



Projeto Final

IA368DD_2023S1: Deep Learning aplicado a Sistemas de Buscas

Student: Marcus Vinícius Borela de Castro

IndIR

Indexing Improving Information Retrieval in the Juris Dataset
(Jurisprudence Statements and Thesaurus of the
Federal Court of Accounts of Brazil - TCU)

Leonardo Pacheco & Marcus Borela

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




IndIR - Datasets and other prepared files

JURIS_TCU

Collection of statements from the [Selected Jurisprudence of the Federal Court of Auditors](#).

Files:

- [doc.csv](#) - each document contains a statement from the [Selected Jurisprudence](#).
- [query1.csv](#) - 50 queries generated from the access log of the [Integrated Research of the Federal Court of Auditors](#) (most executed queries).
- [query2.csv](#) - transformation of questions in [query3](#) to search expression, removing part of the words used.
- [query3.csv](#) - 50 queries, each generated by LLM from a statement among the most accessed in the access log of the [Integrated Research of the Federal Court of Auditors](#).
- [qrel.csv](#) 



O enunciado procura retratar o entendimento contido na deliberação da qual foi extraído, não constituindo, todavia, um resumo oficial da decisão proferida pelo Tribunal. Tampouco objetiva representar o posicionamento prevalecente no TCU sobre a matéria.

ACÓRDÃO:

[Acórdão 602/2008-Plenário](#)

DATA DA SESSÃO:

11/03/2008

RELATOR:

BENJAMIN ZYMLER

ÁREA:

Desestatização

TEMA:

Concessão pública

SUBTEMA:

Estudo de viabilidade

OUTROS INDEXADORES:

Requisito, Regulação, **Leilão**, Preço máximo

TIPO DO PROCESSO:

ACOMPANHAMENTO

ENUNCIADO:

A definição do valor máximo de **leilão**, em processo de regulação, assenta toda a viabilidade do empreendimento, de forma que seu estabelecimento deve representar o correto balanceamento entre a modicidade tarifária e a atratividade comercial.

JURIS_TCU_INDEX

Indexing database for the statements of the [Selected Jurisprudence](#) using terms from the [External Control Vocabulary of the Federal Court of Auditors \(VCE\)](#).

Files:

- [doc.csv](#) - each document contains the definition of a term from the [VCE](#).
- [query.csv](#) - each query is a statement from the [Selected Jurisprudence](#).
- [qrel.csv](#) - each record corresponds to an indexing of a statement from the [Selected Jurisprudence](#) by a term from the [VCE](#). This indexing was performed by operators of the Jurisprudence system and can be observed in the [Integrated Research of the Federal Court of Auditors](#) - [Selected Jurisprudence database](#).

5. INDEXAÇÃO DO ENUNCIADO JURISPRUDENCIAL

5.1 Etapas do Processo de Indexação

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). Exemplo:

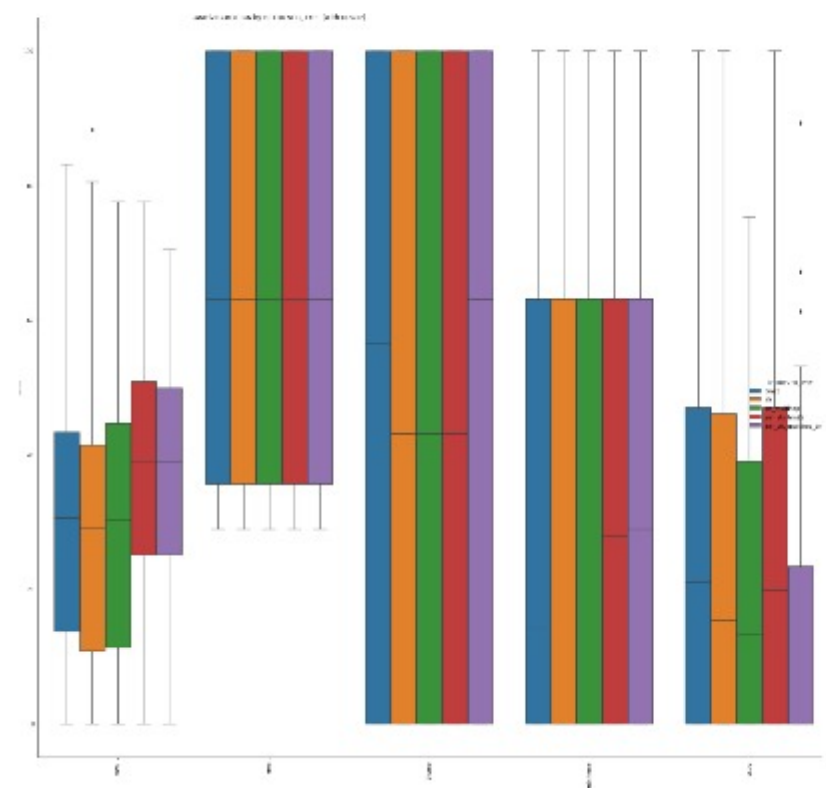
Contrato Administrativo. Aditivo. Serviço novo. Preço. Referência. Orçamento estimativo. BDI. Equilíbrio econômico-financeiro. Desconto.

