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:

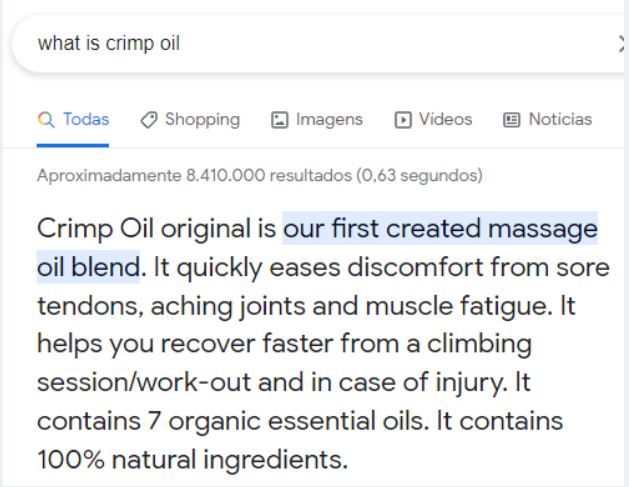
Query: what is crimp oil

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).



Resultados interessantes/inesperados

tokenizer(f"{query} [SEP] {passage}")

80% treinamento, 20% validação

seed 42: 0.605/0.605
seed 123: 0.566/0.566

70% treinamento, 30% validação

seed 42: 0.5849/X
seed 123: 0.5919/X

tokenizer.encode_plus(query, passage, ...)

80% treinamento, 20% validação

seed 42: 0.5852/X
seed 123: 0.6198/0.6212

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

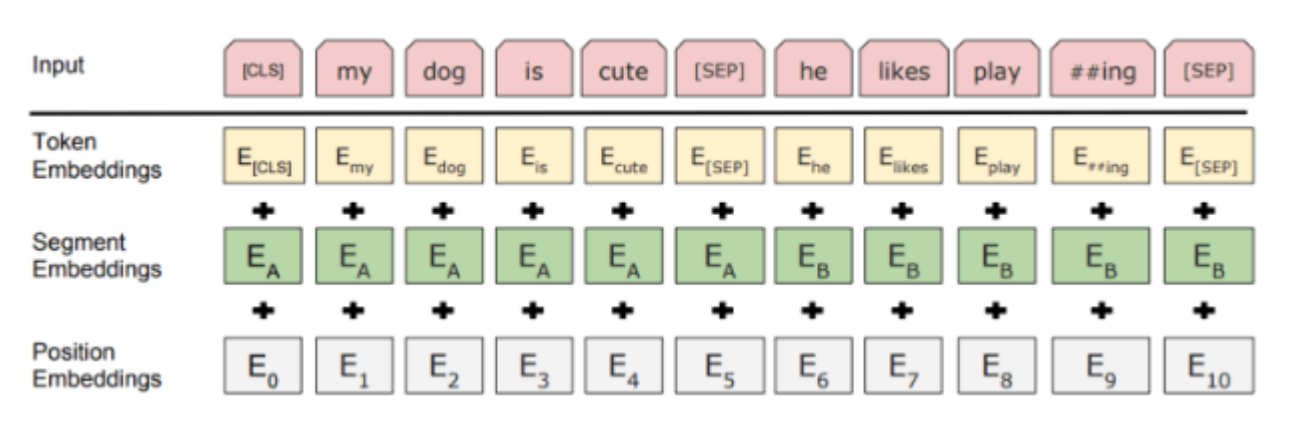
Dúvidas

1. Posso chamar tokenizer(f’{query} [SEP] {passage}’) assim mesmo ou preciso chamar como uma lista de strings? Qual a diferença?

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

3. 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?

Obrigado

Leandro Carísio

carisio@gmail.com