que, para os modelos usados, um treinamento com dados representativos pode promover ótimos resultados, levando modelos bem menores (MINILM e PTT5-BASE com 106 milhões e 222 milhões de parâmetros respectivamente) a superarem modelos bem maiores (MT5-3B com 3 bilhões).

3.4. Uso do INDIR e indexações em buscas no JURIS-TCU

Foram experimentadas três abordagens de uso do INDIR e de dados de indexação para se melhorar as buscas: em expansões de consultas, em expansões de documentos e como rerranqueador.

⁵ Mais detalhes sobre a configuração do treinamento podem ser obtidos em [finetune_reranker_mt5_seq2seq_colab.ipynb](https://app.neptune.ai/marcusborela/IA386DD/runs/details?viewId=standard-view&shortId=IAD-106&type=run) (PTT5) e [finetuning_reranker_sequence_classification_bert_like.ipynb](https://app.neptune.ai/marcusborela/IA386DD/runs/details?viewId=standard-view&shortId=IAD-108&type=run) (MINILM). Rastros em <https://app.neptune.ai/marcusborela/IA386DD/runs/details?viewId=standard-view&shortId=IAD-106&type=run> (MINILM) e <https://app.neptune.ai/marcusborela/IA386DD/runs/details?viewId=standard-view&shortId=IAD-108&type=run> (PTT5, parcial; na segunda etapa, única rastreada, houve um acréscimo no número de registros de validação para 4028 registros).

⁶ Os resultados podem ser encontrados em `data\search\juris-tcu-index`.

As consultas foram expandidas em até dez termos preditos pelo PTT5-INDIR-400. O termo EXPQ_CNT indica nas tabelas que se seguem a quantidade de expansões. também foi experimentada expansão das consultas com metadados extras dos termos: preferencialmente sinônimos, ou, na ausência desses, termos relacionados, caso existam. A Tabela 2 demonstra as expansões realizadas para a primeira consulta no query.csv (JURIS-TCU). Cada coluna representa um tipo de expansão (EXPD_TYPE). Na primeira, só com o Termo, usou-se no primeiro estágio do pipeline um JOIN entre BM25 e STS, 30 registros de cada. Na segunda, Termo com extra, mudou-se os números do pipeline JOIN usado para 10 registros do BM25 e 30 do STS, que nos testes visuais realizados demonstraram trazer termos mais relevantes.

SEARCH FILTER	RANKER TYPE	NDCG	RANK1	TIME SPENT
area	----	50,06	5,48	0,01
	MINILM	51,97	4,61	0,04
	PTT5_BASE	60,55	3,52	0,14
	MT5_3B	62,09	3,75	0,69
	MINILM_INDIR_400	74,19	2,34	0,04
	PTT5_INDIR_106	87,26	1,48	0,16
	PTT5_INDIR_400	87,93	1,46	0,13
theme	----	40,08	29,40	0,13
	MINILM	32,50	53,04	0,76
	PTT5_BASE	35,47	51,62	3,04
	MT5_3B	53,78	26,91	20,43
	MINILM_INDIR_400	48,59	15,90	0,78
	PTT5_INDIR_106	63,51	6,13	3,89
	PTT5_INDIR_400	64,84	5,12	3,05
subtheme	----	23,53	53,07	0,13
	MINILM	22,94	74,42	0,79
	PTT5_BASE	26,14	36,70	3,07
	MT5_3B	35,30	24,51	20,91
	MINILM_INDIR_400	29,20	14,36	0,80
	PTT5_INDIR_106	46,18	3,52	3,99
	PTT5_INDIR_400	47,86	3,15	3,05
extra	----	19,34	31,91	0,13
	MINILM	11,50	57,77	0,80
	PTT5_BASE	14,09	31,13	3,13
	MT5_3B	26,72	18,38	21,22
	MINILM_INDIR_400	33,85	6,45	0,81
	PTT5_INDIR_106	43,89	3,42	3,87
	PTT5_INDIR_400	42,85	3,08	3,10
total	----	26,74	9,15	0,13
	MINILM	23,08	16,73	0,80
	PTT5_BASE	23,09	6,29	3,12
	MT5_3B	37,52	4,48	21,16
	MINILM_INDIR_400	43,71	2,36	0,81
	PTT5_INDIR_106	57,67	1,51	4,02
	PTT5_INDIR_400	59,60	1,36	3,09
total_gte_5	----	28,33	9,61	0,13
	MINILM	24,76	16,86	0,80
	PTT5_BASE	24,56	8,32	3,16
	MT5_3B	37,57	6,90	21,32
	MINILM_INDIR_400	43,49	2,46	0,81
	PTT5_INDIR_106	57,61	1,66	4,03
	PTT5_INDIR_400	59,82	1,43	3,13

Tabela 1 - Performance dos indexadores (JURIS-TCU-INDEX) com retriever BM25, top-k 300

Cabe ressaltar que os INDIR's foram treinados para indexar enunciados e não consultas como as expressões da primeira linha da tabela. Mas, em uma verificação visual, os primeiros termos acrescentados parecem ser relevantes para cada consulta.

