Trade-offs de eficiência e qualidade

Monique Monteiro –
moniquelouise@gmail.com

Conceitos importantes

- Importância do uso de:
  - Pipelines em geral
  - Medição de custos temporais e econômicos
  - Aferição de métricas a cada componente adcionado/removido
    - Um único modelo excelente ou mediano pode fazer toda a diferença no resultado final

Problemas no desenvolvimento

- Alta variância de latência no ambiente do Google Colab em um determinado dia
- Solução:
  - Nenhuma! Apesar de todas as tentativas de otimização, em um determinado horário os tempos subitamente melhoraram

Modelos utilizados

- cross-encoder/ms-marco-MiniLM-L-6-v2 (reranqueamento)
  - Sem finetuning (off-the-shelf)
  - Treinado InPars (dataset coletivo)
- SPLADE (naver/splade-cocondenser-ensembledistil) (reranqueamento)
- MonoT5 = castorini/monot5-3b-med-msmarco
  - Inspiração: resultados obtidos pelo Leonardo Ávila
  - Primeiro modelo *off-the-shelf* adaptado para o domínio médico

Primeiro baseline – BM25 + reranqueador sem finetuning

- Reranqueador cross-encoder/ms-marco-MiniLM-L-6-v2
  - Sem finetuning no TRE – COVID
  - Uso de *sentence-transformers*
  - NDCG@10 = 0,27 (pior que BM25, com NDCG@10 = 0,5947)
  - Biblioteca caixa-preta pode ter levado a algum erro na chamada
  - Eficiência não foi avaliada

Segundo baseline - SPLADE

- NDCG@10 = 0,7268
- Latência = 1,31 segs. p/ *query*
- Máquina = T4
- Tempo de indexação:
  - Tokenização = 41,13 segundos.
  - Geração da matriz de documentos = 18 minutos e 25 segundos
  - Construção do índice invertido = 20 minutos e 53 segundos
  - Total = aprox. 40 minutos

Segundo baseline – Estimativa de custos

USD p/ query	100 queries por dia em 1 mês	Custo de indexação
“0” (0,000076)	$ 0,23	$ 0,14

Pipeline 1 – Pyserini BM25 + SPLADE (top 1000)

- NDCG@10 = 0,7419
- Latência = 1,49 segs. p/ query
- Máquina = T4
- Tempo de indexação:
  - Tokenização = 41,13 segundos.
  - Geração da matriz de documentos = 18 minutos e 25 segundos
  - Construção do índice invertido = 20 minutos e 53 segundos
  - Total = aprox. 40 minutos
- Não foi contabilizada a indexação pelo Pyserini
  - Uso do “prebuilt index” *beir-v1.0.0-trec-covid.flat*

Pipeline 1 – Estimativa de custos

USD p/ query	100 queries por dia em 1 mês	Custo de indexação
“0” (0,000087)	$ 0,26	$ 0,14

Pipeline 2 – Doc2Query + BM25 + SPLADE (top 100)

- NDCG@10 = 0,7443
- Latência = 0,5 segs. p/ query
- Máquina = A100 para indexação, T4 para inferência
- Tempo de indexação (documentos previamente explandidos):
  - Tokenização = 34,33 segundos.
  - Geração da matriz de documentos = 36 minutos e 36 segundos
  - Construção do índice invertido = 3 horas, 35 minutos e 34 segundos
  - Total = aprox. 4 horas
  - OBS. 1: Período de extrema latência no Google Colab
  - OBS. 2: Pendente cálculo do tempo para expansão dos documentos

Pipeline 2 – Estimativa de custos

USD p/ query	100 queries por dia em 1 mês	Custo de indexação
“0” (0,000029)	$ 0,09	$ 6,00

Pipeline 3 – Doc2Query + SPLADE (top 1000)

- NDCG@10 = 0,7319
- Latência = 1,07 segs. p/ query
- Máquina = A100 para indexação, T4 para inferência
- Tempo de indexação = idem pipeline anterior

USD p/ query	100 queries por dia em 1 mês	Custo de indexação
“0” (0,000062)	$ 0,19	$ 6,00

Pipeline 4 - Doc2Query + BM25 + SPLADE + InPars

- InPars
  - Reúso da versão treinada no TREC-COVID
  - Ajustes levaram NDCG@10 de 67% para 70%
    - Uso de *token type ids* na inferência (dicas Leandro/Mirelle)
- NDCG@10 = 0,6768
- Latência = 1,82 segs. p/ query
- Máquina = A100 para indexação, T4 para inferência
- Tempo de indexação = idem pipeline anterior

Pipeline 4 – Estimativa de custos

USD p/ query	100 queries por dia em 1 mês	Custo de indexação
“0” (0,000106)	$ 0,32	$ 6,00

Pipeline 5 – Doc2Query + SPLADE + InPars

- Remoção da etapa BM25
- NDCG@10 = 0,7602
- Latência = 1,91 segs. p/ query
- Máquina = A100 para indexação e inferência
- Tempo de indexação = idem pipeline anterior

USD p/ query	100 queries por dia em 1 mês	Custo de indexação
“0” (0,000796)	$ 2,39	$ 6,00

Pipeline 6 – BM25 + SPLADE + MonoT5 (top 1000)