Quando houver a celebração de aditivos contratuais para a inclusão de novos serviços, tanto nos regimes baseados em preço global quanto nos regimes de empreitada por preço unitário e tarefa, o preço desses serviços deve ser calculado considerando as referências de custo e taxa de BDI especificadas no orçamento- base da licitação, subtraindo desse preço de referência a diferença percentual entre o valor do orçamento-base e o valor global obtido na licitação, com vistas a garantir o equilíbrio econômico-financeiro do contrato e a manutenção do percentual de desconto oferecido pelo contratado (art. 37, inciso XXI, da Constituição Federal e arts. 14 e 15 do Decreto 7.983/2013).

A indexação compreende duas etapas distintas: a análise conceitual do assunto e a tradução dos conceitos em descritores.

Experimentados alguns pipelines de busca como indexador

Indexador



2 exemplos não relevantes (os primeiros 2 retornados daquela categoria) para cada relevante

Criados dados de treinamento para monotS a partir do pipe escolhido

```
[ ] 1 train_examples['query'][0], train_examples['text'][0], train_examples['label'][0]
```

```
{'SÚMULA TCU 1: Não se compreendem como vencimento, para efeito de concessão de pensão especial com fundamento na Lei nº 3.738, de 04/04/60, as vantagens previstas no art. 184 da Lei nº 1.711, de 28/10/52.',  
  'O termo é "Pessoal".\nPessoal tem nota de escopo: "Designação genérica de todos os servidores ou funcionários civis pertencentes ao quadro de pessoal de um órgão ou entidade".\nPessoal tem nota de escopo: "Tema agrupador para área de atuação do Controle Externo".\nPessoal é uma generalização de: "Servidor público", "Funcionário público", "Pessoal civil", "Pessoal militar", "Pessoal temporário" e "Colaborador".\nPessoal tem termo relacionado: "Sisac", "Despesa com pessoal" e "HCAAF".\nPessoal tem tradução em espanhol: "Recursos humanos".\nPessoal tem tradução em inglês: "Personal" e "Human resources".',  
  1)
```

Tentando treinar unicamp-dl/mt5-3B-mmarco-en-pt

T **Thiago Soares Laitz** 8h17
@Marcus Borela vocês estão treinando o t5 pra classificar se o documento é ou não relevante, é isso mesmo ?

 **Marcus Borela** 8h37
Isssso **@Thiago Soares Laitz**

O reranker que está no hugging face: unicmp-dl/mt5-3B-mmarco-en-pt

T  **3 respostas** Última resposta hoje à(s) 8h57

T **Thiago Soares Laitz** 12h26
Oi **@Marcus Borela**, vou te mandar o código aqui, mas eu tive que adaptar bastante coisa para ficar mais facil para vocês (eu estava treinando o t5 para uma tarefa diferente) (editado)



Hoje a noite vou testar esse código no colab pra ver se não quebrei nada

T **Thiago Soares Laitz** 12h27
Estou mandando agora caso queira ir dando uma olhada

GPU A100; Batch size=2

OutOfMemoryError: CUDA out of memory.

Tried to allocate 16.00 MiB (GPU 0; 39.56 GiB total capacity; 36.46 GiB already allocated; 16.56 MiB free; 38.23 GiB reserved in total by PyTorch)

