Trabalho Prático
Máquinas de Busca

Anna Paula Figueiredo Gonçalves
PCC547 – Recuperação de Informação - Denilson Alves Pereira

UFLA
UNIVERSIDADE FEDERAL DE LAVRAS

Sumário

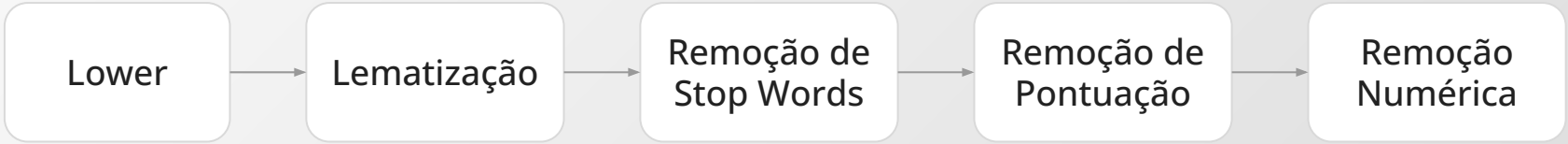
Objetivo

- Explorar a base de dados CFC (Cystic Fibrosis)
- Utilizar técnicas de Vetorização
  - Embeddings
  - TF-IDF
- Criar uma máquina de busca utilizando o FAISS
- Avaliar os resultados

