

IA368-DD

Deep Learning aplicado a Sistemas de Buscas

1o. semestre 2023

Professores: Roberto Lotufo e Rodrigo Nogueira

Assunto da primeira aula

- Questionário expectativa
- Introdução: Regras
- Dinâmica de grupo - Socialização
- Exercícios Próxima Aula:
 - Notebook/Apresentação
 - Leitura do artigo

Roteiros das Aulas

3 horas no total, sendo:

- Colab Notebook: 4 apresentações ***informais*** de 15 minutos cada (1.5 horas no total, considerando discussões)
- Leitura do artigo: 2 apresentações de 15 minutos cada sobre os conceitos e contribuições mais importantes do artigo. (1 hora no total, considerando discussões)
- Discussão exercício da próxima semana (30 min)

Alunos serão escolhidos no dia pelos professores para a apresentação

Avaliação

15% - Leitura de artigos

- Apresentação

50% - Programas semanais feitos em PyTorch no Jupyter/Colab

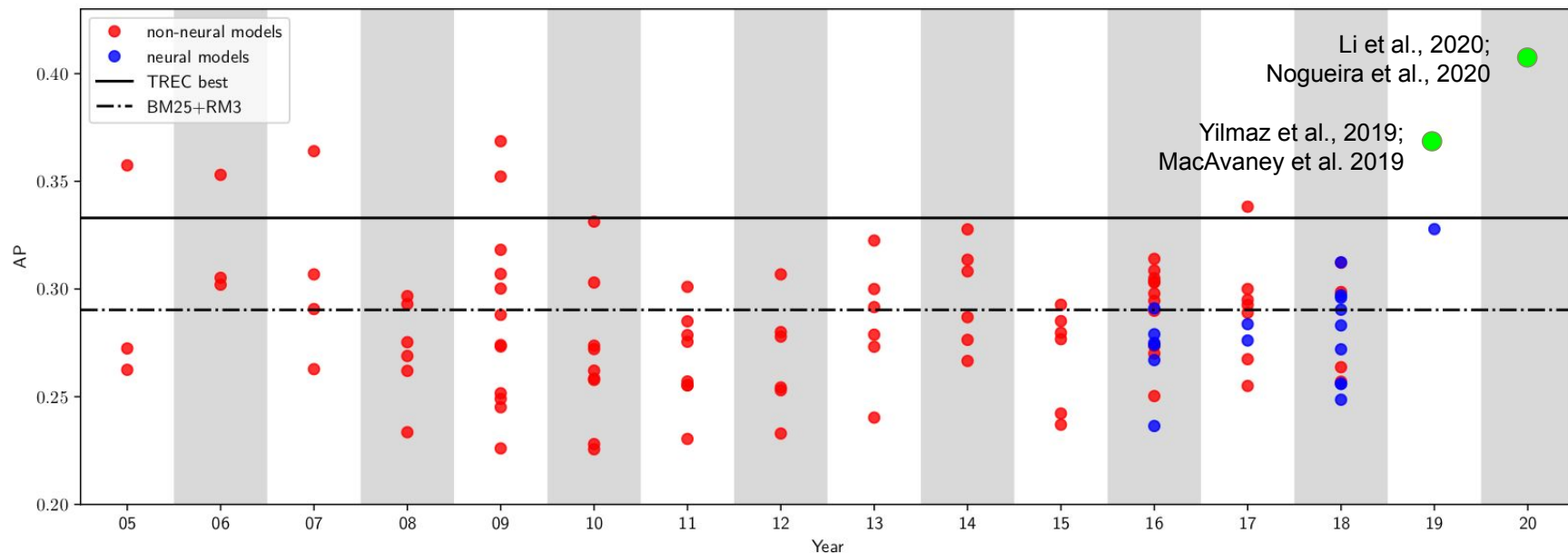
- Jupyter notebook
- Apresentação

35% - Projeto Final (4-5 semanas)