Mostrando os recursos de 21:19 a 21:32

RAM do sistema
6.4 / 83.5 GB



RAM da GPU
39.5 / 40.0 GB



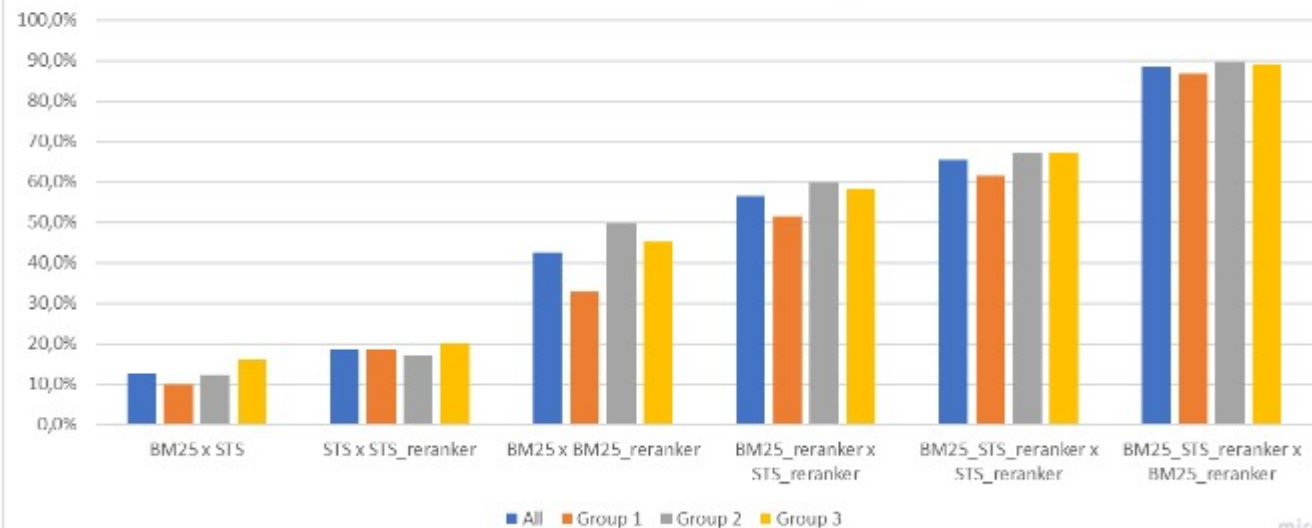
Disco
38.2 / 166.8 GB



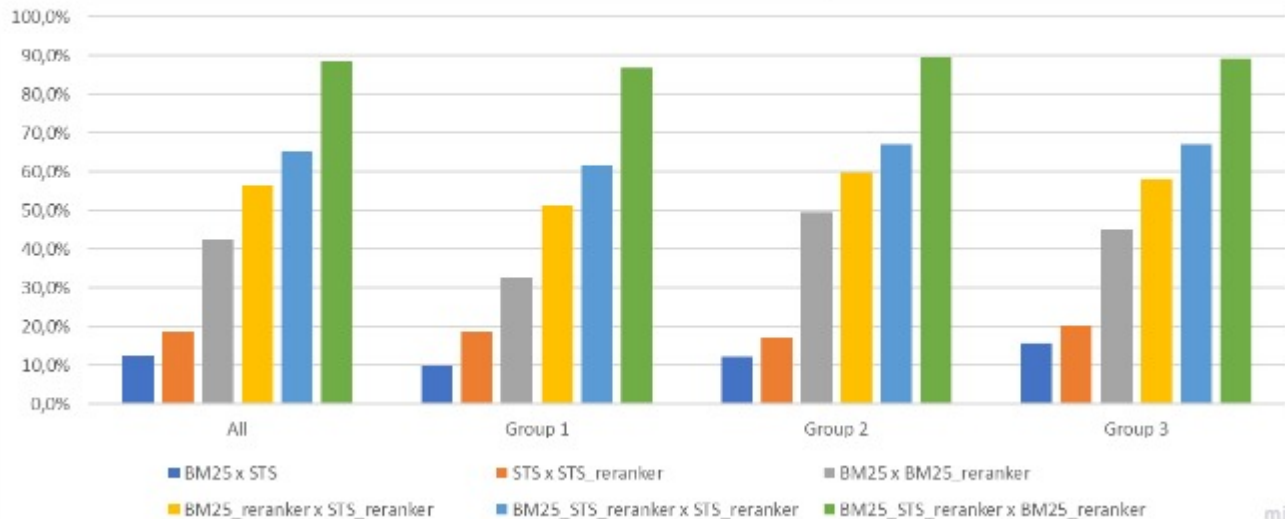


Busca

Similaridade de resultados entre Pipelines



Similaridade de resultados entre Pipelines



Avaliação por LLM

prompt

Human: 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 consulta.

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

0 - irrelevante - o enunciado não responde a consulta;

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

2 - relevante - o enunciado responde parcialmente a consulta;

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

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

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

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

o valor de Razão deve ser a motivação da escolha do score.

Consulta: concessão remunerada de uso de bens públicos modalidade

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.

