INPT

Projet de fin d'année

Data extraction and analysis of Co-author network analysis

Fait par:

Abdessamd akharaz

Houdayfa Housny

Encadré par:

Ikram el asri

Contents

Table des matières

Chapitre 1

Introduction

La recherche scientifique repose sur la collaboration entre chercheurs, facilitée par la publication d'articles en co-auteur. Analyser ces collaborations peut révéler des informations précieuses sur la structure et la dynamique des réseaux de recherche. Ce projet vise à extraire des données sur les auteurs et leurs co-auteurs à partir de la plateforme Semantic Scholar, puis à construire et analyser un réseau de collaboration.
Semantic Scholar est une base de données académique qui indexe des millions d'articles de recherche et fournit des informations détaillées sur les auteurs, les publications, et les relations de co-auteur. En utilisant des techniques de web scraping et d'analyse de données, nous avons collecté des informations sur les auteurs et leurs collaborations. Ces données ont ensuite été utilisées pour construire un réseau de co-auteur, où les nœuds représentent des auteurs et les liens représentent des collaborations entre eux.
L'analyse de ce réseau de co-auteur permet d'identifier des tendances et des motifs dans les collaborations scientifiques. Par exemple, il est possible de repérer des clusters de chercheurs travaillant sur des sujets similaires, de mesurer la centralité des auteurs influents, et d'analyser l'évolution des réseaux de collaboration au fil du temps. Ces informations peuvent être utilisées pour comprendre comment les idées se propagent dans la communauté scientifique, identifier des opportunités de collaboration, et évaluer l'impact de certains chercheurs ou institutions.
Ce rapport présente les méthodes utilisées pour l'extraction des données et la construction du réseau, les résultats de l'analyse, et les implications de ces résultats pour la compréhension des réseaux de recherche scientifique. En fournissant une vue d'ensemble des collaborations académiques, ce projet contribue à la littérature existante sur l'analyse des réseaux sociaux et offre des outils pour une gestion plus efficace des collaborations scientifiques.

Les technologies :

Selenium :

```python
from selenium import webdriver
from selenium.webdriver.chrome.service import Service
from selenium.webdriver.common.by import By
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
```

Selenium est un outil puissant en Python qui permet d'automatiser les navigateurs web. Il est couramment utilisé pour le test d'applications web, la navigation automatisée, ainsi que pour le scraping de données à partir de pages web. Grâce à Selenium, les scripts peuvent interagir avec les éléments d'une page web de la même manière qu'un utilisateur humain, ce qui inclut la soumission de formulaires, la navigation entre les pages, et l'extraction de contenu dynamique.

Networkx :

NetworkX est une bibliothèque Python destinée à la création, la manipulation et l'étude des structures, dynamiques et fonctions complexes des réseaux. Elle permet de travailler avec différents types de graphes (non orientés, orientés, multigraphes), de calculer des métriques de réseau (comme le degré, la centralité, les chemins les plus courts), et de visualiser les réseaux de manière efficace. NetworkX est largement utilisée dans les domaines de la science des données, de l'analyse des réseaux sociaux, et de la recherche opérationnelle, entre autres.

Pyvis :

Pyvis est une bibliothèque Python qui facilite la visualisation interactive des réseaux et des graphes. Elle s'appuie sur la bibliothèque JavaScript vis.js pour générer des graphes dynamiques et interactifs pouvant être visualisés dans les navigateurs web. Pyvis permet aux utilisateurs de créer des visualisations attrayantes et intuitives des réseaux, où l'on peut zoomer, faire glisser les nœuds et explorer les connexions. Cette bibliothèque est particulièrement utile pour ceux qui souhaitent présenter des réseaux complexes de manière visuellement engageante et facile à comprendre.

Matplotlib :

Matplotlib est une bibliothèque Python largement utilisée pour la création de visualisations statiques, animées et interactives. Elle offre une grande flexibilité pour générer divers types de graphiques, tels que des courbes, des histogrammes, des diagrammes en barres, des scatter plots, et bien d'autres. Matplotlib permet de personnaliser presque tous les aspects des visualisations, y compris les axes, les légendes, les annotations et les couleurs. C'est un outil essentiel pour les scientifiques des données, les chercheurs et les ingénieurs, facilitant l'exploration et la communication des données de manière claire et informative.

