PROJET DE SYNTHÈSE : BIG DATA AVEC SPARK ET BD VECTORIELLES

DURÉE : 2 SEMAINES À COMPTER DE LA DATE DE REMISE DU PROJET

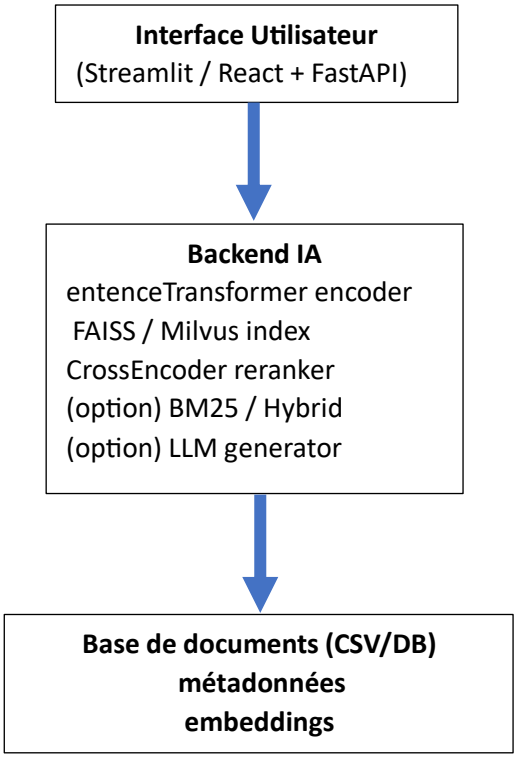
Les moteurs de recherche modernes reposent sur des bases de données vectorielles. Elles permettent d'effectuer des recherches sémantiques c'est-à-dire de trouver les documents les plus proches *en sens* d'une requête, grâce à des représentations denses qui sont les *embeddings*.

Ce projet prolonge le TP en classe sur FAISS. L'objectif est de construire une application de recherche sémantique interactive, capable de répondre à des requêtes en langage naturel en retrouvant les documents les plus pertinents, à partir d'un corpus textuel, au choix(scientifique, légal, médical, ou institutionnel).

Choisir un de ces domaines au choix et prendre jeu de données s'y référant sur kaggle .

- Recherche d'articles scientifiques (PubMed, arXiv, etc.)
- Recherche de jurisprudence (jugements, lois, textes réglementaires)
- FAQ médicale (OMS, santé publique)
- Questions pédagogiques (cours d'IA, documents d'école)
- Moteur d'exploration d'actualités (BBC, Reuters)

L'architecture visée par ce travail est comme ci-dessous présentée :



Etapes Clés

Étape 1 : Construction du corpus

- Chaque groupe choisit un domaine et collecte ou télécharger un jeu de données dont la taille est comprise dans l'intervalle 500–2000 documents (articles, textes, etc.).
- Nettoyage du texte.
- **Sauvegarde du corpus (docs.csv).**

Étape 2 : Vectorisation et Indexation

- Générer les embeddings avec sentence-transformers/all-MiniLM-L6-v2.
- Stocker dans FAISS (IndexFlatIP ou IndexIVFPQ).
- **Sauvegarder l'index (.faiss).**

Étape 3 : API Backend

- Créer une API REST avec **FastAPI** :
  - *POST /query* qui retourne les top documents
  - *GET /docs/{id}* qui retourne le texte d'un document
- Ajouter le re-ranking CrossEncoder.

Étape 4 : Interface Web

Vous avez deux options pour cette étape :

- Option 1 : **Streamlit**
- Option 2 : **Frontend React + FastAPI backend**

Étape 5 : Évaluation et visualisation

- Ajouter un tableau Streamlit de métriques (Recall@10, MRR@10, latence moyenne).
- Visualiser les embeddings (indice : UMAP ou t-SNE).

Étape 6 : Extension libre

Cette dernière partie est votre espace de liberté totale.
Je veux votre signature, votre touche personnelle.

Je vous connais :

- *vous êtes talentueux,*
- *vous êtes capables d'aller loin,*
- *et vous êtes prêts à surprendre.*

Alors ,

Impressionnez-vous vous-mêmes.

Parce que moi, je sais déjà que vous pouvez créer quelque chose d'exceptionnel !!!

Critères d'évaluation

Critère	**Points**
Qualité du pipeline IA	4
Performance du moteur (Recall/MRR/latence)	3
Qualité de l'interface utilisateur	3
Clarté du code et documentation	3
Extension ou innovation	4
Petite vidéo de démo	3
Total	**20 points**