INDIR
Indexing Improving Information Retrieval

(Jurisprudence Statements and 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 by providing datasets for model training and evaluation in Portuguese. The datasets, JURIS-TCU-INDEX and JURIS-TCU, represent search activities and offer high-quality content. The study proposes an automatic indexer model named INDIR, which is trained using indexing data from JURIS-TCU-INDEX. The search pipeline based on INDIR achieved excellent results in evaluation metrics. The INDIR-106 (T5-BASE model) outperformed the MT5-3B model, demonstrating the effectiveness of a smaller yet well-trained model. The use of INDIR for query and document expansion improved search performance in the JURIS-TCU dataset, particularly benefiting sparse searches in all analyzed metrics. Even searches with rankers showed improvements in precision@50 and recall@100. The results underscore the importance of incorporating semantic similarity into search pipelines in the first retrieval stage, alongside bm25, to address vocabulary mismatch cases while maintaining bm25's effectiveness.

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 dentro de grandes coleções. 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 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) propôs um modelo em termos de codificadores que mapeiam consultas e documentos em um espaço de representação, e uma função de comparação que calcula as pontuações de consulta-documento. Ele abordou a temática de indexação, em que o indexador humano desempenha o papel do codificador de documento Nd, e a saída pode ser vista como um vetor *multi-hot* em que cada dimensão representa um descritor de conteúdo. E afirma que no momento da busca a necessidade de informação precisa ser mapeada para o mesmo espaço de representação desses descritores de conteúdo Nq. Os bibliotecários foram os que primeiro desempenharam o papel de codificação de documentos e de consultas com a indexação.

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 (termos descritores) e compreende duas etapas distintas: a análise conceitual do assunto e a tradução dos conceitos em descritores.

Lancaster (2004) afirma que um mesmo texto pode ser indexado de formas distintas por indexadores diferentes, uma vez que a análise conceitual é moldada para ajustar os conceitos a uma clientela (usuários de um sistema de busca). E porque há uma dependência do vocabulário controlado utilizado, que disponibiliza os termos para indexação (etapa de tradução) [1].

Para o aperfeiçoamento do par IR e NLP, é necessária a existência de *datasets* (conjuntos de dados) tanto para treinamentos de modelos quanto para avaliação de resultados quanto à relevância dos documentos retornados para uma busca, inclusive em português.

Um importante conjunto de informações para a Administração Pública é a base da Jurisprudência Selecionada do Tribunal de Contas da União (TCU), que permite a pesquisa aos enunciados elaborados pela Diretoria de Jurisprudência a partir de deliberações selecionadas do TCU, com base em critérios de relevância jurisprudencial, definidos em Brasil (2016).

São objetivos do presente trabalho:

1. construir um indexador automático para o sistema de enunciados de jurisprudências do TCU a partir de dados de indexações realizadas;
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.

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 (qrels.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: 10 termos fixos do VCE
- Tema: termo livre do VCE

[1] Muitas instituições governamentais e privadas possuem tesauros. 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 e Steven Spielberg, sempre na busca por relíquias e pelo seu uso para o bem (e não por nazistas).

- Subtema: também termo livre do VCE.

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 Selecionada;
- 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, anomizados, 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 – scores de relevância para determinado enunciado responder a uma consulta.

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

A partir do arquivo de log contendo buscas efetuadas (query.csv), 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 (grupo 1), correspondendo aos IDs 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, para que produzisse 5 perguntas que pudessem ser respondidas por cada um destes enunciados. Exemplo:

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.

A partir das 5 perguntas produzidas foi construída manualmente uma pergunta final. Cada pergunta passou por revisão e ajuste final por especialista humano em jurisprudência do TCU, chegando a um grupo de 50 perguntas completas (grupo 3), correspondendo aos IDs 101 a 150. Os arquivos gerados estão registrados em pasta do projeto LLM-JURIS-TCU em https://github.com/marcusborela/ind-ir/tree/main/data/llm_juris_tcu.

As 50 perguntas completas foram transformadas em expressões de busca que pudessem gerar resultado na atual versão da 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), correspondendo aos IDs de 51 a 100.

c. Produção de scores de relevância

Como base em cada uma das 150 consultas produzidas, foram realizadas diversas buscas sobre enunciados de jurisprudência, com resultados registros em pasta do projeto JURIS-TCU-BASIC (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, para que produzisse um score de relevância (de 0 a 3) para o enunciado em relação à consulta (no caso do grupo 3, relacionada como pergunta). Exemplo:

Prompt:
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 score 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 score.

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}

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

Conforme visto na introdução, soluções para problemas de relevância costumam envolver IR e NLP. No caso específico do JURIS-TCU-INDES, dada a impossibilidade de se usar um classificador, será desenvolvido um pipeline de busca por termos mais relevantes.

A impossibilidade de se usar um classificador se dá 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. Conforme Fibura

Os documentos de JURIS-TCU-INDEX são na verdade os termos do VCE. E dado um enunciado de jurisprudência, deseja-se obter os termos mais relevantes para o enunciado. A Figura 1 demonstra que há grande parte dos termos não têm mais do que 5 indexações. A Figura 2 mostra uma visão quantitativa das indexações para cada uma das quatro categorias (área, tema, subtema e extra): total de indexações, termos distintos usados, e de documentos indexados.

