



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

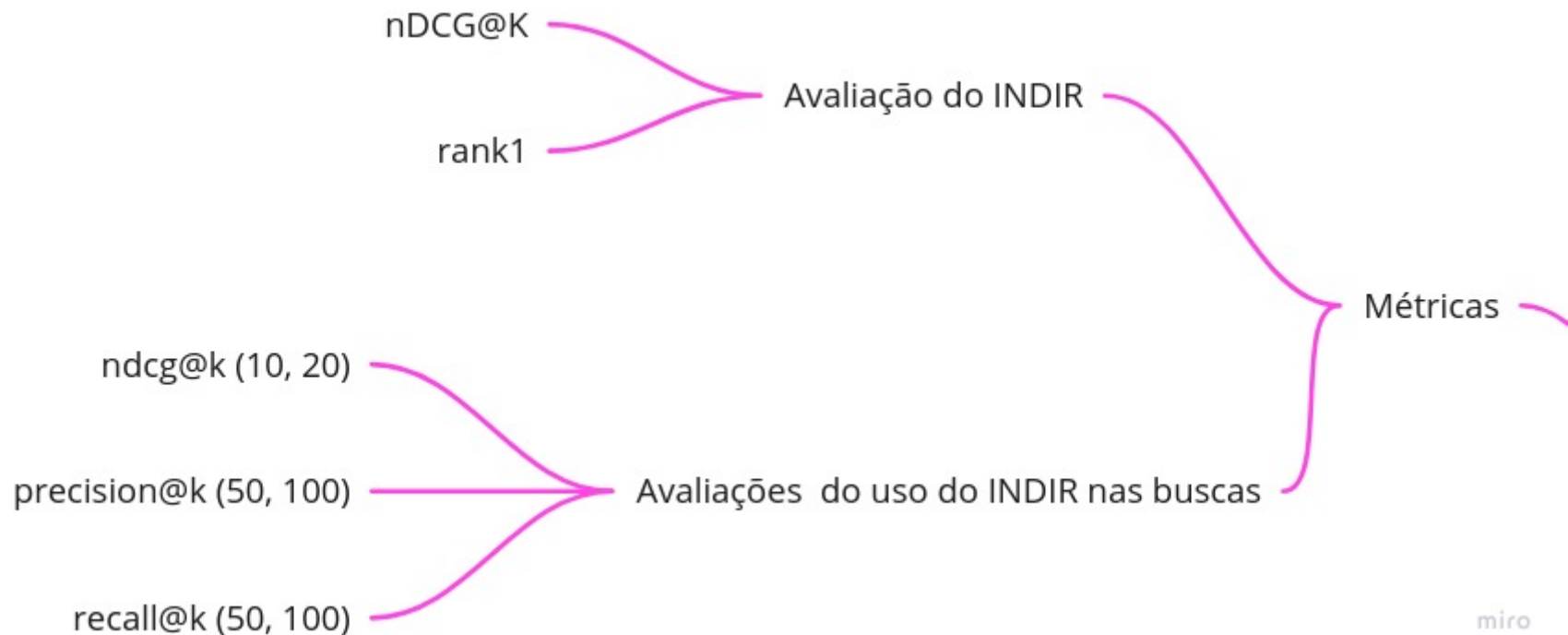
JURIS-TCU

Gerar datasets com relevância em português

Desenvolver um indexador INDIR (como um pipeline de busca) - JURIS_TCU_INDEX

Avaliar possível uso do indexador na busca

Objetivos do Projeto





Cadastro e indexações

Sistema de Enunciados de Jurisprudências

Extração e tratamento de dados

Metodologia
miro

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



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)

Dados dos 2 sistemas

query: JURIS

doc: VCE

qrel: JURIS x VCE

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

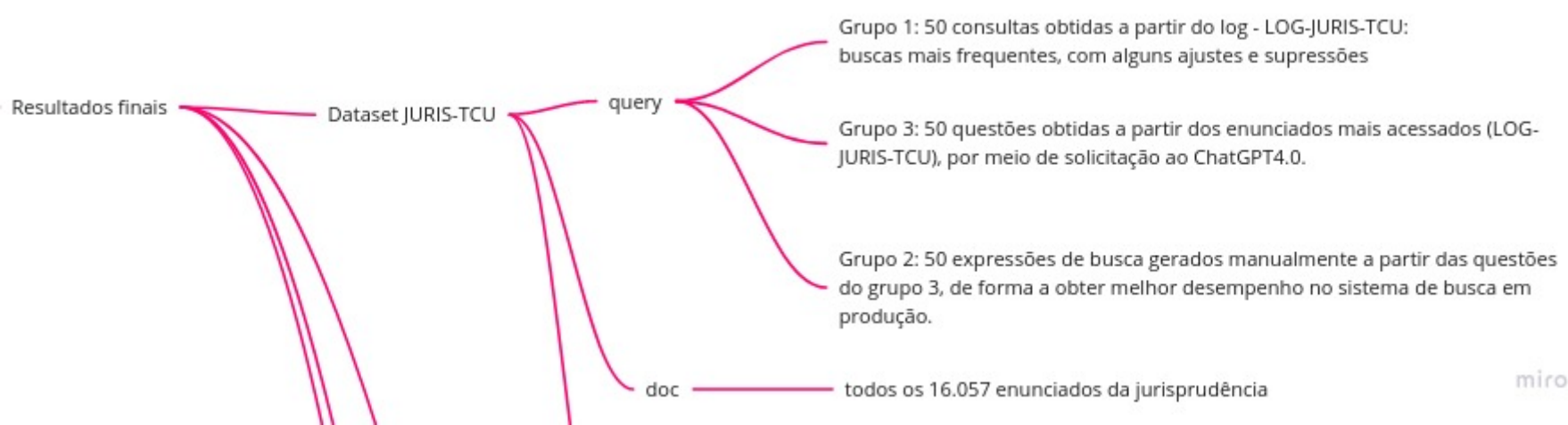
doc: JURIS

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



qrel

Avaliação de todas as 150 queries, com 15 documentos cada.

