Notebook – Qualidade vs Eficiência

Leandro Carísio

O que foi avaliado

Relação entre SPLADE (S), doc2query (d2q) e InPars (I)

Modelos usados:

S: naver/splade-cocondenser-selfdistil
d2q: Expansão gerada na aula 5
I: cross-encoder/ms-marco-MiniLM-L-6-v2 com fine-tuning feito na aula 8

Variações:

1. S
2. S + I
3. S + I (d2q)
4. S (d2q)
5. S (d2q) + I
6. S (d2q) + I (d2q)
7. S + apenas expansão d2q
8. S + apenas expansão d2q + I (doc original)

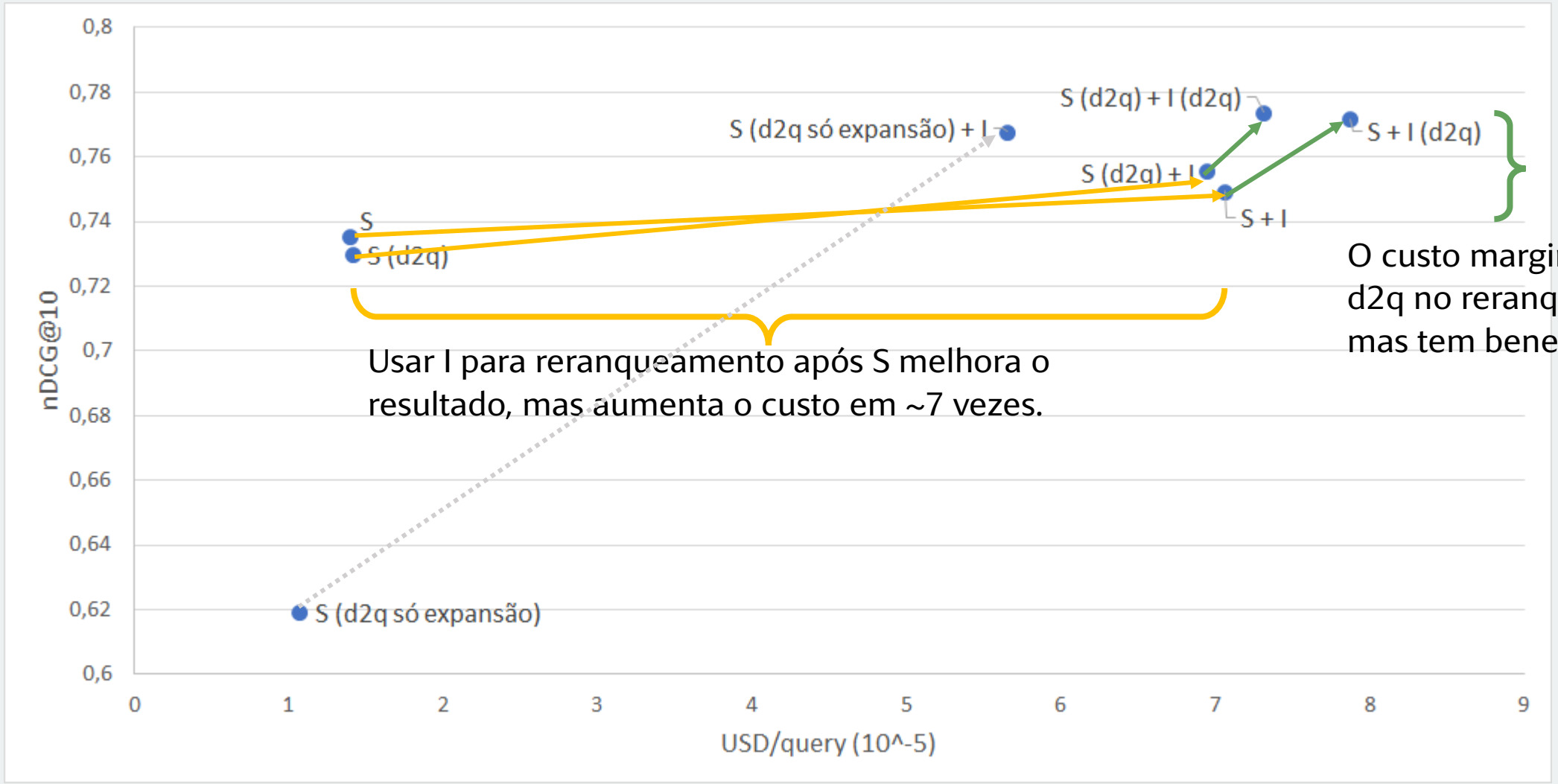
Resultados (usando Colab - V100)

Pipeline	nDCG@10	Tam índice S (pickle, MB)	seg/query	Tempo CPU (seg/query)	Tempo GPU (seg/query)	USD/query (x 10^{-5})
S	0.7354	294.5 MB 12m33s	1.35	1.33	0.02	1.39
S + I	0.7493		1.73	1.30	0.43	7.06
S + I (d2q)	0.7718		1.92	1.44	0.48	7.87
S (d2q)	0.7298	294.2 MB 13m26s	1.38	1.36	0.02	1.41
S (d2q) + I	0.7559		1.75	1.33	0.42	6.94
S (d2q) + I (d2q)	0.7737		1.72	1.27	0.45	7.31
S (d2q só expansão)	0.6193	145.4 MB 6m40s	0.97	0.95	0.02	1.07
S (d2q só expansão) + I	0.7678		1.30	0.95	0.35	5.65

Em todos os casos, o índice SPLADE foi gerado com max_seq_length = 256 (se mostrou melhor do que com 512) e batch_size = 32

Resultados (usando Colab - V100)

Para medir o tempo/query, fiz as 50 queries sequencialmente e medi o tempo. Fiz 3x e tirei a média.



Obrigado

Leandro Carísio

carisio@gmail.com