Expansão	Termo	Termo + extra (sinônimo ou termo relacionado)
0	técnica e preço	técnica e preço
1	técnica e preço	técnica e preço
2	técnica e preço - Licitação de técnica e preço	técnica e preço - Licitação de técnica e preço - Técnica e preço e Licitação técnica e preço
3	técnica e preço - Licitação de técnica e preço, Proposta técnica	técnica e preço - Licitação de técnica e preço - Técnica e preço e Licitação técnica e preço, Proposta técnica -- Proposta de preço
4	técnica e preço - Licitação de técnica e preço, Proposta técnica, Preço de mercado	técnica e preço - Licitação de técnica e preço - Técnica e preço e Licitação técnica e preço, Proposta técnica -- Proposta de preço, Proposta de preço - Proposta comercial

Tabela 2 - Exemplos de expansões para consultas (JURIS-TCU) usando PTT5-INDIR-400 (Parcial, sem expansões de 5 a 10)

Os documentos foram expandidos com os dados dos termos do VCE considerando-se duas origens (EXPD TYPE): indexação feita pelo usuário (dados do qrel em JURIS-TCU-INDEX) e indexação com predições, de um a cinco, pelo próprio PTT5-BASE-INDIR-400. Foram criadas versões diferentes de documentos conforme os metadados do VCE considerados na expansão (EXPD_VAL): só termos (*term*), termos e seus sinônimos quando existentes (*+syn*), termos e os termos com ele relacionados quando existentes (*+rel*), e a união desses dois últimos, termos e seus sinônimos e termos relacionados (*+syn+rel*). A Tabela 3 demonstra as 4 versões de índices criados com conteúdo expandido a partir de indexações feitas pelo usuário (EXPD TYPE = USER).

Expansão	Texto
Sem	SÚMULA TCU 2: Configura-se como vencimento, para efeito da concessão da pensão especial com fundamento na Lei nº 3.738, de 04/04/60, o valor do símbolo correspondente ao cargo em comissão exercido pelo funcionário, à época do seu falecimento.
Term	SÚMULA TCU 2: (...) à época do seu falecimento. Cargo em comissão, Doença, Pessoal, Viúvo, Pensão especial.
Term + synonym (+syn)	SÚMULA TCU 2: (...) à época do seu falecimento. Cargo em comissão - Ocupante de cargo em comissão, Cargo de direção, chefia e assessoramento, Cargo comissionado, Cargo de confiança, Exercente de cargo em comissão e Cargo de direção, chefia ou assessoramento. Doença - Enfermidade e Moléstia. (...)
Term + related term (+rel)	SÚMULA TCU 2: (...) à época do seu falecimento. Cargo em comissão -- Destituição de cargo, Livre nomeação, Livre exoneração, Nepotismo, Faixa etária, Função de confiança e Afastamento para servir a outro órgão ou entidade; Doença -- Medicina, Auxílio-doença, Doente, Perícia médica e Doença preexistente; (...)
Term + synonym + related term (+syn+rel)	SÚMULA TCU 2: (...) à época do seu falecimento. Cargo em comissão - Ocupante de cargo em comissão, Cargo de direção, chefia e assessoramento, Cargo comissionado, Cargo de confiança, Exercente de cargo em comissão e Cargo de direção, chefia ou assessoramento. -- Destituição de cargo, Livre nomeação, Livre exoneração, Nepotismo, Faixa etária, Função de confiança e Afastamento para servir a outro órgão ou entidade. Doença - Enfermidade e Moléstia. -- Medicina, Auxílio-doença, Doente, Perícia médica e Doença preexistente. (...)

Tabela 3 - Índices com expansões: quatro variações nos metadados dos termos indexadores

Como rerranqueador, utilizou-se o PTT5-BASE-INDIR-400 que se saiu melhor do que o MINILM-INDIR-400, simplesmente INDIR nas tabelas que se seguem, foi empregado o modelo original, PTT5-BASE. Como primeiro estágio foram usados, com top-k de 300, BM25, STS e JOIN. Este é uma união dos resultados dos outros dois *retrievers*, STS e BM25, ambos com top-k de 150, totalizando também 300. Também foram experimentados pipelines sem rerranqueamento. Foi experimentada uma inversão no prompt entre o texto do documento e o texto da consulta passado a cada um desses ranqueadores: 'Query: {document} Document:

{query} Relevant:'. Essas versões receberam nas tabelas de resultado um sufixo "i": INDIRi e BASEi. A ideia foi verificar se o INDIR que foi treinado tendo enunciados como consulta no JURIS-TU-INDEX iria ter um resultado melhor uma vez que no JURIS-TCU os enunciados são os documentos. Porém, como mostram os resultados, houve um decréscimo nas métricas apuradas. Todavia, o decréscimo foi bem inferior ao encontrado no PTT5-BASE, provavelmente por ele não tratar bem consultas com textos grande e bem maiores do que os textos dos documentos, que na inversão se tornaram as consultas, que são bem menores. Percebe-se também que o INDIRi, ao contrário dos outros ranqueadores, se beneficiou com as expansões nas consultas, considerando as métricas PRECISION@50 e RECALL@100, conforme Apêndice B. Provavelmente porque com as expansões, as consultas enriquecidas com metadados dos termos do VCE, se aproximavam de conteúdos usados no seu treinamento.

Foram realizados diversos experimentos de buscas para todas as 150 consultas (JURIS-TCU). A Tabela 4 ilustra as variações experimentadas nos parâmetros⁷. O gradiente da cor verde indica o quanto o valor foi superior ao valor base de cada métrica na busca esparsa, usando-se apenas BM25 e sem expansões nas consultas nem nos índices (o valor na interseção entre a primeira linha e a primeira coluna de cada métrica, em cinza).

