Faculté des Sciences et Techniques
Master Intelligence Artificielle et Ingénierie Informatique

Rapport de Projet: Analyse de Données par Web Scraping des Profils de l'Université UCA

Student:
Hamza Bousalih
Anas Bouzina

30 décembre 2023

Table des matières

Liste des figures

1 Introduction

Dans le cadre de ce projet, notre objectif principal était de mener une collecte de données approfondie grâce à une recherche sur le Web dans les dossiers de l'Université Cadi Ayyad (UCA). L'Université Cadi Ayyad, en tant qu'établissement d'enseignement supérieur renommé, abrite une population diversifiée d'universitaires, de chercheurs et d'étudiants.
Cette méthode de collecte de données vise à rassembler des informations diverses et pertinentes, permettant ainsi une analyse approfondie des dynamiques au sein de l'université.
L'importance de cette initiative réside dans sa capacité à obtenir des données riches et à jour, permettant ainsi une compréhension détaillée de la composition académique de l'UCA.
Ces données seront importantes pour les analyses ultérieures afin de faire la lumière sur des aspects tels que la répartition des domaines de recherche, la collaboration interdisciplinaire et le profil de la recherche au sein de l'université.
Ce projet d'exploration du Web sur les archives de l'Université Cadi Ayyad revêt une importance significative du point de vue de l'exploration approfondie des données académiques, ouvrant la voie à une meilleure compréhension des dynamiques académiques et de leur signification.

2 Objectifs du Projet

2.1 Collecte de Données :

L'objectif principal de ce projet est d'effectuer une collecte complète de données à l'aide de techniques avancées de web scraping.
En nous concentrant sur les profils individuels de l'Université Cadi Ayyad (UCA), nous visons à extraire des informations précises et complètes.
Ce processus comprend la collecte de données telles que la formation universitaire, les domaines d'études, les publications et d'autres éléments essentiels d'un profil académique.

2.2 Analyse de Données :

Une fois les données collectées, l'accent sera mis sur l'application de méthodes d'analyse des données pour découvrir des tendances significatives, des corrélations pertinentes et des informations clés dans les enregistrements extraits.Rephrase
Cette phase d'analyse s'articulera autour d'une compréhension approfondie de la dynamique universitaire, mettant en lumière des aspects tels que la répartition des domaines d'expertise, les collaborations potentielles entre chercheurs et les indicateurs liés aux acquis d'apprentissage.

3 Méthodologie

3.1 Choix des Outils :

Pour mener à bien le processus de web scraping, nous avons opté pour une combinaison d'outils robustes et largement utilisés dans le domaine. Les langages de programmation principaux pour ce projet sont Python, appuyé par les bibliothèques requests et BeautifulSoup.
Pour l'analyse de données, d'autres bibliothèques spécialisées telles que pandas seront utilisées. Pandas offre des structures de données flexibles et des outils de manipulation efficaces pour explorer, nettoyer et analyser les données extraites. Matplotlib et Seaborn seront également

déployés pour créer des visualisations informatives permettant une interprétation visuelle des résultats de l'analyse. Cette approche holistique, combinant le scraping avec une analyse approfondie, vise à fournir des résultats significatifs pour les objectifs du projet.

3.2 Éthique du Scraping :

La réalisation de ce projet a été guidée par des principes éthiques rigoureux afin de garantir le respect des règles et politiques de l'Université Cadi Ayyad. Nous avons pris des précautions pour éviter toute perturbation indue des serveurs, en espaçant les requêtes pour minimiser l'impact sur les performances du site. De plus, nous avons examiné attentivement les conditions d'utilisation du site web de l'université pour nous assurer que notre activité de web scraping était en conformité avec leurs directives.
Il est important de souligner que notre intention est uniquement de collecter des données publiques disponibles sur les profils universitaires, en évitant toute intrusion dans des données sensibles ou privées. Nous sommes ouverts à coopérer avec l'université pour répondre à toute préoccupation éthique et assurer la transparence de nos méthodes.

