Notebook – doc2query

Leandro Carísio

Conceitos do exercício

1. Indexar TREC-COVID e testar com BM25

2. Treinar um modelo seq2seq usando MS MARCO
   (X, Y) => (DOCUMENTO, QUERY)

3. Usar o modelo treinado para inferir queries para os doc do TREC-COVID e enriquecer a base (usei 10 queries, top_p = 0.9)

4. Indexar a base enriquecida e testar com BM25

Problemas e soluções no desenvolvimento

Principal dificuldade foi com a biblioteca (HF)

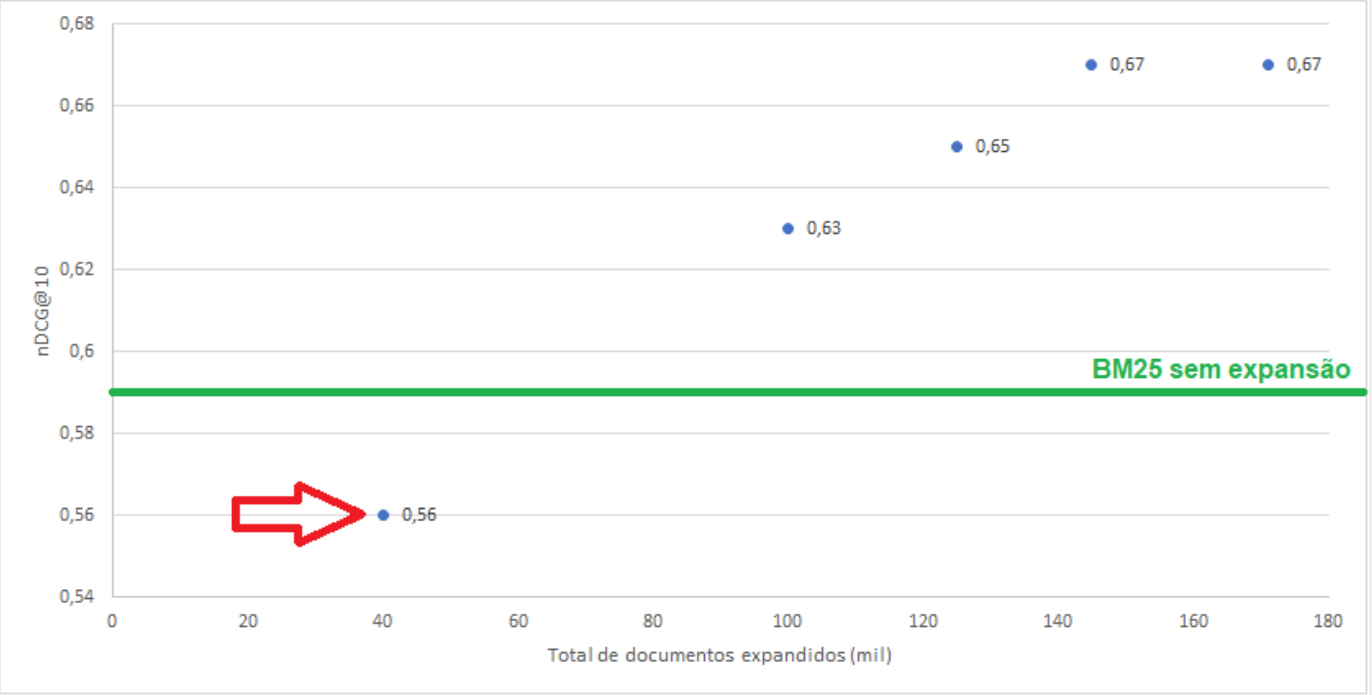
=> Solução: caderno da Monique

Tempo de treinamento

Resultados interessantes/inesperados

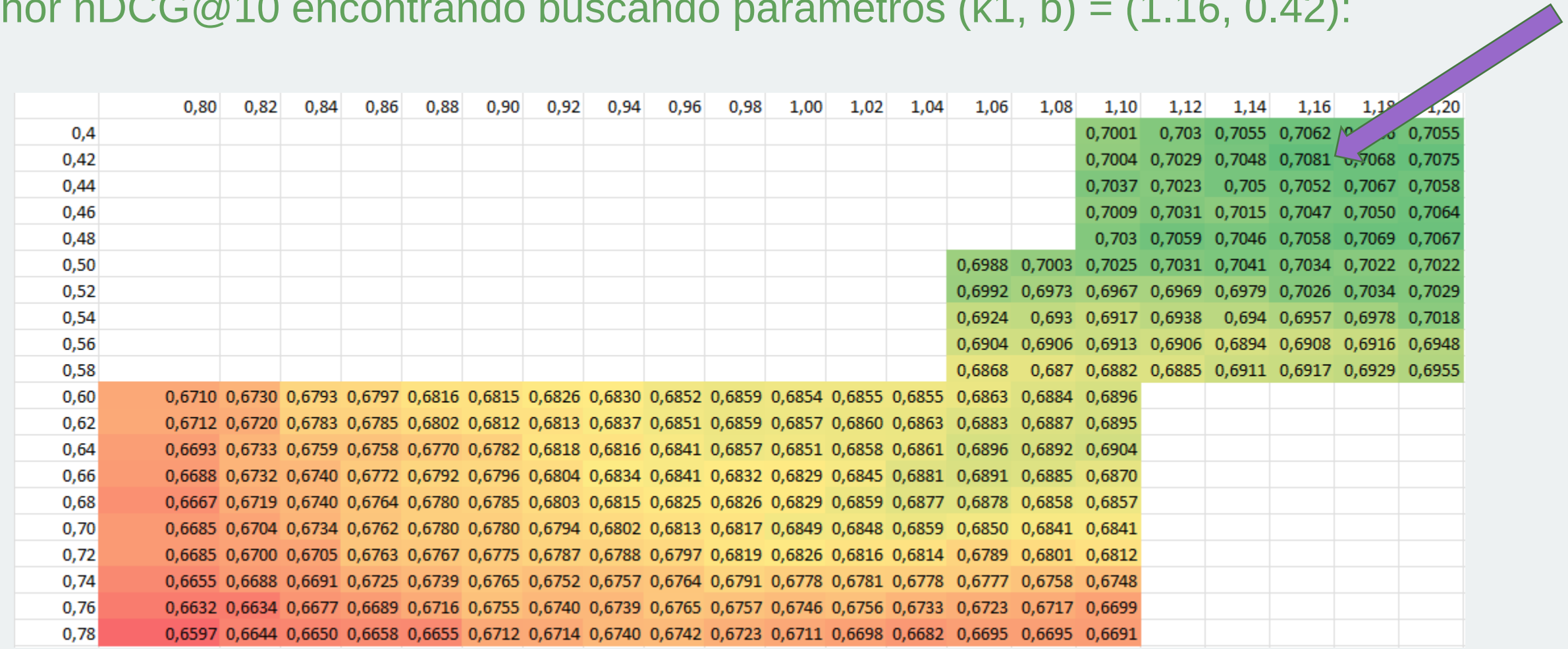
Método	nDCG@10
BM25	0,5956
BM25 doc. original + expansão	0,6719
BM25 sem doc. original + expansão	0,5225

Resultados para
k1 = 0.82
b = 0.68



Resultados interessantes/inesperados

Melhor nDCG@10 encontrando buscando parâmetros (k1, b) = (1.16, 0.42):



	0,80	0,82	0,84	0,86	0,88	0,90	0,92	0,94	0,96	0,98	1,00	1,02	1,04	1,06	1,08	1,10	1,12	1,14	1,16	1,18	1,20
0,4																0,7001	0,703	0,7055	0,7062		0,7055
0,42																0,7004	0,7029	0,7048	0,7081		0,7075
0,44																0,7037	0,7023	0,705	0,7052	0,7067	0,7058
0,46																0,7009	0,7031	0,7015	0,7047	0,7050	0,7064
0,48																0,703	0,7059	0,7046	0,7058	0,7069	0,7067
0,50														0,6988	0,7003	0,7025	0,7031	0,7041	0,7034	0,7022	0,7022
0,52														0,6992	0,6973	0,6967	0,6969	0,6979	0,7026	0,7034	0,7029
0,54														0,6924	0,693	0,6917	0,6938	0,694	0,6957	0,6978	0,7018
