
Projeto Final

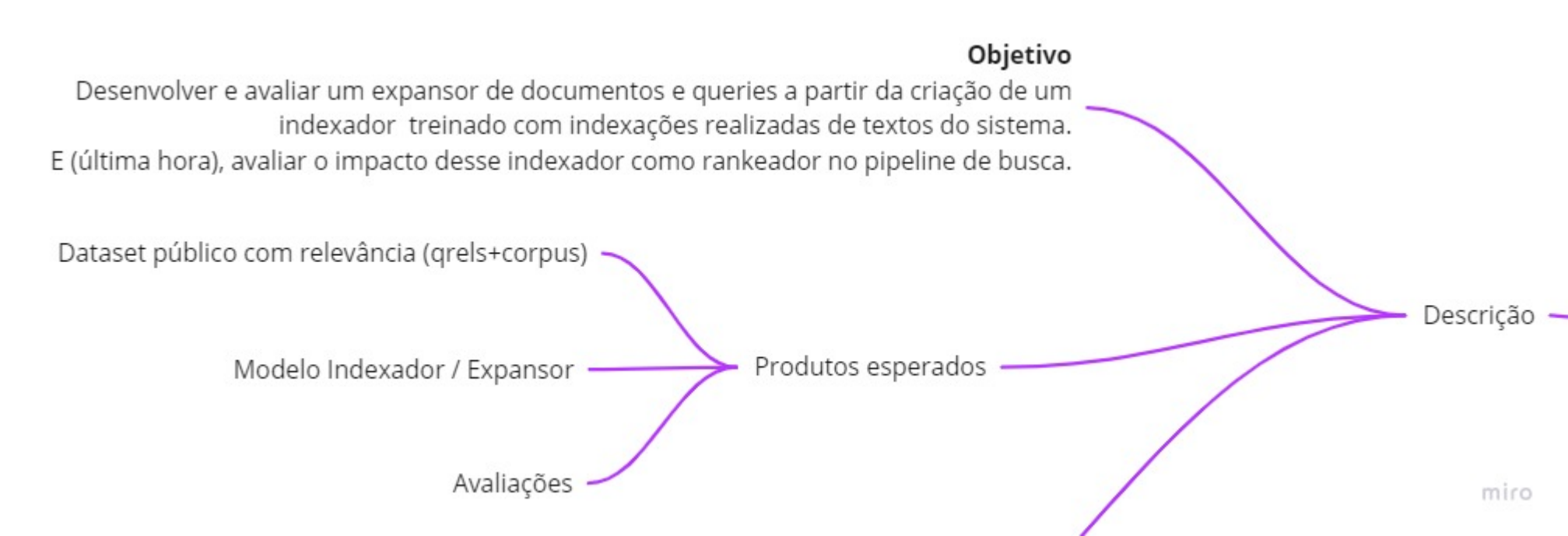
IA368DD_2023S1: Deep Learning aplicado a Sistemas de Buscas

Student:  Marcus Vinícius Borela de Castro

**Ind-ER** (Indexer: Expander & Ranker)

Leonardo Pacheco & Marcus Borela

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


Objetivo
Desenvolver e avaliar um expansor de documentos e queries a partir da criação de um indexador treinado com indexações realizadas de textos do sistema.
E (última hora), avaliar o impacto desse indexador como rankeador no pipeline de busca.
Dataset público com relevância (qrels+corpus)
Modelo Indexador / Expansor
Avaliações
Produtos esperados
Descrição
miro

