

# Buscador Denso



Monique Monteiro

[moniquelouise@gmail.com](mailto:moniquelouise@gmail.com)

Conceitos  
importantes

## *Dense retriever*

Uso de representações  
contínuas e densas

Expectativa de  
preservação da semântica



DPR (*Dense Passage  
Retrieval*)

# Problemas e soluções encontrados no desenvolvimento

- Alta sensibilidade do *finetuning* do modelo a hiperparâmetros como:
  - Tamanho do *batch*
  - Tamanho máximo da sequência
  - Taxa de aprendizagem
  - *Learning rate scheduler*
- Solução:
  - hiperparâmetros utilizados pelo Leandro

# Hiperparâmetros comuns a todos os experimentos

- Número de épocas: 18
- Tamanho do *batch*: 32
- *Pooling*: [CLS] token



# Resultados – similaridade por produto escalar

LR	<i>Scheduler</i>	<i>Validation loss</i>	NDCG@10	<i>Max length</i>	<i>Approx. KNN</i>	K
1e-5	Linear	0,0590	0.28552	512	-	-
2e-5	Cosine	0,0372	0.4186	256	-	-
2e-5	Cosine	0,0372	0.4186	256	hnswlib	-
2e-5	Cosine	0,0372	0.403	256	Manual	5
2e-5	Cosine	0,0372	0.334	256	Manual	50



# Resumo dos principais resultados

## Esperados

- *Approximate Nearest Neighbors*
  - Quanto maior o número de clusters, pior o resultado final

## Interessantes/inesperados

- Variações mínimas na *loss* de validação impactam diretamente métrica final
- Clusterização com 400 clusters não concluiu
  - Tempo proibitivo de processamento (Kmeans – scikit-learn)
- Implementação do HNSWLIB por padrão é busca exaustiva
  - Implementação “caixa-preta” com parâmetros pouco intuitivos