Chapitre 2

Méthodologies

2.1 Collecte des données

2.1.1 Semantic Scholar

Semantic Scholar est une plateforme de recherche académique avancée, développée par l'Allen Institute for AI (AI2) en 2015. Son objectif principal est de faciliter l'accès et la découverte d'articles scientifiques pour les chercheurs, étudiants, et professionnels du monde entier. Contrairement aux moteurs de recherche académiques traditionnels, Semantic Scholar utilise des techniques avancées d'intelligence artificielle et de traitement du langage naturel pour améliorer la pertinence et la qualité des résultats de recherche.

Caractéristiques Principales de Semantic Scholar

1. Indexation Étendue:

- Semantic Scholar indexe des millions d'articles de recherche provenant de diverses disciplines scientifiques, incluant les sciences de la vie, la médecine, les sciences sociales, l'informatique, l'ingénierie, et bien d'autres.

2. Recherche Sémantique:

- Utilisant des algorithmes de traitement du langage naturel (NLP), Semantic Scholar comprend le contexte des requêtes de recherche, permettant ainsi de fournir des résultats plus pertinents. Par exemple, au lieu de se baser uniquement sur les mots-clés exacts, le moteur de recherche peut interpréter l'intention derrière une recherche et proposer des articles liés thématiquement

3. Graphiques de Citations:

- L'une des fonctionnalités distinctives de Semantic Scholar est sa capacité à visualiser les réseaux de citations. Les utilisateurs peuvent voir comment les articles sont interconnectés par les citations, ce qui aide à identifier les travaux les plus influents dans un domaine spécifique.

4. Résumé Automatique:

- Semantic Scholar génère des résumés automatiques des articles, appelés "highly influential citations", permettant aux utilisateurs de comprendre rapidement l'importance et le contenu des travaux de recherche sans avoir à lire chaque article en entier.

5. Support de Données:

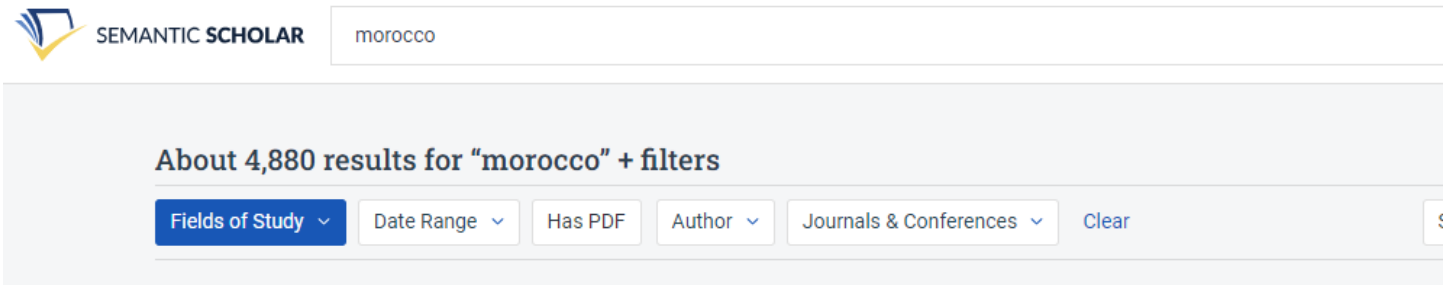
- Semantic Scholar soutient les initiatives de science ouverte en offrant des liens vers les ensembles de données associés aux publications, lorsqu'ils sont disponibles, ce qui encourage la transparence et la reproductibilité dans la recherche scientifique.

Semantic Scholar révolutionne la manière dont les chercheurs accèdent à la littérature scientifique en fournissant des outils plus intelligents et des analyses plus profondes. Les avantages principaux incluent :

- Découverte de Connaissances: La visualisation des réseaux de citations et les filtres avancés aident à découvrir des articles et des relations qui pourraient ne pas être immédiatement apparents avec des moteurs de recherche traditionnels.

- Amélioration de la Recherche Collaborative: Les fonctionnalités de profil d'auteur et de co-auteur encouragent la collaboration et la mise en réseau, ce qui est essentiel dans un paysage de recherche de plus en plus interdisciplinaire.

Nous avons utilisé des outils de recherche académique tels que Semantic Scholar pour extraire un ensemble de données d'articles d'ingénierie pertinents pour le Maroc. En utilisant des mots-clés spécifiques comme "Maroc" et "ingénierie", nous avons ciblé les articles qui sont directement liés au développement et aux initiatives dans le domaine de l'ingénierie dans ce pays.