EXPQ CNT	EXPQ TYPE	EXPD TYPE	EXPD VAL	RAN KER	NDCG@10			PRECISION@50			RECALL@100		
					BM25	STS	JOIN	BM25	STS	JOIN	BM25	STS	JOIN
1	Termo	----	----	----	52,27	18,04	24,06	12,76	6,09	7,55	82,65	46,26	59,05
				INDIR	58,65	52,61	59,04	14,25	10,68	14,65	87,42	64,55	90,64
				INDIRi	51,65		52,76	13,76		14,27	87,12		90,82
				BASE	65,97	57,21	67,40	14,77	10,89	15,17	88,89	64,90	92,03
				BASEi			33,94			12,47			85,51
		USER	term	----	53,69	21,04	26,60	13,35	6,73	8,36	86,85	52,14	63,93
				INDIR	58,23	53,38	58,19	14,39	11,57	14,65	90,08	70,40	92,74
				INDIRi	45,49		46,72	13,08		13,53	86,36		90,23
				BASE	64,19	58,59	64,59	15,17	11,84	15,36	92,43	70,70	94,30
				BASEi			22,24			10,24			76,73
			+syn	----	51,85		26,43	13,29		7,95	86,38		61,65
				INDIR	52,58		53,21	13,43		13,71	87,74		90,42
		INDIR-5	term	----	49,24	18,58	23,00	12,51	6,52	7,91	84,89	51,77	62,39
				INDIR	55,58	52,28	56,16	13,80	11,17	14,01	88,07	68,45	89,98
				INDIRi	44,54		46,32	13,04		13,45	85,89		88,39
				BASE	64,70	58,28	65,01	14,91	11,61	15,16	91,29	69,47	92,21
				BASEi			26,87			10,57			80,02
			+syn	----	47,07		22,17	12,16		7,56	81,33		59,02
				INDIR	45,80		46,81	12,28		12,56	83,38		85,33
	Termo + extra	----	----	----	41,64	17,66	20,04	11,08	5,91	7,00	74,18	45,33	54,98
				INDIRi			50,27			13,19			85,47
		USER	term	----	44,05	20,55	20,04	11,61	6,40	6,51	77,49	49,62	53,90
				INDIRi			44,96			12,56			84,07
			+syn	----	41,87	20,05	17,10	11,55	6,71	5,97	77,15	52,45	50,11
		INDIR-5	term	----	38,63	18,65	17,32	10,91	6,43	6,45	75,79	48,86	52,01
				INDIRi			42,92			12,68			83,38

Tabela 4 - Visão parcial com variações de experimentações realizadas

A Tabela 5 mostra as combinações que, mesmo sem ranqueamento, se beneficiaram das expansões realizadas. As expansões de termos ("term") e de termos e de seus sinônimos ("+syn") feitas pelo próprio usuário ("USER") superaram os valores base nas 3 métricas

⁷ Não foram realizadas combinações exaustivas dos parâmetros mas algumas que poderiam indicar se o caminho poderia levar a melhoras nas métricas apuradas. Ilustrando o BASE e BASEi foram experimentados para: quanto à expansão de consultas: sem e com expansão de um termo (sem dados extras); quanto à expansão no documento: sem e com expansão de termos, tanto de indexações do usuário quanto cinco preditas pelo INDIR. BASE foi testado em combinações com todos os 3 tipos de retrievers. Já o BASEi só com JOIN. Todas as combinações experimentadas e seus resultados podem ser encontrados em data\search\juris-tcu\search_experiment_juris_tcu.csv.

analisadas: NDCG@10 (+0,73), RECALL@100 (+2.85) e PRECISION@50 (+0.81). Praticamente todas as expansões somente de termos em consultas promoveram um acréscimo no RECALL@100⁸. O maior valor de RECALL@100, 86.85, se deu com expansão dos 2 lados: consultas (com um termo) e documento (somente com os termos indexados pelo usuário), combinação que também alcançou PRECISION@50 superior. O maior NDCG@12, 55.72, foi com a expansão somente no documento de termos e a maior precisão, 13.73, foi com expansão no documento de termos e seus sinônimos.

EXPQ TYPE	EXPD TYPE	EXPQ CNT	EXPD VAL	NDCG@10	RECALL@100	PRECISION@50
				bm25	bm25	bm25
----	----	----	----	54,99	82,94	12,92
	INDIR-1		term		83,19	13,07
	INDIR-3		term		84,30	13,07
	INDIR-5		+syn		83,29	
			term		85,05	12,93
	USER		+syn	55,58	85,79	13,73
			term	55,72	85,74	13,53
Termo	INDIR-1	1	term		83,79	
	INDIR-3		term		84,76	
	INDIR-5		term		84,89	
	USER	1	+syn		86,38	13,29
			term		86,85	13,35
		2	term		85,31	13,00
		3	term		84,38	
		4	term		84,16	
		5	term		83,42	
		6	term		83,55	
		7	term		82,99	
		9	term		83,06	
		10	term		83,28	

Tabela 5 - Combinações que favoreceram a busca esparsa

A Tabela 7 traz as combinações que alcançaram resultados entre os top-40 superiores aos valores base considerados em alguma das métricas. Há algumas promissoras que alcançaram valores entre os superiores em todas as métricas e em todos os *retrievers* testados (BM25, STS e JOIN).

Percebe-se que, nas buscas sem ranqueamento, o *retriever* BM25 alcançou melhores resultados do que os outros dois *retrievers*. Porém nas buscas com rerranqueamento, a configuração com *retriever* JOIN alcançou melhores resultados do que com BM25 e este melhores resultados do que STS. Talvez um indicativo de que a similaridade semântica precisa ser considerada, até para se tratar casos de *vocabulary mismatch*, mas sem se deixar de lado a efetividade do BM25.

