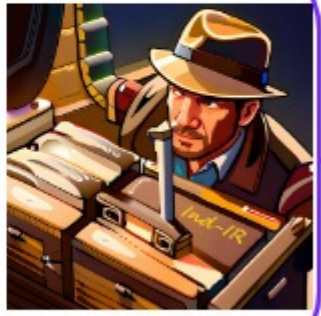
Projeto Final
IA368DD_2023S1: Deep Learning aplicado a Sistemas de Buscas
Student: Marcus Vinícius Borela de Castro

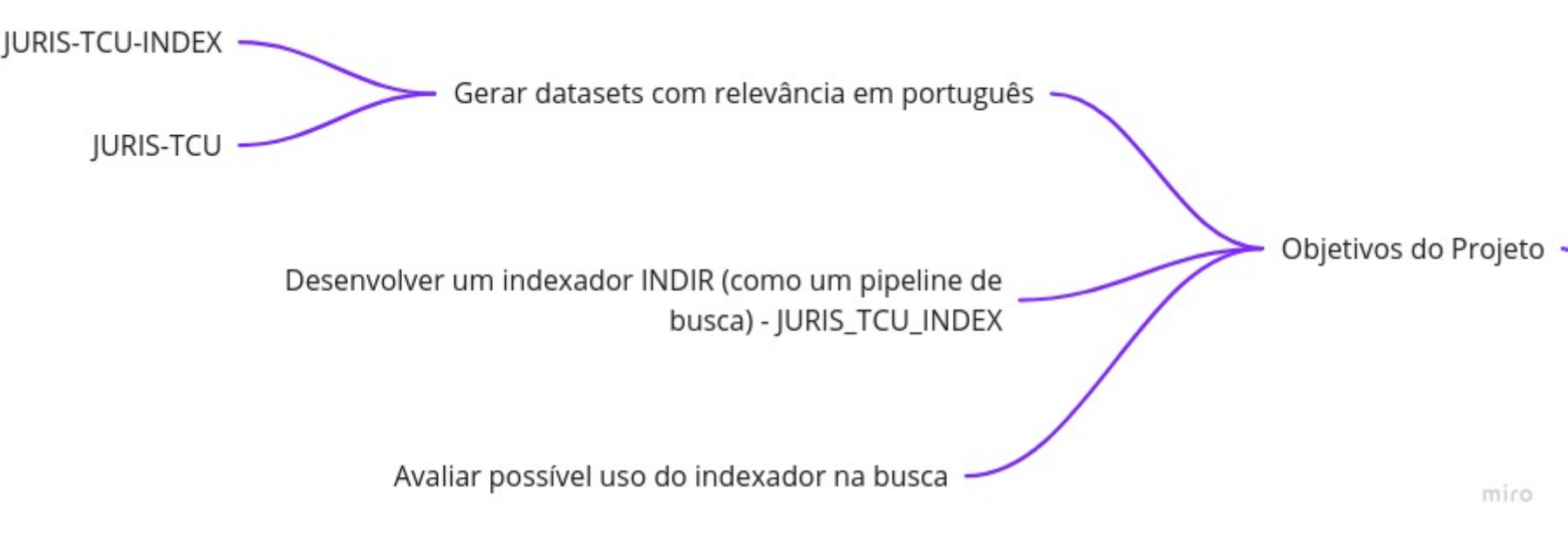
IndIR

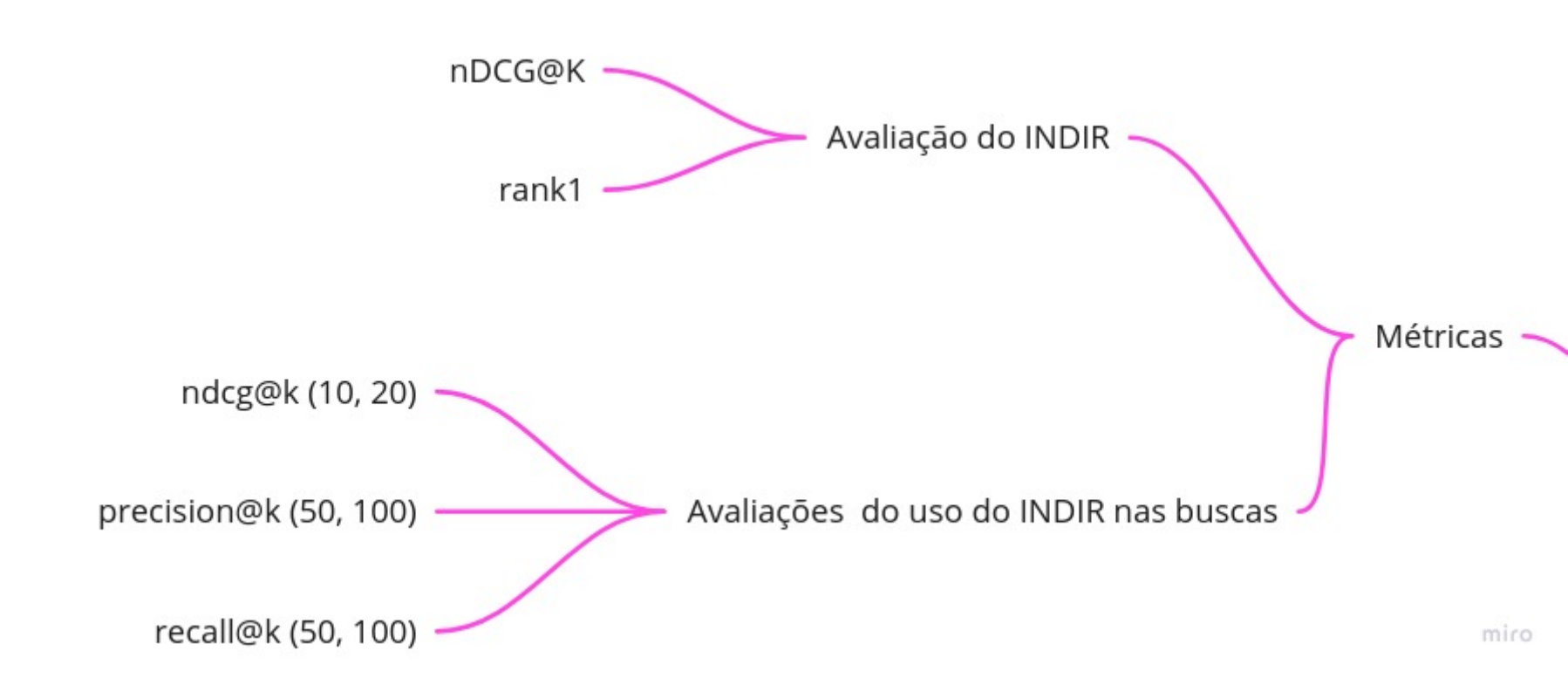
Indexing Improving Information Retrieval in the Juris Dataset
(Jurisprudence Statements and Thesaurus of the
Federal Court of Accounts of Brazil - TCU)

Leonardo Pacheco & Marcus Borela

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




JURIS-TCU-INDEX
JURIS-TCU
Gerar datasets com relevância em português
Objetivos do Projeto
Desenvolver um indexador INDIR (como um pipeline de busca) - JURIS_TCU_INDEX
Avaliar possível uso do indexador na busca
miro

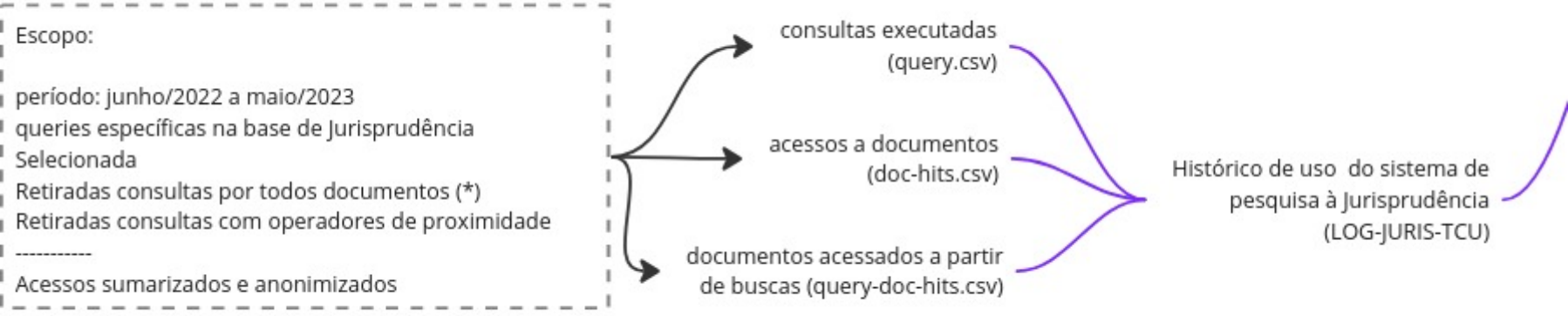

nDCG@K
rank1
Avaliação do INDIR
Métricas
ndcg@k (10, 20)
precision@k (50, 100)
recall@k (50, 100)
Avaliações do uso do INDIR nas buscas
miro

Cadastro e indexações

Sistema de Enunciados de Jurisprudências

Extração e tratamento de dados

Metodologia


Escopo:
período: junho/2022 a maio/2023
queries específicas na base de Jurisprudência Selecionada
Retiradas consultas por todos documentos (*)
Retiradas consultas com operadores de proximidade

Acessos sumarizados e anonimizados
consultas executadas (query.csv)
acessos a documentos (doc-hits.csv)
documentos acessados a partir de buscas (query-doc-hits.csv)
Histórico de uso do sistema de pesquisa à Jurisprudência (LOG-JURIS-TCU)

Sistema Vocabulário de Controle Externo (VCE)

Ação de controle

DEF: Ação de controle é uma generalização dos atos do TCU, conduzida por uma de suas unidades, tendo um ministro relator, buscando investigar aspectos de um objeto de controle, podendo levar a uma decisão (acórdão) e a uma sanção.