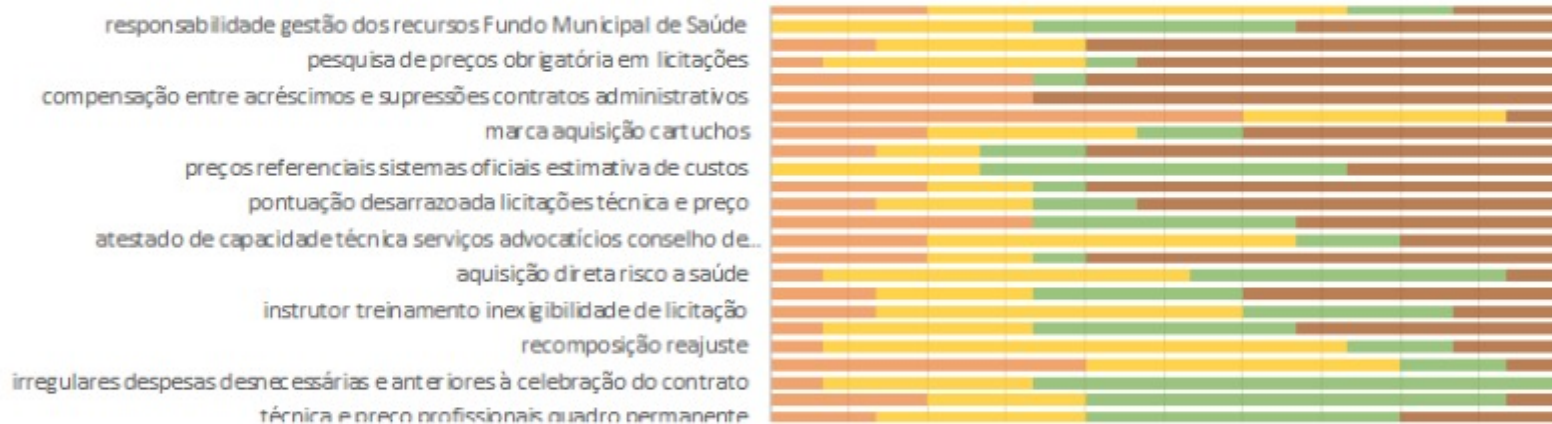
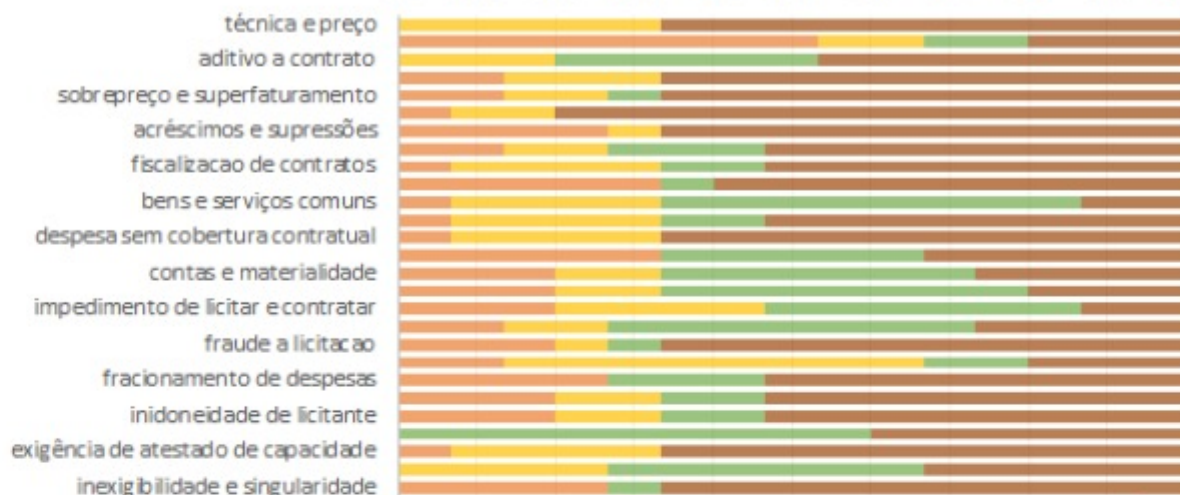
Seleção de documentos mais relevantes: pipeline de busca completo (BM25+STS => rerank) visando obter top-10

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)

