TP Noté : Web Scraping, Prétraitement et Analyse sur des Données de PubMed (Durée : 1h30)

Objectif

L'objectif de ce TP est d'extraire les informations liées à des articles de PubMed, de sauvegarder cette extraction dans un dataset, puis de les prétraiter et analyser. Le TP est à réaliser sur un notebook où vous commenterez votre code. Le rendu se fera sur GitHub.

Web Scraping

Recherchez des articles scientifiques sur PubMed en utilisant un mot-clé spécifique et extrayez les informations suivantes :

- **Titre de l'article.**
- **Résumé** (abstract).
- **Auteurs.**

Adaptez votre script pour itérer sur plusieurs pages afin de récupérer davantage d'articles. Stockez ensuite ces informations dans un dataframe.

Analyse Exploratoire et Prétraitement

Une fois les informations dans un dataframe, réalisez les tâches suivantes :

- Appliquez une pipeline de prétraitement (comme vu dans les TD) sur les résumés et les titres des articles. Les pipelines peuvent être différentes pour les deux champs.
- Après avoir appliqué ces pipelines :
  - Créez un nuage de mots à partir des résumés.
  - Calculez la fréquence des mots présents dans les résumés et réalisez une visualisation pour représenter cette distribution.