Collecte et stockage des données

Dans cette partie, nous créons un fichier CSV dans lequel nous stockons nos données sous forme de deux colonnes : la première pour les auteurs et la deuxième pour la liste des co-auteurs.

```
120 with open('authors_and_coauthors.csv', 'w', newline='', encoding='utf-8') as csvfile:
121     fieldnames = ['Author', 'Co-authors']
122     writer = csv.DictWriter(csvfile, fieldnames=fieldnames)
123     writer.writeheader()
124     for data in authors_data:
125         writer.writerow(data)
```

	A	B
1	Author	Co-authors
2	I. Ourya	S. Abderafi, Nouhaila Nabil, N. Boutammachte, S. Rachidi
3	Nouhaila Na	S. Rachidi, N. Lahboubi, A. Alaoui-Belghiti, Hassan El Bari, Sanae Habchi, I. Ourya, Seddiq Sebbahi, Oussama Bayssi, N. Boutammachte, A. Hajjaji, S. Abderafi, R. Villa, S. Laasri, Yasna Mortezaei
4	S. Rachidi	M. Asbik, Nouhaila Nabil, K. E. Alami, N. Zari, A. Amrani, B. Mitchell, C. Fausser, Dominique Pelca, H. Ait Ousaleh, Yousra Filali Baba, Abdechafik Elharrak, Yassine Nassereddine, H. Bouzekri, P. Fla
5	Soufiane Bal	H. Ezâ€zahraouy, H. Labrim, M. Lakhal, B. Hartiti, M. Bhihi
6	L. Sadek	H. Talibi Alaoui, H. Alaoui, E. Sadek, A. Bataineh, A. Bentbib, I. Hashim, Tania A. LazÃfr, Ishak Hashim, O. Isik, K. Shah, O. Sadek, Mohammed S. Abdo, A. Sami Bataineh, A. AkgÃ¼l, Bouchra Abouzai
7	O. Sadek	S. Touhtouh, A. Hajjaji, M. Rkhis, M. El Jouad, F. Belhora, L. Sadek, R. Anoua, K. Shah, M. Bejar, E. Dhahri, Mohammed S. Abdo, El Mahdi Bouabdalli, El Mahdi Bouabdalli, Aymane Dahbi, L. Sadek, M
8	T. Abdeljawa	K. Shah, F. Jarad, Aziz Khan, Manar A. Alqudah, D. Baleanu, K. Nisar, Bahaaeldin Abdalla, P. Mohammed, G. Rahman, Mohammed S Abdo, H. Khan, Kamal Shah, M. Sarwar, J. Alzabut, Q. Alâ€Mdalla
9	S. Abderafi	T. Bounahmidi, S. Vaudreuil, Mohamed Anouar Kamzon, Ahmed Tgarguifa, A. Brahim, Ahmed Bichri, Jaouad Eddouibi, Yousra Jbari, I. Ndiaye, S. Aboudaoud, M. Chaanaoui, A. Rich, J. Klein, Fayrou
10	Sara El Hassa	A. Mezrhab, M. Moussaoui, M. Charai, Othmane Horma, Aboubakr El hammouti, Y. Admi, D. Santana, J. Benhamou, S. Channouf, Mohammed Amine Moussaoui, Ouassila selhi, Hanane Miri, Mugu
11	F. Oueslati	B. Ben-Beya, T. Lili, Salwa Fezai, N. Ben-Cheikh, H. Beji, A. Belghith, Fezai Salwa, D. Santana, Othmane Horma, Sara El Hassani, M. Moussaoui, B. B. Beya, A. Mezrhab