0,56														0,6904	0,6906	0,6913	0,6906	0,6894	0,6908	0,6916	0,6948
0,58														0,6868	0,687	0,6882	0,6885	0,6911	0,6917	0,6929	0,6955
0,60	0,6710	0,6730	0,6793	0,6797	0,6816	0,6815	0,6826	0,6830	0,6852	0,6859	0,6854	0,6855	0,6855	0,6863	0,6884	0,6896					
0,62	0,6712	0,6720	0,6783	0,6785	0,6802	0,6812	0,6813	0,6837	0,6851	0,6859	0,6857	0,6860	0,6863	0,6883	0,6887	0,6895					
0,64	0,6693	0,6733	0,6759	0,6758	0,6770	0,6782	0,6818	0,6816	0,6841	0,6857	0,6851	0,6858	0,6861	0,6896	0,6892	0,6904					
0,66	0,6688	0,6732	0,6740	0,6772	0,6792	0,6796	0,6804	0,6834	0,6841	0,6832	0,6829	0,6845	0,6881	0,6891	0,6885	0,6870					
0,68	0,6667	0,6719	0,6740	0,6764	0,6780	0,6785	0,6803	0,6815	0,6825	0,6826	0,6829	0,6859	0,6877	0,6878	0,6858	0,6857					
0,70	0,6685	0,6704	0,6734	0,6762	0,6780	0,6780	0,6794	0,6802	0,6813	0,6817	0,6849	0,6848	0,6859	0,6850	0,6841	0,6841					
0,72	0,6685	0,6700	0,6705	0,6763	0,6767	0,6775	0,6787	0,6788	0,6797	0,6819	0,6826	0,6816	0,6814	0,6789	0,6801	0,6812					
0,74	0,6655	0,6688	0,6691	0,6725	0,6739	0,6765	0,6752	0,6757	0,6764	0,6791	0,6778	0,6781	0,6778	0,6777	0,6758	0,6748					
0,76	0,6632	0,6634	0,6677	0,6689	0,6716	0,6755	0,6740	0,6739	0,6765	0,6757	0,6746	0,6756	0,6733	0,6723	0,6717	0,6699					
0,78	0,6597	0,6644	0,6650	0,6658	0,6655	0,6712	0,6714	0,6740	0,6742	0,6723	0,6711	0,6698	0,6682	0,6695	0,6695	0,6691					

nDCG@10 0.7081 (com expansão) e 0.6255 (sem expansão)

Obrigado

Leandro Carísio

carisio@gmail.com