AI:

{'Score': 2, 'Razão': 'O enunciado aborda a concessão remunerada de uso de bens públicos e menciona o pregão como modalidade de licitação adequada, mas não trata de todas as nuances possíveis relacionadas ao tema.'}

query x docs relevantes

```
for idx, row in df_join_bm25_stc_reranker.query('RANK <= 10 and QUERY_ID == 101').iterrows():
    score, reasoning = get_llm_score(row['QUERY TEXT'], row['DOC TEXT'])
    df_join_bm25_stc_reranker.loc[idx, 'LLM_SCORE'] = score
    df_join_bm25_stc_reranker.loc[idx, 'LLM_REASONING'] = reasoning
df_join_bm25_stc_reranker.query('RANK <= 10 and QUERY_ID == 101').sort_values('RANK')
```

✓ 49.6%

| QUERY_ID | RANK | DOC ID | ENGINE | QUERY TEXT | DOC TEXT | LLM SCORE | LLM REASONING |
|----------|------|--------|-----------------------------|---|--|-----------|--|
| 43851 | 101 | 1 | 2845 (BM25[STS]+ReRanker) | Qual é a modalidade de licitação adequada para... | Em regra, o prego é a modalidade de licitação... | 3.0 | O enunciado responde diretamente à pergunta, L... |
| 41428 | 101 | 2 | 17360 (BM25[STS]+ReRanker) | Qual é a modalidade de licitação adequada para... | É recomendável a utilização do prego para a c... | 2.0 | O enunciado responde parcialmente à pergunta. ... |
| 54904 | 101 | 3 | 5214 (BM25[STS]+ReRanker) | Qual é a modalidade de licitação adequada para... | A criação das áreas comerciais de centros públ... | 3.0 | O enunciado responde diretamente à pergunta, L... |
| 46534 | 101 | 4 | 58598 (BM25[STS]+ReRanker) | Qual é a modalidade de licitação adequada para... | Na licitação que tem por objeto a concessão re... | 1.0 | O enunciado está relacionado ao tópico da perq... |
| 19717 | 101 | 5 | 14862 (BM25[STS]+ReRanker) | Qual é a modalidade de licitação adequada para... | O contrato administrativo de concessão remuner... | 1.0 | O enunciado aborda o tema de concessão remuner... |
| 60024 | 101 | 6 | 149558 (BM25[STS]+ReRanker) | Qual é a modalidade de licitação adequada para... | Isso impõe legal a utilização do modelo de ... | 1.0 | O enunciado de jurisprudência aborda a utilizac... |
| 74510 | 101 | 7 | 10865 (BM25[STS]+ReRanker) | Qual é a modalidade de licitação adequada para... | É cabível a utilização do prego para concessõ... | 2.0 | O enunciado aborda a modalidade de licitação p... |
| 41179 | 101 | 8 | 17029 (BM25[STS]+ReRanker) | Qual é a modalidade de licitação adequada para... | Para a aquisição de bens, o manual de Administraç... | 1.0 | O enunciado de jurisprudência aborda a modalid... |
| 41001 | 101 | 9 | 19399 (BM25[STS]+ReRanker) | Qual é a modalidade de licitação adequada para... | Deve ser utilizada a modalidade prego para aq... | 1.0 | O enunciado aborda o tema da modalidade de li... |
| 73636 | 101 | 10 | 34109 (BM25[STS]+ReRanker) | Qual é a modalidade de licitação adequada para... | Nas licitações para a concessão de serviços públ... | 1.0 | O enunciado está relacionado ao tema de licita... |

query x docs ñ relevantes

```
query_df = 'QUERY_ID == 101 and (RANK == ' + ' or RANK == '.join([str(r) for r in list(np.random.choice(700, 3, replace=False) + 301)]) + ')\nquery_df
```

✓ Uu

```
'QUERY_ID == 101 and (RANK == 873 or RANK == 400 or RANK == 557)'
```

```
for i, (idx, row) in enumerate(df_bm25.query(query_df).iterrows()):  
    score, reasoning = get_llm_score(row['QUERY_TEXT'], row['DOC_TEXT'])  
    df_bm25.loc[idx, 'LLM_SCORE'] = score  
    df_bm25.loc[idx, 'LLM_REASONING'] = reasoning  
df_bm25[df_bm25['LLM_SCORE'].notnull()].sort_values(['QUERY_ID', 'RANK'])
```

| | QUERY_ID | RANK | DOC_ID | ENGINE | QUERY_TEXT | DOC_TEXT | LLM_SCORE | LLM_REASONING |
|--------|----------|------|--------|--------|---|---|-----------|---|
| 50273 | 101 | 400 | 16858 | BM25 | Qual é a modalidade de licitação adequada para... | Oiente de contratação que possa levar à negati... | 0.0 | O enunciado de jurisprudência aborda o tema de... |
| 120024 | 101 | 557 | 13775 | BM25 | Qual é a modalidade de licitação adequada para... | Não é necessário desconsiderar a personalidade... | 0.0 | O enunciado de jurisprudência não aborda a con... |
| 76157 | 101 | 873 | 31572 | BM25 | Qual é a modalidade de licitação adequada para... | É causa de impedimento para participar de lic... | 0.0 | O enunciado de jurisprudência aborda o tema de... |