3.3 Analyse de Données :

La phase d'analyse de données impliquera l'utilisation de méthodes statistiques et visuelles pour extraire des insights significatifs. Les outils tels que pandas et matplotlib en Python seront déployés pour explorer les tendances, les corrélations, et les schémas au sein des données extraites. Cette approche analytique contribuera à dégager une compréhension approfondie des profils universitaires de l'UCA, ouvrant ainsi la voie à des interprétations utiles pour les objectifs du projet

4 Google Scholar

Google Scholar est un moteur de recherche spécialisé dans la recherche académique. Il indexe une vaste gamme de documents universitaires, y compris des articles de revues, des thèses, des livres, des résumés de conférences, et d'autres travaux académiques. Conçu pour faciliter l'accès à la littérature universitaire, Google Scholar permet aux chercheurs et aux étudiants de trouver des informations pertinentes pour leurs travaux, offrant ainsi une source précieuse de données pour des analyses approfondies. Dans le cadre de notre projet, nous avons utilisé Google Scholar pour collecter des données spécifiques sur les profils académiques des membres de l'Université UCA, ce qui a contribué de manière significative à l'enrichissement de notre ensemble de données et à la profondeur de notre analyse.

5 Outils Utilisés

5.1 Web Scraping

1. **Python :** En raison de sa polyvalence, de sa simplicité et de sa communauté active, Python a été choisi comme le langage principal pour le développement du web scraping.

2. **Librairies utilisées :**

    - **requests :** Utilisé pour effectuer des requêtes HTTP.

- **BeautifulSoup :** Utilisé pour le parsing HTML.
- **csv :** Utilisé pour la manipulation de fichiers CSV.
- **json :** Utilisé pour la manipulation de fichiers JSON.

5.2 Transformation de données

1. **Python :** De plus, Python est utilisé pour transformer les données extraites en une nouvelle forme adaptée à l'analyse ultérieure.

2. **Librairies utilisées :**

   - **pandas :** Utilisé pour la manipulation et l'analyse des données tabulaires.
   - **json :** Utilisé pour la manipulation de fichiers JSON.
   - **copy :** Utilisé pour la copie d'objets et de données.
   - **re :** Utilisé pour les expressions régulières, facilitant la recherche et la manipulation de motifs dans les chaînes de caractères.

5.3 Analyse de Données

1. **Python :** Python a été choisi comme le langage principal pour l'analyse des données. Il est utilisé pour appliquer différentes méthodes d'analyse aux données collectées lors du web scraping.

2. **Librairies utilisées :**

   - **numpy :** Utilisé pour le support de tableaux multidimensionnels et les opérations mathématiques.
   - **matplotlib :** Employé pour la création de graphiques et de visualisations.
   - **fanalysis :** Utilisé pour l'analyse factorielle des données.
   - **seaborn :** Utilisé pour la création de visualisations statistiques attrayantes.
   - **sklearn :** Une bibliothèque complète pour l'apprentissage automatique et l'exploration des données.
   - **scipy :** Utilisé pour les outils et algorithmes mathématiques avancés.

5.4 Environnements de Développement

1. **Jupyter :** Jupyter est un environnement interactif basé sur des cellules, favorisant l'exploration et l'analyse des données.

2. **VSCode :** Visual Studio Code (VSCode) offre un éditeur de code puissant avec un support intégré pour Python et des fonctionnalités avancées.

3. **PyCharm :** PyCharm est un environnement de développement intégré (IDE) robuste, offrant des outils avancés pour le développement Python.

5.5 Autres Outils

1. **ChatGPT :** Nous avons utilisé ChatGPT pour simplifier le processus de codage et la recherche, facilitant ainsi le développement du projet de manière efficace.

6 Web Scraping

6.1 Objectif

Le projet vise à présenter une méthodologie automatisée de collecte d'informations sur les chercheurs affiliés à l'Université Cadi Ayyad en utilisant le web scraping. L'objectif principal est d'obtenir des données telles que les profils des chercheurs, leurs domaines de spécialité, les statistiques de citation, et les détails de leurs articles de recherche à partir de Google Scholar.

6.2 Méthodologie

1. **Liste d'URLs de profils :**
   Le script commence par définir une liste d'URLs correspondant aux pages de profils des chercheurs de l'Université Cadi Ayyad sur Google Scholar.