TE: Ação de controle externo

UP: Ações de controle

Ação de controle externo

DEF: Toda ação empreendida para a consecução da missão institucional do TCU, no âmbito de suas funções finalísticas. (Fonte: BRASIL. Tribunal de Contas da União. Glossário de Controle Externo. Revisão set.2017)

TG: Ação de controle

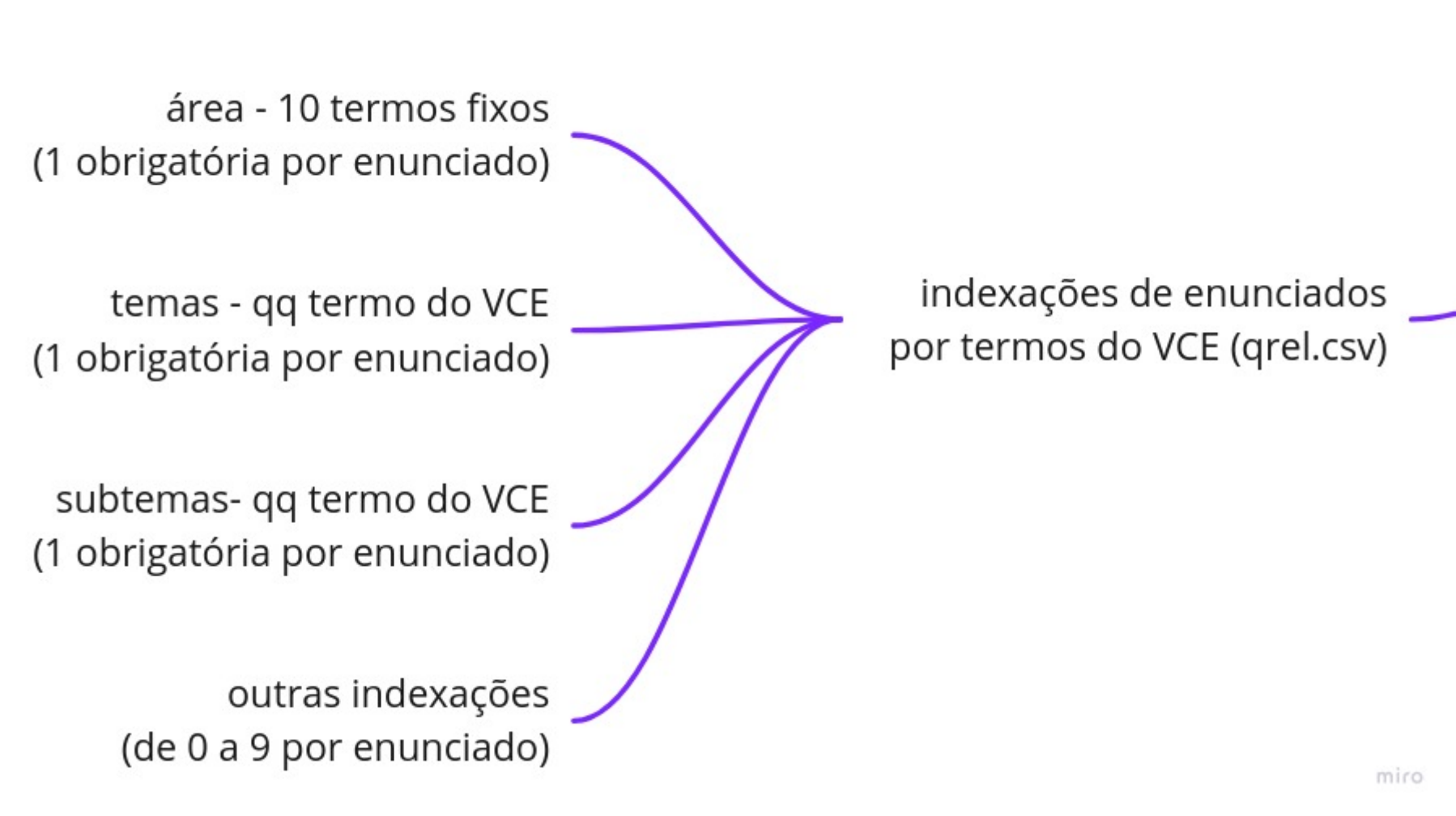
TR: Controle externo
Relatório de monitoramento
Volume de Recursos Fiscalizados
Órgão de controle externo

Ação de descumprimento de preceito fundamental

USE: Arguição de descumprimento de preceito fundamental

Ação de esbulho

USE: Ação possessória


área - 10 termos fixos
(1 obrigatória por enunciado)
temas - qq termo do VCE
(1 obrigatória por enunciado)
subtemas- qq termo do VCE
(1 obrigatória por enunciado)
outras indexações
(de 0 a 9 por enunciado)
indexações de enunciados
por termos do VCE (qrel.csv)
miro

Dados dos 2 sistemas
query: JURIS
doc: VCE
qrel: JURIS x VCE

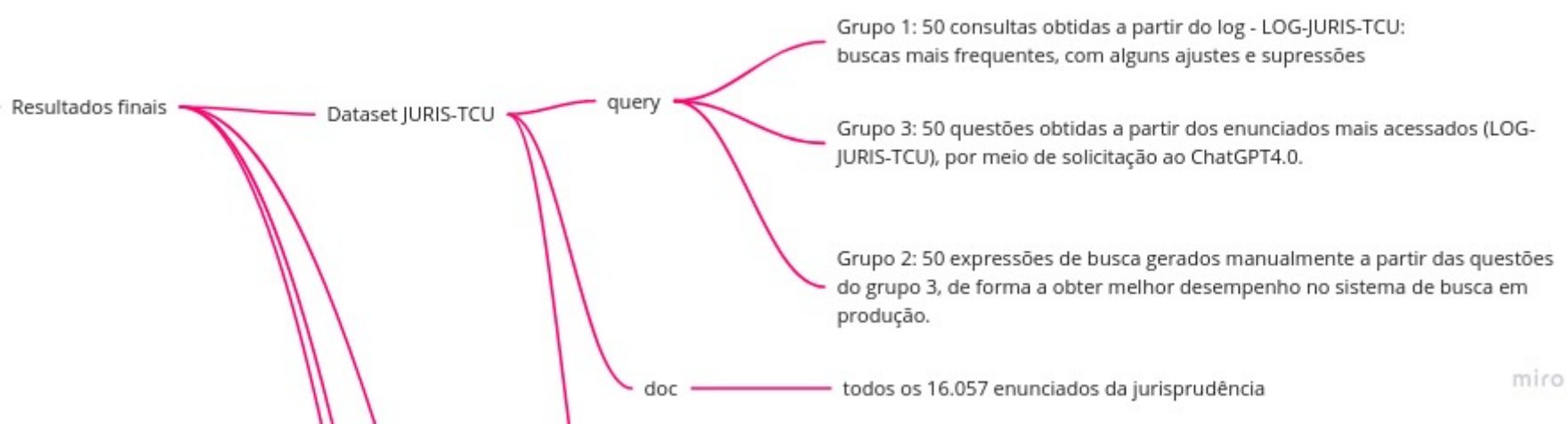
Produção do dataset de indexação
(JURIS-TCU-INDEX)