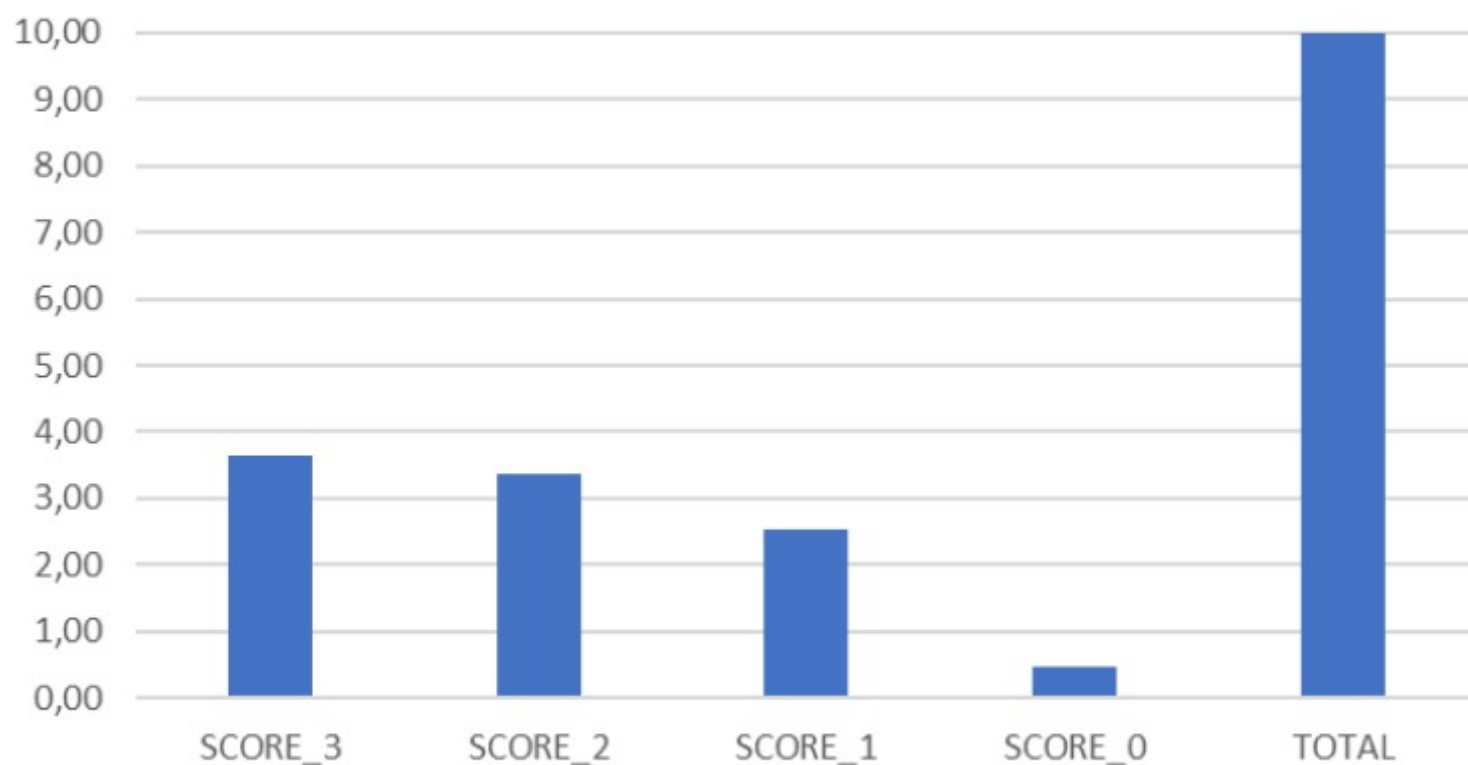
AVALIAÇÕES POR QUERY

0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

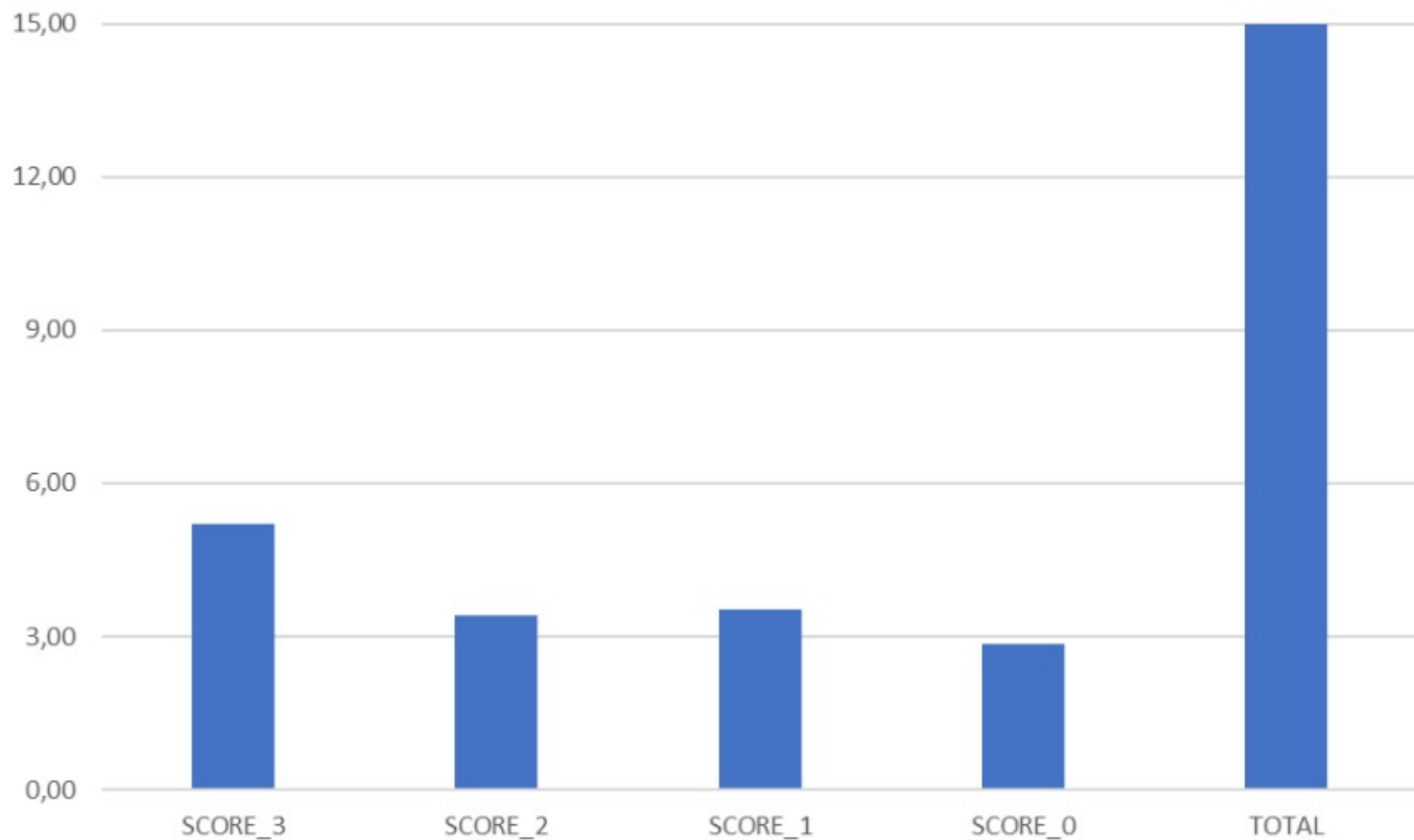


SCORE_0 SCORE_1 SCORE_2 SCORE_3

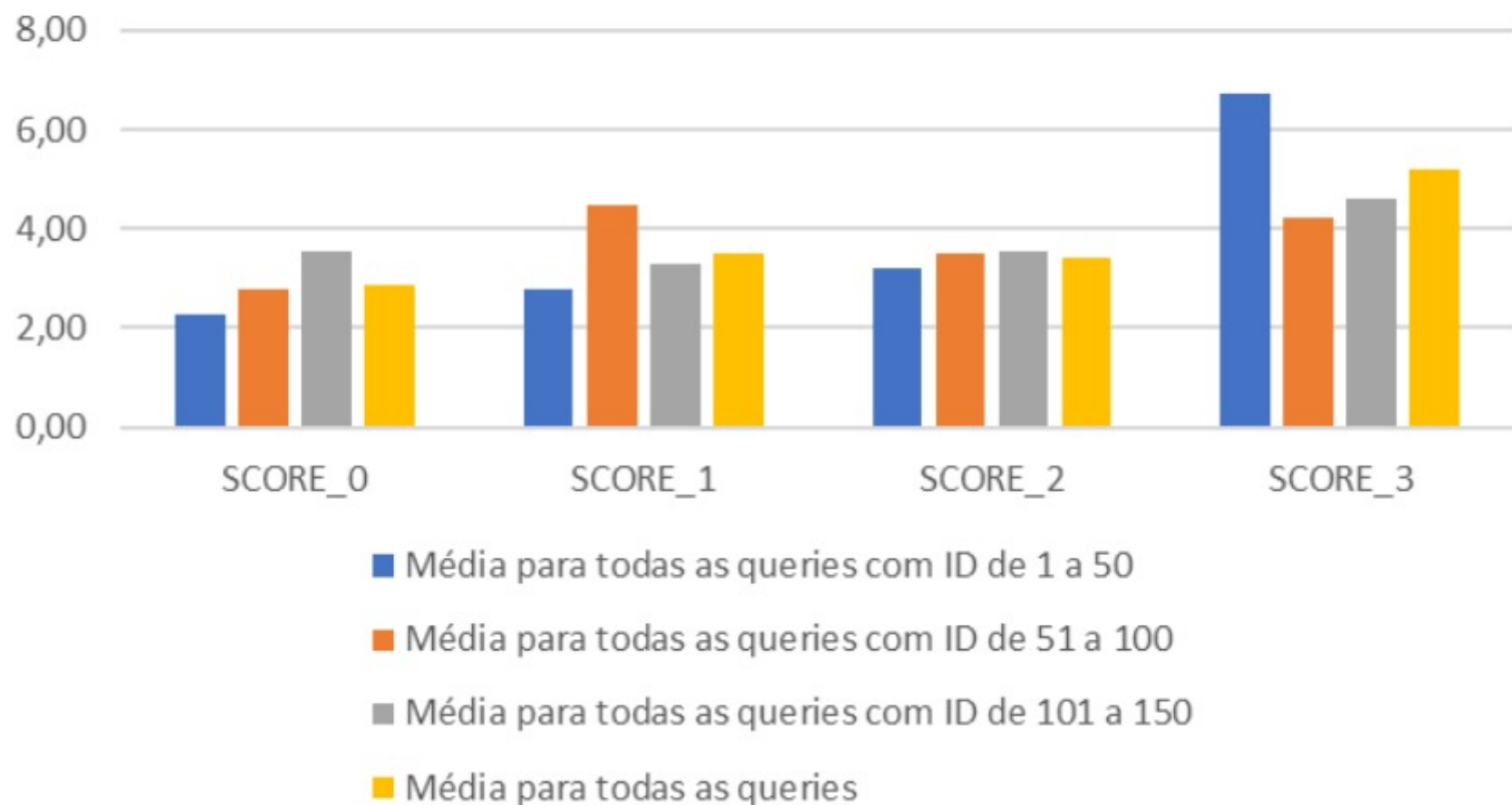
Média - Avaliações geradas pelo LLM



Média - Avaliações geradas pelo LLM (todas)



Média de enunciados por score Comparação entre os grupos de 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)