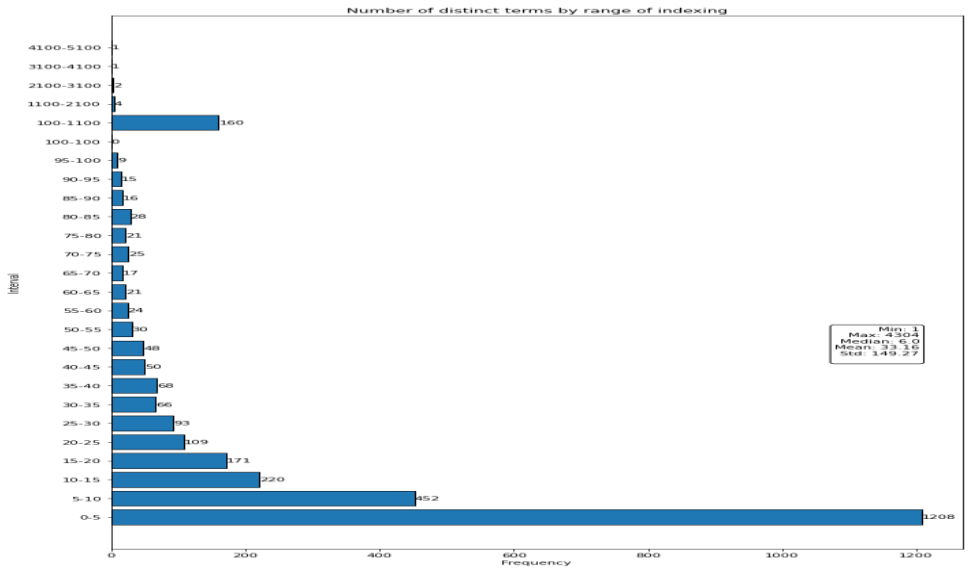

Figura 1 - Número de termos distintos por faixa de número de indexações

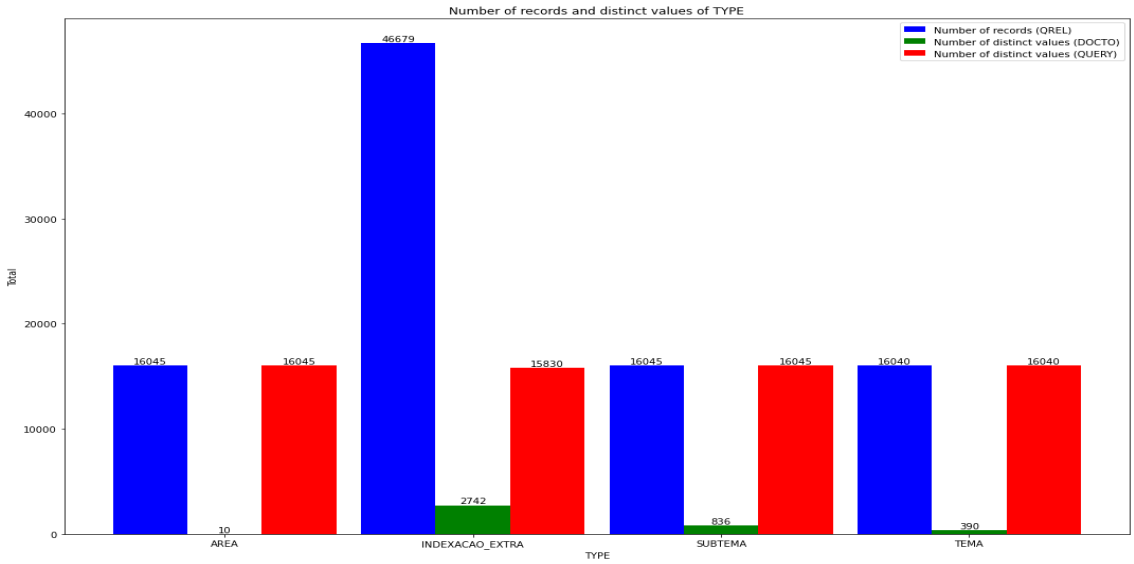

Figura 2 - Visão quantitativa das indexações por categoria.

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.

