Natural Language Processing : First Project

TripAdvisor Recommendation 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 when 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 recommendation system relying only on user's reviews.

Given a review or a set of reviews of a specific place as a request, your recommendation system

will propose the most similar place present in the corpus.

Expectation

Your goal is to develop an original recommendation system for TripAdvisor Reviews. In order to do that, you are allowed to use any kind of pre-treatment and manipulate the vocabulary of the reviews. You can use pre-trained machine learning models or learn your own model. You can mix different approaches, but you aren't allowed to use direct supervised learning (for a given query, learning to predict the best place).

First you need to implement a bm25 baseline on this specific dataset. You can use the Rank-BM25 python library https://pypi.org/project/rank-bm25/.

Secondly you need to propose a model capable of doing better than BM25. (you can mix BM25 with your model/modifications)

In order to evaluate BM25 and your approach, you need to follow this given protocol:

Download the dataset TripAdvisor Hotel Review from Kaggle :

https://www.kaggle.com/datasets/joebeachcapital/hotel-reviews/data

On the table Reviews.csv, keep only the reviews where ratings are composed strictly with this aspects :

“service”, “cleanliness”, “overall”, “value”, “location”, “sleep quality”, “rooms” (not more and not less in order to compare places accurately).

You must concatenate reviews from the same place based on attribute “offering_id”. The rating of a place is just the average of all the reviews ratings on each aspect.

So, in order to evaluate BM25 or your approach, for a given query place, your model return the most related place according to reviews. The difference of ratings between the query and the returned place defines our score. Use a Mean Square Error (MSE) on all aspects between query and place returned. Compute the MSE for all places and average all the results in order to obtain a unique score. In this context for your proposition, beating BM25 means achieving a lower MSE than BM25.

Warning : rating aspects are only used for evaluation. Your model is blind to any rating and rely only on reviews.

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 1 december 2024.**

Please share your colab notebook with me:

Christophe.rodrigues.bento@gmail.com