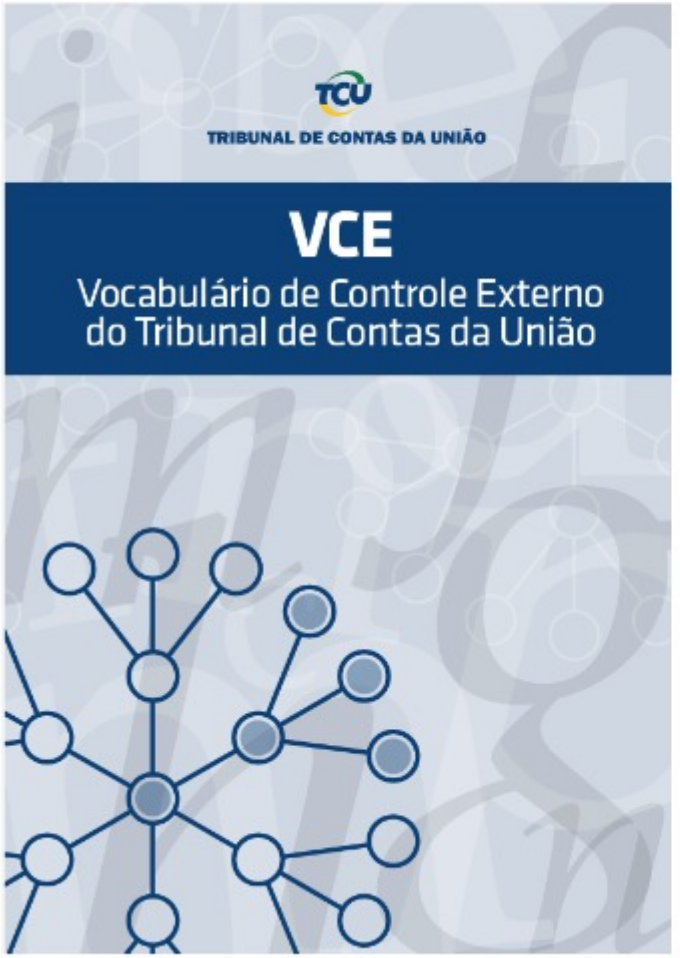
# Metodologia por analogia

# Explicando o nome

## Ind-ER:  Indexer Expander & Ranker


Imagem sem título gerada por IA BlueWillow no Discord

# Vocabulário de Controle Externo (VCE)




## Conceitos

"o **Tesauro** do Tribunal de Contas da União (VCE) foi construído para padronizar o tratamento de informações especializadas e conferir maior agilidade e precisão na recuperação dos conteúdos presentes nos sistemas de informação do TCU."

"Construído a partir do Tesauro do TCU (TECON), de 1992, voltado apenas à informação jurisprudencial, o VCE teve escopo ampliado e objetiva uniformizar a terminologia técnica utilizada em todo o TCU, além de facilitar o intercâmbio de informações entre sistemas e bases de dados, possibilitando melhor integração. Configura-se ainda em importante ferramenta para uso colaborativo por instituições similares, cuja área de atuação envolva o Controle."

Exemplo

**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

A **indexação** é uma operação que consiste em identificar os principais conceitos que caracterizam o conteúdo de um texto para obtençãode uma representação da informação relevante por meio de linguagem controlada e padronizada (termos descritores)

Compreende duas etapas distintas: a análise conceitual do assunto e a tradução dos conceitos em descritores

Manual de Sistematização e Divulgação da Jurisprudência do Tribunal de Contas da União
PORTARIA-TCU No 85, DE 06 DE JUNHO DE 2022

