TP 2
Construire un index minimal

Présentation du sujet

Écrivez un index non positionnel.

Vous partirez de la liste d'urls fournie au format json qui a été généré depuis un crawler.
Il faudra en extraire les titres, les tokenizer et construire un index web en suivant l'algorithme présenté en cours.

Avant de construire l'index web, sortir des statistiques sur les documents telles que:
- Le nombre de documents
- Le nombre de tokens
- La moyenne des tokens par documents
- Et toutes les informations qui vous paraitraient intéressantes à avoir pour la construction d'un index web

Une fois terminé, votre programme écrira dans un fichier **title.non_pos_index.json** l'index ainsi créé et dans un fichier **metadata.json** les informations statistiques que vous aurez calculé.

En bonus

- Appliquer un data processing plus poussé, par exemple en appliquant un stemmer.
  Vous pouvez utiliser cette librairie pour le stemming: https://www.nltk.org/api/nltk.stem.html
  Elle vous propose plusieurs stemmers, vous pouvez les tester sur quelques phrases avant de choisir celui qui vous convient le mieux.
  Ce data processing plus poussé ne devra pas écraser l'index précédent mais créer un nouvel index, **mon_stemmer.title.non_pos_index.json**
- Créer un index positionnel sur ces memes données. Il ne devra pas écraser le ou les précédents index crées mais en créer un nouveau **title.pos_index.json**
- Vous pouvez aussi vous amuser à créer un index pour d'autres informations contenues dans l'HTML (content, h1 etc.)

Ce qui est demandé

Un dossier avec votre index écrit en **python**.
Le code devra s'executer dans un fichier **main.py** à la racine du projet.
Vos fonctions publiques devront être testées dans un fichier **tests.py**.

L'explication du code et de son execution devra être décrite dans un fichier **README.md** à la racine du projet. Ce fichier devra aussi comprendre le nom des contributeurs du projet.

Si vous êtes familier avec GitHub, vous pouvez aussi m'envoyer le lien vers votre repository.

Les librairies dont vous aurez besoin

- Pour lire les fichiers html: https://www.crummy.com/software/BeautifulSoup/bs4/doc/
- Pour sortir des statistiques des documents: https://pandas.pydata.org/