2. **Récupération des liens de profils :**
   + Pour chaque URL de la liste, le script envoie une requête HTTP pour obtenir le contenu de la page.
   + Le contenu de chaque page est analysé avec BeautifulSoup pour extraire les liens vers les profils des chercheurs.

3. **Fonctions pour extraire les informations :**
   Le script définit plusieurs fonctions (par exemple, `getNameProfil`, `getUnevercity`, `getspicialty`, `getArticle`) pour extraire des informations spécifiques à partir du contenu HTML d'une page de profil.

4. **Extraction d'Informations Générales sur les Chercheurs :**
   Les informations générales telles que le nom, l'affiliation universitaire, et la spécialité sont extraites de chaque profil.

5. **Extraction des Articles de Recherche :**
   Les détails de chaque article de recherche, y compris le titre, les auteurs, la description, l'année de publication, et le nombre de citations, sont collectés.

6. **Extraction des Statistiques de Citation :**
   Les statistiques de citation globales et celles depuis 2018 sont extraites, notamment l'indice h, le nombre total de citations, et l'indice h_i10.

7. **Boucle principale de scrapping :**
   + Le script parcourt ensuite la liste de liens de profils.
   + Pour chaque lien, une nouvelle requête HTTP est envoyée pour récupérer le contenu de la page du profil du chercheur.

8. **Stockage des Informations :**
   + Les données extraites sont organisées dans des structures de données appropriées, telles que des listes et des dictionnaires, pour faciliter la manipulation et l'analyse ultérieures.

+ Les informations finales sont sauvegardées dans un fichier JSON nommé "uca.json" pour une référence future.

9. **Création d'une structure de données finale :**
   Les données extraites sont ensuite structurées dans un format JSON qui est écrit dans un fichier "uca.json".

10. **Répétition pour chaque profil :**
    Ce processus de scrapping est répété pour chaque chercheur dans la liste, populant ainsi le fichier JSON avec les informations de tous les chercheurs ciblés.

11. **Utilisation de modules externes :**
    Le script utilise des modules externes tels que `requests` pour effectuer des requêtes HTTP et `BeautifulSoup` pour analyser le contenu HTML.

La méthodologie de web scraping mise en œuvre dans ce projet a permis de collecter de manière efficace des informations clés sur les chercheurs affiliés à l'Université Cadi Ayyad. Ces données peuvent être utiles pour des analyses approfondies, des évaluations de la recherche universitaire, et d'autres applications liées à la recherche scientifique.

7 Transformation de données

Dans cette étape du projet, j'ai effectué la transformation des données extraites depuis leur format JSON d'origine vers des formats plus adaptés à nos besoins, notamment du format JSON vers CSV. Cette transformation a été essentielle pour structurer les données de manière plus compréhensible et faciliter leur analyse ultérieure. Les principales transformations réalisées comprennent :

7.1 Conversion du Format JSON vers CSV

Le processus de transformation des données a débuté par la conversion du format JSON résultant du *web scraping* en fichiers CSV. Cette conversion a permis une meilleure manipulation des données et une préparation plus efficace pour l'analyse.

7.2 Création de Fichiers Multiples

7.2.1 Profils et Détails

Un fichier a été créé pour les profils des membres de l'université avec les détails suivants :

- **Nom et Prénom**
- **ASS - Université**
- **Spécialité**
- **H Indice Global**
- **Citations Globales**
- **H Indice_I10 Global**

- **H Indice depuis 2018**
- **Citations depuis 2018**
- **H Indice_I10 depuis 2018**
- **Article** (Numéro d'article)
- **Faculté**

7.2.2 Articles

Un autre fichier a été généré spécifiquement pour les détails des articles avec les colonnes suivantes :

- **Titre de l'Article**
- **Année**
- **Description**
- **Revue**
- **Volume**
- **Numéro**
- **Pages**

7.2.3 Table de Contingence

Une table de contingence a été créée pour analyser les fréquences entre les facultés et les spécialités des professeurs. Cette table a facilité une compréhension approfondie des relations entre les différentes entités dans l'université.

7.3 Création de Nouvelles Colonnes

Plusieurs nouvelles colonnes ont été générées à partir des colonnes existantes pour améliorer la granularité des données et faciliter l'analyse.