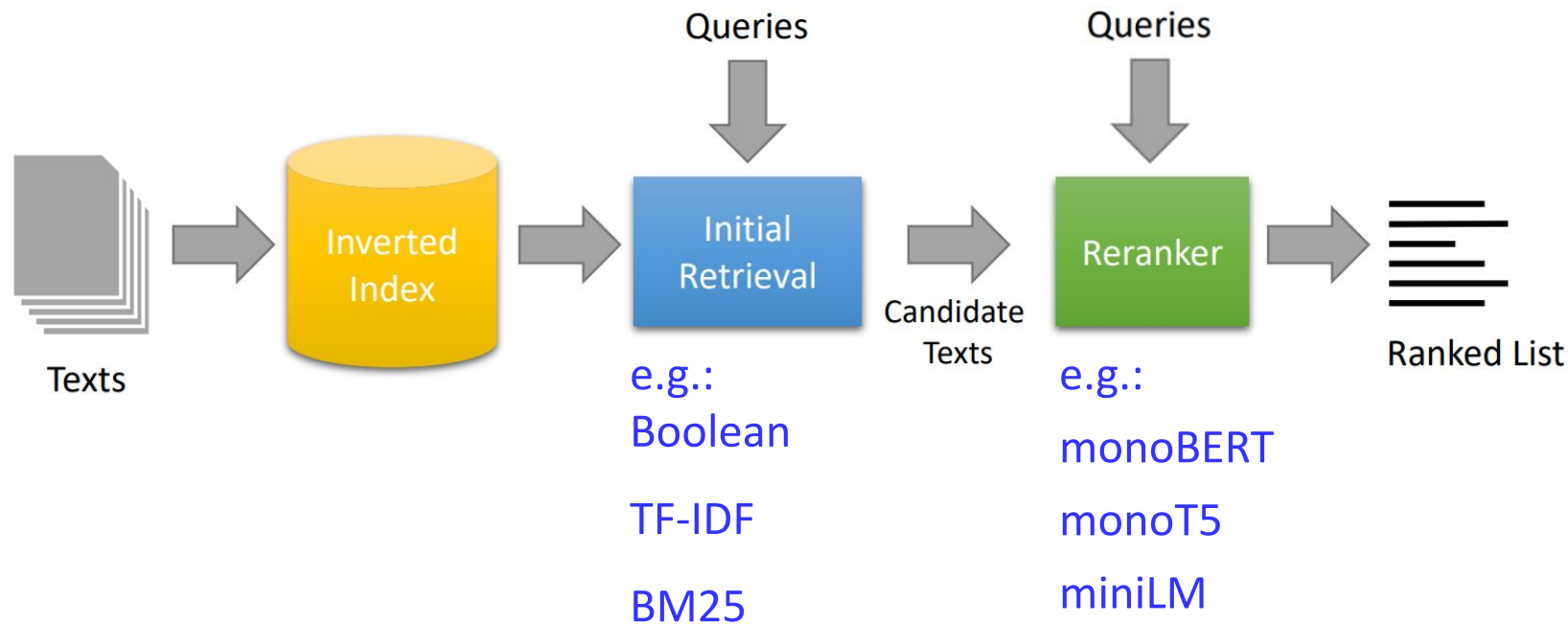
[Código de conduta](#)

Progress in Information Retrieval - Robust04

Some of them are zero-shot!



A Simple Search Engine



Nossa experiência: nsx.ai

neuralsearchx[®]

neuralmind[™]

Login

Search

Web



for

Como plantar tulipas



Found 30 results (1.39 seconds)

Preencha metade do vaso com terra rica em nutrientes. Coloque uma fina camada de areia de construção. Depois destes passos importantes coloque o bulbo na tulipa no centro. A ponta do bulbo deve ficar voltada para cima. Cubra com terra sem encher completamente o vaso. Quando

Plantar Tulipas A tulipa ...

<https://guiadasplantas.com/como-plantar-tulipa/>

Como Plantar Tulipa - Guia Das Plantas

<https://www.vortexmag.net/jardim-como-e-quando-plantar-tulipas/>

Jardim: como e quando plantar tulipas? | VortexMag

Cuidar do local onde onde serão plantadas é de extrema importância no cultivo de **tulipas**. Prepare o solo da seguinte forma: Escolher um local ensolarado com boa drenagem Cavar a área e afrouxar o solo com cerca de 30 cm de profundidade Adicionar um pouco de adubo ou composto seco ao solo

<https://construindodecor.com.br/tulipa/>

Tulipa: Significado, Como Plantar, Como Cultivar e Cores - ConstruindoDECOR

Como **Plantar Tulipas**. Como **plantar** tulipa. A priori, use um vaso que tenha aproximadamente 20 centímetros de diâmetro e com furos. Também é necessário que o vaso tenha entre 15 e 40 centímetros de profundidade e podem ser de cerâmica ou plástico. Desse modo, um vasinho medindo 20 cm de diâmetro comp...