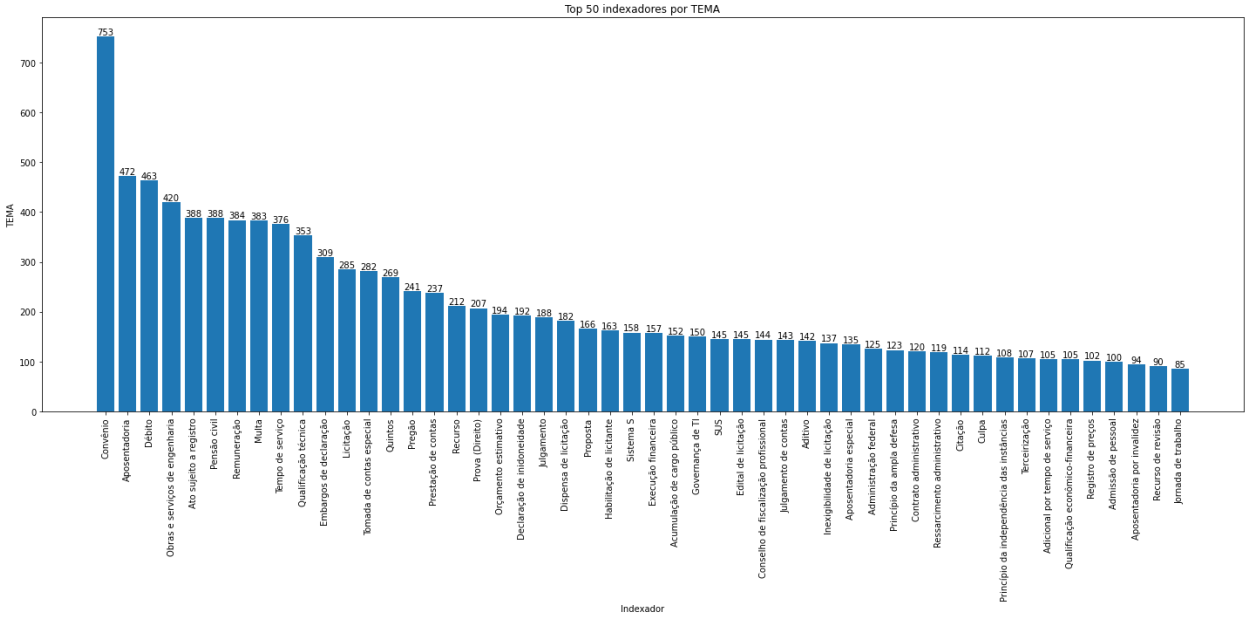

Figura 3 - Indexações por tema (top-50)

A possibilidade de se usar classificação por grupos de termos também foi descartada pelos mesmos motivos anteriores. O número de classes seria 11893, tendo cada agrupamento de 1 a 20 indexações, com média de 1.35 indexações (desvio padrão 1.13).

A ideia então é construir um pipeline de busca com dois estágios retriever e ranker. Serão experimentados como retriever: bm25, similaridade de sentença (sts) e uma combinação entre essas (join). Como ranker, serão experimentados modelos disponíveis para a língua portuguesa e treinados modelos a partir de dados de treinamento a serem gerados.

A qualidade do pipeline será avaliada a partir de duas métricas: rank1 e ndcg@12. Utilizou-se a métrica rank1 pois ela evidencia a primeira posição de um documento relavante 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 se justifica pois 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 2.4 que alcançarem melhor desempenho em JURIS-TCU-INDEX serão usadas para melhoria da performance das buscas em JURIS-TCU.

Serão experimentados dois tipos de expansões: de queries e de documentos. Os documentos serão expandidos com os dados dos termos indexados no sistema. Serão criadas versões diferentes de documentos conforme os metadados considerados na expansão. Já as queries serão expandidas por termos sugeridos pelo pipeline de indexação. Serão avaliadas expansões com um total de 1 a 10 termos.

Serão consideradas as métricas ndcg@10, precision@50 e recall@100. Elas, em conjunto, cercam dimensões complementares de um sistema de busca. Precision indica o grau de acerto do modelo entre os primeiros 50 documentos retornados, enquanto recall indica o percentual entre os documentos relevantes efetivamente retornados entre os 100 primeiros. Ndcg vai mais no detalhe, atentando para a ordem e a posição dos relevantes na lista retornada.

3. Experimentos

3.1. Dataset de Indexação – JURIS-TCU-INDEX

Foi produzido o *dataset* JURIS-TCU-INDEX, formado por 16057 enunciados (query.csv), 13205 documentos do VCE (doc.csv) e 94809 indexações de enunciados por termos do VCE (qrels.csv). As figuras da seção 2.4 permitem uma visão quantitativa de seus dados. A figura abaixo evidencia que o tamanho dos textos (doc.csv) é relativamente baixo.

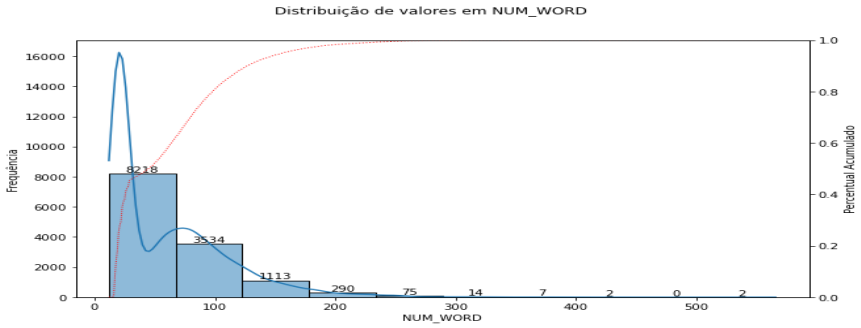

Figura 4 - Histograma de número de palavras em textos do VCE

3.2. Dataset de Avaliação – JURIS-TCU

Foi produzido o *dataset* JURS-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). Conforme figura abaixo, o tamanho dos enunciados também é pequeno.

