TP2 - Développement d'index

Durée : 3 heures

Objectif

Créer différents types d'index à partir d'un jeu de données de produits e-commerce pour préparer la construction d'un moteur de recherche.

Données d'entrée

Fichier JSONL contenant 150 documents avec pour chaque ligne :

- URL
- Titre
- Description
- Features du produit
- Reviews sur le produit
- Liens

Étapes guidées

1. Lecture et traitement de l'url
   - Parser le fichier JSONL
   - Extraire les informations des URLs :
     - ID produit
     - Variante (si présente)

2. Filtrage des documents
   - Créer un index inversé pour le titre :
     - Utiliser la tokenization par espace
     - Supprimer les stopwords et la ponctuation
     - Pour chaque token, stocker uniquement la liste des IDs des documents
   - Créer le même type d'index pour la description

3. Index des reviews
   - Créer un index pour les reviews :
     - Nombre total de reviews
     - Note moyenne
     - Dernière note
   - Cet index ne rassemble pas des informations textuelles, nous les utiliserons pour faire remonter les documents avec les meilleures notes. Il ne doit donc pas être inversé.

4. Index des features
   - Traiter les features de marque et d'origine du produit comme un champ textuel :
     - Tokenization par espace
     - Un index inversé par feature (marque, origine, etc.)
     - Stocker les IDs des documents pour chaquefeature

5. Index de position
   - Pour les champs titre et description, créez une fonction qui génère un index inversé contenant l'information de document et de positions dans chaque document.

6. Tests et optimisation
   - Gérer les erreurs courantes

7. Documenter
   - Rédiger un README détaillant :

    - La structure des index
    - Les choix techniques
    - Les features supplémentaires implémentées (en plus de marque et origine)
-
-

Rappels de programmation:
- Une fonction ne fait qu'une action, si vous avec envie de nommer votre fonction do_something_and_do_something_else -> alors il vous faut deux fonctions
- Le nom d'une fonction commence toujours par un verbe d'action
- Les noms des fonctions/variables doivent être écrit en anglais, tout comme la documentation.

Livrable

- Script Python contenant :
    - Fonctions de création des index
    - Fonctions de sauvegarde/chargement
- README.md détaillant :
    - La structure de chaque index
    - Les choix d'implémentation
    - Les fonctionnalités bonus ajoutées
    - Des exemples d'utilisation
    - Comment lancer votre code

Votre code devra produire cette liste d'index:
- Titre (document + position)
- Description (document + position)
- Un par feature
- Reviews

Critères d'évaluation

- Respect des consignes
- Propreté et lisibilité du code