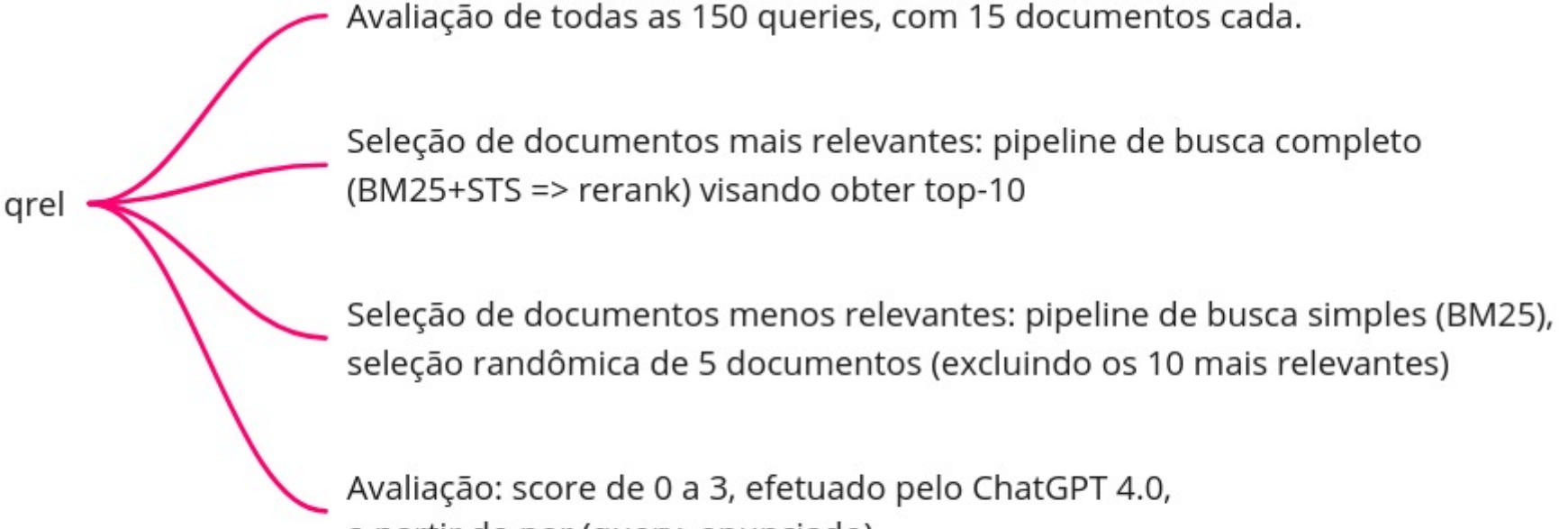
doc: JURIS
queries e qrel construídos

Produção do dataset de avaliação
(JURIS-TCU)

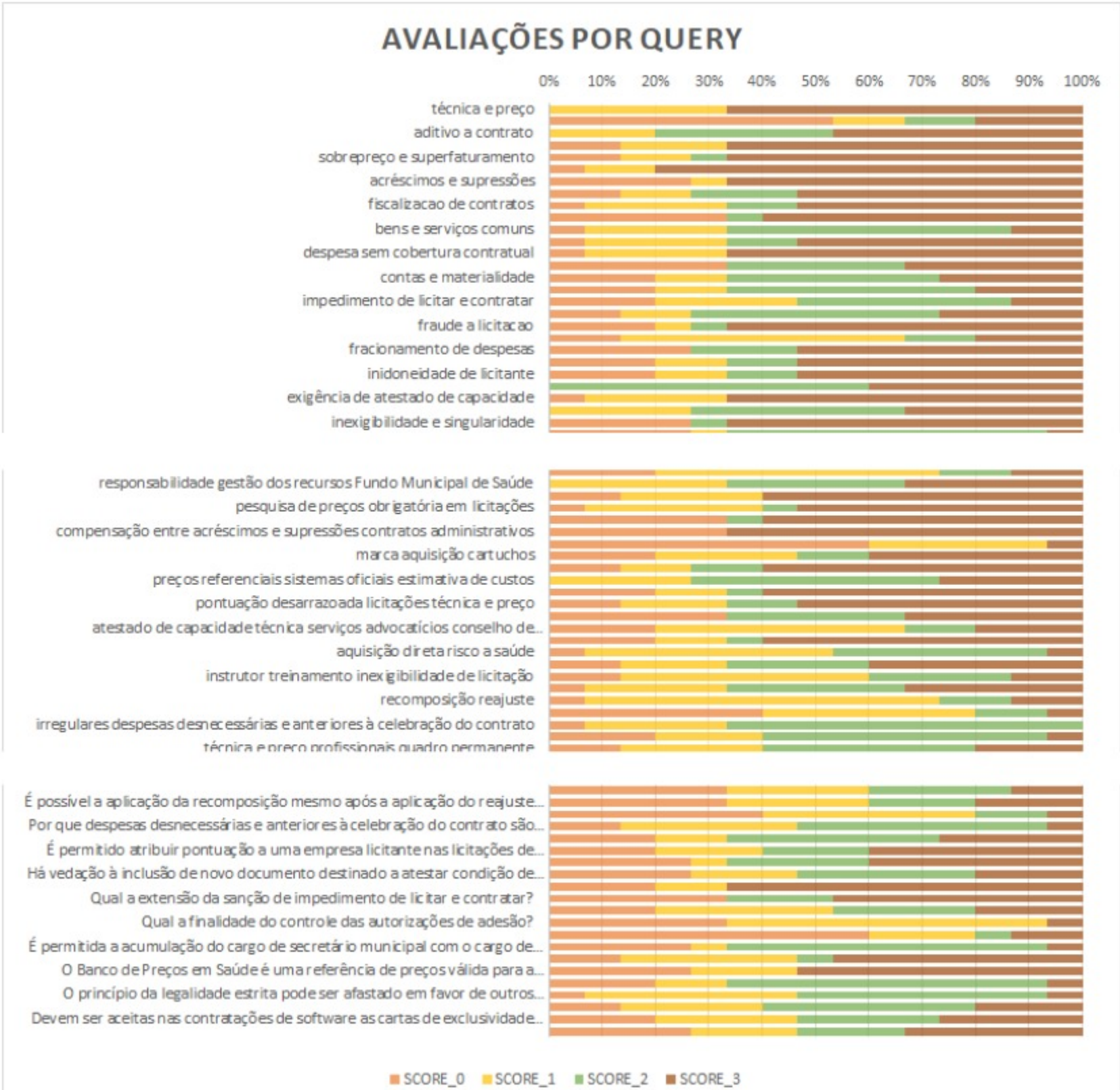
Construção do INDIR como pipeline de busca

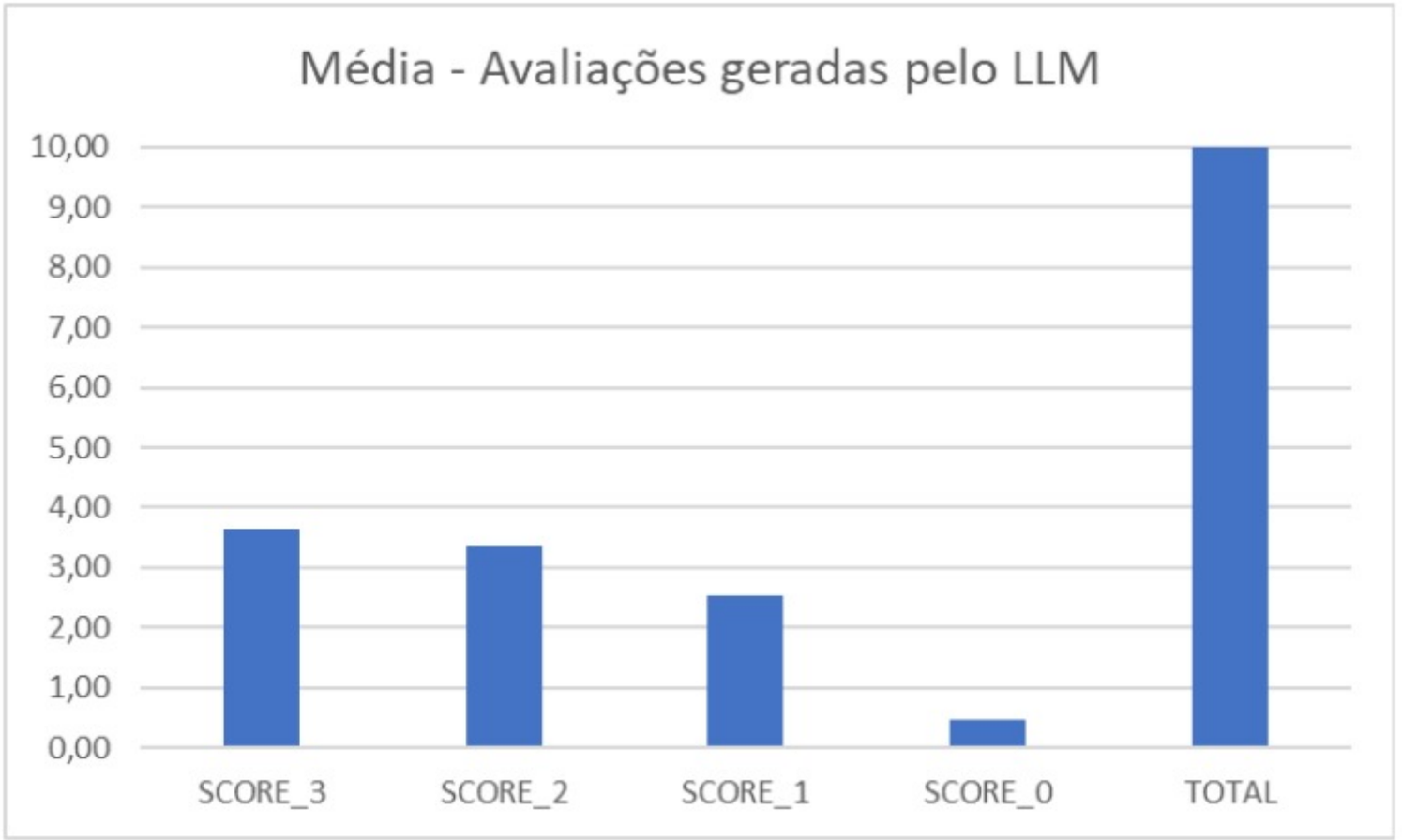
Experimentação de uso do INDIR na busca no JURIS-TCU