- Máquina = T4 para indexação, A100 para inferência
- NDCG@10 = 0,7842
- Latência = 21,1 segs. p/ query (proibitiva!)
  - Motivos:
    - número excessivo de documentos para reranqueamento
    - biblioteca baseada em geração de arquivos intermediários
- Tempo de indexação = aprox. 40 minutos

USD p/ query	100 queries por dia em 1 mês	Custo de indexação
“0” (0,008792)	$ 26,38	$ 0,14

Pipeline 7 – BM25 + SPLADE + MonoT5 (top 100)

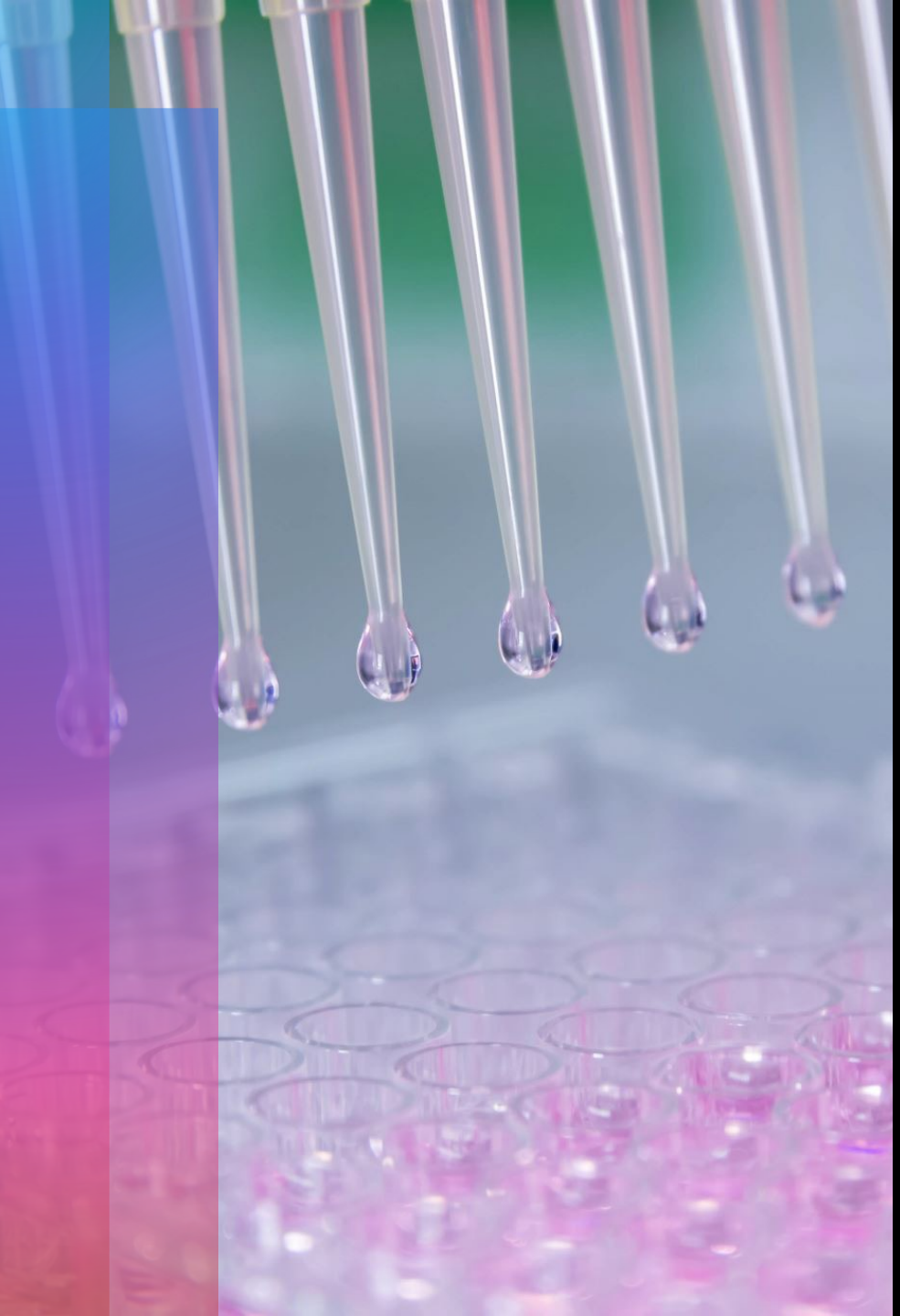
- Máquina = A100
- NDCG@10 = 0,8098
- Latência = 3,02 segs. p/ query
  - Melhorou consideravelmente, porém ainda alta!
- Tempo de indexação = idem pipeline anterior

USD p/ query	100 queries por dia em 1 mês	Custo de indexação
“0” (0,001258)	$ 3,78	$ 0,14

Pipeline 8 – BM25 + SPLADE + MonoT5 (top 50)

- Máquina = A100
- **NDCG@10 = 0,8128**
- Latência = 2,54 segs. p/ query
- Tempo de indexação = idem pipeline anterior

USD p/ query	100 queries por dia em 1 mês	Custo de indexação
“0” (0,001058)	$ 3,18	$ 0,14



Tópico avançado

- Diante dos avanços recentes em LLMs, até quando vamos precisar de “pipelines” de busca?