TP: Indexation

Partie 1:

1. Créer un petit corpus de 20 documents textes.
2. Faire le pré-traitement du corpus
3. Construire l'index inversé : on se contente du numéro du document.

Partie 2 :

1. Implémenter les 3 modules : compression, maintenance, parallélisation.
2. Mesurer :
    - Temps d'indexation avant/après parallélisation.
    - Taille mémoire avant/après compression.
    - Discuter les compromis espace/temps.

Partie3 :

1. Créer un index Elasticsearch avec un analyzer personnalisé pour le corpus.
2. Visualiser et commenter :
    - Le résultat du `_analyze`
    - Le contenu de `_segments`
    - Les statistiques de `_stats`
3. Mesurer :
    - Temps d'indexation avec 1, 2, 4 shards
    - Taille disque avant/après `_forcemerge`
4. Comparer les résultats avec l'indexation manuelle en Python.
5. Discuter comment Elasticsearch gère la compression, la maintenance et la parallélisation de manière plus efficace que votre implémentation manuelle ?