Geração de dados de treinamento
(uma história à parte)



Palavras motivacionais na
Palestra do Prof Rodrigo em 23/6



último slide: estória criada pelo LLM

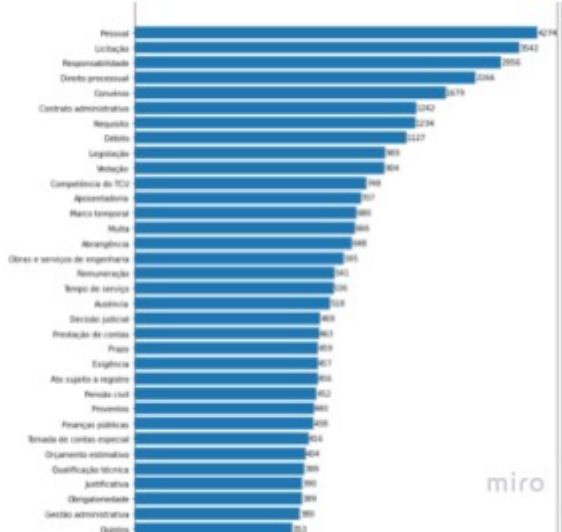
miro

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)

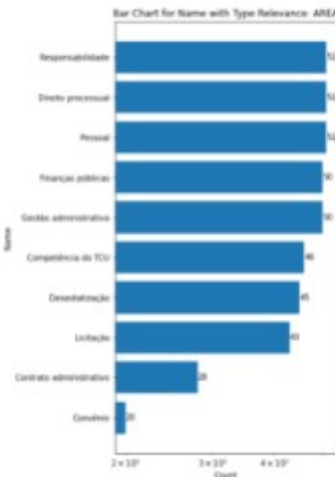
v3: 1+, 1- (limitado a 50 por termo, ordem decrescente de data - mais recentes, mais pesquisados)