⁸ Exceção feita a EXPQ_CNT = 8 expansões

EXPQ TYPE	EXPQ CNT	EXPD TYPE	EXPD VAL		NDCG@10			RECALL@100		PRECISION@50		
				RANKER	BM25	STS	JOIN	BM25	JOIN	BM25	JOIN	
----	----	----	----	BASE	72,95	61,12	73,49	91,08	93,92	15,51	15,77	
				INDIR	64,09	56,67	64,16	90,74	93,67	15,16	15,51	
				INDIRi				88,19	92,15		14,57	
		INDIR-1	+syn	INDIR				87,96	89,87			
			term	INDIR	60,91		60,62	90,81	92,35	14,83	14,92	
		INDIR-3	+syn	INDIR							14,11	
			term	INDIR	60,40		60,70	90,52	93,08	14,83	15,01	
		INDIR-5	+rel	INDIR							14,11	
			term	BASE	70,82	62,68	70,76	91,92	93,68	15,60	15,79	
				INDIR	60,26	55,11	60,50	90,51	92,21	14,69	14,92	
		USER	+rel	INDIR	57,65		57,96	89,91	90,45	14,81	14,95	
			+syn	INDIR	59,25		59,41	90,76	93,12	14,35	14,68	
			+syn+rel	INDIR				89,19	90,07			
			term	BASE	70,66	63,54	71,25	93,00	95,42	15,79	16,08	
		INDIR		63,46	57,85	63,16	92,19	94,74	15,15	15,39		
Termo	1	----	----	BASE	65,97	57,21	67,40	88,89	92,03	14,77	15,17	
				INDIR	58,65	52,61	59,04	87,42	90,64	14,25	14,65	
				INDIRi				87,12	90,82		14,27	
		INDIR-1	term	INDIR				87,81	90,28		14,40	
		INDIR-5	term	BASE	64,70	58,28	65,01	91,29	92,21	14,91	15,16	
				INDIR				88,07	89,98			
		USER	+syn	INDIR				87,74	90,42			
			term	BASE	64,19	58,59	64,59	92,43	94,30	15,17	15,36	
				INDIR	58,23	53,38	58,19	90,08	92,74	14,39	14,65	
				INDIRi				86,36	90,23			
		2	----	----	INDIR							14,09
					INDIRi							14,23
		USER	term	INDIR				88,15	89,89			
	3	----	----	INDIRi							14,28	
	4	----	----	INDIR	57,36		57,67					
				INDIRi							14,31	
	5	----	----	INDIR	57,47		57,07					
	6	----	----	INDIR	57,03		57,40					
	9	USER	term	INDIRi				87,85	89,95			
10	USER	term	INDIRi				87,74	90,01				

Tabela 6 - Combinações com resultado favorável (visão parcial)

Todas as métricas foram impactadas com as expansões em alguma combinação experimentada. Considerando-se o critério de maior número de casos de superação relacionados no Apêndice B, que traz todas as 284 combinações que tiveram algum valor acima, a métrica mais impactada foi o RECALL@100 (97%: 276 de 284), seguida pela precisão@50 (70%: 199 de 284) e NDCG@10 (31%, 88 de 284),

Considerando buscas com ranqueamento: o PTT5-BASE alcançou melhores resultados do que o INDIR-400, com diferenças maiores no NDCG@10 e pequenas nas outras métricas. Isso talvez seja explicado pelo INDIR ter se especializado no domínio JURIS-TCU-INDEX e é sabido que esses modelos muito ajustados em um domínio não alcançam resultados similares

fora do domínio do treino. No NDCG@10 o PTT5-BASE alcançou melhores resultados sem expansões (+18.5). Já nas métricas PRECISION@50 e RECALL@100, os melhores resultados se deram com a expansão no índice com termo indexado pelo usuário: +3.16 e +12.48, respectivamente.

4. Conclusão

Os *datasets* em português construídos e publicados JURIS-TCU-INDEX e JURIS-TCU são representativos da atividade de buscas. Os dois são administrados cuidadosamente por pessoal especializado no TCU, o que implica uma qualidade de conteúdo. Seus textos pequenos facilitam seu uso por modelos de *Machine Learning* (ML).

Na formação do *dataset* JURIS-TCU, os três grupos contendo 50 consultas cada foram selecionados de forma distinta, podendo apresentar resultados diferentes dependendo da tarefa empregada, portanto, pode ser recomendável recortar apenas o grupo que melhor atenda ao fim desejado. As consultas do grupo 1 são buscas genéricas construídas para desempenhar bem no sistema de busca atual, com foco na recuperação de todos os documentos relevantes sobre um tema, mas com baixo desempenho para responder a perguntas completas em português. Já as do grupo 3 são completas, mas têm baixo desempenho na busca atual. São geradas a partir dos enunciados mais populares, o que garante maior representatividade. Os pipelines de busca construídos têm um desempenho melhor para esse grupo, devido à maior cobertura dos recuperadores e ao uso de mecanismos de combinação e rerank. No entanto, os resultados nas figuras 6, 7 e 10 podem indicar um desempenho superior da busca BM25, mas também podem ser influenciados pelo viés do LLM em gerar perguntas com termos encontrados no enunciado original. As características das do grupo 2 equilibram as do grupo 1 e do grupo 3, visando melhorar o desempenho das buscas no sistema em operação, em relação ao grupo 3. Apesar de apresentarem piora de desempenho em relação ao grupo 3 nos pipelines utilizados, como indicado nas figuras 6 e 9, elas são úteis para análises comparativas.

