Natural Language Processing : First Project

Information Retrieval Challenge

Beating BM25

Context

As showned on the BeiR paper : https://arxiv.org/pdf/2104.08663.pdf, BM25 remains one of the best approaches on average tested on different datasets.

BM25 is a popular improvement of TF-IDF:

Given a query *Q*, containing keywords qi,…,qn , the BM25 score of a document *D* is :

$$\text{score}(D,Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i,D)\cdot(k_1+1)}{f(q_i,D)+k_1\cdot\left(1-b+b\cdot\frac{|D|}{\text{avgdl}}\right)}$$

Where only 3 elements are important :

- a local one : the frequency of the word inside the document (TF)
- a global one : the scarcity of the word inside the complete corpus (IDF)
- a preference selection : for 2 documents with same TF-IDF, the shortest one will be prefered.

The goal of the project is to develop our own information retrieval system on a specific corpus (NFCorpus). It's a medical corpus. The document are abstracts of medical publication from PubMed and the queries are scraps of vulgarisation on the topics linked to some PubMed articles.

It's a very complex corpus where modern deep learning approaches fail to perform better than BM25.

Expectation

Your goal is to develop an original information retrieval system on NFCorpus. In order to do that, you are allowed to use any kind of pre-treatment and manipulate the vocabulary of the documents. You can use a pre-trained word2vec model or learn your own word2vec model. You can mix everything but you aren't allowed to use direct supervised learning (for a given query predicting the best document).

BM25 is your baseline and you need to find a way to improve the result. The metric used is the ndcg @5 (it evaluate the top 5 results returned by the model).

The notebook NFCorpusBM25.ipynb is available here:

https://colab.research.google.com/drive/1zOsx2n_JdHAwvtLRNdsaaShoVIr5yCLq?usp=sharing

It allows to load NFCorpus dataset. It runs a bm25 model and finally it evaluate with ndcg metric the model. Your code will be also a colab notebook and need to compare to this code on BM25.

Details

The deliverables are your colab of your model and a small report with explanations.

Your report must explain what technics/approaches you use, how you use them and the results obtained. If an approach doesn't work as planned you can show and explain (It will be very appreciated).

You can work in pairs of students. Your report must contain the names of students involved. Your report must explain the logic of your approaches and results. You can write in English or French. Your report must contain your link to your Colab Notebook.

Your report must be deposited on DeVinciLearning before 20 november 2023.

Please share your colab notebook with your group teacher:

shikezhan@gmail.com

Christophe.rodrigues.bento@gmail.com