Session Été 2020 – IFT3225
Examen N° 04
Date limite : **09 Août 2020 à mi-nuit sur Studium**
Travail individuel

Important !
Vous **devez** indiquer vos sources en entête des fichiers concernés. De même, toute récupération de code doit être indiquée. Si vous omettez d'indiquer une source de code et que nous le découvrons lors de la correction, vous serez dans une situation de plagiat.

Tâche – Moteur de recherche :

Vous devez créer un prototype de moteur de recherche (par exemple l'interface de Google).

Vous avez une liste de documents à indexer.

1ère étape (50%): indexation en utilisant la technique vue en cours TF-IDF. Vous pouvez utiliser un langage de programmation de votre choix (Python, Java, etc.) pour créer l'index (matrice de terme/document) des documents.

Il n'y a aucune restriction par rapport à la sauvegarde de l'index (JSON, texte, XML, SQL, …)

Il est permis d'utiliser des APIs déjà existantes comme StanfordNLP, word2vec, CoreNLP.

2ème étape (40%): Calcul de similarité entre le vecteur de la requête et celui des documents avec la mesure Cosinus. (Vous devez implémenter cette métrique)

3ème étape (10%): affichage des documents triés par ordre de pertinence par rapport à la requête.

Vous pouvez utiliser n'importe quel langage de programmation **vu en cours** pour créer l'interface utilisateur ainsi que la réalisation de la 2ème et 3ème étape.

Remise :

Vous devez déployer votre travail sur les serveurs du DIRO, ainsi que sur Studium en un fichier compressé .zip.

Vous devez rendre un fichier readme décrivant les détails de processus d'indexation et de recherche utilisés.