Análises manuais dos escores produzidos pelo ChatGPT (qrel do JURIS-TCU) indicam excelente qualidade. Os resultados demonstrados na figura 7, referentes a enunciados não reencontrados, indicam que a restrição para um *dataset* de avaliação de melhor qualidade parece estar nos mecanismos de busca empregados.

A construção do indexador como um pipeline de busca com ranqueador treinado a partir de dados gerados se mostrou viável com ótimos resultados demonstrados nas métricas NDCG@12 e RANK1. Resultados demonstram que um modelo com cerca de 30 vezes menos parâmetros pode alcançar com dados adequados de treinamento resultados superiores ao de modelos bem maiores, caso do MINILM-INDIR contra o MT5-3B.

O uso do INDIR como rerranqueador e de indexações em expansões de consultas e de documentos impactaram positivamente as configurações de buscas realizadas. Com as expansões, buscas esparsas (só BM25) alcançaram melhores resultados com a expansão de documentos com termos indexados pelo usuário: NDCG@10: +0.73; PRECISION@50: +0.61 e RECALL@100: +2.80. Buscas com ranqueamento se beneficiaram das expansões nas métricas de PRECISION@50 e RECALL@100 alcançando +3.16 e +12.48 respectivamente. Já para a métrica NDCG@10 os valores máximos foram alcançados sem expansões.

O ranqueador PTT5-BASE superou o INDIR nas buscas no JURIS-TCU, ainda que com diferenças menores nas métricas PRECISION@50 e RECALL@100. Esse é um indicativo de que o INDIR por ter se especializado no domínio JURIS-TCU-INDEX (*in-domain*) não consiga resultados similares fora do domínio do treino (*out-of-domain*).

Também se constatou a importância de se usar o prompt na ordem e com textos de tamanho próximos aos usados no treinamento do modelo. A inversão experimentada, provavelmente por levar a um texto de consulta bem superior ao do documento, deteriorou a performance do modelo PTT5-BASE teve impactos até positivos (em certas combinações de expansões às consultas) no PTT5-BASE-INDIR-400 que foi treinado com textos maiores na consulta (os enunciados de jurisprudência).

O fato do pipeline com ranqueamento e *retriever* JOIN ter alcançado melhores resultados do que com *retriever* BM25 parece ser um indicativo de que a similaridade semântica

precisa ser considerada nas buscas para tratar casos de *vocabulary mismatch* (não trazidos no BM25), mas sem deixar de lado a efetividade do BM25.

5. Trabalhos futuros

Vislumbram-se alguns trabalhos futuros. Quanto ao *dataset* JURIS-TCU, aumentar o número de consultas e de informações de relevância com novas técnicas e prompts. Sendo maior, pode viabilizar até mesmo uma divisão entre dados de avaliação e dados de treinamento para permitir avaliação do uso de modelos treinados “*in-domain*” no JURIS-TCU.

Quanto ao indexador no JURIS-TCU-INDEX: treinar versões do INDIR com mais parâmetros (com 3 bilhões ou mais) visando se obter um desempenho ainda superior aos alcançados tanto no JURIS-TCU-INDEX e, principalmente, com as expansões no JURIS-TCU.

Também novas análises podem ser realizadas, novas combinações experimentadas, outras métricas consideradas, outros filtros utilizados nas buscas. Cabe também experimentar um LLM (*Large Language Model*) como estágio 3 de ranqueamento para os pipelines de busca tanto do JURIS-TCU-INDEX quanto do JURIS-TCU.

Espera-se a aplicação do processo aqui desenvolvido de Indexação apoiando *Information Retrieval* (INDIR) em outros sistemas que fazem uso de indexações, do próprio TCU ou de outras instituições governamentais ou privadas.

6. Agradecimentos

Ao TCU, não só pela disponibilização da nuvem com GPT 4.0, equipamentos, mas à toda sua equipe: pelas soluções desenvolvidas para o trato dos *datasets* aqui disponibilizados.

Aos colegas de disciplina na Unicamp, pela interação e crescimento mútuo conquistados durante o semestre. Em particular ao Leandro Carísio Fernandes, também servidor do TCU, pela ajuda na extração de dados do histórico de buscas, e ao colega Thiago Soares Laitz por partilhar o código para treinamento do modelo PTT5-BASE.

Aos gestores (e suas equipes) das bases aqui partilhadas, pelo cuidado com os dados dos sistemas e o apoio ao presente trabalho: Sérgio Ricardo de Mendonça Salustiano (Enunciados de Jurisprudência) e Beatriz Pinheiro de Melo Gomes (VCE).

Aos professores Roberto Lotufo e Rodrigo Nogueira por cada conexão de conhecimento e de emoção produzida nesses anos de convívio.

7. Referências bibliográficas

BRASIL. Pesquisa de Jurisprudência: Guia rápido <https://portal.tcu.gov.br/tcucidades/publicacoes/detalhes/pesquisa-de-jurisprudencia-guia-rapido.htm>. Acesso em 7 julho 2023. 2016.

_____. Tribunal de Contas da União. Vocabulário de controle externo do Tribunal de Contas da União – 3.ed. rev. e ampl. – Brasília : TCU, Instituto Serzedello Corrêa, Centro de Documentação. Disponível em: https://portal.tcu.gov.br/data/files/F8/04/8E/5E/A0B3071068A7C107F18818A8/VCE_TCU.pdf. Acesso em 25 maio 2023. 2019.

