Natural Language Processing:

First Project Information Retrieval _ Challenge Beating BM25

GROUP : RUNAVOT Paul, MONTOYA Adèle _ ESILV DIA02 Année 5

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).

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.

L’algorithme BM25 est une version améliorée du TF-IDF prenant compte de la longueur des documents lors du classement. Cet algorithme étant déjà très efficace sur le corpus, il est compliqué d’améliorer ses performances.

Parmi les différentes approches testées (différent style de tokenisation, modification de BM25, vectorisation des mots), une seule s’est réellement montrée convaincante et a eu des performances supérieures à BM25.

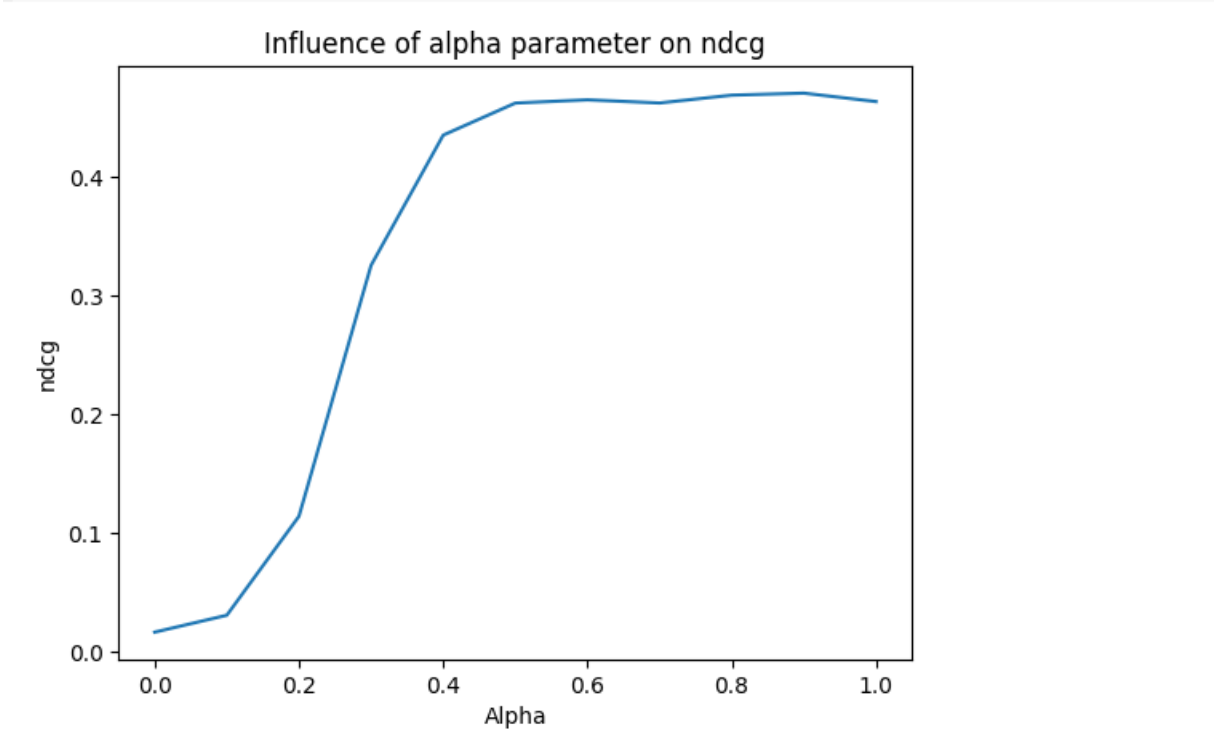
Cette approche est inspirée d’un article scientifique [1] démontrant l’efficacité de l’association du modèle BERT et de l’algorithme dans certains cas. Les résultats des deux algorithmes sont combinées à l’aide de la formule suivante :

$$s(p) = \alpha \hat{s}_{BM25}(p) + (1 - \alpha) s_{BERT}(p)$$

Cette formule est utilisée pour calculer le score final s(p) d'un passage p en interpolant deux scores différents : le score normalisé donné par le modèle BM25 sBM25(p) et le score donné par un modèle BERT sBERT(p).

Il a donc fallu implémenter BERT, convertir le corpus et les requêtes en vecteurs grâce au word embedding et créer la fonction de calcul de score qui est une similarité cosinus.

Ci-dessous, le ndcg sur l'entièreté du corpus en fonction du paramètre alpha.



Lorsque alpha est à 0, cela signifie que seuls les scores BERT sont utilisés pour le classement. À mesure qu'alpha augmente, la contribution des scores BM25 devient plus importante.
On observe que le NDCG augmente avec alpha jusqu'à un certain point, indiquant qu'introduire les scores BM25 améliore la qualité des résultats de classement par rapport à l'utilisation exclusive de BERT.

Cependant, à un certain niveau d'alpha (autour de 0.4 à 0.5), l'augmentation du NDCG ralentit et semble atteindre un plateau, ce qui suggère qu'au-delà de ce point, l'ajout de davantage de poids au score BM25 n'améliore pas significativement la qualité du classement.
Ce plateau peut indiquer qu'un équilibre optimal entre les scores BERT et BM25 est atteint pour le NDCG. Le NDCG maximum est atteint pour alpha = 0.9 et est de 0.47.

1. Conclusion

Lien de notre Google Collab :
https://colab.research.google.com/drive/14H_Q5CsSoXo7Vflo1omwPSgdd3IxEMJ7?usp=sharing

Niveau performance sur le premier code fourni (BM25) :

- Sur l'ensemble des documents (nb_docs=3192), on a un socre ndcg de 0.46

Niveau de performance suite aux modifications :

- Sur l'ensemble des documents (nb_docs=3192), on a un socre ndcg de 0.47

Notre code introduit donc une légère avancée par rapport au système fourni qui se basait uniquement sur le modèle BM25 pour le classement des documents. En intégrant les embeddings BERT, nous avons enrichi le processus de récupération d'informations en y ajoutant une dimension de contexte. Les embeddings BERT offrent des représentations vectorielles qui capturent les nuances sémantiques et les relations entre les mots, permettant ainsi d'aller au-delà de la simple fréquence des termes qu'on avait avec BM25.
Ce calcul est désormais fondé sur une combinaison linéaire des scores BM25 et des scores issus de BERT, avec le paramètre alpha servant à ajuster finement leur contribution relative. Ce paramètre est crucial car il permet de trouver l'équilibre optimal entre la précision de la correspondance des termes de BM25 et la compréhension du langage naturel apportée par BERT.

En conclusion, notre méthode vise à exploiter à la fois la fiabilité de BM25 pour la correspondance exacte des termes et l'aptitude de BERT à évaluer la pertinence dans un contexte donné. Cette combinaison offre une approche de classement des documents plus nuancée et permet une meilleure performance.

REFERENCES

[1] Shuai Wang, Shengyao Zhuang, and Guido Zuccon. 2021. BERT-based Dense Retrievers Require Interpolation with BM25 for Effective Passage Retrieval. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval (ICTIR '21)*.