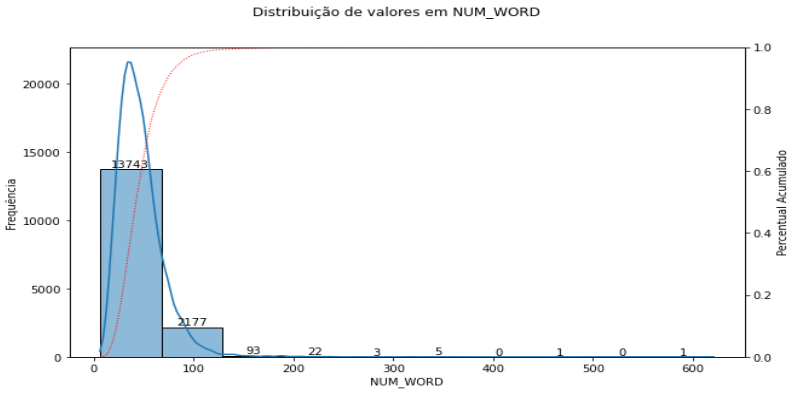

Figura 5 - Histograma de número de palavras em enunciados de jurisprudência

O LLM empregado para geração das perguntas e geração do score das avaliações foi o ChatGPT 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+Rerank: busca BM25, seguida de rerank dos top-300;
3. STS: busca densa por similaridade de sentença;
4. STS+Rerank: busca STS, seguida de rerank dos top-300;
5. (BM25 | STS) + Rerank: buscas BM25 e STS, join (união) dos top-300 de cada e rerank.

Para todos os pipelines com rerank, foi empregado um modelo MonoT5 não ajustado para os dados do domínio. Nas buscas por similaridade, os embeddings forem gerados pelo modelo Bert rufimelo/Legal-BERTimbau-sts-large-ma-v3.

Na figura a seguir, é possível observar que pipelines que empregam BM25 e que empregam reranker tendem a mostrar resultados mais similares entre todos os grupos de consultas. Observa-se também maior diferença de resultados para consultas do grupo 1, em especial quanto ao uso do reranker após BM25, mas também as que comparam o pipeline 4.

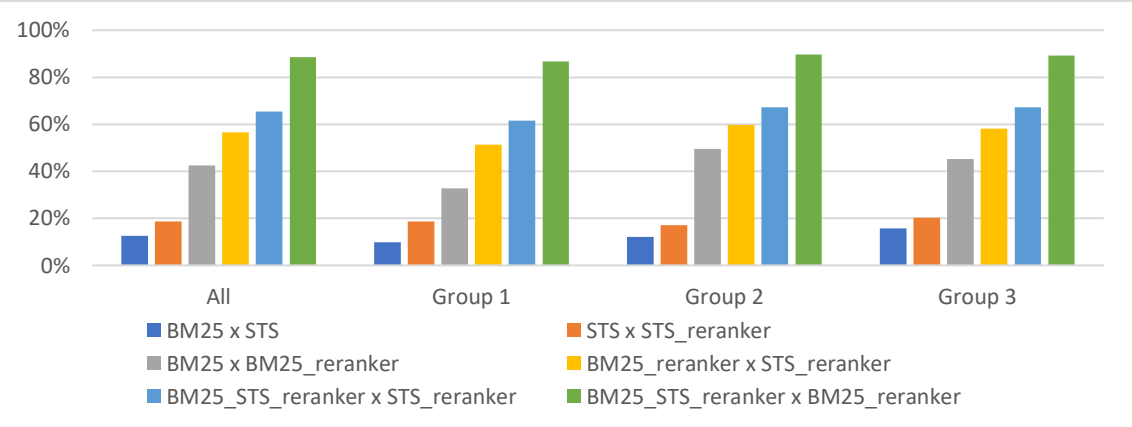

Figura 6 - Similaridade de resultados entre pipelines de busca

As consultas do grupo 3 são gerados a partir de certos enunciados 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 reranker melhora bastante a colocação desses enunciados. Já nos pipelines que possuem busca esparsa BM25, há um percentual muito maior de enunciados reencontrados e em primeira posição no ranking,