_____. Portaria-TCU - 85, de 06 de junho de 2022. Aprova o *Manual de Sistematização e Divulgação da Jurisprudência do Tribunal de Contas da União*. Disponível em: <https://pesquisa.apps.tcu.gov.br/#/documento/ato-normativo/Ac%25C3%25B3rd%25C3%25A3o%2520n%25C2%25BA%25202800%252F2022/%2520/score%2520desc/1/%2520>. Acesso em 25 maio 2023. 2022.

FURNAS, George W.. et al. The vocabulary problem in human-system communication. **Communications of the ACM**, v. 30, n. 11, p. 964-971, 1987.

LANCASTER, Frederic Wilfrid. Indexação e resumos: Teoria e Prática. **Tradução de AA Briquet de Lemos. Brasília**, 2004.

LIN, Jimmy; NOGUEIRA, Rodrigo; YATES, Andrew. **Pretrained transformers for text ranking: Bert and beyond**. Springer Nature, 2022.

LIN, Jimmy. A proposed conceptual framework for a representational approach to information retrieval. In: **ACM SIGIR Forum**. New York, NY, USA: ACM, 2022. p. 1-29.

MANNING, Christopher D. **An introduction to information retrieval**. Cambridge university press, 2009.

NOGUEIRA, Rodrigo et al. Document expansion by query prediction. **arXiv preprint arXiv:1904.08375**, 2019.

Apêndice A - Escores de avaliação por consulta no *dataset* JURIS-TCU

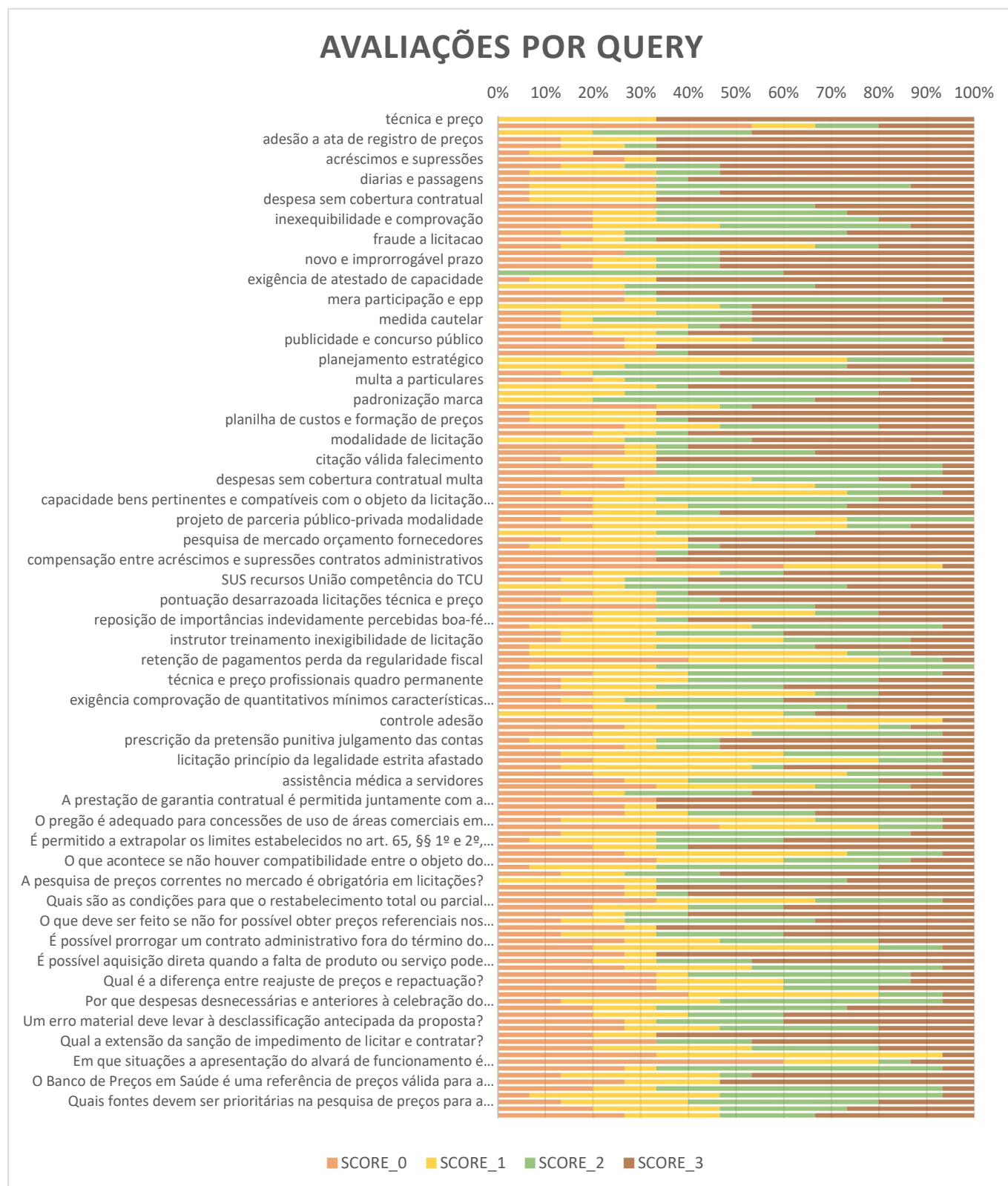


Figura 11 - Escores de avaliação por consulta no *dataset* JURIS-TCU

Apêndice B – Combinações com algum resultado superior ao base de cada métrica em consultas no JURIS-TCU