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



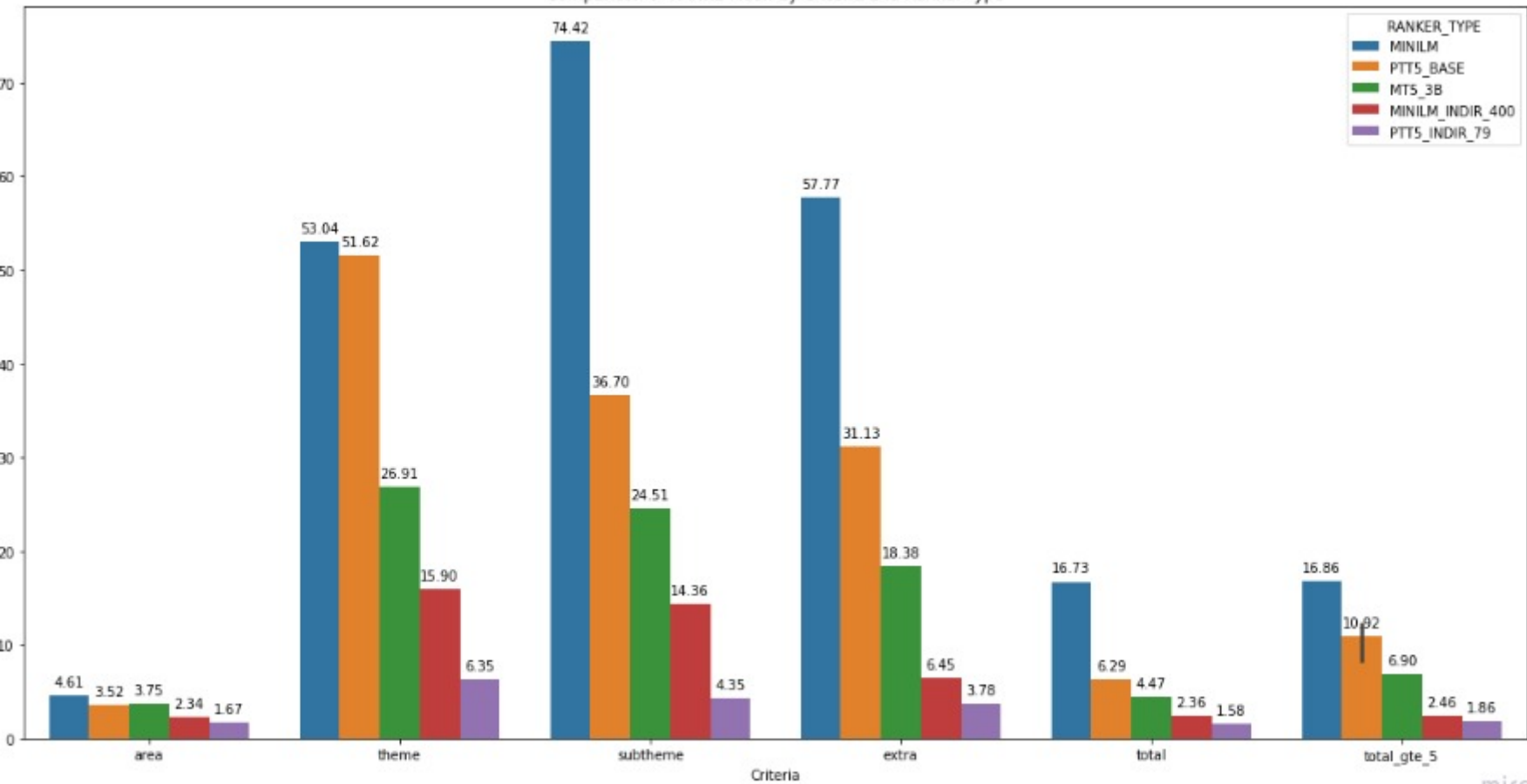
v3: 1+, 1- (sample bm25@1000; balanceamento limitando quantidade por tipo de indexação)

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

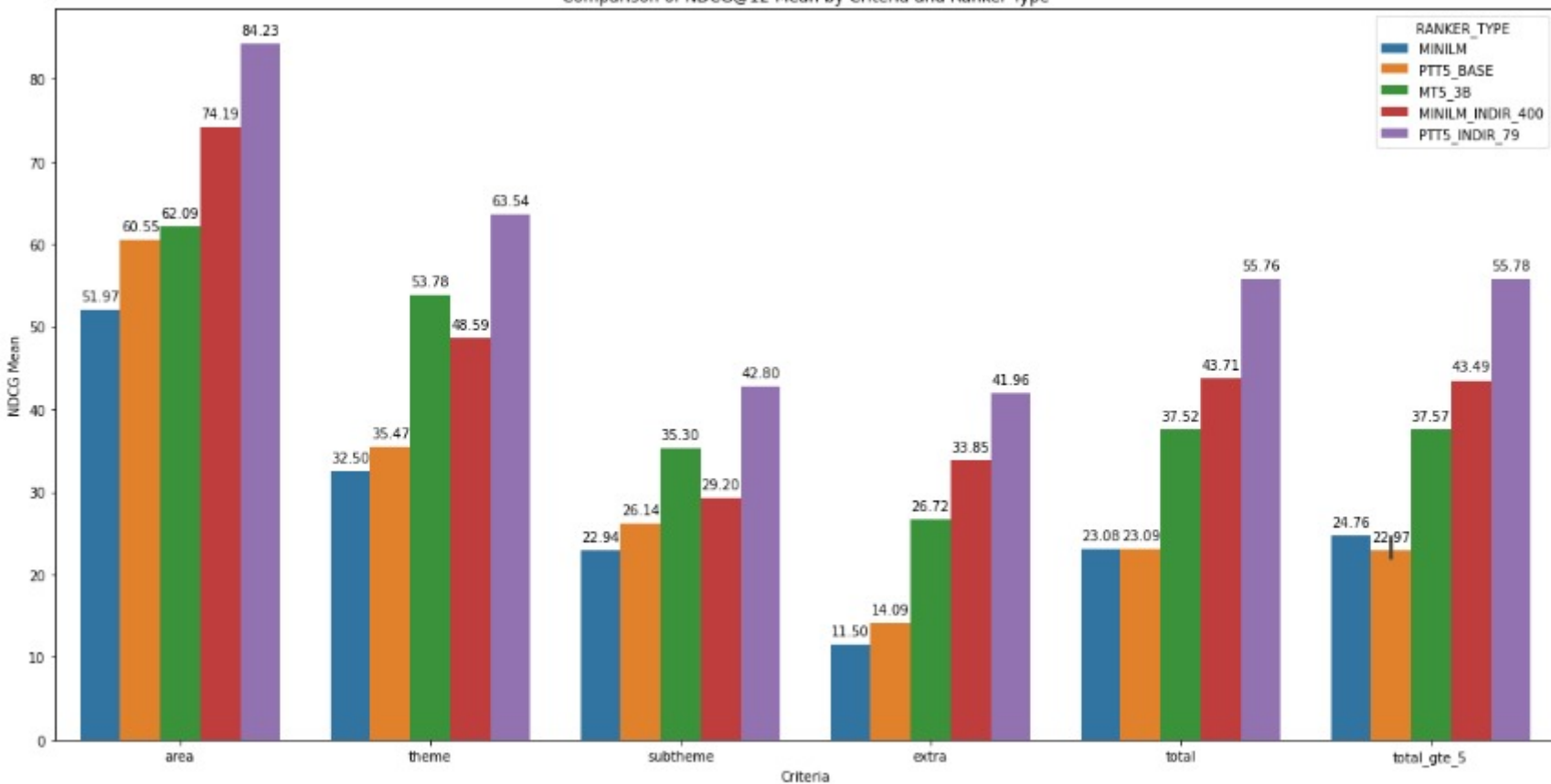
PTT5_INDIR_83
unicamp-dl/ptt5-base-pt-msmarco-100k-v2

