DEVOIR DE WEB SÉMANTIQUE
MASTER INFORMATIQUE INTERNET DONNÉES ET CONNAISSANCES (IDC)

Réaliser par:
BAH Mamadou Alpha
&
REKIK Manel

Introduction:

Ce devoir consiste à optimiser un moteur de recherche en rajoutant une couche de **word2vec** pour enrichir la requête d'un utilisateur. Il est subdivisé en deux (2) grands parties dont la première qui implémente word2vec pour trouver tous les mots proches de la requête contenus dans le fichier d'apprentissage et la deuxième qui construit le moteur de recherche avec **lucene**.

Première partie:

Nous avons implémenté la classe **DocumentIndexer** dans laquelle nous avons défini la méthode qui indexe les documents issus du fichier LA_TRANSITION_ECOLOGIQUE.csv et la méthode qui effectue la recherche en fonction de la requête fournie. Les documents sont récupérés et traités dans la classe **MyFileReader** avant d'être utilisés par la méthode qui indexe. Enfin, une fois que les documents sont indexés, nous pouvons alors effectuer des recherches.

Deuxième partie:

Dans cette partie, nous avons créé la classe **Traitement** pour générer les tokens issus de sample.txt. Ensuite nous avons construit le modèle word2vec sur les tokens avec quelques paramètres et enfin pour avoir les mots proches d'une requête, nous la passons à la méthode **wordsNearest** du modèle. Nous avons choisi la colonne 11 du fichier LA_TRANSITION_ECOLOGIQUE.csv comme elle contient plus de données pour l'apprentissage de word2vec.

Mise en relation des parties:

Une fois que le modèle nous fournit les mots proches d'une requête, nous reformulons la requête en la rajoutant le résultat des mots proches (requête + mots proches). Ensuite, nous faisons la recherche avec cette requête et le moteur nous renvoie la réponse que nous enregistrons dans le fichier resultats_recherche.txt

Ressource :

Nous avons travaillé sur les données du fichier LA_TRANSITION_ECOLOGIQUE au format csv. Nous avons dû le supprimer ainsi que les autres dans le projet car ils prennent trop d'espace mémoire. Avec eux la taille compressée du projet est supérieure à la taille requise sur webmail et github. Pour tester le projet, il faudra télécharger ce fichier et créer les fichiers sample.txt, resultats_recherche.txt et vectorRes.txt dans le repertoire du projet.