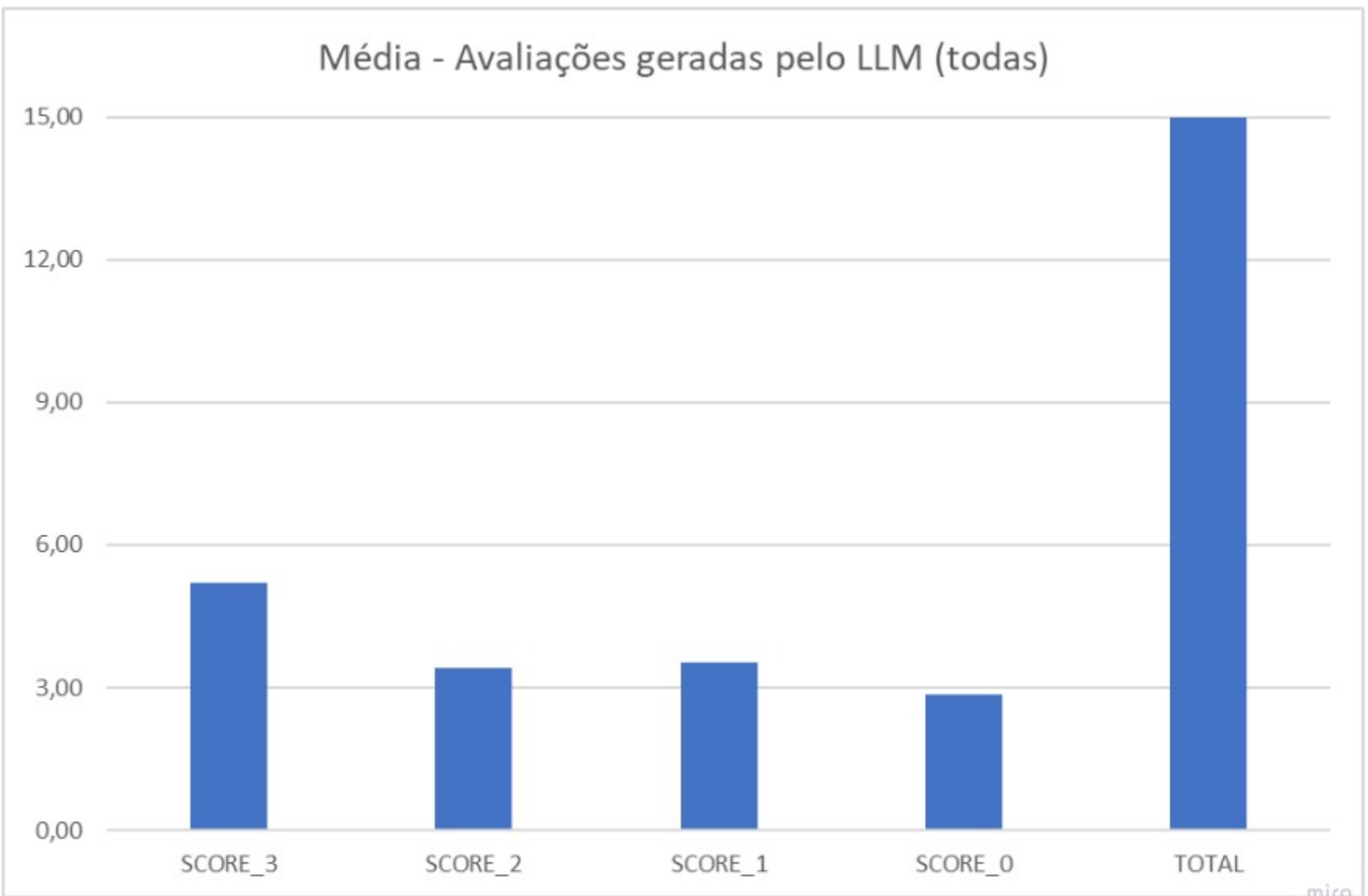

Resultados finais
Dataset JURIS-TCU
query
Grupo 1: 50 consultas obtidas a partir do log - LOG-JURIS-TCU: buscas mais frequentes, com alguns ajustes e supressões
Grupo 3: 50 questões obtidas a partir dos enunciados mais acessados (LOG-JURIS-TCU), por meio de solicitação ao ChatGPT4.0.
Grupo 2: 50 expressões de busca gerados manualmente a partir das questões do grupo 3, de forma a obter melhor desempenho no sistema de busca em produção.
doc
todos os 16.057 enunciados da jurisprudência
miro

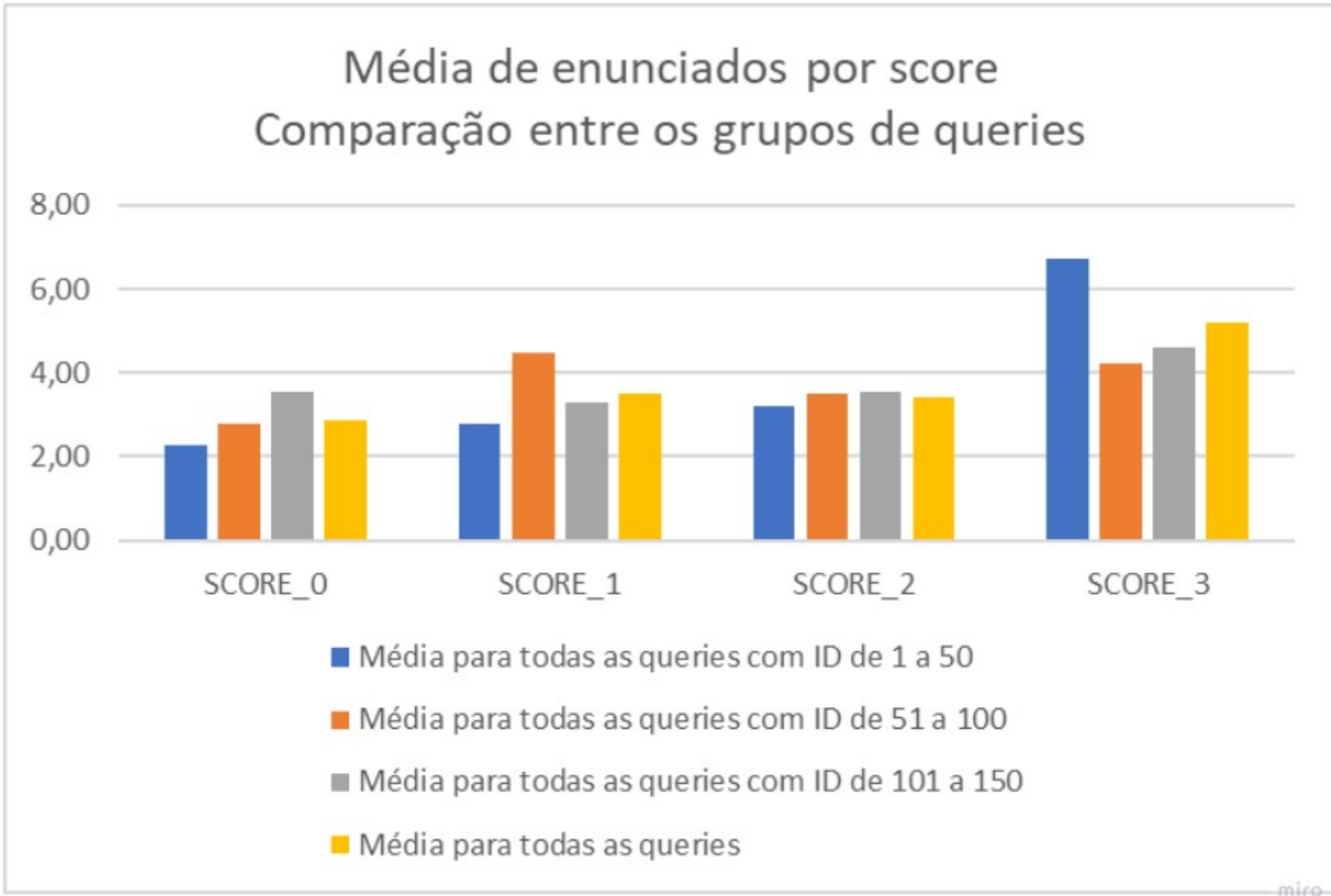

qrel
Avaliação de todas as 150 queries, com 15 documentos cada.
Seleção de documentos mais relevantes: pipeline de busca completo (BM25+STS => rerank) visando obter top-10
Seleção de documentos menos relevantes: pipeline de busca simples (BM25), seleção randômica de 5 documentos (excluindo os 10 mais relevantes)
Avaliação: score de 0 a 3, efetuado pelo ChatGPT 4.0, a partir do par (query, enunciado)