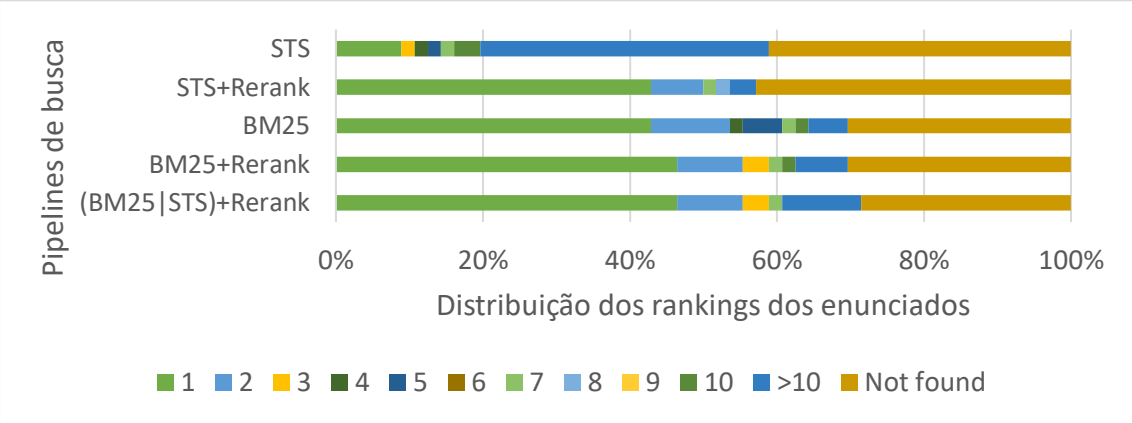

Figura 7 - Enunciados reencontrados - rankings

mesmo sem o reranker, o qual tem um impacto relativamente menor.

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

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

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, 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

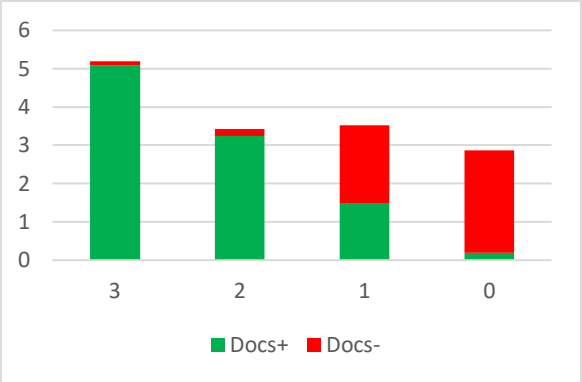

Figura 8 - Média de avaliações por score

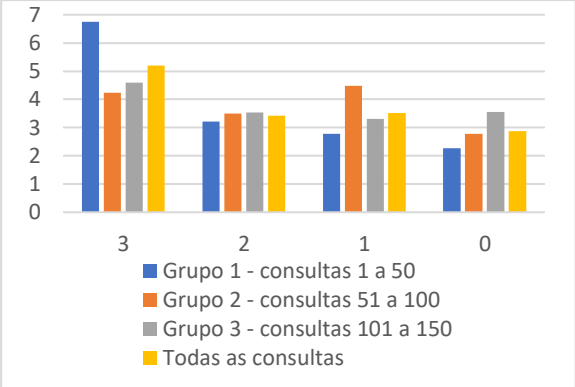

Figura 9 - Média de scores por grupo de consultas

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.

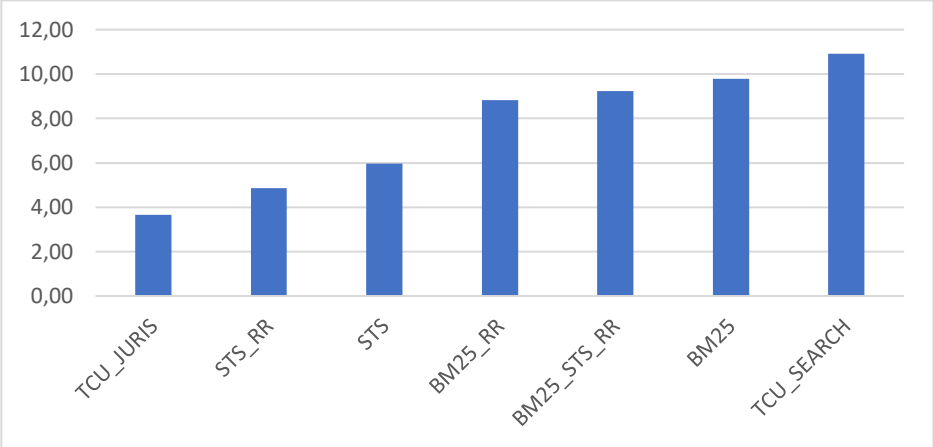

Figura 10 - Docs encontrados entre os top-20 da Pesquisa do TCU

3.3. Indexador INDIR

Os dados dos textos dos termos VCE foram carregados em um índice local Elastic Search. Diversos pipelines de busca foram experimentados usando-se em python a biblioteca haystack. 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). Utilizou-se o tamanho máximo de 350 tokens nos experimentos.

Foram construídos 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.

As configurações experimentadas de retriever envolveram bm25, similaridade de sentença (sts, embeddings gerados pelo modelo rufimelo/Legal-BERTimbau-sts-large-ma-v3, BertModel) e uma combinação entre essas (join). Também foram experimentados pipes em dois estágios com modelos de ranqueadores 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). Essas configurações, todas com trazendo 300 registros (top k), 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 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 categorias[2] área, tema, subtema e extra) optou-se por se treinar um modelo específico para a missão, denominado INDIR.