IndIR - Indexing Improving Information Retrieval in the Juris Dataset (Jurisprudence Statements and Thesaurus of the Brazil Federal Court of Accounts - TCU)

Leonardo Pacheco e Marcus Borela

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

| | | | | | Se m | Semana 6 | | | | | | | |
|---|-------------------|-----------|---------|---------|---------|----------|----|----|----|----|----|----|----|
| | | | | | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 |
| TAREFA | ATRIBUÍDO PARA | PROGRESSO | INÍCIO | TÉRMINO | q | q | s | s | d | s | t | q | q |
| Planejamento | | | | | | | | | | | | | |
| Plano básico | LM | 100% | 18/5/23 | 2/6/23 | | | | | | | | | |
| 1a Apresentação | LM | 100% | 22/5/23 | 24/5/23 | | | | | | | | | |
| Definir base de dados | LM | 100% | 23/5/23 | 31/5/23 | | | | | | | | | |
| Definir métricas | LM | 100% | 25/5/23 | 18/6/23 | | | | | | | | | |
| Construção do Dataset de Indexação (Juris-TCU-Index) | | | | | | | | | | | | | |
| Carga dos dados do sistema Juris (incluindo indexações) | LM | 100% | 27/5/23 | 2/6/23 | | | | | | | | | |
| Explorar dados do sistema Juris | LM | 100% | 1/6/23 | 5/6/23 | | | | | | | | | |
| Explorar dados de indexações | LM | 100% | 5/6/23 | 7/6/23 | | | | | | | | | |
| Tratar dados | LM | 100% | 27/5/23 | 13/6/23 | | | | | | | | | |
| Gerar dataset (com qrels) | M | 100% | 7/6/23 | 14/6/23 | | | | | | | | | |
| Validar formato dataset | M | 100% | 7/6/23 | 9/6/23 | | | | | | | | | |

| Construção do Dataset de Avaliação de Busca (Juris-TCU) | | | | |
|---|----|------|---------|---------|
| Explorar dados de log da busca | L | 100% | 1/6/23 | 14/6/23 |
| Criar queries | L | 100% | 5/6/23 | 15/6/23 |
| Definir prompt LLM para qrel | LM | 90% | 15/6/23 | 17/6/23 |
| Montar base de avaliação (qrel) | L | 40% | 16/6/23 | 25/6/23 |
| Construção do pipeline de indexação | | | | |
| Criar estrutura busca (elastic search, etc) | M | 100% | 4/6/23 | 7/6/23 |
| Criar estrutura de registro e avaliação (métrica) | M | 100% | 4/6/23 | 12/6/23 |
| Realizar experimentos: indexador como pipeline de busca | LM | 100% | 12/6/23 | 21/6/23 |
| Treinar modelo | M | 5% | 16/6/23 | 23/6/23 |
| (ou: experimentar LLM como estágio 3 dos pipes) | M | 0% | 16/6/23 | 26/6/23 |
| Frente pipeline de busca | | | | |
| Criar estrutura busca (elastic search, etc) | M | 90% | 17/6/23 | 26/6/23 |
| Criar estrutura de registro e avaliação (métrica) | M | 90% | 17/6/23 | 26/6/23 |
| Experimentar: indexadores como expansão (cadastro) | LM | 0% | 20/6/23 | 25/6/23 |
| Frente Avaliação e Fechamento | | | | |
| Relatório | LM | 0% | 16/6/23 | 30/6/23 |
| Preparar apresentação final | LM | 0% | 24/6/23 | 30/6/23 |
| Publicar datasets com qrels (indexação e busca da jurisprudência) | LM | 60% | 4/6/23 | 30/6/23 |

Referências

R.Nogueira,Z.Jiang,R.Pradeep,andJ.Lin. Document ranking with a pretrained sequence- to sequence model. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 708–718. Disponível em: <https://arxiv.org/pdf/2010.06467.pdf>. Acesso em 25 maio 2023. 2020.

BRASIL. 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.