Project 1: Information Retrieval Challenge Beating BM25

Aurélie Martin – DIA2

Project description:

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](mailto:shikezhan@gmail.com)

[Christophe.rodrigues.bento@gmail.com](mailto:Christophe.rodrigues.bento@gmail.com)

Information Retrieval Challenge Report:

Introduction:

In this report, I detail the enhancements I made to an initial information retrieval system based on the BM25 algorithm. My objective was to augment the system's ability to effectively handle the NFCorpus, which comprises medical abstracts from PubMed publications and associated queries. I aimed to achieve this by integrating the Word2Vec model and experimenting with various text preprocessing techniques, including stemming, lemmatization, and the use of synonyms.

Initial Setup:

The starting point of my project was a system leveraging the BM25 algorithm, known for its effectiveness in ranking documents based on the frequency of query terms, adjusted by their inverse document frequency. This provided me with a baseline performance measured using the normalized discounted cumulative gain (ndcg).

Enhancements:

1. Experimentation with Text Preprocessing Techniques

I experimented with various text preprocessing techniques:

* **Stemming**: Reducing words to their root form, which helped in generalizing different forms of a word.
* **Lemmatization**: This involved reducing words to their base or dictionary form. While more contextually accurate, it was computationally intensive.
* **Synonyms**: I explored using synonyms to expand query terms, aiming to capture a broader range of relevant documents.

After thorough testing and evaluation, I decided to implement stemming. This decision was based on a balance between computational efficiency and the ability to generalize different forms of words, which proved beneficial for the medical corpus at hand.

1. Word2Vec Integration

I trained a Word2Vec model on our corpus to create word embeddings, which are adept at capturing the semantic context of words. This methodology enabled the system to retrieve documents that are semantically relevant to queries, beyond simple keyword matching.

Training Word2Vec

Using the Gensim library, I trained the Word2Vec model on the tokenized documents. I carefully selected parameters such as vector size and window size to optimize the model for our specific dataset.

Vectorizing Documents and Queries

After training, I vectorized both the documents and the queries using the Word2Vec model. This process involved converting text into numerical vector representations, encapsulating the semantic essence of the words.

1. Retrieval and Ranking Enhancement

Using cosine similarity, I calculated the similarity between query vectors and document vectors from the Word2Vec model. This provided similarity scores for each document-query pair.

1. Combining BM25 and Word2Vec Scores

To leverage the benefits of BM25's keyword matching and Word2Vec's semantic understanding, I combined the scores from both models. This combination involved a weighted sum of the scores from BM25 and Word2Vec for each document-query pair.

1. Evaluation Using NDCG

I used the ndcg metric to evaluate the effectiveness of the enhanced retrieval system. This evaluation showed an improvement over the initial BM25-only setup, indicating better retrieval of relevant documents, especially where semantic context is crucial.

Conclusion:

The integration of Word2Vec with BM25, combined with the choice of stemming as the primary text preprocessing technique, resulted in a more nuanced information retrieval system. It now efficiently balances keyword relevance and semantic similarity.

This project demonstrated the potential of combining traditional information retrieval algorithms with modern NLP techniques, especially in complex and specialized domains like medical literature. The improvements in retrieval performance underscore the importance of continuous experimentation and optimization in the field of information retrieval.