Por falta de disponibilidade e 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 por um pipe de busca (foram experimentados mt5-3B e bm25) que já não fossem positivos. 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 conseguiram bons resultados (ndcg@12 acima de 50%). Então veio o pulo do gato, algo já sabido na literatura, mas em que estávamos 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 semanticamente. Gerou-se então uma nova versão de dados de treinamento (em data\train_juris_tcu_index\train_data_juris_tcu_index.7z) 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 categoria, e 200 no total das 4 classesi[3]. Desses registros, foram retirados 2460 registros de indexação das 100 queries de avaliação. E, durante o treinamento, 2014 registros para uma base de validação. O treinamento do PTT5 foi realizado no google colab com uma gpu A100 de 40gb e o MINILM em uma máquina local com gpu 3090 de 24gb com rastro no Neptune.ai[4].

Os resultados estão relacionados na Tabela 1. A segunda coluna apresenta os ranqueadores comparados. Possuem sufixo INDIR_XX os modelos treinados com os dados gerados. O sufixo numeral 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.

Como pode ser observado, os resultados alcançados tanto pelo MINILM-INDIR e principalmente pelos PTT5-INDIR foram bem superiores do que os outros modelos (ou 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

[2] As buscas, por opção de projeto, tiveram um filtro que restringia a termos do VCE que já foram usados na classe em questão. Por exemplo, para a classe área, que na prática é um problema de classificação, eram buscados sempre os únicos dez termos já usados. Adicionalmente, usou-se o critério "total_gte_5" para uma busca mais restritiva (gte, "greater than or equal"). Também houve um filtro adicional restringindo a termos do VCE que não fossem "localização", "verbo" e "organização", que foram usados apenas 894 vezes das 94809 indexações realizadas (0.94%) e que 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. No critério total, houve um caso. No tema, 2 (sts) e 4 (bm25), etc.

[3] Ver detalhes no código da função generate_train_data_per_relevance_type em code\generate_train_data_juris_tcu_index.ipynb.

[4] Mais detalhes sobre a configuração do treinamento podem ser obtidos em finetune_reranker_mt5_seq2seq_colab.ipynb (PTT5) e finetuning_reranker_sequence_classification_bert_like.ipynb (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&detailsTab=charts&shortId=IAD-108&type=run (PTT5, parcial).

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 maiores (MT5-3B com 3 bilhões).

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

A Figura 9 permite uma melhor visualização das diferenças alcançadas entre os modelos INDIR treinados e os seus concorrentes na métrica Rank1. 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.

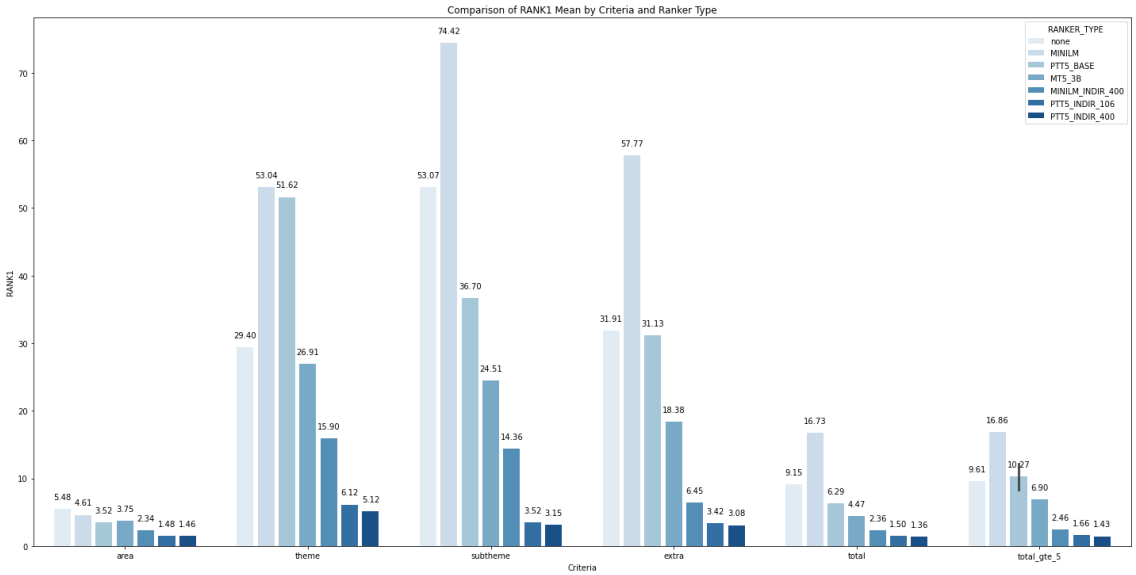

Figura 8 - Comparativo da métrica RANK1 em JURIS-TCU-INDEX

3.4. Uso do INDIR em buscas

Considerando que Nogueira et al. (2019) melhoraram as buscas com a expansão de documentos com termos, queries geradas automaticamente, representativos do documento e que, como visto na introdução, também há a abordagem de se expandir as consultas, implementamos essas duas opções.

