Rapport sur le Scraping de Google Scholar

Introduction :

Le projet vise à effectuer une analyse de données à partir des profils académiques des chercheurs de Université Cadi Aayad disponibles sur Google Scholar.

Étapes du projet :

1.Scraping de Données :

1. Choix de la Méthode de Scraping :

Le scraping de données est la première étape cruciale du projet. Nous avons choisi d'utiliser les bibliothèques Python telles que BeautifulSoup et Selenium pour extraire les informations des profils académiques. Cette décision a été motivée par la structure HTML des pages de profils académiques, qui nécessite un traitement sophistiqué.

2.Librairies utilisées :

- Requests :Utilisé pour effectuer des requêtes HTTP et récupérer le contenu des pages web.
- Beautiful Soup : Utilisé pour extraire des informations structurées à partir du code HTML des pages web.
- Selenium : Employé pour automatiser l'interaction avec des pages web, particulièrement utile pour les pages avec un contenu dynamique généré par JavaScript.
- Pandas : Utilisé pour la manipulation et l'analyse de données, notamment pour stocker les résultats du scraping dans des DataFrames.
- CSV (pandas) : Utilisé pour lire et écrire des données au format CSV.
- Re(expressions régulières) : Offre des outils puissants pour identifier et extraire des motifs spécifiques dans les données textuelles, facilitant ainsi le processus de récupération d'informations structurées à partir de pages web lors du scraping.

2.Le processus de scrapping

2.1. Scraping des Résultats de Recherche

Nous avons initié le processus en effectuant une recherche sur Google Scholar avec le terme "university caddy ayyad".

- Extraction des informations de la première page : ID, lien, titre, auteurs, citations, versions, liens associés, PDF (si disponible).
- Extension du scraping à toutes les pages jusqu'à 600 articles.
- Stockage des données dans "articles.xlsx".

=>Le code de cette partie est dans le fichier articles.ipynb

2.2.Scraping des Auteurs et de Leurs Articles

En utilisant le fichier obtenu précédemment, nous avons isolé le nom et le lien de chaque auteur dans un nouveau fichier authors_links.

- Bouclage sur chaque lien et Extraction des informations des auteurs : nom, poste, e-mail, départements, h-index, i-index, citations. Stockage dans Author_data.csv
- Extraction des détails de chaque article de l'auteur : citations, volume, date de publication, coauteurs, revue, pages, numéro.
- Stockage des données dans "articles_info.csv".

2.3.Scraping des Coauteurs

- Collecte des données sur les coauteurs.
- Structuration dans "COAUTHORS.csv" : auteur principal, nom, titre, lien, e-mail, miniature.

=>Le code de ces deux parties est dans le fichier authors.ipynb

2. Stockage des Données:

Les données extraites sont stockées dans des fichiers CSV et XLSX distincts pour une analyse ultérieure.

3. Application des Méthodes d'Analyse de Données :

3.1.ACP

Ensemble de Données :

Chaque entrée du jeu de données représente un enseignant-chercheur.

Variables :

"Nombre de Citations" : Total des citations attribuées aux publications du professeur.
"Indice h" : Mesure la productivité académique et l'impact des publications.
"Indice 10H" : Variant de l'indice h considérant les publications avec au moins 10 citations.
"Nombre d'Articles" : Total des articles publiés par le professeur.

Objectifs

Comment les professeurs sont-ils positionnés les uns par rapport aux autres dans l'espace défini par ces variables ?

Peut-on regrouper les professeurs en clusters présentant des caractéristiques similaires dans ces variables ?

Quelles sont les variables clés qui influent le plus sur la variabilité observée parmi les professeurs ?

Résultats

Une première essaie et dans le fichier acp.ipynb .

3.1. ANOVA à mesures répétées

Ensemble de Données

- Ensemble de Données
- Les colonnes comprennent le nombre de citations et le nombre d'articles pour chaque domaine.

Objectifs

Y a-t-il une différence significative entre les moyennes des citations et des articles pour chaque domaine au fil des années ?

Y a-t-il une évolution significative du nombre de citations et d'articles dans chaque domaine au fil des années ?

Y a-t-il une interaction significative entre le domaine et l'année, indiquant que l'impact temporel n'est pas uniforme entre les domaines ?

Résultats

En cours .

Tâche à faire

1. Propositions d'Autres Méthodes d'Analyse.
2. Structuration du Code.
3. Ajout de Commentaires.
4. Interprétation des Résultats.