INPARS

Monique Monteiro –
moniquelouise@gmail.com

CONCEITOS IMPORTANTES

- LLMs (*Large Language Models*) raramente são usados em IR (Information Retrieval), com notáveis exceções
- Natureza intensiva computacionalmente das tarefas de IR (ex.: rereanqueamento)
- Alto custo financeiro para uso de APIs (ex.: OpenAI)
- *Overhead* mesmo para geração de vetores densos
- Escassez de datasets rotulados
- Alternativa: uso de geradores sintéticos baseados em LLMs

PROBLEMAS E SOLUÇÕES NO DESENVOLVIMENTO

- Dificuldade de uso de few-shot learning com GPT 3.5 Turbo
  - *Solução: uso de zero-shot learning*
- Dificuldade de treinamento do modelo cross-encoder/ms-marco-MiniLM-L-6-v2
  - *Uso de modelo alternativo (microsoft/MiniLM-L12-H384-uncased)*
  - *Uso de dataset maior e mais diverso*
  - *Aumento da taxa de aprendizado*
  - *Modificação na forma como a saída do modelo é obtida (em relação ao reranqueador da aula 2)*
    - Thanks to @Mirelle!
  - *Salvamento do melhor checkpoint*

RESULTADOS

Modelo	Dataset	LR	Épocas	Batch size	Acurácia	NDCG@10	BM25
microsoft/MiniLM-L12-H384-uncased	Próprio (1000 exemplos, balanceados)	5e-5	5	32	98%	61,26%	64,1%
cross-encoder/ms-marco-MiniLM-L-6-v2	Próprio (1000 exemplos, balanceados)	1e-3	10	32	86,5%	52,69%	64,1%
cross-encoder/ms-marco-MiniLM-L-6-v2	União dos datasets gerados pela turma (com exemplos negativos)	1e-3	10	32	96,8%	62,04%	64,1%
cross-encoder/ms-marco-MiniLM-L-6-v2	União dos datasets gerados pela turma (com exemplos negativos, balanceados)	5e-5	10	32	88,71%	**67,52%**	64,1%
microsoft/MiniLM-L12-H384-uncased	União dos datasets gerados pela turma (com exemplos negativos, balanceados)	5e-5	5	16	WIP	**WIP**	64,1%