7.3.1 Extraction des Détails des Articles

Les détails tels que la revue, le volume, le numéro et les pages ont été extraits de la colonne de description pour chaque article, permettant une analyse plus détaillée.

7.3.2 Création de la Colonne de Faculté

Une nouvelle colonne de "Faculté" a été créée à partir de la colonne "ASS - Université" existante. Cela a été fait pour normaliser les noms des unités, car certains professeurs utilisaient des noms non uniformes pour spécifier leur affiliation à l'université.

7.3.3 Regroupement des Spécialités

Les spécialités ont été regroupées pour réduire leur nombre, car certains professeurs utilisaient des termes différents ou des langues différentes pour décrire la même spécialité.

La phase de transformation des données a été cruciale pour préparer les informations extraites du *web scraping* à une analyse plus approfondie. En structurant les données dans des fichiers multiples et en créant de nouvelles colonnes pertinentes, nous avons optimisé notre ensemble de données pour tirer des insights significatifs dans les étapes suivantes de notre projet d'analyse de données.

8 Analyse de Données

8.1 Régression Linéaire

La régression linéaire est une méthode statistique utilisée pour modéliser la relation linéaire entre une variable dépendante et une ou plusieurs variables indépendantes.

8.1.1 Appliquer sur les données des profils

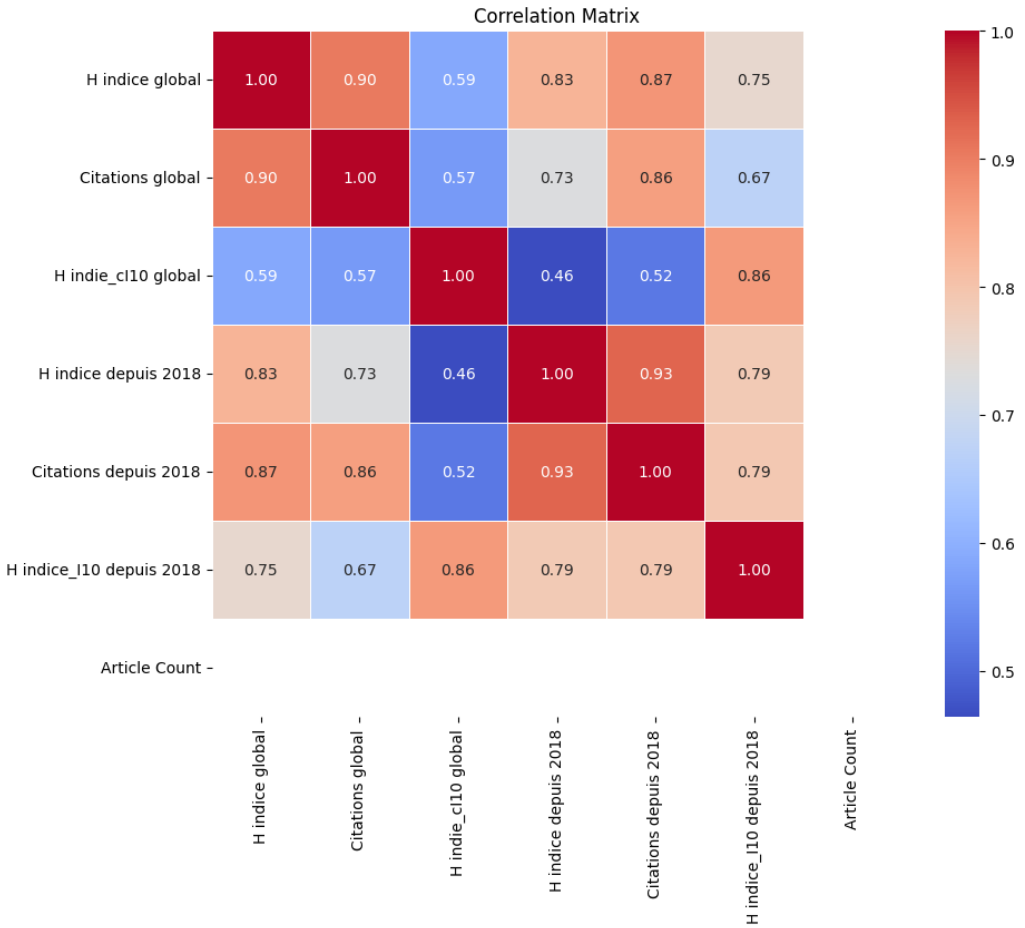

Figure 1: matrice de correcation des profils

Chaque entrée dans la matrice de corrélation représente le coefficient de corrélation entre deux variables. Le coefficient de corrélation est une mesure statistique qui quantifie le degré selon lequel deux variables évoluent l'une par rapport à l'autre. Il varie de -1 à 1, où:

- 1 indique une corrélation positive parfaite (quand une variable augmente, l'autre variable augmente également proportionnellement),
- -1 indique une corrélation négative parfaite (quand une variable augmente, l'autre variable diminue proportionnellement),
- 0 indique aucune corrélation linéaire.

Interprétation

- Après avoir examiné la matrice de corrélation, on peut conclure qu'il existe une relation linéaire entre la variable dépendante `Citations globales` et les variables indépendantes `H Indice global`, `Citations depuis 2018` et `H Indice depuis 2018`. Cela est indiqué par des coefficients de corrélation supérieurs à 0.7, s'approchant de 1.
- De plus, il existe également une forte corrélation entre les variables `Citations depuis 2018` et `H Indice depuis 2018` avec un coefficient de 0.93.
- Il existe une relation faible entre les variables `Citations depuis 2018` et `H indice_cl10 global` avec un coefficient de 0.52, ainsi qu'entre les variables `H Indice depuis 2018` et `H indice_cl10 global` avec un coefficient de 0.4.

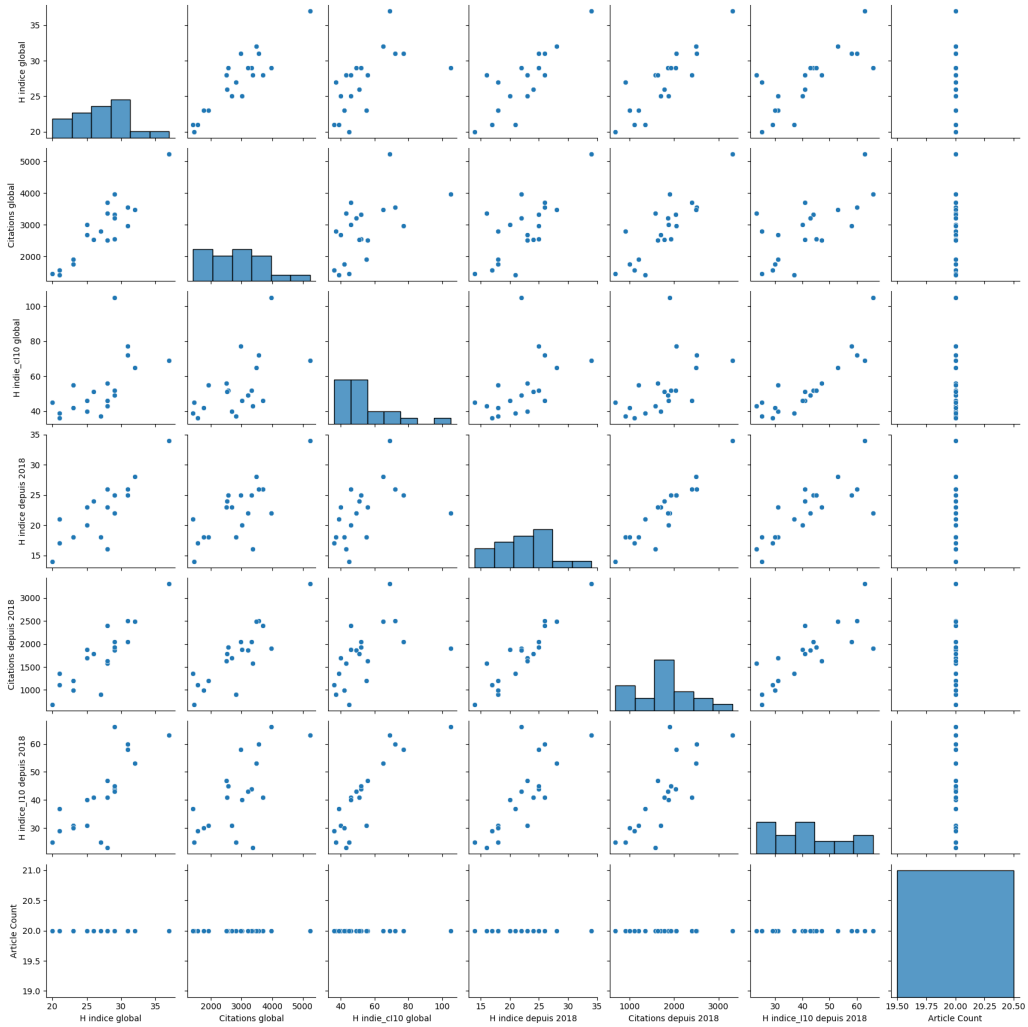

Figure 2: Graphique de la Matrice de Paires des Variables des Profils

Le Pair Plot montre des scatterplots de toutes les paires de variables, illustrant les relations et distributions dans un ensemble de données. La direction et la forme des points indiquent la nature de la relation:

- `Pente Positive` : À mesure qu'une variable augmente, l'autre a tendance à augmenter.
- `Pente Négative` : À mesure qu'une variable augmente, l'autre a tendance à diminuer.
- `Aucun Modèle Clair` : Il pourrait ne pas y avoir de relation linéaire forte.

Interprétation

- La densité et la dispersion des points dans les graphiques de dispersion indiquent la force de la relation entre les variables.
- Si les points se regroupent étroitement autour d'une ligne, cela suggère une corrélation forte.
  - Par exemple, entre `Citation globale` et `H Indice global`.

- Lorsque les points dans un graphique de dispersion sont dispersés sans tendance ou motif clair, cela suggère l'absence de relation linéaire. Les variables peuvent être faiblement corrélées ou avoir une relation non linéaire.
  - Par exemple, entre `H Indice_Cl10 global` avec `Citation globale` et `H Indice global`.

Interprétation du Coefficient de Détermination (R^2):

Le coefficient de détermination (R^2) est très proche de 1 (0,8586833570479674), ce qui signifie que votre modèle explique pratiquement toute la variance de la variable dépendante (Ladder score).

En d'autres termes, environ 85,86% de la variabilité de la variable dépendante est expliquée par les variables indépendantes (H indice depuis 2018, Citations depuis 2018, `H indice global`). Il s'agit d'un ajustement extrêmement précis.