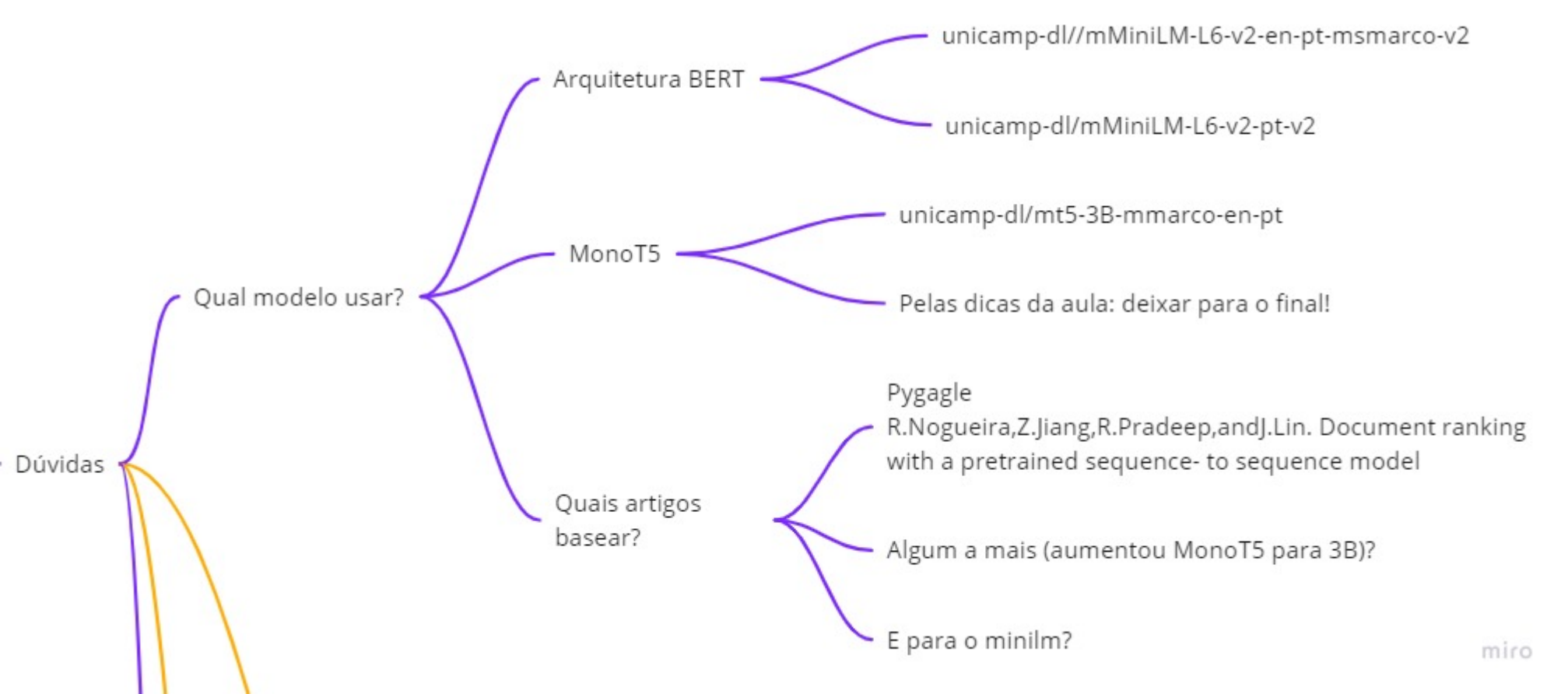
Indexações

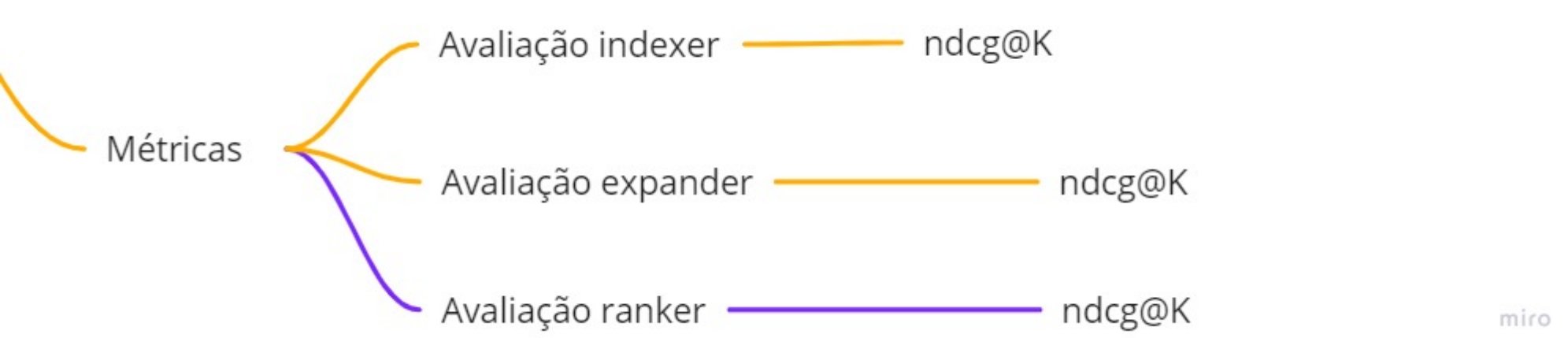
Indexer: um sujeito de princípios, alto grau no big five