12	A. Mezrhab	M. Moussaoui, S. Amraqui, M. Charai, Benyounes Raillani, H. Naji, M. Jami, M. Karkri, Dounia Chaatouf, M. Salhi, J. Benhamou, M. Bouzidi, Y. Admi, F. Moufekkir, S. Channouf, H. Sghiouri, El Bachir
13	M. Mahdavi	K. Schmitt, H. Shayeghi, R. Romero, Hassan Haes Alhelou, Manohar Chamana, F. Jurado, Augustine Awaafo, H. Monsef, A. Bagheri, S. Jalilzadeh, F. Jurado, P. Siano, Stephen B. Bayne, J. CatalÃ£o, A
14	D. Vera	F. Jurado, D. A. LÃ³pez-GarcÃ-a, L. FernÃ¡ndez-Lobato, B. Ruiz-Carrasco, Manuel SÃ¡nchez-Raya, J. Montesdeoca, M. Tostadoâ€VÃ©liz, Y. LÃ³pez-SÃ¡nchez, P. Xu, Mohamed H. Hassan, Hoda Abd El-
15	Hamza El Hal	A. Khallaayoun, K. Ouazzani, Faissal Jelti, A. Jamil, Hamza El Alaoui, Ibtissam Bouarfa, Ahmed Bazzi, A. Khaldoun, M. El Ydrissi, Kedar Mehta, Anas Temouden, Salma Mahidat, Zakaria El Harmouzi,
16	A. Khallaayo	Hamza El Hafdaoui, R. Lghoul, K. Ouazzani, Yikun Huang, Faissal Jelti, Imane L'hadi, A. Jamil, Reda El Makroum, Kedar Mehta, Hamza El Alaoui, Hamza El Hafdaoui, D. Benhaddou, Ahmed Bazzi, Sara
17	K. Ouazzani	M. Benslimane, Hamza El Hafdaoui, Mehdi Tmimi, A. Khallaayoun, M. Berrada, A. Jamil, J. Bentama, Ouissal Drissi El Bouzaidi, A. Allouhi, Faissal Jelti, A. A. Mana, A. Khaldoun, A. Elgarouani, P. Sch
18	Daniel SÃ¡nc	M. Tostadoâ€VÃ©liz, Antonio EscÃ¡mez, F. Jurado, Roque Aguado, Paul ArÃ©valo, L. I. Minchala-Ãvila, D. Vera, D. Benavides, D. Vera, AdriÃ¡n Criollo, A. Ghadimi, M. Miveh, R. Hadria, S. Oulbi
19	Antonio EscÃ	M. Tostadoâ€VÃ©liz, D. Vera, F. Jurado, Daniel SÃ¡nchez-Lozano, R. Aguado, F. Jurado, Roque Aguado, Paul ArÃ©valo, F. Jurado, S. Mansouri, J. Aguado, D. Benavides, D. Vera, Yahya Z. Alharthi, A.
20	Samir Idrissi	Mohamed Oualid Mghazli, Jamal Brigui, Niima Es-sakali, Mohammed Ahachad, J. Brigui, F. El Mansouri, Ibtihal Ait Abdelmoula, Imad Ait Laasri, Moha Cherkaoui, Jens Pfafferott, Mohamed El Manl
21	Mohamed O	Niima Es-sakali, Samir Idrissi Kaitouni, Imad Ait Laasri, Moha Cherkaoui, Samir Idrissi Kaitouni, M. Charai, Imad Ait Laasri, Abdellah Nait-Taour, Abdelkader Outzourhit, Jens Pfafferott, Mohamed E
22	J. Brigui	F. Cacciola, F. Mansouri, M. Palma, M. P. Lovillo, A. Liazid, C. Barroso, L. Mondello, H. E. Cadi, F. El Mansouri, G. F. Barbero, Asmae El Cadi, H. E. Farissi, Yassine Oulad El Majdoub, B. Ramdan, R. Rod
23	F. Z. Echogda	S. Boutaleb, M. Abioui, M. Ouchchen, K. Abdelrahman, M. Ikirri, M. Id-Belqas, T. Abu-Alam, B. Dadi, H. El Ayady, M. Fnais, E. Abia, A. Bendarma, F. Faik, K. Mickus, R. B. Kpan, S. Essoussi, K. S. Sajinl