8.1.2 Appliquer sur les données d'articles

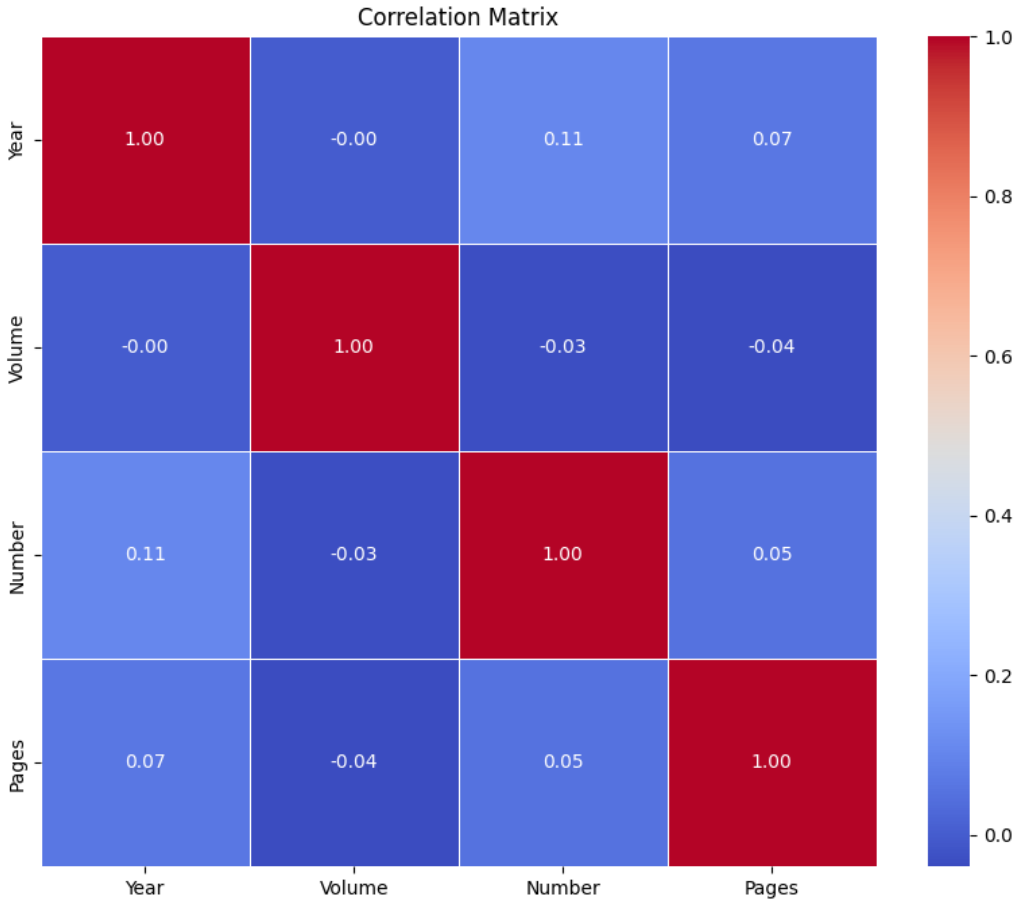

Figure 3: matrice de correcation des articles

Les coefficients de corrélation proches de 0 dans la matrice de corrélation suggèrent une faible corrélation linéaire entre les variables. Cela indique que les variables ne présentent pas de relation linéaire significative les unes avec les autres.

8.2 Analyse de la Variance (ANOVA)

L'ANOVA est une technique statistique permettant de comparer les moyennes de plusieurs groupes afin de déterminer s'il existe des différences significatives entre eux.

8.3 Analyse Factorielle Correspondance (AFC)

L'AFC est une méthode d'analyse des données catégorielles qui permet d'explorer les relations entre deux ensembles de variables.

8.4 Analyse Correspondence Multiple (ACM)

L'ACM est une extension de l'AFC qui permet de traiter plusieurs tableaux de données catégorielles simultanément.

8.5 Analyse en Composantes Principales (ACP)

L'ACP est une technique d'analyse multivariée qui permet de réduire la dimensionnalité des données en identifiant les principales composantes qui en expliquent la variance.

8.6 Analyse Factorielle Discriminante

L'analyse factorielle discriminante est utilisée pour déterminer les combinaisons linéaires de variables qui permettent de discriminer entre différents groupes prédéfinis.

8.7 Classification

La classification est une méthode d'apprentissage supervisée qui vise à attribuer des classes prédéfinies à de nouvelles observations en se basant sur un modèle préalablement construit.

9 Conclusion

Ce projet de web scraping et d'analyse de données a permis une collecte approfondie des profils académiques à l'Université Cadi Ayyad. Grâce à des techniques éthiques de scraping, les données extraites offrent une vision précieuse des domaines de recherche, des collaborations et des statistiques de citation. L'utilisation de Python et d'outils comme BeautifulSoup a facilité le processus. Les résultats fournissent une base pour des analyses plus poussées, contribuant ainsi à une compréhension approfondie de la dynamique académique au sein de l'institution.

10 Références

1. **Outils Utilisés:**

Python	https://www.python.org/
Requests	https://docs.python-requests.org/en/latest/
BeautifulSoup	https://www.crummy.com/software/BeautifulSoup/bs4/doc/
Pandas	https://pandas.pydata.org/
JSON	https://docs.python.org/3/library/json.html
Matplotlib	https://matplotlib.org/
Fanalysis	https://pypi.org/project/fanalysis/
Seaborn	https://seaborn.pydata.org/
Scikit-learn	https://scikit-learn.org/
SciPy	https://www.scipy.org/

2. **Environnements de Développement:**

Jupyter	https://jupyter.org/
VSCode	https://code.visualstudio.com/
PyCharm	https://www.jetbrains.com/pycharm/

3. **Autres Outils:**

ChatGPT	https://openai.com/

4. **Google Scholar:** https://scholar.google.com/