UFLA
UNIVERSIDADE FEDERAL DE LAVRAS

Base de Dados

- 1.239 instâncias
- 100 consulta



UNIVERSIDADE FEDERAL DE LAVRAS

Técnicas de Vetorização

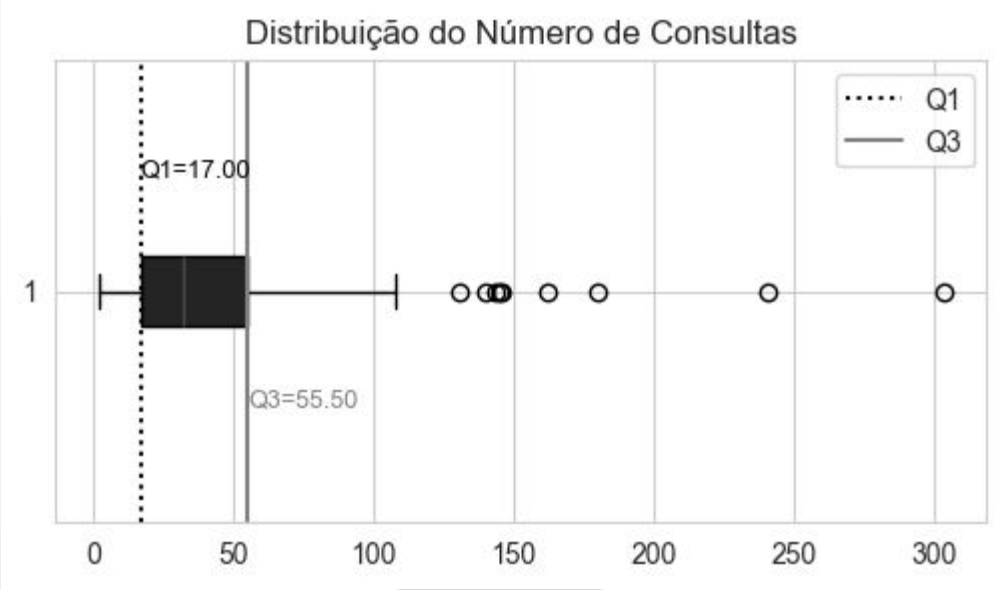
Embeddings

- paraphrase-MiniLM-L6-v2
- 48 segundos

TF-IDF

- 0.15 segundos

Aplicação

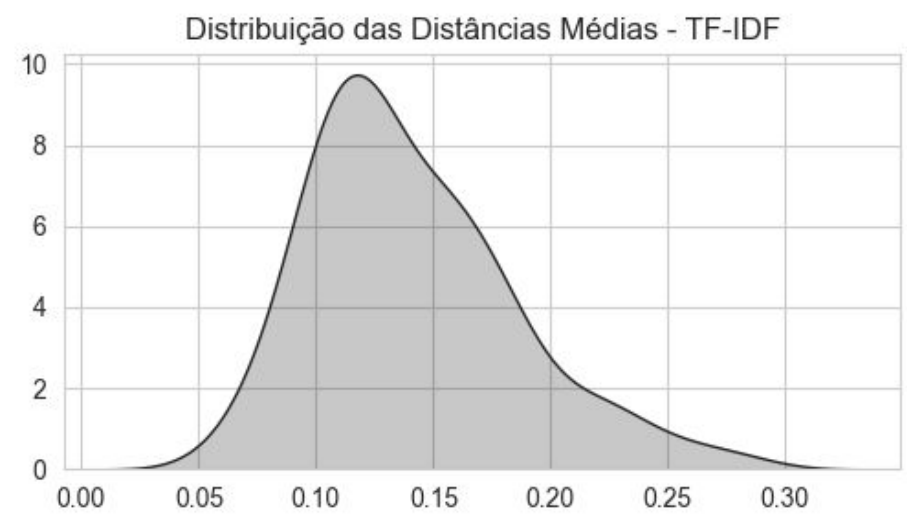


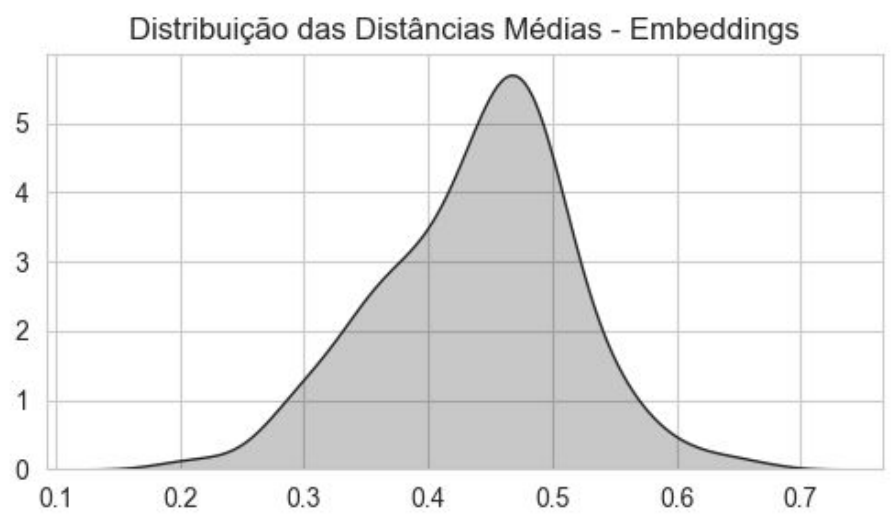
IQR = 38

UNIVERSIDADE FEDERAL DE LAVRAS

Resultados

Análise das Distâncias





Métricas	*Embeddings*	TF-IDF
mínimo	0.1833	0.0330
máximo	0.7553	0.6535
média	0.4384	0.1408
desvio padrão	0.0401	0.0573