Foram realizadas até 10 expansões (o termo EXPQ_CNT indicará a quantidade de expansões) às perguntas usando o modelo PTT5-INDIR-400. 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). Nas duas foi usado o PTT5-INDIR-400 para estimar o termo para a expressão. Porém, na primeira (EXPD_TYPE=indir), usou-se um pipeline de join entre bm25 e sts, 30 registros de cada. No segundo (EXPD_TYPE=indir-extra), além de ter informações extras (preferencialmente sinônimos, ou, na ausência desses, termos relacionados, caso existam) o pipeline join usado considerou 10 registros do bm25 e 30 do sts, que nos testes visuais realizados demonstrou trazer termos mais relevantes para a consulta.

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	Texto + termo	Texto + termo + 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

Também foi realizada uma expansão nos documentos com enunciados de jurisprudências (JURIS-TCU). Para essa expansão, foram usados dois critérios (EXPD TYPE): termos indexados pelos próprios usuários (qrels em JURIS-TCU-INDEX) e de 1 a 5 termos previstos pelo PTT5-INDIR-400. Também houve uma variação nos metadados adicionados em conjunto com os termos indexadores (EXPD_VAL). 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	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	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	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

Foram realizados diversos experimentos de buscas para todas as 150 consultas (JURIS-TCU) com top-k de 300. A Tabela 4 ilustra as variações experimentadas nos parâmetros considerados. As quatro primeiras colunas indicam as expansões realizadas (em consultas e nos documentos) e foram discutidas acima. A quinta coluna indica os modelos usados no estágio de ranqueamento: indir (PTT5-INDIR-400) e base (PTT5-BASE). Também foi experimentada inversão no formato do texto passado ao modelo tanto para o INDIR (indiri) quanto para o PTT5-BASE (basei): 'Query: {document} Document: {query} Relevant:'. A ideia era verificar se o INDIR que foi treinado com enunciados como consulta (query) poderia 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.

Nas tabelas que se seguem há um gradiente na cor verde indicando 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	RANKER	NDCG@10			PRECISION@50			RECALL@100		
					bm25	sts	join	bm25	sts	join	bm25	sts	join
1	indir	----	----	----	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
	indir extra	----	----	----	41,64	17,66	20,04	11,08	5,91	7,00	74,18	45,33	54,98
				indir			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
				indir			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
				indir			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 analisadas: NDCG@10 (+0,73), RECALL@100 (+2.85) e PRECISION@50 (0.81). Praticamente todas as expansões somente de termos em consultas (EXPQ_TYPE="indir") promoveram um acréscimo no RECALL@100[5]. 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 uma precisão 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.

[5] Exceção feita a EXPQ_CNT = 8 expansões

EXPQ TYPE	EXPQ CNT	EXPD TYPE	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
indir	1	indir-1	term		83,79	
		indir-3	term		84,76	
		indir-5	term		84,89	
		user	+syn		86,38	13,29
			term		86,85	13,35
	10	user	term		83,28	
	2	user	term		85,31	13,00
	3	user	term		84,38	
	4	user	term		84,16	
	5	user	term		83,42	
	6	user	term		83,55	
	7	user	term		82,99	
	9	user	term		83,06	

Tabela 5 - Combinações que favoreceram a busca esparsa

A Tabela 6 traz as combinações que alcançaram resultados entre os top-40 superiores aos valores base considerados em alguma das métricas. O Apêndice 2 traz todas as 284 combinações que tiveram algum valor superior, não só os top-40. Há algumas promissoras que alcançaram valores entre os melhores em todas as métricas e com todos os retrievers considerados (bm25, sts e join). No retriever sts, os embeddings foram gerados pelo modelo rufimelo/Legal-BERTimbau-sts-large-ma-v3. O tipo join é uma união de resultados de 2 retrievers: sts com top-k 150 e bm25 com top-k 150, totalizando um top-k do estágio 1 de 300, mesmo valor usado para os demais retrievers.

Percebe-se que, nas buscas sem ranqueamento, o retriever bm25 alcançou melhores resultados do que os outros dois retrievers Porém nas buscas com ranqueamento, a configuração com retriever join alcançou melhores resultados do que com bm25. Talvez um indicativo de que a similaridade semântica precisa ser considerada, até para tratar casos de *vocabulary mismatch*, mas sem deixar de lado a efetividade do bm25.

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
indir	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. No critério de maior número de casos de superação, a métrica mais impactada foi o recall@100 (96,8%: 276 de 284), seguida pela prescision@50 (69.7%: 199 de 284) e NDCG@10 (30.6%, 88 de 276),

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 (+17.80). Já nas métricas precision@50 e recall@100, os melhores resultados se deram com a expansão no índice com termo: +3.16 e +12.48.

TODO

Conclusão:

importância da ordem no prompt: query x document

basei: afetado com inversão, provavelmente porque o texto da consulta ficou grande e bem maior do que o de documento

indiri: melhor com mais expansões (mais próximo do documento VCE para o qual foi treinado)

comentar indiri: melhora com expansões das queries, mas ainda abaixo do indir

apontar resultados para csv experimentos (e detalhe por query)

tirar trabalho futuro: treinar mais

três abordagens: expansor queries; expansão de docs; re-ranqueamento

comentar que usar o indir como expansor do documento não trouxe melhor resultado. do que as expansões do próprio usuário, embora alcancem métricas superiores (colocar apêndice tudo de recall)

Para o relatório:

Avaliar sts com expansões? Melhor ou pior do que bm25/join?