Resposta

Para plantar tulipas, primeiro é necessário preparar o solo, cavando a área e adicionando adubo ou composto seco ao solo. Em seguida, escolha um vaso com pelo menos 20 cm de diâmetro e furos embaixo, com entre 15 e 40 cm de profundidade. Adicione uma mistura de adubo para vasos de boa qualidade e encha-o de meio a dois terços. Coloque um grupo de bulbos no vaso, deixando apenas um pequeno espaço. Enterre os bulbos no substrato, com uma distância de mais ou menos 3 centímetros entre eles e aproximadamente 12 centímetros de profundidade. Quando chegar o inverno, o bulbo da tulipa deve ser plantado em um vaso com terra, mas sem que fique totalmente enterrado e com os brotos virados para cima. A terra deve ser regada e o vaso colocado em um saco plástico para seguir para a geladeira, onde permanecerá nos próximos seis meses.

Explicação

Os [Documentos 1], [Documento 3], [Documento 4], [Documento 5], [Documento 8], [Documento 9], [Documento 11], [Documento 12], [Documento 14] e [Documento 15] fornecem informações sobre como plantar tulipas. Os [Documentos 2], [Documento 6] e [Documento 13] fornecem informações sobre como preparar o solo para a plantação. O [Documento 7] e o [Documento 10] fornecem informações sobre quando plantar tulipas.

Conteúdo do curso

Aprendizado Profundo	PLN	RI
MLP Backpropagation Transformers Mecanismo de atenção Treinamento Supervisionado Auto-supervisão	Text embeddings Modelos de linguagem Modelos seq2seq In-context learning (few-shot) Perguntas e respostas Tradução de máquina Classificadores de texto Teacher-forcing vs predição Greedy decoding, Beam Search TF-IDF/Bag-of-Words Tokenização Geração de Dados sintéticos	BM25 Indexação Buscadores Densos (e Approximate Nearest Neighbor) Buscadores Esparsos Expansão de Documentos Expansão de Queries Rerankeadores Metricas (nDCG, MRR, MAP) Anotação de datasets Trade-offs qualidade vs velocidade Sumarização de múltiplos documentos baseado em perguntas

Artigos Relevantes

PLN	RI
A Neural Probabilistic Language Model (Bengio et al, 2003)	Lecture Notes on IR (Tonellotto)
Word2Vec	Pretrained Transformers for Text Ranking (2020)
Attention is All you Need (2017)	monoBERT (2019)
BERT (2018)	doc2query (2019) e docT5query (2020)
T5 (2019)	monoT5 (2020)
GPT-3 (2020)	DPR (2020)
Scaling Laws (2020)	CoBERT (2020)
Distilling the Knowledge in a NN (2015)	UniCOIL (2021)
PALM (2022)	SPLADE (2021)
Chain of Thought	NeuralSearchX (2022)
	InPars (2022)
	Visconde (2022)
	Transformer Memory as a Differentiable Search Index

Programação dos Exercícios (sujeito à mudanças)

Aula	Exercício	Artigo	Tópicos
1	Buscador Simples: Booleano, TF-IDF, BM25	Pretrained Transformers for Text Ranking (até capítulo 1)	Indexação, Bag-of-Words, TF-IDF, BM25
2	Classificador binário: Análise de Sentimento e Ranqueamento		MLP, Treinamento Supervisionado, Classificadores de texto, Rerankeadores, Métricas (nDCG, MRR, MAP)
3	Aplicar LLM's Zero e Few-shot (aplicação escolhida pelo aluno)		Tokenização, Modelos de linguagem, In-context learning (few-shot), Auto-supervisão
4	Transformer avançado: Implementação e treinamento (modelagem de linguagem)		Transformers, Mecanismo de atenção, Backpropagation
5	Modelo seq2seq: T5 para expansão de documentos (doc2query)		Modelos seq2seq, Teacher-forcing vs predição, Tradução de máquina, Expansão de Documentos
6	Buscadores Densos: DPR		Buscadores Densos, Text Embeddings, Approximate Nearest Neighbor
7	Buscadores Esparsos: SPLADE		Buscadores Esparsos, Expansão de Queries, Tokenização
8	InPars: Adaptação de modelos para novas tarefas		Anotação de datasets, Geração de Dados sintéticos
9	Destilação		Trade-offs qualidade vs velocidade
10	Multi-document QA: Visconde		Perguntas e respostas, Sumarização de múltiplos documentos baseado em perguntas

Importante!

