Eléonor KIOULOU **DIA2 - A5**
Khadija MOKHTARI **15/11/2023**

PROJET NLP

Rapport sur le projet de Traitement Automatique du Langage Naturel (TALN) - Première Partie du Projet d'Information Retrieval Challenge : objectif meilleur résultat que le BM25

Introduction

Le projet consiste à développer un système original de recherche d'informations sur le corpus NFCorpus, un corpus médical composé d'extraits d'articles de PubMed. Le modèle de recherche de base utilisé est le BM25, qui est une amélioration populaire de TF-IDF. L'objectif principal est de surpasser les performances du BM25 en utilisant des méthodes et des approches novatrices.

Approches Techniques

1. Utilisation de BERT

- Nous avons utilisé le modèle BERT (Bidirectional Encoder Representations from Transformers) pour générer des embeddings pour les documents et les requêtes.
- Les embeddings sont calculés en moyennant les embeddings des jetons.
- Le modèle BERT est pré-entraîné sur le modèle 'distilbert-base-uncased' pour garantir des résultats de haute qualité.

2. Calcul de Similarité:

- La similarité cosinus a été utilisée pour mesurer la similitude entre les embeddings des documents et des requêtes.
- Nous avons calculé la similarité pour toutes les combinaisons possibles de requêtes et de documents.

3. Classement des Résultats

- Les résultats ont été classés en fonction des scores de similarité, et les cinq meilleurs résultats ont été sélectionnés.

4. Calcul de NDCG@5

- NDCG@5 (Normalized Discounted Cumulative Gain at 5) a été utilisé comme métrique d'évaluation.
- Nous avons comparé les résultats obtenus avec notre modèle par rapport à ceux du BM25 en utilisant la métrique NDCG@5.

Résultats Obtenus

Les résultats obtenus montrent que notre modèle surpasse le BM25 en termes de NDCG@5. Cela indique que notre modèle améliore la pertinence des documents récupérés par rapport au BM25.

Voici notre résultat final : `NDCG@5: 0.869957234097763`

Le score du BM25 est d'environ 0.81. Notre résultat est donc meilleur que celui du BM25.

Difficultés Rencontrées

- **Sélection de Plage de Documents**

Une des difficultés était de sélectionner une plage spécifique de documents, mais cela a été résolu en itérant à travers les documents dans l'ordre.

- **Optimisation des Paramètres**

Nous avons rencontré des difficultés lors de l'optimisation des paramètres du modèle BERT, mais des résultats significatifs ont été obtenus après quelques essais.

- **Trouver la "relevance" associée**

Après avoir sélectionné nos 5 meilleurs scores grâce au calcul de la similarité, une des difficultés était de trouver la relevance associée à ces scores pour ensuite pouvoir déterminer le NDCG@5.

Conclusion

En conclusion, notre approche utilisant le BERT pour générer des embeddings de documents et de requêtes a montré une assez bonne amélioration par rapport au BM25, comme en témoigne la métrique NDCG@5. Ce projet a fourni une opportunité précieuse pour explorer des méthodes avancées de recherche d'informations dans des domaines médicaux complexes. Les résultats obtenus démontrent le potentiel des modèles de langage profond dans le domaine de l'information médicale.

Liens Colab Notebook
https://colab.research.google.com/drive/1jpV883jGbRenQyYeAcc62kSk1lEfcG2e#scrollTo=XzllVjitik8f