EXPQ TYPE	EXPQ CNT	EXPD TYPE	EXPD VAL	RANKER	NDCG@10			RECALL@100		PRECISION@50	
					bm25	sts	join	bm25	join	bm25	join
----	----	----	----	----	54,99	18,47	21,48	82,94	51,82	12,92	6,72
				BASE	72,95	61,12	73,49	91,08	93,92	15,51	15,77
				BASEi					86,40		
				INDIR	64,09	56,67	64,16	90,74	93,67	15,16	15,51
				INDIRi				88,19	92,15	13,85	14,57
		INDIR-1	+rel	INDIR	55,43		56,19	87,17	88,55	13,72	13,73
				+syn	56,52		56,11	87,96	89,87	13,67	13,95
				term	----			83,19	55,88	13,07	7,16
				INDIR	60,91		60,62	90,81	92,35	14,83	14,92
		INDIR-3	+rel	INDIR	54,64		55,53	85,61	87,05	13,64	13,95
				+syn				86,66	89,59	13,65	14,11
				term	----			84,30	57,87	13,07	7,40
				INDIR	60,40		60,70	90,52	93,08	14,83	15,01
		INDIR-5	+rel	INDIR	55,41		56,12	86,23	87,51	13,68	14,11
				+syn	----			83,29	56,89		
			term	INDIR				87,80	89,36	13,57	13,95
				----				85,05	57,68	12,93	7,27
				BASE	70,82	62,68	70,76	91,92	93,68	15,60	15,79
				INDIR	60,26	55,11	60,50	90,51	92,21	14,69	14,92
				INDIRi				84,73	89,20	12,75	13,41
		USER	+rel	INDIR	57,65		57,96	89,91	90,45	14,81	14,95
				+syn	----	55,58	20,59	85,79	60,49	13,73	7,92
			term	INDIR	59,25		59,41	90,76	93,12	14,35	14,68
				+syn+rel	INDIR	55,54	55,89	89,19	90,07	13,96	14,07
				----	55,72	21,63	25,91	85,74	61,56	13,53	8,12
				BASE	70,66	63,54	71,25	93,00	95,42	15,79	16,08
				INDIR	63,46	57,85	63,16	92,19	94,74	15,15	15,39
Termo + extra	1	----	----	INDIRi					85,47		13,19
		INDIR-5	term	INDIRi					83,38		
		USER	term	INDIRi					84,07		
	5	USER	term	INDIRi					83,71		
Termo	1	----	----	BASE	65,97	57,21	67,40	88,89	92,03	14,77	15,17
				BASEi					85,51		
				INDIR	58,65	52,61	59,04	87,42	90,64	14,25	14,65
				INDIRi				87,12	90,82	13,76	14,27
		INDIR-1	+rel	INDIR				82,58	84,47		
				+syn				84,37	86,12	12,59	12,93
				term	----			83,79	59,39		
				INDIR	55,99		56,33	87,81	90,28	13,96	14,40

EXPQ TYPE	EXPQ CNT	EXPD TYPE	EXPD VAL	RANKER	NDCG@10			RECALL@100		PRECISION@50	
					bm25	sts	join	bm25	join	bm25	join
		INDIR-3	+rel	INDIR				82,79	83,72		
				INDIR				82,04	83,94		
			term	----				84,76	61,46		
				INDIR	55,26		55,71	88,23	89,74	13,83	14,08
		INDIR-5	+rel	INDIR				83,23	85,15	12,65	13,03
				INDIR				83,38	85,33		
			term	----				84,89	62,39		
				BASE	64,70	58,28	65,01	91,29	92,21	14,91	15,16
				INDIR	55,58	52,28	56,16	88,07	89,98	13,80	14,01
				INDIRi				85,89	88,39	13,04	13,45
		USER	+rel	INDIR				87,95	87,74	13,85	14,04
				----				86,38	61,65	13,29	7,95
			+syn+rel	INDIR				87,74	90,42	13,43	13,71
				INDIR				85,37	86,63	13,12	13,32
				term				86,85	63,93	13,35	8,36
			term	----				86,85	63,93	13,35	8,36
				BASE	64,19	58,59	64,59	92,43	94,30	15,17	15,36
				INDIR	58,23	53,38	58,19	90,08	92,74	14,39	14,65
				INDIRi				86,36	90,23	13,08	13,53
	2	----	----	BASEi					84,17		
				INDIR	56,36		56,83	85,63	88,64	13,72	14,09
				INDIRi				85,85	89,68	13,57	14,23
		INDIR-1	+syn	INDIR				81,70	83,32		
			term	INDIR	54,51		55,41	86,06	87,36	13,43	13,69
		INDIR-3	term	INDIR				85,27	86,56	13,29	13,35
		INDIR-5	term	INDIR				85,15	86,70	13,04	13,31
				INDIRi				84,49	87,04	13,05	13,35
		USER	+rel	INDIR				85,28	86,06	12,97	13,11
			+syn	INDIR				85,52	86,76	13,01	13,17
			+syn+rel	INDIR				82,72	83,48		
			term	----				85,31	64,85	13,00	8,17
				INDIR	55,82		55,83	88,15	89,89	13,72	14,01
				INDIRi				86,67	89,67	13,09	13,45
	3	----	----	BASEi					83,21		
				INDIR	56,20		56,71	85,11	87,63	13,32	13,53
				INDIRi				86,01	88,49	13,76	14,28
		INDIR-1	term	INDIR				84,72	85,76	13,09	13,27
		INDIR-3	term	INDIR				85,14	85,66	12,99	13,03
		INDIR-5	term	INDIR				83,98	85,51	12,85	13,05
				INDIRi				84,73	86,56	13,19	13,55
		USER	+rel	INDIR				84,97	84,48	12,80	12,95
			+syn	INDIR				85,14	86,03	12,88	12,99
			term	----				84,38	64,08		