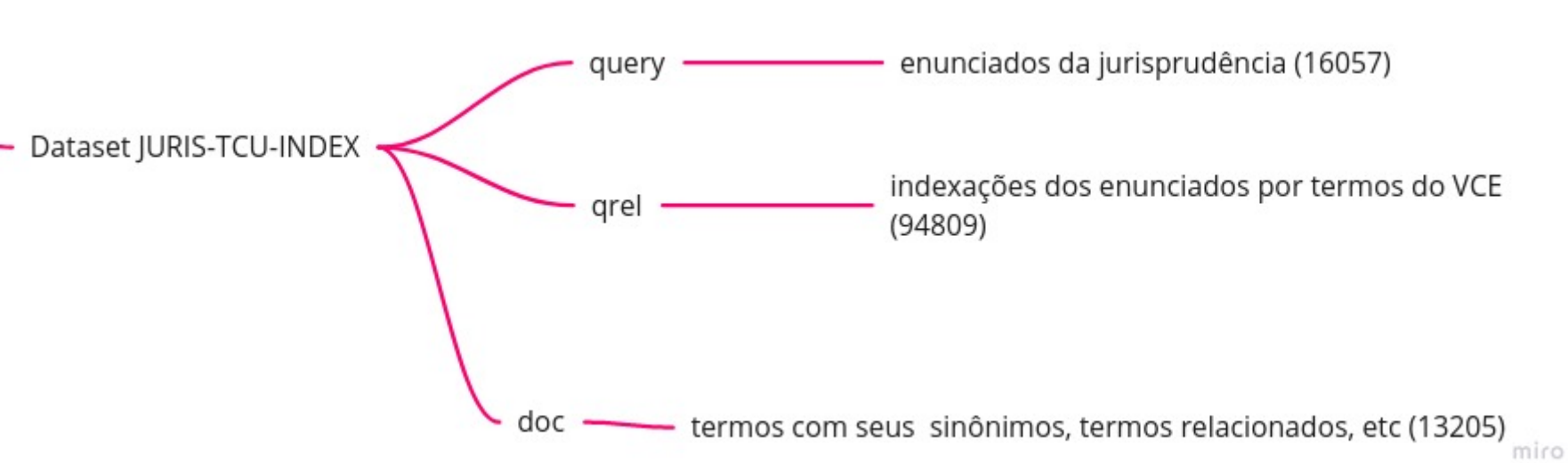
AVALIAÇÕES POR QUERY



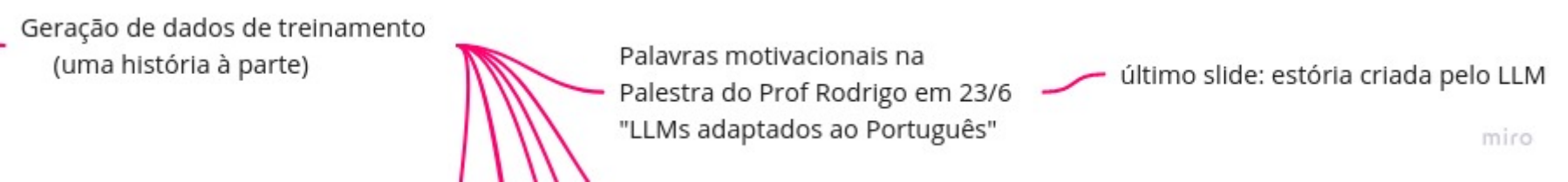

Média - Avaliações geradas pelo LLM
10,00
9,00
8,00
7,00
6,00
5,00
4,00
3,00
2,00
1,00
0,00
SCORE_3
SCORE_2
SCORE_1
SCORE_0
TOTAL


Média - Avaliações geradas pelo LLM (todas)
15,00
12,00
9,00
6,00
3,00
0,00
SCORE_3
SCORE_2
SCORE_1
SCORE_0
TOTAL
miro


Média de enunciados por score
Comparação entre os grupos de queries
8,00
6,00
4,00
2,00
0,00
SCORE_0
SCORE_1
SCORE_2
SCORE_3
Média para todas as queries com ID de 1 a 50
Média para todas as queries com ID de 51 a 100
Média para todas as queries com ID de 101 a 150
Média para todas as queries


Dataset JURIS-TCU-INDEX
query
enunciados da jurisprudência (16057)
qrel
indexações dos enunciados por termos do VCE (94809)
doc
termos com seus sinônimos, termos relacionados, etc (13205)

Indexador INDIR
Pipelines de indexação criados e experimentados
retrievers
bm25, sts, sts_multihop, join_sts_bm25
rankers
none, minilm, ptt5-base, mt5-3b
miro

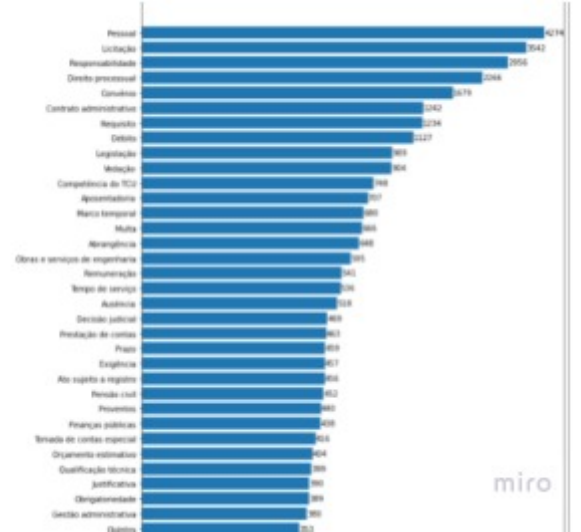

Geração de dados de treinamento
(uma história à parte)
Palavras motivacionais na
Palestra do Prof Rodrigo em 23/6
"LLMs adaptados ao Português"
último slide: estória criada pelo LLM
miro

v1: 1+, 5- (obtidos nos mais próximos)

v2: 1+, 1- (limitado a 100 por termo, já que há um grande desbalanceamento)

Termos distintos: 2859 (indexações: mediana 6.0 , média 33.16)

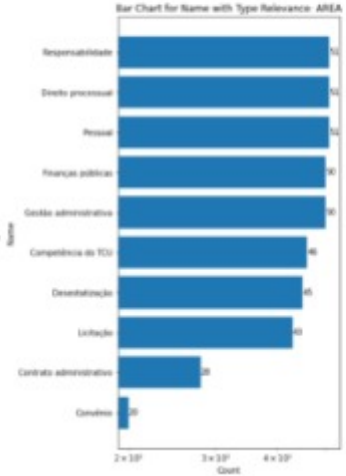
v3: 1+, 1- (limitado a 50 por termo, ordem decrescente de data - mais recentes, mais pesquisados)



Único com bons resultados (v1 a v3):
indexação com critério "AREA"

São só 10 termos
(problema de classificação, na verdade)