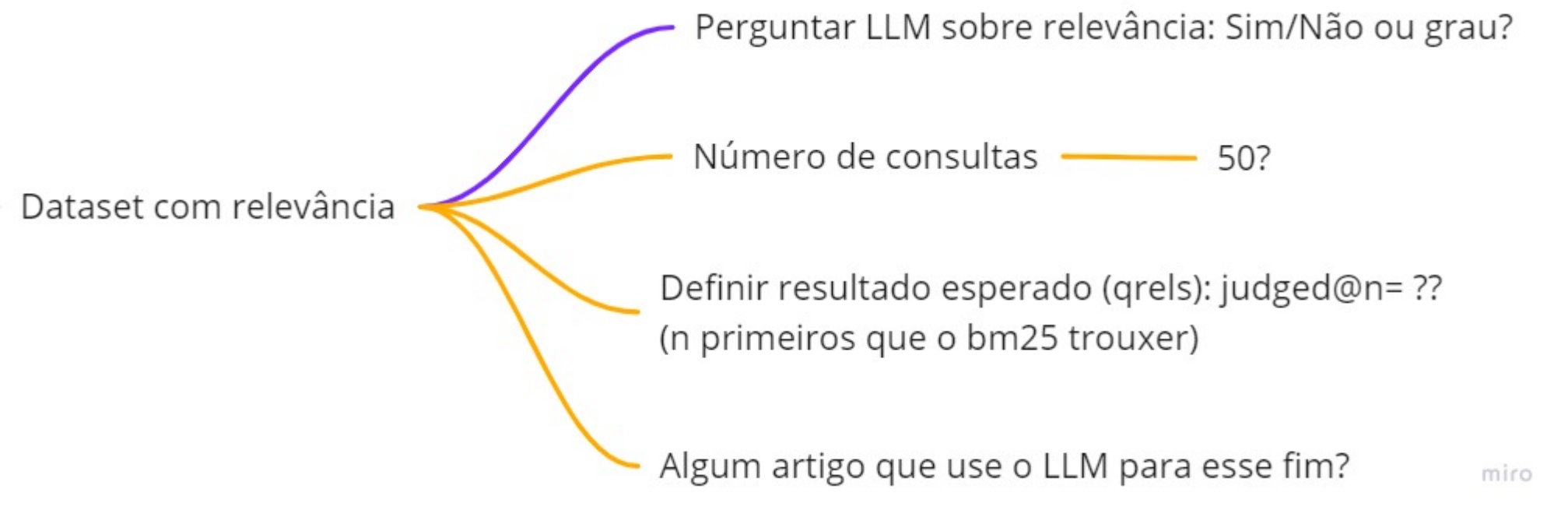
## 5.4 Princípios da Indexação

A indexação deve ser caracterizada sempre pelos seguintes aspectos:

- Especificidade: o indexador tem o compromisso de atingir o maior grau de especificidade possível.
- Exaustividade: o indexador deve atribuir a cada dispositivo tantos descritores quantos forem necessários para descrever todos os conceitos importantes da tese. A exaustividade possibilita a representação de um maior número de informações relevantes, aumentando a capacidade de recuperação do sistema.
- Concordância: o indexador deve limitar-se fielmente ao conteúdo do dispositivo. Não devem ser utilizados descritores para conceitos que, apesar de constarem do inteiro teor ou do excerto do acórdão, não aparecem no texto do dispositivo do enunciado.
- Coerência: o indexador deve aplicar consistentemente as regras de indexação. Os descritores deverão ser usados sempre da mesma forma, por diferentes indexadores ou pelo mesmo indexador, em épocas diversas. Devem ser utilizados descritores idênticos para documentos que tratam de um mesmo assunto. Assim, o usuário poderá confiar que, utilizando os mesmos descritores, localizará os mesmos conceitos, com um índice de previsibilidade razoável.
- Imparcialidade: o indexador deve abster-se de incluir descritores que representem avaliações ou opiniões pessoais, enfocando os principais conceitos contidos no dispositivo de forma imparcial e sem preconceitos.
- Fidelidade: os descritores escolhidos pelo indexador devem reproduzir fielmente o conteúdo do documento. Assim, o usuário encontrará facilmente o documento de seu interesse, pois o descritor selecionado o conduziu até a informação relevante que procurava.
- Bom Senso: o indexador não deve incluir descritores para conceitos acessórios ou informações não relevantes.


Dúvidas
Qual modelo usar?
Arquitetura BERT
unicamp-dl//mMiniLM-L6-v2-en-pt-msmarco-v2
unicamp-dl/mMiniLM-L6-v2-pt-v2
MonoT5
unicamp-dl/mt5-3B-mmarco-en-pt
Pelas dicas da aula: deixar para o final!
Quais artigos basear?
Pygagle
R.Nogueira,Z.Jiang,R.Pradeep,andJ.Lin. Document ranking with a pretrained sequence- to sequence model
Algum a mais (aumentou MonoT5 para 3B)?
E para o minilm?
miro


Métricas
Avaliação indexer
ndcg@K
Avaliação expander
ndcg@K
Avaliação ranker
ndcg@K


Perguntar LLM sobre relevância: Sim/Não ou grau?
Número de consultas
50?
Dataset com relevância
Definir resultado esperado (qrels): judged@n= ??
(n primeiros que o bm25 trouxer)
Algum artigo que use o LLM para esse fim?
miro