2.2 Analyse des données

Une fois les données collectées, plusieurs étapes d’analyse sont réalisées :

- **<u>Nettoyage des données pour éliminer les doublons et corriger les erreurs de formatage.</u>**

Le nettoyage des données est une étape cruciale dans toute analyse de données, particulièrement lorsqu'il s'agit de données collectées via des techniques de web scraping. Cette étape assure que les données utilisées pour la construction et l'analyse du réseau sont précises et fiables. Voici un détail des processus impliqués dans le nettoyage des données pour éliminer les doublons et corriger les erreurs de formatage.

Les doublons peuvent se produire lorsque la même information est collectée plusieurs fois, soit à partir de différentes sources soit à partir de la même source mais sous différentes formes. L'élimination des doublons est essentielle pour garantir que chaque auteur et chaque co-auteur ne soient comptés qu'une seule fois dans le réseau.

```
author_name = driver.find_element(By.XPATH, value: "//h1[@data-test-id='author-name']")

if author_name.text not in processed_authors:
    co_authors_list = get_coauthors()
    authors_data.append({'Author': author_name.text, 'Co-authors': ', '.join(co_authors_list)})
    processed_authors.add(author_name.text)
```

- **Construction d'un graphe où les nœuds représentent des auteurs et les arêtes représentent des collaborations.**

La construction d'un graphe est une étape clé dans l'analyse de réseau, permettant de visualiser et de comprendre les relations entre les auteurs. Voici un guide détaillé pour construire un graphe où les nœuds représentent des auteurs et les arêtes représentent des collaborations.

a) Étapes de Construction du Graphe

b) Chargement des Données Nettoyées

c) Initialisation du Graphe

d) Ajout des Nœuds

e) Ajout des Arêtes

f) Calcul des Attributs des Nœuds

g) Sauvegarde et Visualisation du Graphe

```
file = "authors_and_coauthors.csv"



G = nx.Graph()



with open(file, newline='', encoding='utf-8') as csvfile:
    reader = csv.reader(csvfile)
    next(reader)
    for row in reader:
        author = row[0]
        co_authors = row[1].split(", ")
        for co_author in co_authors:
            G.add_edge(author, co_author)

```

- **Calcul des métriques de réseau telles que la centralité, le degré, et les communautés.**

Dans l'analyse des réseaux, plusieurs métriques sont utilisées pour comprendre la structure et les dynamiques du réseau. Les métriques couramment calculées incluent la centralité, le degré, et l'identification des communautés. Voici un guide détaillé pour calculer ces métriques.

```
with open(file, newline='', encoding='utf-8') as csvfile:
    reader = csv.reader(csvfile)
    next(reader)
    for row in reader:
        author = row[0]
        co_authors = row[1].split(", ")
        for co_author in co_authors:
            G.add_edge(author, co_author)

communities = community_louvain.best_partition(G)
nx.set_node_attributes(G, communities, name: 'group')
# Create the network with Pyvis
node_degree = dict(G.degree)
nx.set_node_attributes(G, node_degree, name: 'size')
nx.set_node_attributes(G, node_degree, name: 'degree_centrality')
G1 = Network()
G1.from_nx(G)
G1.show( name: 'mygraph.html', notebook=False)

```

- **Utilisation de l'algorithme de Louvain pour la détection des communautés.**

L'algorithme de Louvain est une méthode populaire pour détecter les communautés dans les réseaux complexes. Il est basé sur l'optimisation de la modularité, une mesure de la densité des connexions à l'intérieur des communautés par rapport aux connexions entre les communautés. Voici un guide détaillé pour utiliser l'algorithme de Louvain pour détecter les communautés dans un réseau de collaborations entre auteurs.

```
20
21    communities = community_louvain.best_partition(G)
22    nx.set_node_attributes(G, communities, name: 'group')
23    # Create the network with Pyvis
24    node_degree = dict(G.degree)
```

2.3 Visualisation

La visualisation des données joue un rôle crucial dans l'interprétation des résultats de l'analyse de réseau. Nous avons utilisé la bibliothèque Pyvis pour générer des graphes interactifs qui permettent une exploration détaillée du réseau de collaboration entre auteurs. De plus, nous pouvons également utiliser la bibliothèque Matplotlib pour créer des visualisations statiques et des tracés qui complètent notre analyse. Voici comment nous avons utilisé ces deux bibliothèques pour visualiser les données et interpréter les résultats.