MINILM_INDIR_400
unicamp-dl/mMiniLM-L6-v2-pt-v2

Comparison of RANK1 Mean by Criteria and Ranker Type



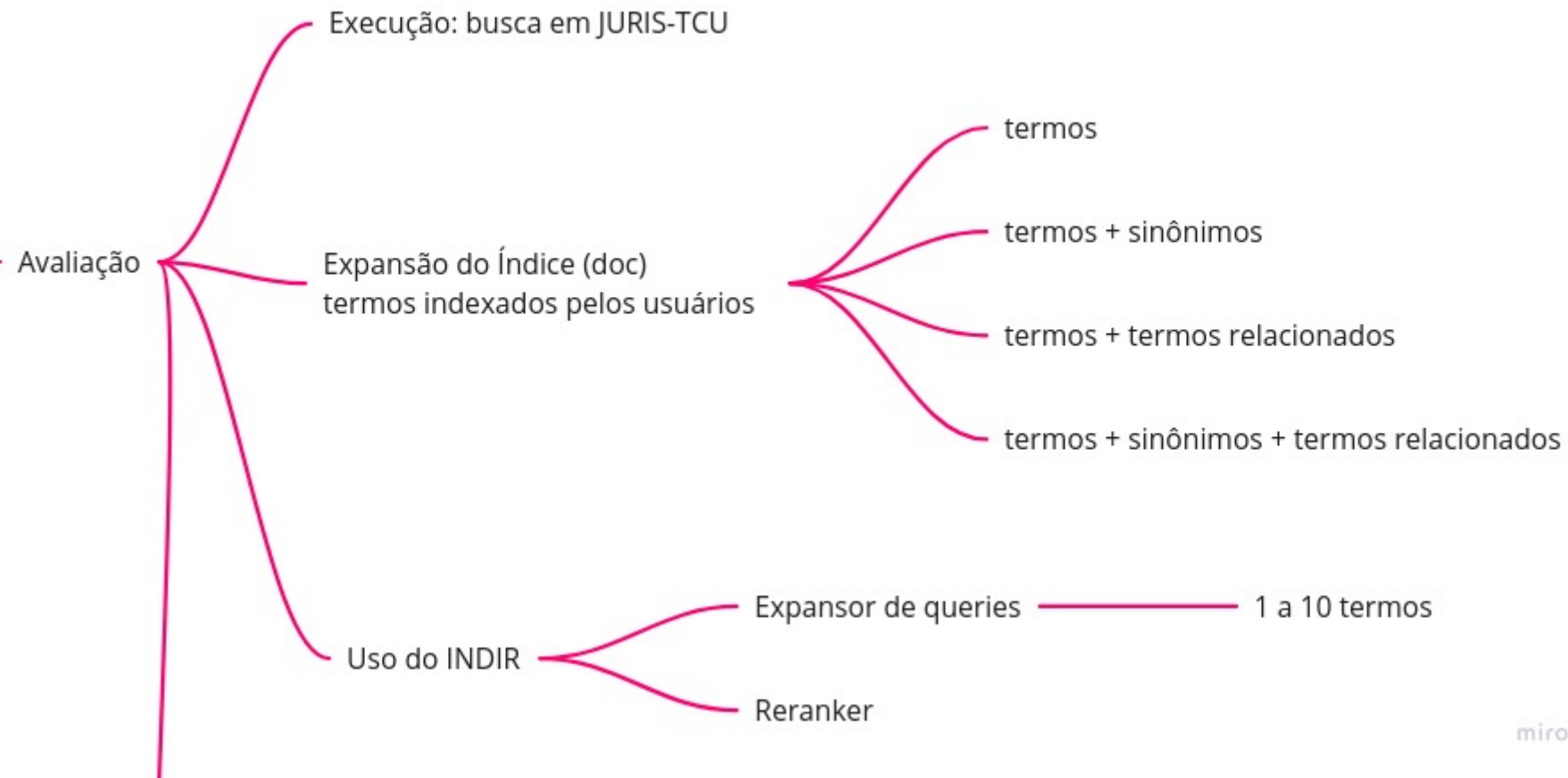
Comparison of NDCG@12 Mean by Criteria and Ranker Type



Se pudermos complementar
(para o relatório)

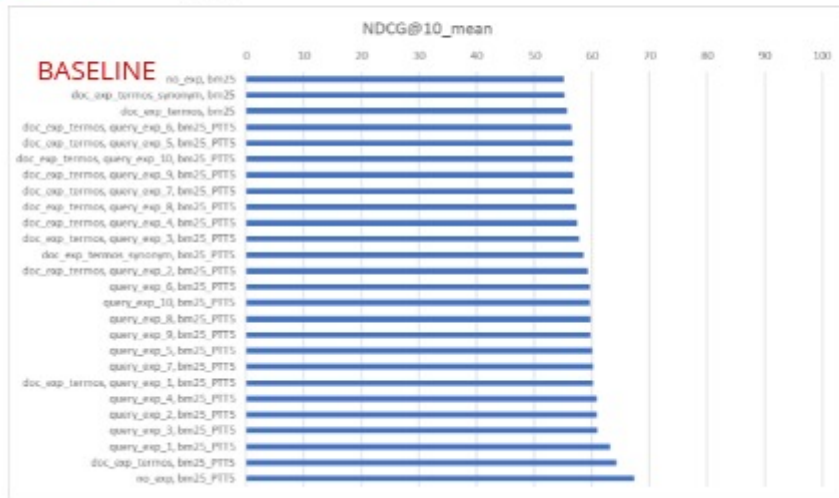
Novo INDIR em treinamento (138%)