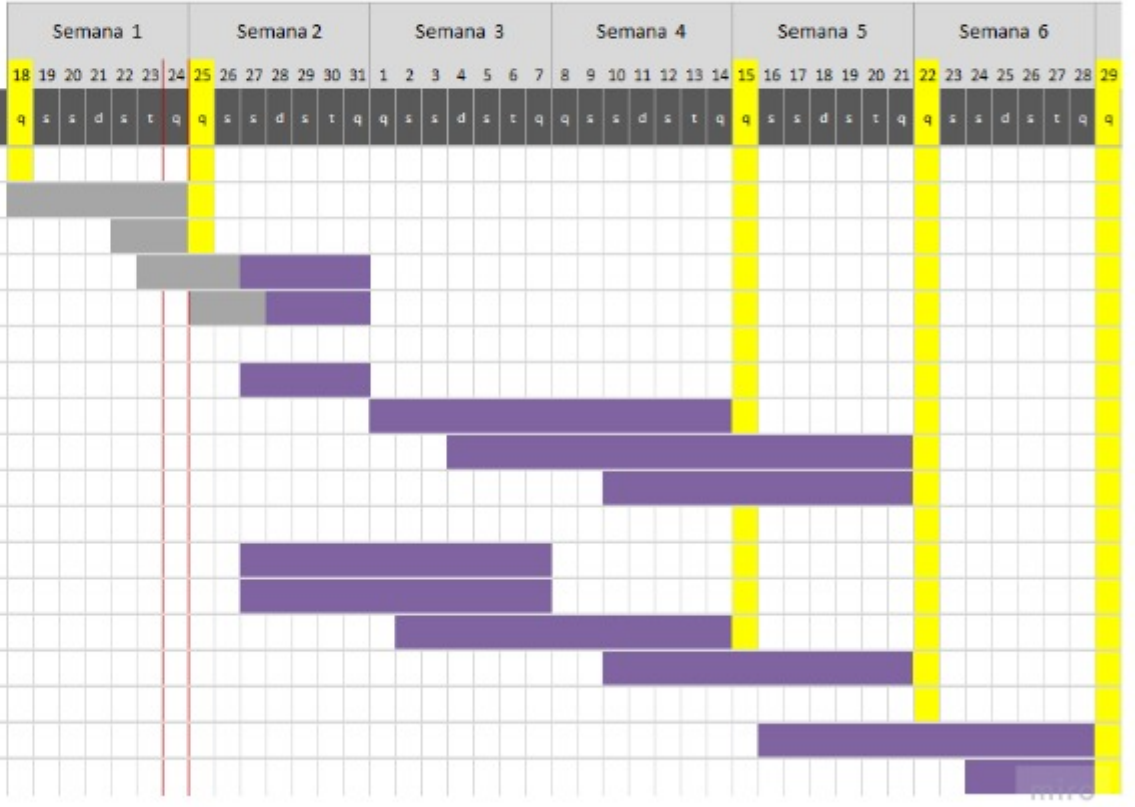
# Ind-E - Indexer Expander

**Projeto Final da disciplina IA368DD_2023S1: Deep Learning Aplicado a Sistemas de Buscas**

Leonardo Pacheco e Marcus Borela

https://miro.com/app/board/uXjVO-OAf1w=/

| TAREFA | ATRIBUÍDO PARA | PROGRESSO | INÍCIO | TÉRMINO |
|---|---|---|---|---|
| **Planejamento** | | | | |
| Plano básico | LM | 100% | 18/5/23 | 24/5/23 |
| Apresentação | LM | 100% | 22/5/23 | 24/5/23 |
| Definir base de dados | LM | 50% | 23/5/23 | 31/5/23 |
| Definir métricas | LM | 50% | 25/5/23 | 31/5/23 |
| **Frente Modelo** | | | | |
| Preparar dados de indexações | M | 0% | 27/5/23 | 31/5/23 |
| Treinar modelo | M | 0% | 1/6/23 | 14/6/23 |
| Expandir documentos (várias bases) | M | 0% | 4/6/23 | 21/6/23 |
| Expandir consultas | M | 0% | 10/6/23 | 21/6/23 |
| **Frente Avaliação** | | | | |
| Preparar dados do sistema alvo | L | 0% | 27/5/23 | 7/6/23 |
| Criar queries | L | 0% | 27/5/23 | 7/6/23 |
| Montar base de avaliação (qrel) | L | 0% | 2/6/23 | 14/6/23 |
| Realizar experimentos | LM | 0% | 10/6/23 | 21/6/23 |
| **Fechamento** | | | | |
| Relatório | LM | 0% | 16/6/23 | 28/6/23 |
| Apresentação final | LM | 0% | 24/6/23 | 28/6/23 |



Referências


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.