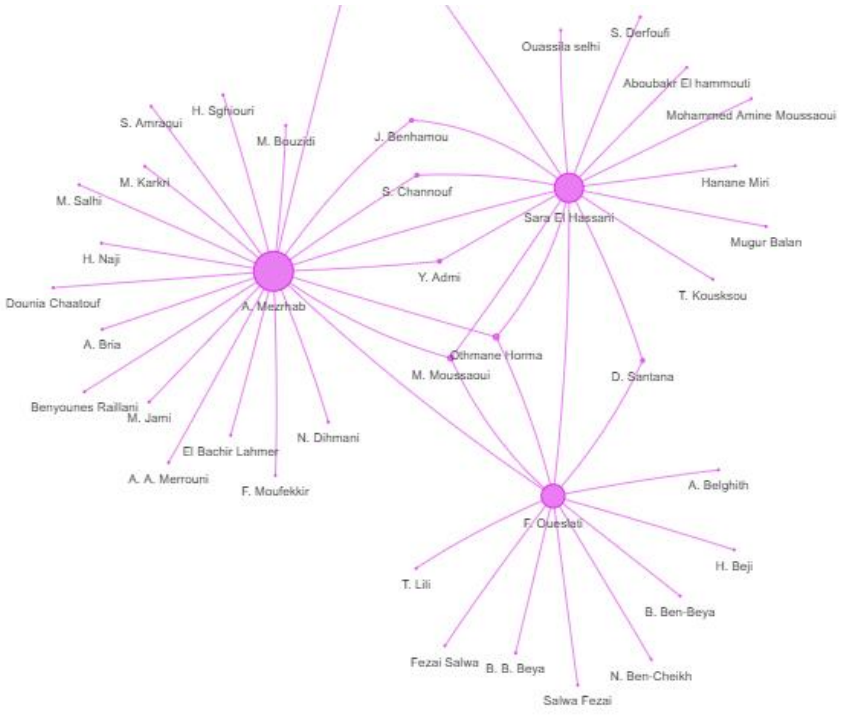
Visualisation avec Pyvis

```
6   file = "authors_and_coauthors.csv"
7
8
9   G = nx.Graph()
10
11
12  with open(file, newline='', encoding='utf-8') as csvfile:
13      reader = csv.reader(csvfile)
14      next(reader)
15      for row in reader:
16          author = row[0]
17          co_authors = row[1].split(", ")
18          for co_author in co_authors:
19              G.add_edge(author, co_author)
20
21  communities = community_louvain.best_partition(G)
22  nx.set_node_attributes(G, communities, name: 'group')
23  # Create the network with Pyvis
24  node_degree = dict(G.degree)
25  nx.set_node_attributes(G, node_degree, name: 'size')
26  nx.set_node_attributes(G, node_degree, name: 'degree_centrality')
27  G1 = Network()
28  G1.from_nx(G)
29  G1.show( name: 'mygraph.html', notebook=False)
30
```

Pyvis est une bibliothèque Python qui permet de créer des graphes interactifs basés sur les données de NetworkX. Nous avons utilisé Pyvis pour créer des visualisations interactives des réseaux de collaboration entre auteurs. Dans ces visualisations :

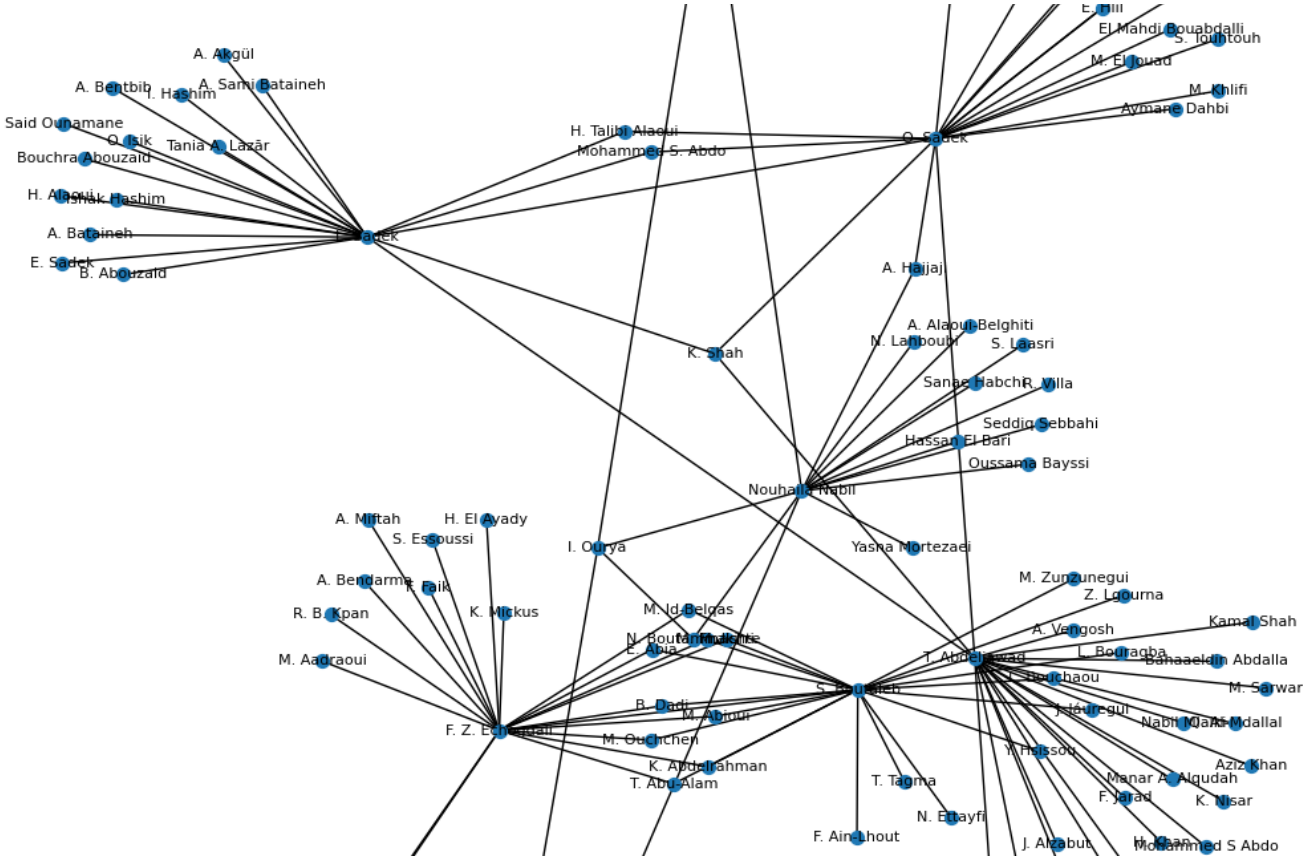
- Les nœuds représentent les auteurs, et les arêtes représentent les collaborations.
- Les nœuds sont colorés et dimensionnés en fonction des différentes métriques calculées, telles que le degré et la centralité.
- En survolant un nœud, des informations détaillées sur l'auteur sont affichées, y compris ses métriques calculées.