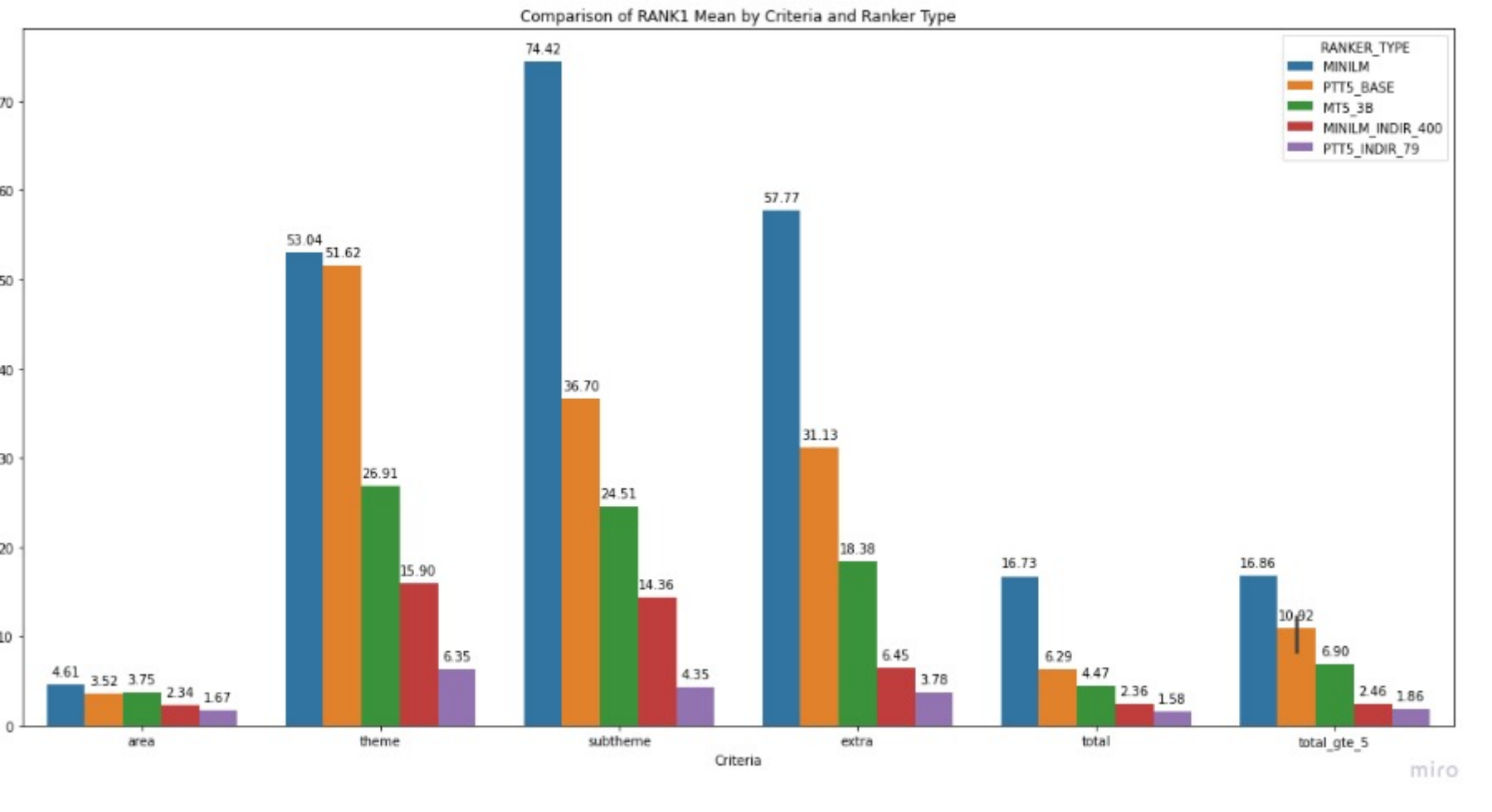
10 termos bem diferentes; pessoal, contrato, etc

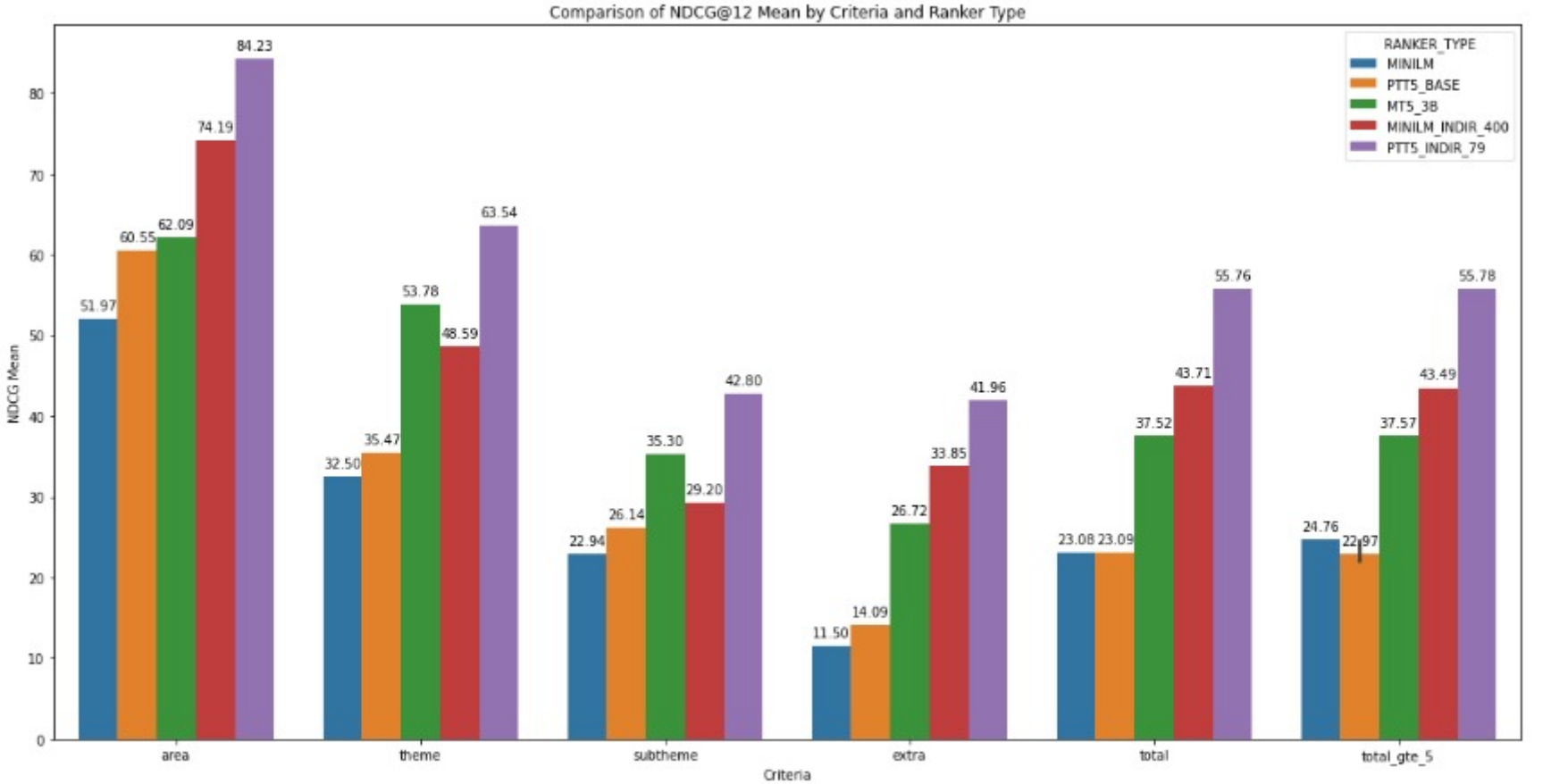


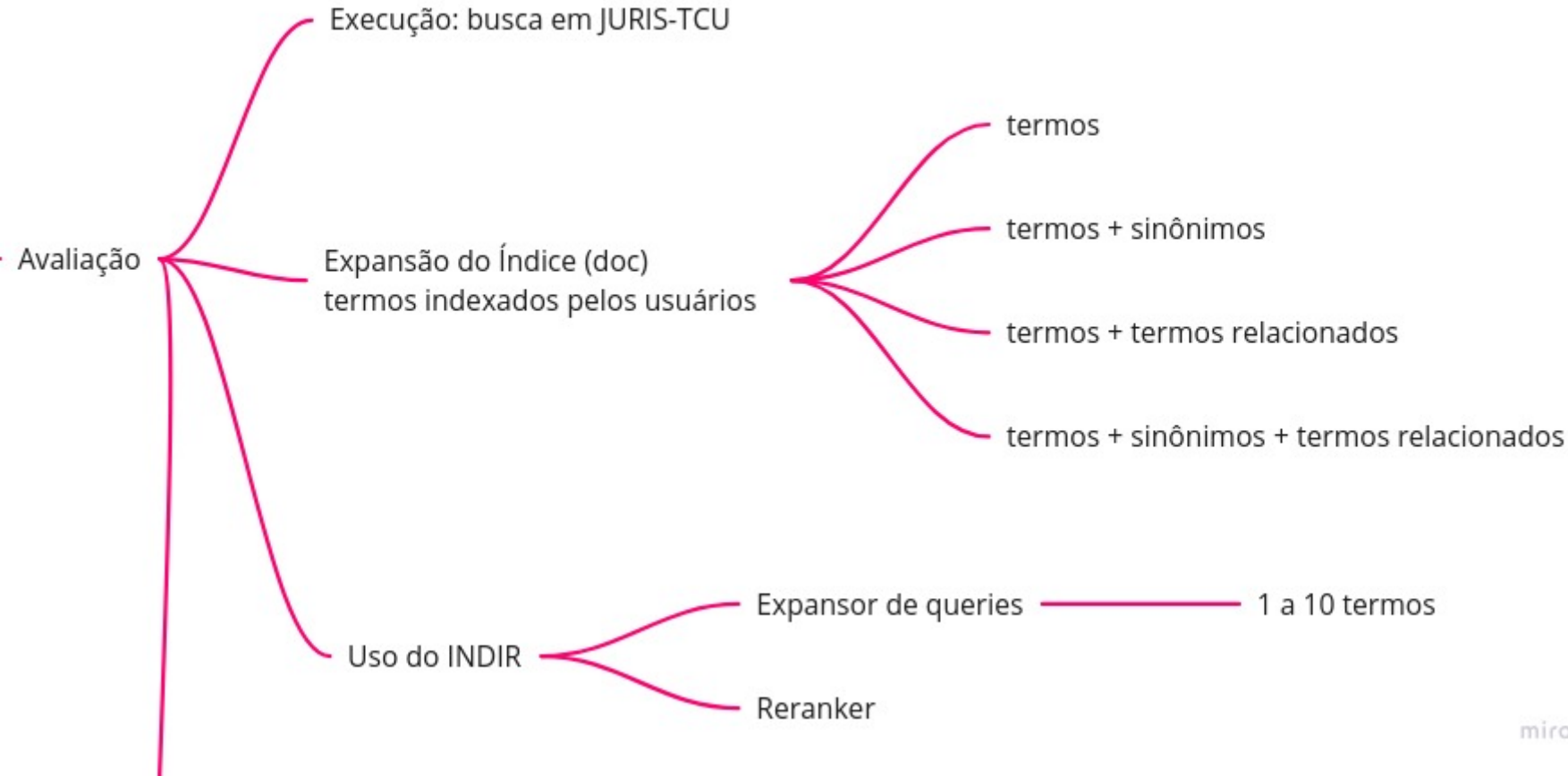
v3: 1+, 1- (sample bm25@1000; balanceamento limitando quantidade por tipo de indexação)