EXPQ TYPE	EXPQ CNT	EXPD TYPE	EXPD VAL	RANKER	NDCG@10			RECALL@100		PRECISION@50	
					bm25	sts	join	bm25	join	bm25	join
	4	----	----	INDIR				87,11	88,67	13,49	13,65
				INDIRi				87,37	89,47	13,51	13,81
				BASEi					83,74		
				INDIR	57,36		57,67	84,50	87,26	13,36	13,65
				INDIRi				86,10	89,01	13,75	14,31
				INDIR-1	term			84,34	85,06	12,93	13,12
				INDIR-3	term			84,17	85,76	12,84	13,03
				INDIR-5	term			83,96	85,25	12,80	13,09
								85,56	86,28	13,32	13,53
				USER	+rel			83,80	84,44	12,77	12,93
					+syn			84,38	85,57		
					term			84,16	60,36		
				INDIR	55,11		55,40	86,84	88,15	13,48	13,63
				INDIRi				88,00	89,53	13,75	14,04
	5	----	----	INDIR	57,47		57,07	84,30	85,77	13,32	13,36
				INDIRi				86,06	88,04	13,61	14,07
				INDIR-1	term			82,79	84,27		
				INDIR-3	term			84,01	84,73		
				INDIR-5	term			83,41	84,16		
								84,82	86,29	13,21	13,39
				USER	+rel			82,34	83,93		
					+syn			83,47	85,84		
					term			83,42	58,13		
				INDIR	55,57		55,49	85,85	87,16	13,39	13,52
				INDIRi				87,07	89,03	13,48	14,07
	6	----	----	INDIR	57,03		57,40	83,18	83,68	13,15	13,19
				INDIRi				85,37	86,79	13,67	13,84
				INDIR-1	term			82,09	83,97	12,71	13,01
				INDIR-3	term			83,39	84,57		
				INDIR-5	term			83,61	83,92		
								85,73	86,37	13,40	13,51
				USER	+rel			82,21	83,03		
					+syn			83,31	84,87		
					term			83,55	55,73		
				INDIR	55,14		55,69	86,33	86,85	13,44	13,51
				INDIRi				88,05	89,47	13,76	14,01
	7	----	----	INDIR	56,17		56,83	82,95	83,44	13,21	13,25
				INDIRi				84,89	86,82	13,53	13,73
				INDIR-1	term			82,47	83,03		
				INDIR-3	term			83,70	84,09		
				INDIR-5	term			83,20	83,12		
								84,71	85,48	13,23	13,35

EXPQ TYPE	EXPQ CNT	EXPD TYPE	EXPD VAL	RANKER	NDCG@10			RECALL@100		PRECISION@50	
					bm25	sts	join	bm25	join	bm25	join
		USER	+rel	INDIR				81,55	83,54		
			+syn	INDIR				82,13	82,98		
			term	----				82,99	55,68		
				INDIR				86,04	86,78	13,31	13,44
				INDIRi				87,68	89,14	13,75	14,00
	8	----	----	INDIR	56,27		56,51	83,86	84,63	13,28	13,44
				INDIRi				85,68	87,13	13,64	13,85
		INDIR-1	term	INDIR				83,25	84,48	12,77	13,00
		INDIR-3	term	INDIR				83,89	83,92		
		INDIR-5	term	INDIR				83,53	83,57		
				INDIRi				85,40	85,97	13,28	13,49
		USER	+rel	INDIR				82,03	83,85		
			+syn	INDIR				82,33	82,99		
			term	INDIR	54,75		55,05	86,44	87,66	13,39	13,60
				INDIRi				87,97	89,79	13,84	13,99
	9	----	----	INDIR	55,85		56,38	84,09	86,09	13,31	13,64
				INDIRi				85,31	87,93	13,63	14,04
		INDIR-1	term	INDIR				83,68	83,89	12,89	12,95
		INDIR-3	term	INDIR				83,96	83,93		
		INDIR-5	term	INDIR				83,58	83,79	12,91	13,01
				INDIRi				85,52	86,14	13,23	13,44
		USER	+rel	INDIR				81,82	83,51		
			term	----				83,06	54,17		
				INDIR	55,20		55,44	86,78	87,90	13,41	13,61
				INDIRi				87,85	89,95	13,87	14,01
	10	----	----	INDIR	55,87		56,33	84,16	85,75	13,23	13,53
				INDIRi				85,38	87,65	13,63	13,99
		INDIR-1	term	INDIR				83,74	84,02	12,88	13,01
		INDIR-3	term	INDIR				83,89	83,70		
		INDIR-5	term	INDIR				83,61	83,81	12,89	13,00
				INDIRi				85,40	86,22	13,20	13,45
		USER	+rel	INDIR				81,84	83,04		
			term	----				83,28	54,32		
				INDIR	54,77		55,18	86,66	87,86	13,35	13,60
				INDIRi				87,74	90,01	13,85	14,01

Tabela 7 - Experimentações com resultado superior em pelo menos uma métrica