Ces visualisations interactives nous permettent d'explorer le réseau de collaboration de manière intuitive, en identifiant les auteurs les plus influents, les communautés de collaboration, et les tendances générales dans le réseau.



Interprétation avec Matplotlib

En plus des visualisations interactives, nous pouvons utiliser la bibliothèque Matplotlib pour créer des tracés statiques qui complètent notre analyseUn histogramme des degrés des nœuds peut nous donner un aperçu de la distribution de la connectivité des auteurs dans le réseau.


A. Akgül
A. Bentbib
A. Sami Bataineh
I. Hashim
Said Ounamane
O. Isik
Tania A. Lazār
Bouchra Abouzaid
H. Alaoui
Ishak Hashim
A. Bataineh
E. Sadek
B. Abouzaid
H. Talibi Alaoui
Mohammed S. Abdo
O. Sadek
El Mahdi Bouabdalli
S. Ioualtouh
M. El Jouad
M. Khlifi
Aymane Dahbi
A. Haijaj
A. Alaoui-Belghiti
N. Lahboubi
S. Lmasri
K. Shah
Sanae Habchi
R. Villa
Seddiq Sebbahi
Hassan El Bari
Oussama Bayssi
Nouhaila Nabil
A. Miftah
H. El Ayady
S. Essoussi
I. Ourya
Yasna Mortezaei
A. Bendarma
F. Faik
K. Mickus
R. B. Kpan
M. Id-Belqas
M. Aadraoui
E. Apia
M. Zupzunegui
Z. Lgourna
A. Vengosh
Kamal Shah
L. Boutaqba
Bahaaeldin Abdalla
M. Sarwar
B. Dadi
M. Aboud
F. Z. El
M. Ouchchen
K. Abdelrahman
T. Abu Alam
T. Tagma
N. Ethayfi
F. Ain Lhout
J. Arzabut
Nabil Mdallal
Aziz Khan
Manar A. Alqudah
F. Jarad
K. Nisar
Mohammed S Abdo