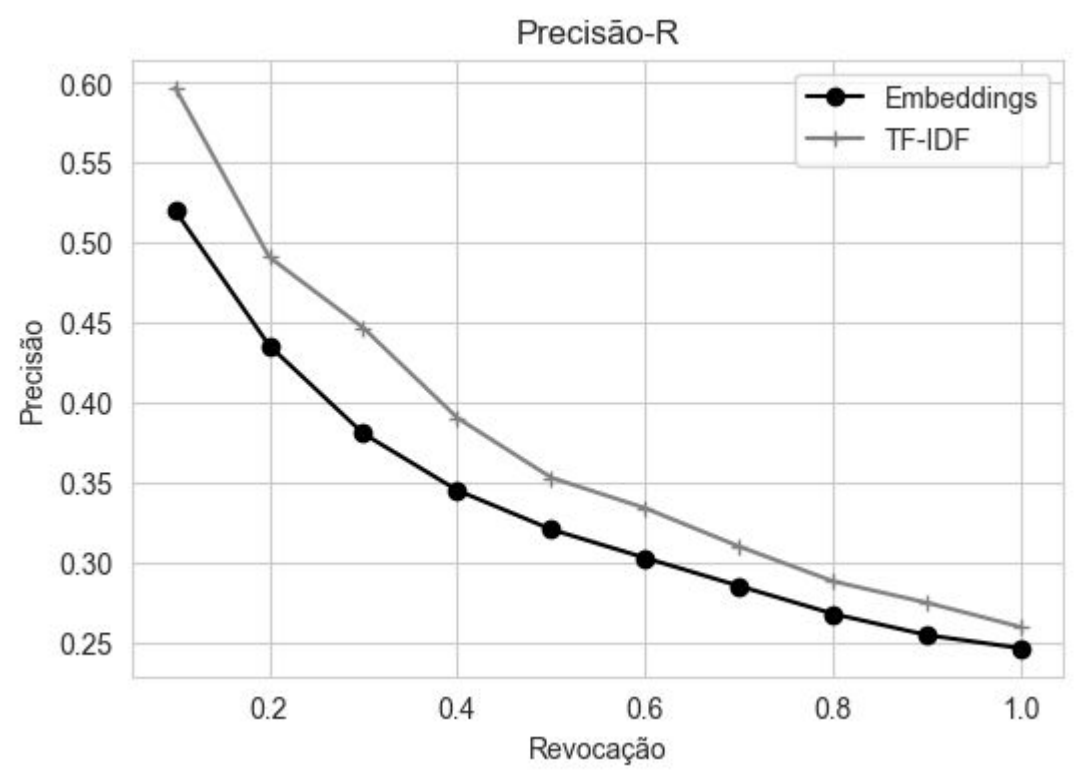
UNIVERSIDADE FEDERAL DE LAVRAS

Resultados

Métricas	*Embeddings*	TF-IDF
precisão média	0.2461	0.2592
revocação média	0.2670	0.2927
precisão-R	0.3359	0.3745
p@5	0.47	0.52
p@10	0.39	0.46
MRR(5)	0.88	0.93