Modelos treinados
(gratidão Hugo e Thiago Laitz
pelo apoio no slack e código)

- PTT5_INDIR_83
  unicamp-dl/ptt5-base-pt-msmarco-100k-v2
- MINILM_INDIR_400
  unicamp-dl/mMiniLM-L6-v2-pt-v2

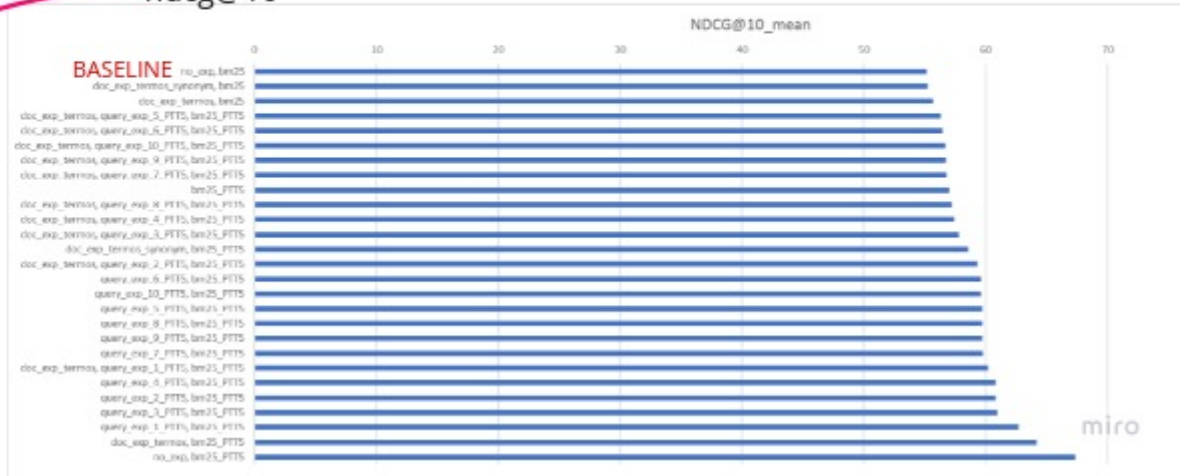

Comparison of RANK1 Mean by Criteria and Ranker Type
RANKER_TYPE
MINILM
PTT5_BASE
MT5_3B
MINILM_INDIR_400
PTT5_INDIR_79
4.61
3.52
3.75
2.34
1.67
53.04
51.62
26.91
15.90
6.35
74.42
36.70
24.51
14.36
4.35
57.77
31.13
18.38
6.45
3.78
16.73
6.29
4.47
2.36
1.58
16.86
10.92
6.90
2.46
1.86
0
10
20
30
40
50
60
70
area
theme
subtheme
extra
total
total_gte_5
Criteria
miro


Comparison of NDCG@12 Mean by Criteria and Ranker Type
RANKER_TYPE
MINILM
PTT5_BASE
MT5_3B
MINILM_INDIR_400
PTT5_INDIR_79
NDCG Mean
Criteria
area
theme
subtheme
extra
total
total_gte_5
51.97
60.55
62.09
74.19
84.23
32.50
35.47
53.78
48.59
63.54
22.94
26.14
35.30
29.20
42.80
11.50
14.09
26.72
33.85
41.96
23.08
23.09
37.52
43.71
55.76
24.76
22.97
37.57
43.49
55.78

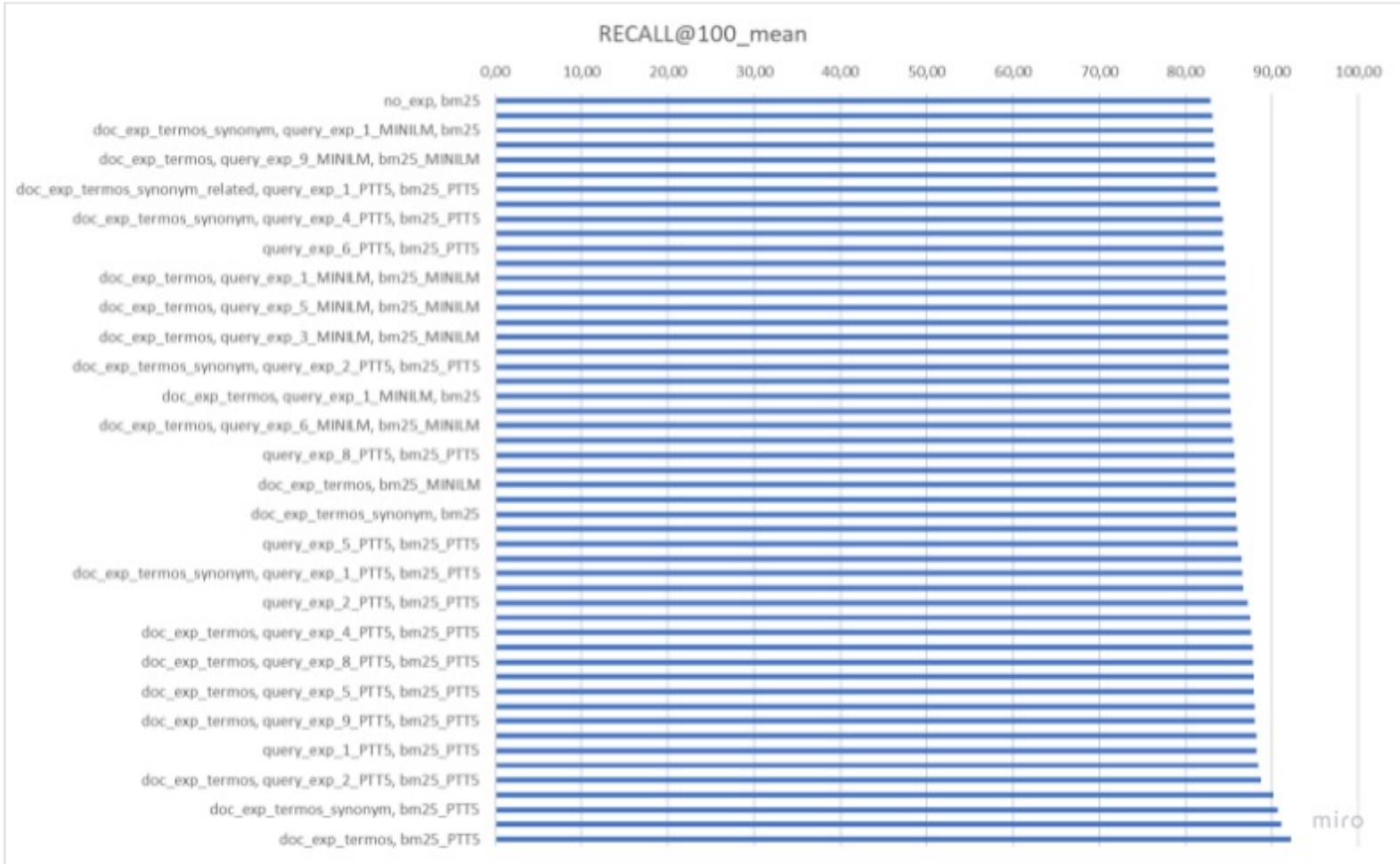

Avaliação
Execução: busca em JURIS-TCU
Expansão do Índice (doc)
termos indexados pelos usuários
termos
termos + sinônimos
termos + termos relacionados
termos + sinônimos + termos relacionados
Uso do INDIR
Expansor de queries
1 a 10 termos
Reranker
miro