Comparar pesquisa com PTT5-BASE



Comparação
(casos vencedores)

ndcg@10

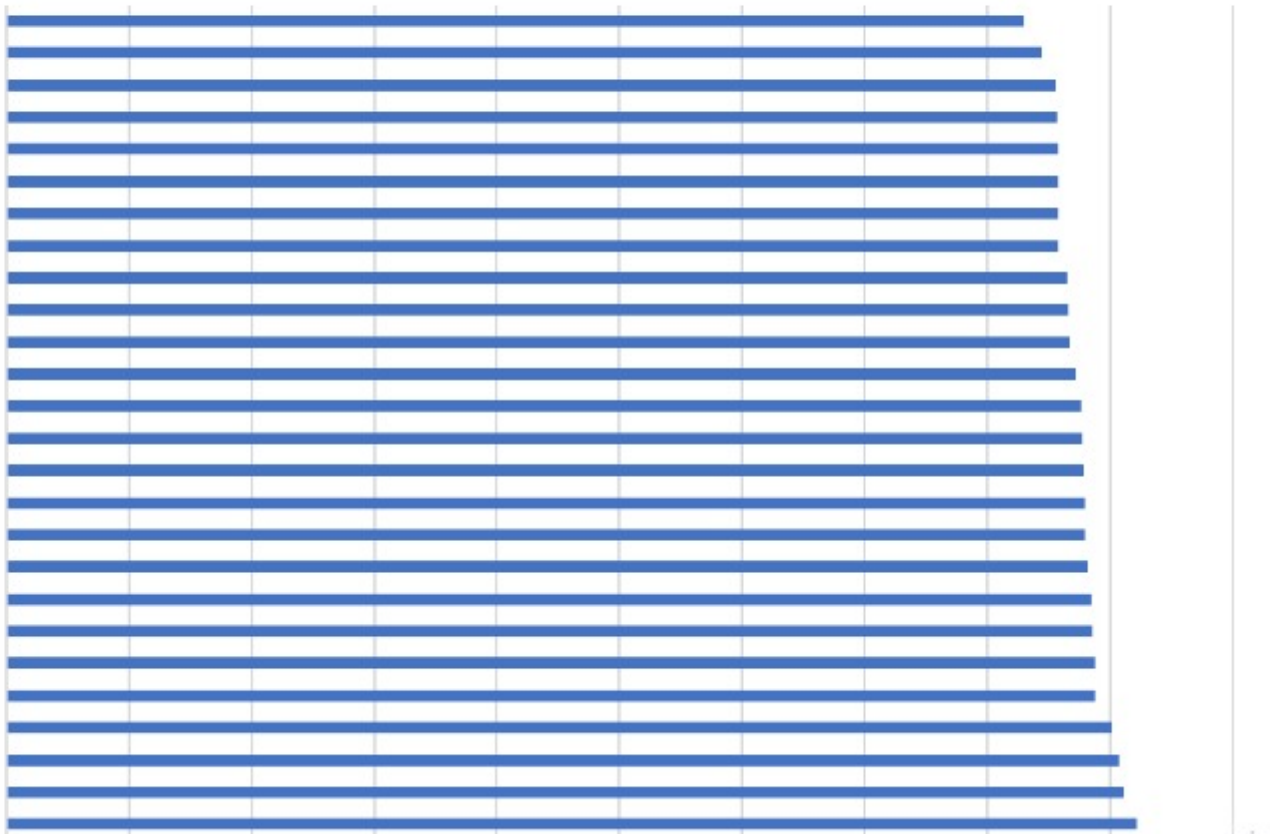


recall@100

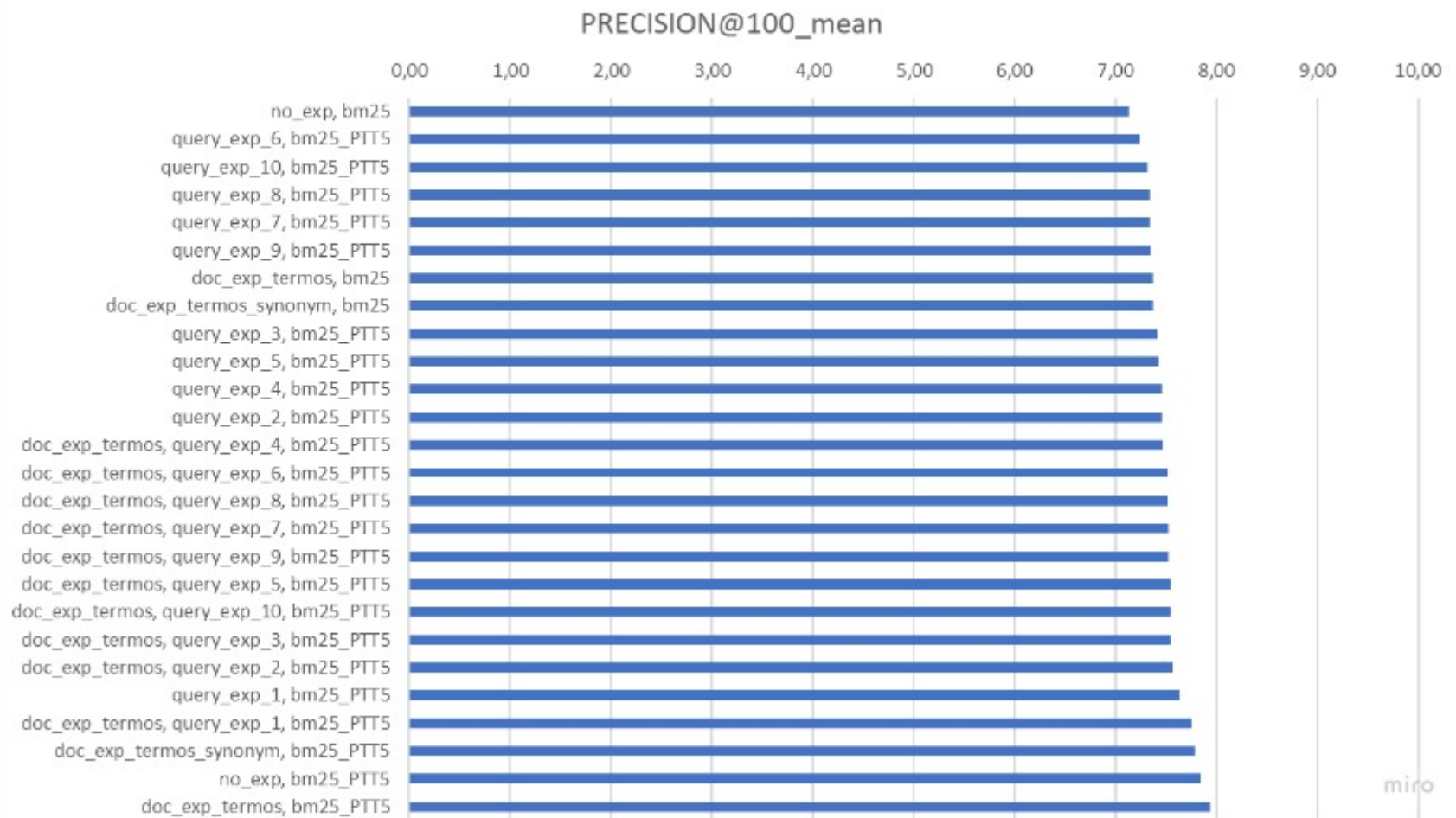
RECALL@100_mean

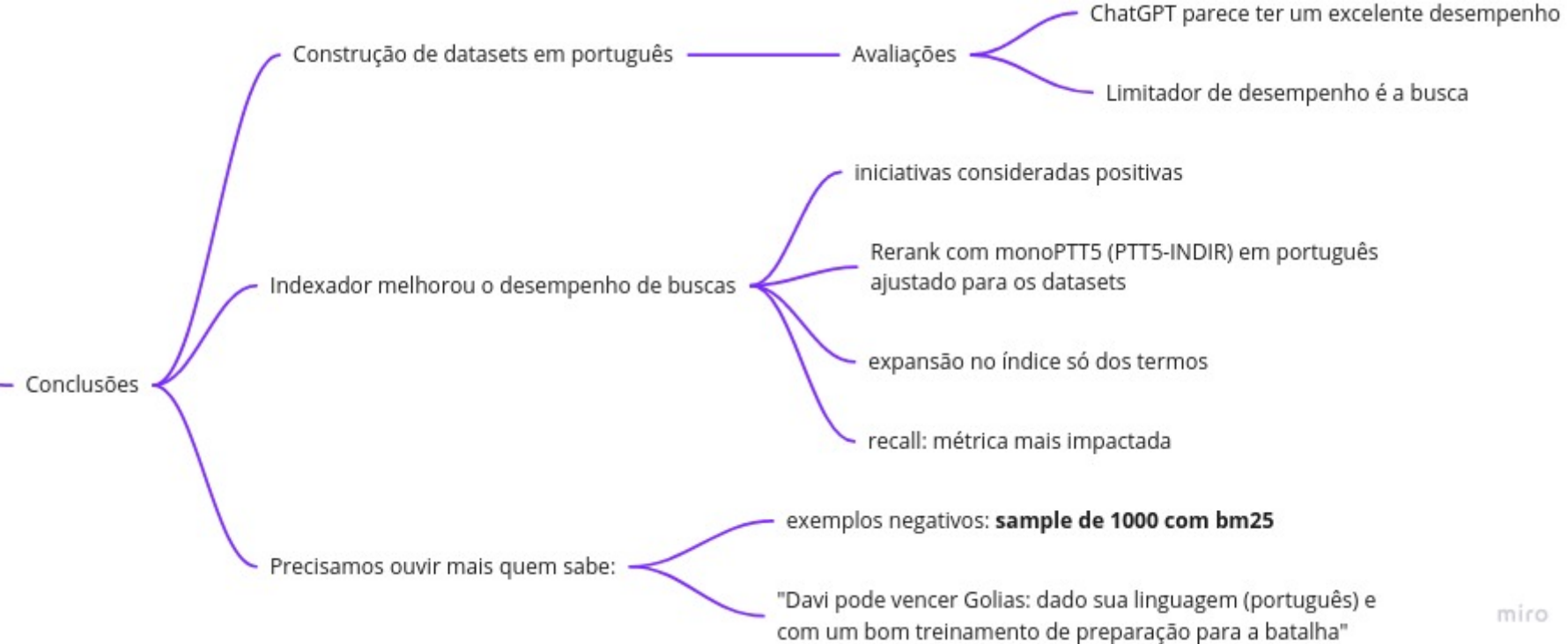
0,00 10,00 20,00 30,00 40,00 50,00 60,00 70,00 80,00 90,00 100,00

no_exp, bm25
query_exp_6, bm25_PTT5
query_exp_10, bm25_PTT5
query_exp_8, bm25_PTT5
doc_exp_termos, bm25
query_exp_7, bm25_PTT5
query_exp_9, bm25_PTT5
doc_exp_termos_synonym, bm25
query_exp_4, bm25_PTT5
query_exp_5, bm25_PTT5
query_exp_3, bm25_PTT5
query_exp_2, bm25_PTT5
doc_exp_termos, query_exp_4, bm25_PTT5
doc_exp_termos, query_exp_6, bm25_PTT5
doc_exp_termos, query_exp_8, bm25_PTT5
doc_exp_termos, query_exp_7, bm25_PTT5
doc_exp_termos, query_exp_9, bm25_PTT5
doc_exp_termos, query_exp_10, bm25_PTT5
doc_exp_termos, query_exp_3, bm25_PTT5
doc_exp_termos, query_exp_5, bm25_PTT5
doc_exp_termos, query_exp_2, bm25_PTT5
query_exp_1, bm25_PTT5
doc_exp_termos, query_exp_1, bm25_PTT5
doc_exp_termos_synonym, bm25_PTT5
no_exp, bm25_PTT5
doc_exp_termos, bm25_PTT5



precision@100








Trabalhos futuros

Imagina um ptt5_3B? (mais lento)

LLM como estágio 3?



Referências (WIP)

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.

INDIR: nasceu uma hipótese... agora: muita água para se beber!!!