Falta mudar o número do Apêndice

Falta revisão do texto

4. Conclusão

Os datasets em português contruídos 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 claro. 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 scores 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 bem gerados se mostrou viável com ótimos resultados demonstrados nas métricas ndcg@12 e rank1.

Um modelo pequeno bem treinado pode alcançar resultados melhores do que modelos bem maiores modelos não treinados. O MINILM com 106 milhões de parâmetros superou o MT5-3B com 3 bilhões no dataset JURIS-TCU-INDEX.

O uso do indexador como expansor de consultas ou como expansor de documentos impactou todas as configurações de buscas realizadas. Com as expansões, buscas esparsas (só bm25) alcançaram melhores resultados: ndcg@10: +0.55; precision@50: +0.61 e recall@100: +3.88. Buscas com ranqueamento se beneficiaram das expansões nas métricas de precision@50 e recall@100 alcançando +3.16 e +12.48 respectivamente (comparações com o valor alcançado com bm25).

O ranqueador PTT5-BASE superou o INDIR-106 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-106 por ter se especializado no domínio JURIS-TCU-INDEX (in-domain) não consiga resultados similares fora do domínio do treino (out-of-domain).

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 como possíveis 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 o modelo PTT5-INDIR com mais épocas e modelos maiores (com 3 bilhões ou mais de parâmetros) visando 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. Outras métricas, outros filtros (sem restringir aos termos já usados) e outras análises (casos não encontrados). 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.

Quanto às expansões, cabe experimentar o próprio indexador INDIR gerando as expansões nos documentos dos índices. Tendo o mesmo critério as expansões dos dois lados (consultas e documentos), estimam-se resultados ainda melhores. Também espera-se a aplicação do processo aqui desenvolvido de INDexação apoiando Information Retrieval (INDIR) em outros sistemas, do próprio TCU ou de outras instituições governamentais e privadas que fazem uso de tesauros como indexação.

6. Agradecimentos

Ao TCU, não só pela disponibilização da nuvem com GPT3.5 e 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. Brasilia**, 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.

8. Apêndices

8.1. Dataset JURIS-TCU - Scores de avaliação por consulta

Figura 11 - Quantidade de avaliações por score em cada query

8.2. Experimentações com resultado favorável

EXPQ CNT	EXPQ TYPE	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	indir	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	indir				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		
				indir				87,80	89,36	13,57	13,95
			term	----				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	25,81	85,79	60,49	13,73	7,92
				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
			term	----	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
				indiri				85,19	89,23	12,60	13,45
1	indir	----	----	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	indir				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
		indir-3	+rel	indir				82,79	83,72		
			+syn	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
			+syn	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
			+syn	----				86,38	61,65	13,29	7,95
				indir				87,74	90,42	13,43	13,71
			+syn+rel	indir				85,37	86,63	13,12	13,32
			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

EXPQ CNT	EXPQ TYPE	EXPD TYPE	EXPD VAL	RANKER	NDCG@10			RECALL@100		PRECISION@50	
					bm25	sts	join	bm25	join	bm25	join
	indir_extra	----	----	indir					85,47		13,19
		indir-5	term	indir					83,38		
		user	term	indir					84,07		
2	indir	----	----	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	indir	----	----	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		
				indir				87,11	88,67	13,49	13,65
				indiri				87,37	89,47	13,51	13,81
4	indir	----	----	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	indir				84,34	85,06	12,93	13,12
		indir-3	term	indir				84,17	85,76	12,84	13,03
		indir-5	term	indir				83,96	85,25	12,80	13,09
				indiri				85,56	86,28	13,32	13,53
		user	+rel	indir				83,80	84,44	12,77	12,93
			+syn	indir				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	----	----	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	indir				82,79	84,27		
		indir-3	term	indir				84,01	84,73		
		indir-5	term	indir				83,41	84,16		
				indiri				84,82	86,29	13,21	13,39
		user	+rel	indir				82,34	83,93		
			+syn	indir				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
	indir_extra	user	term	indir					83,71		
6	indir	----	----	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	indir				82,09	83,97	12,71	13,01
		indir-3	term	indir				83,39	84,57		

EXPQ CNT	EXPQ TYPE	EXPD TYPE	EXPD VAL	RANKER	NDCG@10			RECALL@100		PRECISION@50	
					bm25	sts	join	bm25	join	bm25	join
		indir-5	term	indir				83,61	83,92		
				indiri				85,73	86,37	13,40	13,51
		user	+rel	indir				82,21	83,03		
			+syn	indir				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	----	----	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	indir				82,47	83,03		
		indir-3	term	indir				83,70	84,09		
		indir-5	term	indir				83,20	83,12		
				indiri				84,71	85,48	13,23	13,35
		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	----	----	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	----	----	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	----	----	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 favorável em pelo menos uma métrica