Comparação
(casos vencedores)

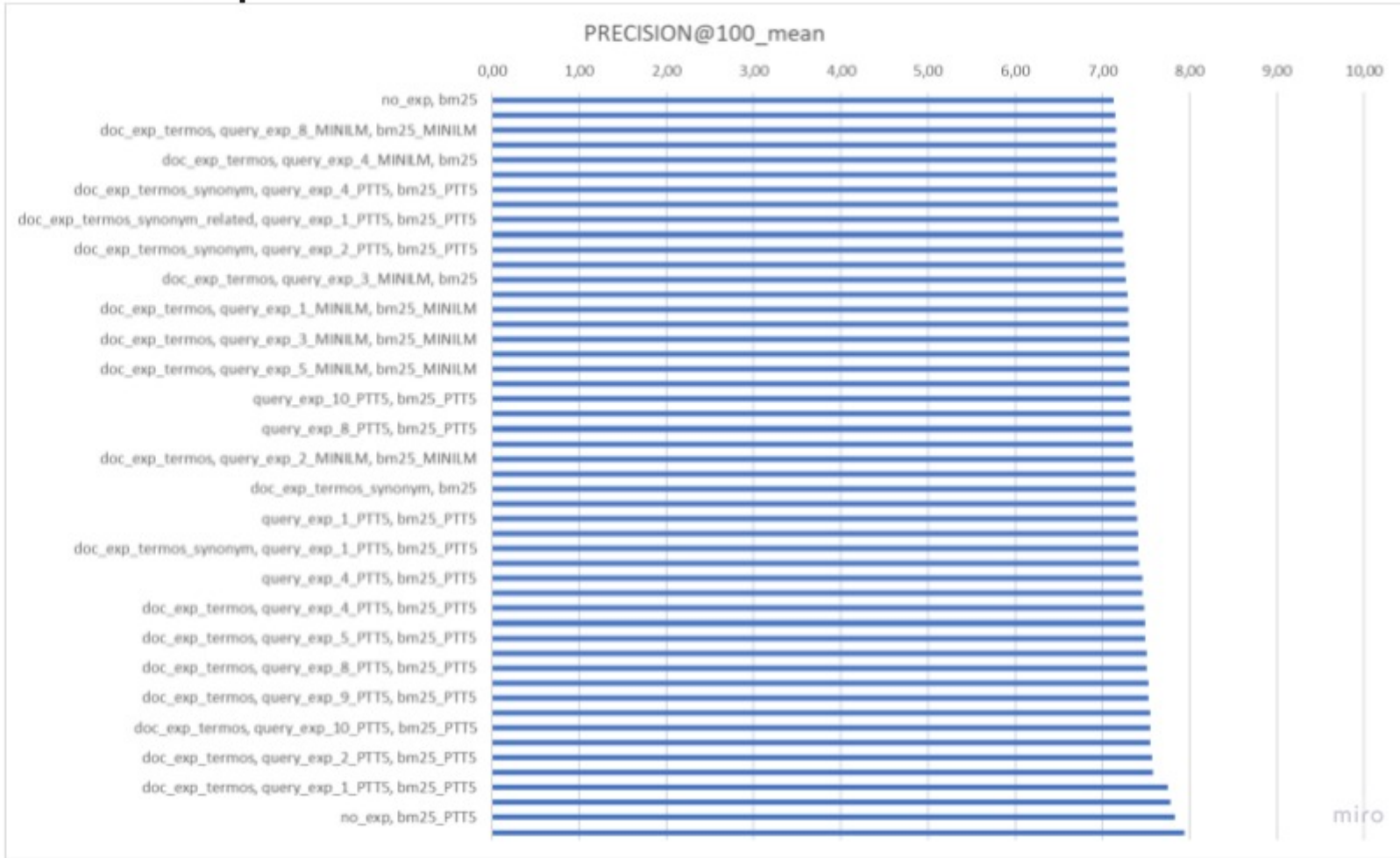
ndcg@10



recall@100


RECALL@100_mean
0,00
10,00
20,00
30,00
40,00
50,00
60,00
70,00
80,00
90,00
100,00
no_exp, bm25
doc_exp_termos_synonym, query_exp_1_MINILM, bm25
doc_exp_termos, query_exp_9_MINILM, bm25_MINILM
doc_exp_termos_synonym_related, query_exp_1_PTT5, bm25_PTT5
doc_exp_termos_synonym, query_exp_4_PTT5, bm25_PTT5
query_exp_6_PTT5, bm25_PTT5
doc_exp_termos, query_exp_1_MINILM, bm25_MINILM
doc_exp_termos, query_exp_5_MINILM, bm25_MINILM
doc_exp_termos, query_exp_3_MINILM, bm25_MINILM
doc_exp_termos_synonym, query_exp_2_PTT5, bm25_PTT5
doc_exp_termos, query_exp_1_MINILM, bm25
doc_exp_termos, query_exp_6_MINILM, bm25_MINILM
query_exp_8_PTT5, bm25_PTT5
doc_exp_termos, bm25_MINILM
doc_exp_termos_synonym, bm25
query_exp_5_PTT5, bm25_PTT5
doc_exp_termos_synonym, query_exp_1_PTT5, bm25_PTT5
query_exp_2_PTT5, bm25_PTT5
doc_exp_termos, query_exp_4_PTT5, bm25_PTT5
doc_exp_termos, query_exp_8_PTT5, bm25_PTT5
doc_exp_termos, query_exp_5_PTT5, bm25_PTT5
doc_exp_termos, query_exp_9_PTT5, bm25_PTT5
query_exp_1_PTT5, bm25_PTT5
doc_exp_termos, query_exp_2_PTT5, bm25_PTT5
doc_exp_termos_synonym, bm25_PTT5
doc_exp_termos, bm25_PTT5
miro

precision@100



- Conclusões*
  - Construção de datasets em português
    - Avaliações
      - ChatGPT parece ter um excelente desempenho
      - Limitador de desempenho é a busca
  - Treinamento de um Indexador com ótimos resultados alcançados (INDIR-TCU-INDEX)
  - Indexador melhorou o desempenho de buscas
    - iniciativas consideradas positivas
    - expansão no índice só dos termos
    - recall: métrica mais impactada
  - Reranker melhorou desempenho das buscas no JURIS-TCU
  - avaliação "out of domain" não superou modelo não treinado "in-domain"
    - reranker com PTT5-INDIR ajustado no domínioJURIS-TCU-INDEX não superou os resultados de sua versão PTT5-BASE no dataset JURIS-TCU
  - Expansões com **PTT5-INDIR** alcançaram melhores resultados do que outras expansões (MINILM-INDIR e **PTT5-BASE**)

num_query_exp = 1; ranker = PTT5_BASE

expansor	doc_exp	ndcg@10	prec@50	recall@100
PTT5_BASE	sem	64.01	14.37	88.74
PTT5_INDIR	sem	66.11	14.65	91.09
PTT5_BASE	termo	62.03	14.88	91.09
PTT5_INDIR	termo	65.40	15.02	91.23

  - Precisamos ouvir mais quem sabe:
    - exemplos negativos: **sample de 1000 com bm25**
    - "Davi pode vencer Golias: dado sua linguagem (português) e com um bom treinamento de preparação para a batalha"

- Dúvida
  - Se inverter query / doc faria diferença em um PTT5?
- Gostaríamos de complementar (relatório)
  - Novo INDIR em treinamento (parou em 128%, atual: 83%)
  - Terminar experimentos PTT5-BASE (com expansores: PTT5-BASE, PTT5-INDIR)
- Trabalhos futuros
  - Imagina um ptt5_3B? (mais lento)
  - LLM como estágio 3?
  - Aumentar o dataset de avaliação JURIS-TCU com novas técnicas
  - Criar um dataset de treino JURIS-TCU

Referências (WIP)

R.Nogueira,Z.Jiang,R.Pradeep,andJ.Lin. Document ranking with a pretrained sequence- to sequence model. In Findingsof 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.

Agradecimentos

- Ao TCU, pela disponibilização da nuvem com GPT3.5 e 4.0;
  (e toda sua equipe: pelas soluções desenvolvidas no trato dos datasets aqui disponibilizados)
- Leandro Carísio, pela ajuda na extração de dados de histórico das buscas
- código treino: Hugo e Thiago Laits (representando toda a turma)
- Professores
- Sérgio Salustiano
  - pelo cuidado com os dados de jurisprudência (e toda sua equipe)
  - e apoio na revisão das questões
- Beatriz Pinheiro — pelo cuidado com os dados do tesauro VCE (e toda sua equipe)

Estorinha final:
"Indir: nasceu uma hipótese... Saiu do seu domínio e bebeu muita água, graças às extensões que ofertou e, principalmente, à salutar presença de seu amigo de base Reranker"

Fonte: LLM V.5.4
(Louco na Linguagem Marcus?)