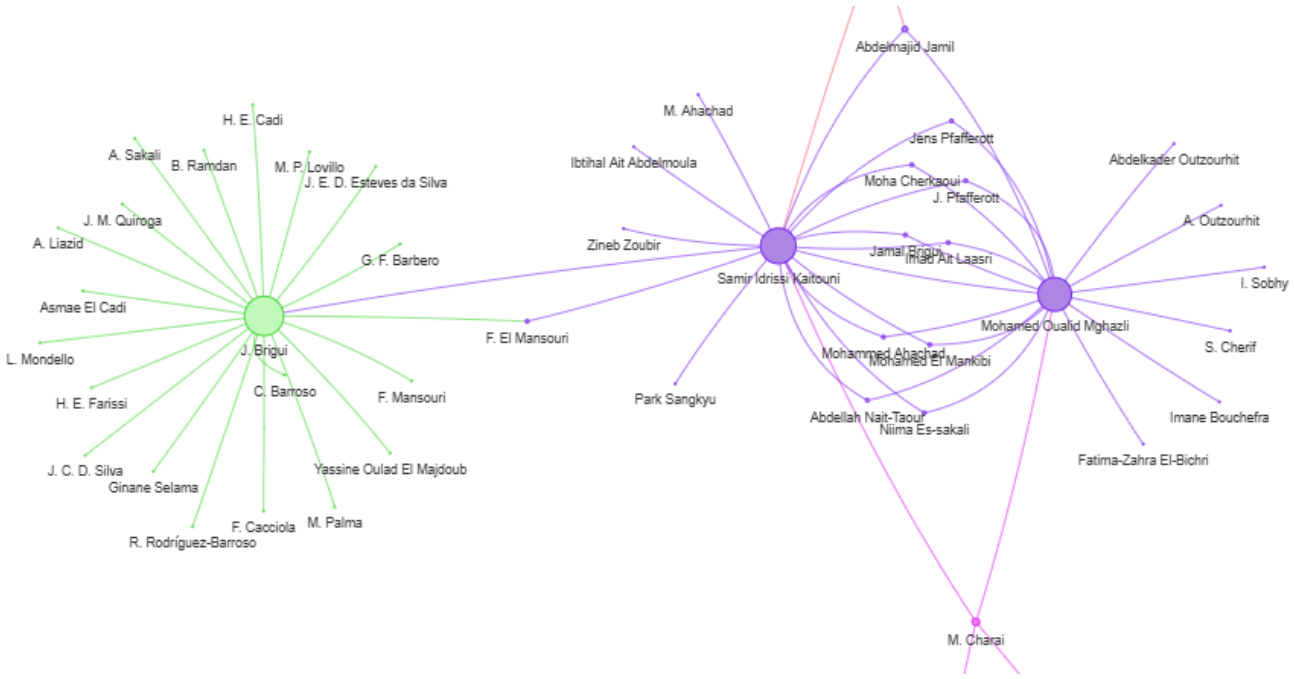
Chapitre 3

Résultats

Les résultats de cette étude révèlent plusieurs tendances intéressantes dans les réseaux de co-auteurs :

- Identification des auteurs les plus influents par leur centralité.
- Détection des communautés de chercheurs travaillant sur des thèmes similaires.
- Analyse des dynamiques de collaboration à travers différentes périodes.

Des visualisations interactives permettent de naviguer dans le réseau et d'explorer les relations individuelles entre les chercheurs.



Chapitre 4

Discussion

Les résultats obtenus soulèvent plusieurs points de discussion :

- Les structures de collaboration peuvent varier significativement selon les disciplines.
- L'impact des collaborations internationales sur la diffusion des connais-sances.
- Les limitations de l'approche utilisée, notamment en termes de couver- ture des données et de biais possibles.

La discussion aborde également des résultats qu'on n'arrive pas à simuler, notamment en ce qui concerne la visualisation du réseau :

- Dans la liste des co-auteurs, il peut y avoir des auteurs travaillant sur d'autres recherches. Il est important d'établir des liens entre eux.
- De plus, dans la partie extraction des données, il est nécessaire d'extraire les citations pour chaque auteur afin de les utiliser dans le calcul du degré.
- Aussi dans la partie d'identification des communistes dans veut le faire selon le domaine ou bien la spécialité de chaque auteur

Chapitre 5

Conclusion

Cette étude sur les réseaux de co-auteurs démontre l'utilité des analyses de réseau pour comprendre les dynamiques de collaboration scientifique. Les techniques utilisées, incluant le scraping de données et l'analyse de réseau, offrent des outils puissants pour explorer et

visualiser les interactions entre les chercheurs. Des pistes de recherche futures incluent l'intégration de données supplémentaires et l'amélioration des algorithmes de détection de communautés.

Reference :

https://pyvis.readthedocs.io/en/latest/

https://networkx.org/

https://matplotlib.org/

https://selenium-python.readthedocs.io/waits.html