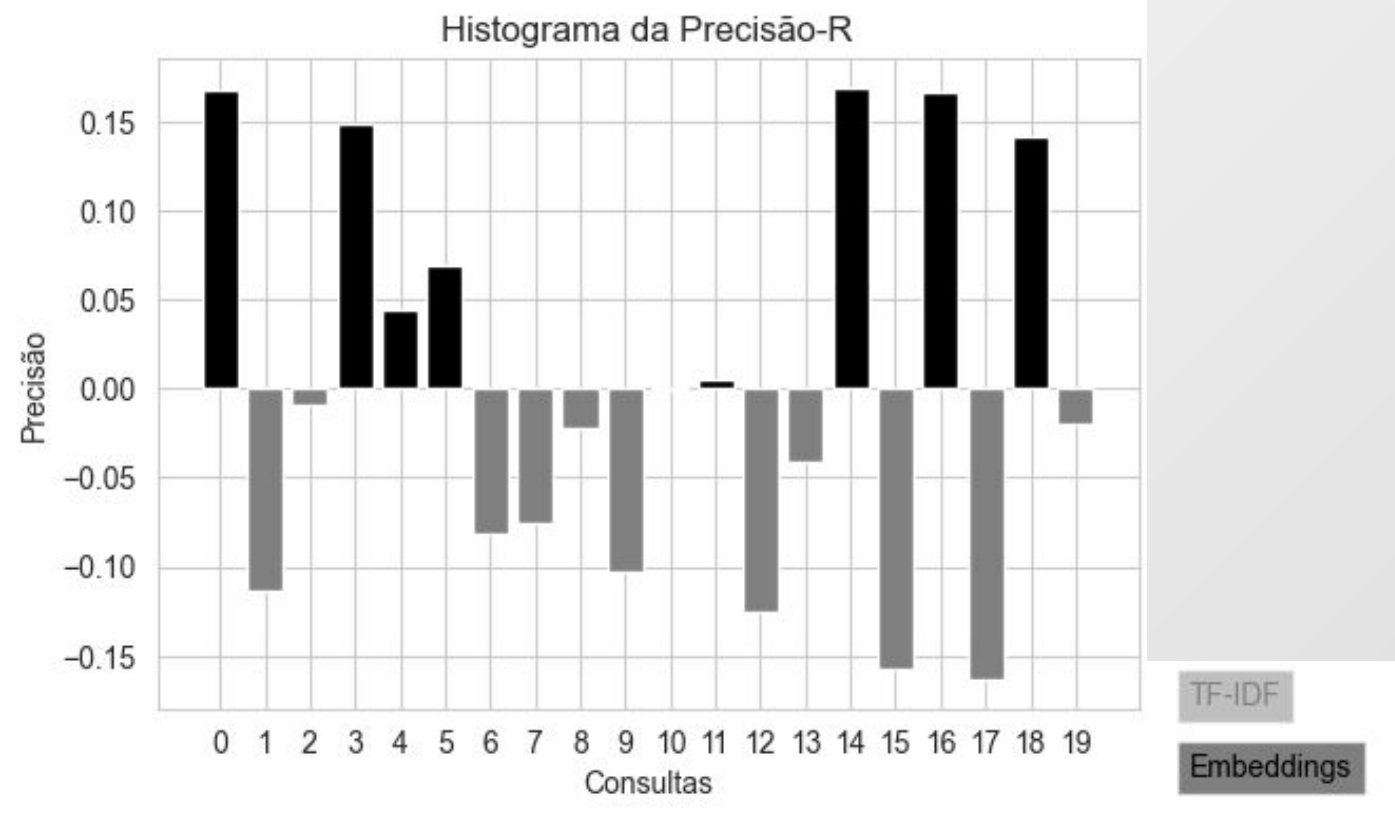
Resultados

Precisão - R



UNIVERSIDADE FEDERAL DE LAVRAS

Resultados



UNIVERSIDADE FEDERAL DE LAVRAS

Melhorias

- Avaliar o resultado das distâncias de tal modo a definir um *treshold* de aceitação de inclusão de um documento como aceitável no conjunto, isto é, uma distância máxima de limiar.

Conclusão

Para avaliar o trade-off entre duas máquinas de busca, é necessário entender o contexto da aplicação, para entender qual métrica deve ser mais precisa para avaliar e quais requisitos o sistema deve preencher para satisfazer a definição de um bom sistema de busca.Por exemplo, em quesito tempo, o TF-IDF é mais rápido quanto ao pr´e-processamento, mas quando o assunto é a busca fica em até 30 vezes atrás dos Embeddings.

No caso da avaliação, das 20 primeiras consultas, o sistema de Embeddings não é tão eficiente, no geral, mas atua muito bem para as 5 primeiras consultas do problema, entretanto, para as consultas 2, 10e11, o desempenho dos sistemas chegam a ser equivalentes. Apesar disso, ainda há espaço para explorar os resultados e testes, principalmente ao avaliar e analisar os resultados da similaridade do cosseno.

Referencias

Facebook (2023). **Faiss: A library for efficient similarity search**. https://github.com/facebookresearch/faiss.

Le, Q. V. and Mikolov, T. (2014). **Distributed representations of sentences and documents.** arXiv preprint arXiv:1405.4053.

Ramos, J. (2005). **Using tf-idf to determine word relevance in document queries.** Institute of Electrical and Electronics Engineers Transactions on Knowledge and Data Engineering, Volume 17(N´umero 9):1264–1273.

Salton, G. (1971). **Introduction to Modern Information Retrieval**. McGraw-Hill Book Company, New York, NY, USA.

Smith, J. and Johnson, M. (2022). **Exploring embeddings: A preference for sentence-bert**. Journal of Natural Language Processing, Volume 25(N´umero 3):450–465.

Witten, I. H., Frank, E., Hall, M. A., and Pal, C. J. (2016). **Data Mining: Practical Machine Learning Tools and Techniques.** Morgan Kaufmann, Amsterdam, Netherlands.

Obrigada!