Notebook – Reranking pós BM25

Leandro Carísio

Técnicas para garantir que a implementação está correta

Pesquisar algumas queries e ver se os resultados mais altos/baixos no ranking faz algum sentido => não garante, mas é uma evidência. Exemplo:

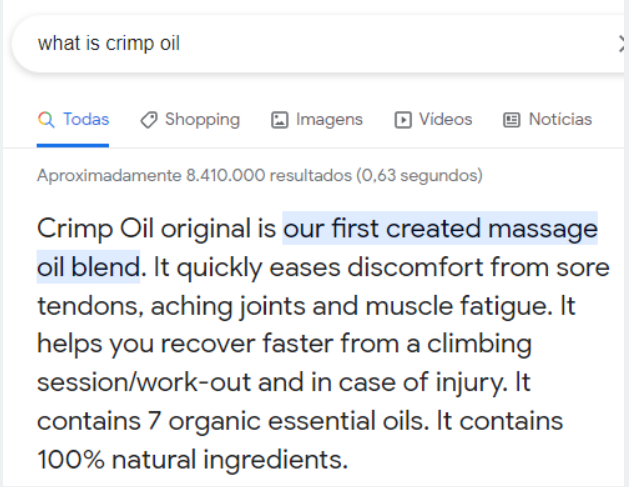
Query: what is crimp oil

Alto no ranking

- Crimp Oil is a 100% natural blend of essential oils and plant extracts that aids in the recovery of climbing-related injuries. 8626887
- Metolius Crimp Oil. 1 A healing massage oil for climbers hands and muscles. 2 Reduces pain and swelling in tendons, joints and muscles. 3 100% natural blend of essential oils and plant extracts. 4 Designed by climbers. 5 Available in 10 and 30 ml bottles.

Baixo no ranking

- Edible Oil Smoke & Flash Points [TEMPERATURE CHART] The smoke points of oils are important. These temperatures indicate at what temperature a particular type of oil will begin to smoke at, and they are key for allowing manufacturers to choose the right oils for their production process.
- Oil is found in underground pools of oil called reservoirs. This oil location is not what one might typically expect when considering the term pool. It is impossible to go swimming in these pools! Industry experts use the term pool to define accumulations of hydrocarbon in zones of subsurface rock. (see oil accumulation).



Técnicas para garantir que a implementação está correta

Inserir documentos fake que sabe-se que não tem relação com a pesquisa e ver se o score está muito baixo.

```
Query 735922: what is crimp oil: 100% ████████████████████ 32/32 [00:05<00:00, 6.22it/s]
Alto no ranking...
8626892, 0.997641921043396: Crimp Oil is a 100% natural blend of essential oils and plant extracts that aids in the recovery
8626887, 0.9969774484634399: Metolius Crimp Oil. 1  A healing massage oil for climbers hands and muscles. 2  Reduces pain an
7307866, 0.9961373209953308: crimp (plural crimps) 1  A fastener or a fastening method that secures parts by bending metal a
8765058, 0.9916137456893921: Definition of crimp. 1 : something produced by or as if by crimping: such asa : a section of ha
8765056, 0.9842262864112854: Definition of crimp. 1  1 : something produced by or as if by crimping: such asa : a section of

Baixo no ranking...
4672571, 0.0013112531742081046: what is the general crude oil commodity (Crude Oil per Barrel) trading name on NYSE..... thi
4403334, 0.0013111595762893558: Note: Not all essential oils are created equally. Many of the essential oils on the market t
3435309, 0.0013110919389873743: Essential oils are evaluated using a process called Gas Chromatography. This test tells the
7176467, 0.0013109034625813365: Edible Oil Smoke & Flash Points [TEMPERATURE CHART] The smoke points of oils are important.
3680623, 0.0013107211561873555: Oil is found in underground pools of oil called reservoirs. This oil location is not what on

Retirando o texto "feijão tropeiro is one of the best Brazilian foods. there isn't a single soul that doesn't like it"...


O texto fake ficou na posição 382 com score 0.001322504598647356
```

Apesar de ter ficado numa posição relativamente alta do ranking, o score é compatível com as últimas posições

Resultados interessantes/inesperados

Resultados de nDCG@10. O primeiro número é com o BM25 Pyserini, o segundo, quando existir, é implementação da Aula 1

tokenizer(f"{query} [SEP] {passage}")	tokenizer.encode_plus(query, passage, ...)
80% treinamento, 20% validação seed 42: 0.605/0.605 seed 123: 0.566/0.566	80% treinamento, 20% validação seed 42: 0.5852/X seed 123: 0.6198/0.6212
70% treinamento, 30% validação seed 42: 0.5849/X seed 123: 0.5919/X	70% treinamento, 30% validação seed 42: 0.5474/X seed 123: 0.5981/X

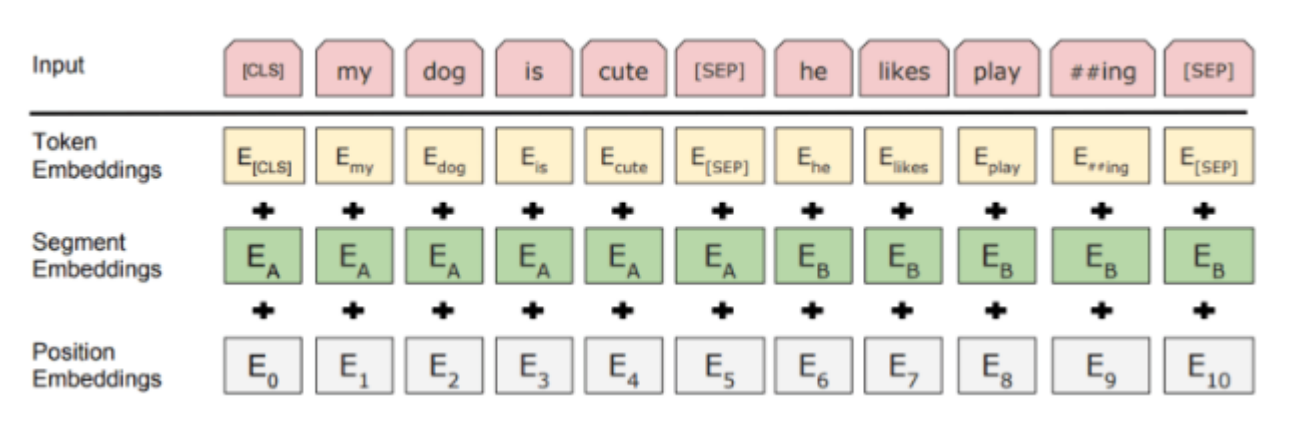
Para essa quantidade de dados, eu não esperava que a seed tivesse um efeito tão grande no nDCG@10

Dúvidas

1. O que exatamente é esse attention_mask e onde é usado?

2. O eval() no retorno de 1.000 passagens do BM25 levou uns 5 segundos. Como fazer pra isso ser prático numa pesquisa de documentos grandes?

Dúvidas



1. Porque todos os embeddings (token, segment e position) são somados? Como a rede sabe o que é o que?

2. Porque os segment embeddings são necessários? O [SEP] já não é suficiente?

Obs: o artigo chegou a testar o efeito de não usar segment embeddings, mas e se tirasse o segment embeddings e o [SEP]?

Obrigado

Leandro Carísio

carisio@gmail.com