- É esperado que cada aluno assine o Colab Pro (~60 reais/mês) ou equivalente
- Cursos passados mostraram que é muito difícil realizar os exercícios usando a versão grátis do Colab.
- Caso precisem de ajuda financeira, contate os professores.

Dicas

- Usar ChatGPT para tudo
- Ser crítico nas respostas do ChatGPT
- Notebook com bastante documentação (texto explicando as células ou bloco de células)

Roteiro para Apresentação do Notebook

A apresentação deve cobrir *pelo menos* 3 dos 7 itens abaixo:

1. Explicação de conceitos importantes do exercício feito
2. Técnicas para garantir que a implementação está correta
3. Truques de código que funcionaram
4. Problemas e soluções no desenvolvimento
5. Resultados interessantes/inesperados
6. Uma dúvida "básica" que você ou os colegas possam ter
7. Um tópico "avançado" para discutirmos

Roteiro para Apresentação do Artigo

A apresentação deve cobrir *pelo menos 2* dos 5 itens abaixo:

1. Explicação de conceitos importantes do artigo
2. A contribuição do artigo
3. Resultados interessantes/inesperados
4. Uma dúvida "básica" que você ou os colegas possam ter
5. Um tópico "avançado" para discutirmos

Exercício desta semana

1. Usar o BM25 implementado pelo pyserini para buscar queries no TREC-DL 2020
 - Documentação referencia:
<https://github.com/castorini/pyserini/blob/master/docs/experiments-msmarco-passagere.md>
2. Implementar um buscador booleano/bag-of-words.
3. Implementar um buscador com TF-IDF
4. Avaliar implementações 1, 2, e 3 no TREC-DL 2020 e calcular o nDCG@10

Nos itens 2 e 3:

- Fazer uma implementação que suporta buscar eficientemente *milhões* de documentos.
- Não se pode usar bibliotecas como sklearn, que já implementam o BoW e TF-IDF.

Perguntas a serem respondidas

Quais os problemas que sistemas de busca resolvem?

Qual a diferença entre processamento de linguagem natural e recuperação de informações?

Qual o algoritmo de buscas mais usado e por quê?

Quais são as limitações do BM25?

Qual a estrutura de dados usada pelo BM25 que permite que busquemos eficientemente milhões de documentos?

Se aumentarmos em 10x o número de documentos na coleção, de quantas vezes vai aumentar o tempo para responder uma query?

Como avaliamos a qualidade de um sistema de buscas?

Quais metricas são comumente usadas?


Quais os problemas quando usamos um dataset de avaliação de sistemas de buscas

Sparse Retrieval with BM25

Widely used in Academia and Industry

Works in 2 phases:

1. **Indexing (offline):** inverted index construction: a dictionary whose keys are words and values are documents that contain those words;
2. **Retrieval/ranking:** for each word q_i in the query Q , compute a score for each document D that contains the word:



```
{  
  "apple": [doc_32, doc_5],  
  "house": [doc_85, doc_9],  
  ...  
}
```

$$\text{score}(D, Q) = \sum_{i=1}^n \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right)},$$

Suffers from the "Vocabulary mismatch problem": *car* and *automobile* are completely different to BM25

However, a hard-to-beat algorithm

Index construction

doc1: "car banana apple"



```
inverted_index =  
  { "car": [doc1],  
    "banana": [doc1],  
    "apple": [doc1]  
  }
```

doc2: "car girl boy"



```
inverted_index =  
  { "car": [doc1, doc2],  
    "banana": [doc1],  
    "apple": [doc1],  
    "girl": [doc2],  
    "boy": [doc2]  
  }
```

doc3: "house"



```
inverted_index =  
  { "car": [doc1, doc2],  
    "banana": [doc1],  
    "apple": [doc1],  
    "girl": [doc2],  
    "boy": [doc2],  
    "house": [doc3]  
  }
```

Retrieval time

query: "car banana"

Returns:

- doc1: score = 2 -> rank 1
- doc2: score = 1 -> rank 2

query: "automobile"

Returns:

- None

```
inverted_index =  
    { "car": [doc1, doc2],  
      "banana": [doc1],  
      "apple": [doc1],  
      "girl": [doc2],  
      "boy": [doc2],  
      "house": [doc3]  
    }
```