RAPPORT
Projet
Network Analysis

Awa Syr DIAGNE

Professeur :
Julien
Velcin

MARS 2024

Awa Syr DIAGNE
Université Lumière Lyon 2 ,5 Av. Pierre Mendès France, 69500 Bron France
Professeur Julien Velcin ,Professeur d'informatique
Université Lumière Lyon 2
Cours de Network Analysis

Introduction

La théorie des réseaux sociaux conçoit les interactions sociales en termes de nœuds et liens. Les nœuds sont habituellement les acteurs sociaux dans le réseau, mais ils peuvent aussi représenter des institutions, et les liens sont les interactions ou les relations entre ces nœuds. L'analyse des réseaux sociaux permet ainsi de prendre connaissance des acteurs clés du réseau, de l'influence de chacun dans le groupe et de la qualité des interactions et des relations qu'ils entretiennent.

Dans le cadre de notre projet , nous voulons développer un système de recherche d'information qui permet de naviguer efficacement dans un grand corpus de données textuelles. L'idée serait de développer notre propre moteur de recherche et de pouvoir entrer des mots-clefs et retourner le résultat correspondant à cette requête .

Les données ont été prises dans un grand corpus de données textuelles du site de persée www.persee.fr . Le dossier comporte plusieurs bases de données qui représentent plusieurs corpus.
Elle comporte des informations essentielles aux documents . Nous avons le ou les auteurs , le titre , les citations , les dates de publications etc…

Tout d'abord , nous devons prétraiter les données. En effet, pour l'analyse ,il serait judicieux et logique de les regrouper en un seul fichier pour pouvoir créer un seul et unique corpus.

1. Acquisition des données

Une fois que les fichiers soient regroupés en un seul fichier , il serait ensuite utile de nettoyer et de récupérer les colonnes les plus importantes pour l'analyse.
Nous nous retrouvons donc avec une base de données complète avec 421463 documents et 129 colonnes (qui correspond aux informations du document).
Sur l'ensemble de documents du corpus , nous comptons 84 986 auteurs .Ce qui nous donne environ 5 auteurs par documents.
Ces articles ont été publiés entre 1837 et 2021.La majorité des œuvres ont été publiée vers les années 2000 . En 1996 , plus de 4000 auteurs ont publié des articles.

Awa Syr DIAGNE
Université Lumière Lyon 2 ,5 Av. Pierre Mendès France, 69500 Bron France
Professeur Julien Velcin ,Professeur d'informatique
Université Lumière Lyon 2
Cours de Network Analysis

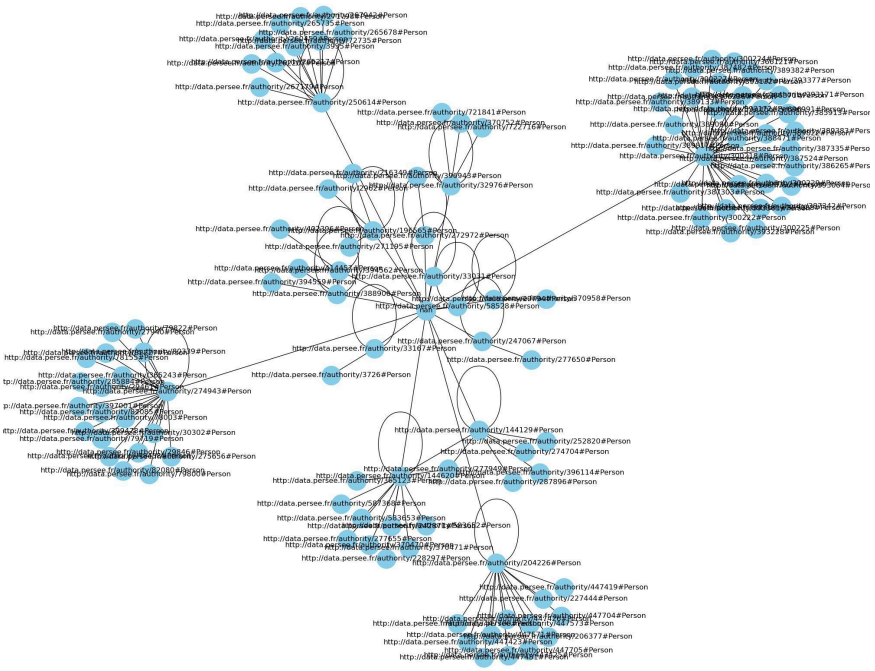
2. Prise en compte de la structure du corpus

Pour commencer notre étude , les nombreuses informations recueillies permettent de structurer notre corpus.

L'analyse de la structure de ce corpus sera concentrée sur les auteurs et les co-auteurs . L'idée est de construire un graphe où les auteurs et les co-auteurs des documents représentent les nœuds et les arêtes représentent les liens que ces auteurs ont entre eux autrement dit, les relations de co-auteurs que peuvent avoir ces auteurs sur un document. Il est bon de préciser que la plupart des documents ou articles présents dans le corpus sont écrits par plusieurs auteurs. Ces auteurs peuvent à la fois être auteurs de plusieurs articles.

Dans un souci de performance computationnelle ,nous sommes rabattus sur 15 colonnes contenant les valeurs de l'auteur.De plus ,nous avons décidé de définir un critère pour pouvoir aussi diminuer la taille de notre jeu de données .

Nous nous sommes focalisés sur les auteurs ayant plus de 750 co-auteurs :



Ce graphe contient 117 noeuds.Ces nœuds sont bien représentés , nous observons aussi une séparation des auteurs selon un critère Pour ce dernier ,nous n'en connaissons pas encore la cause.

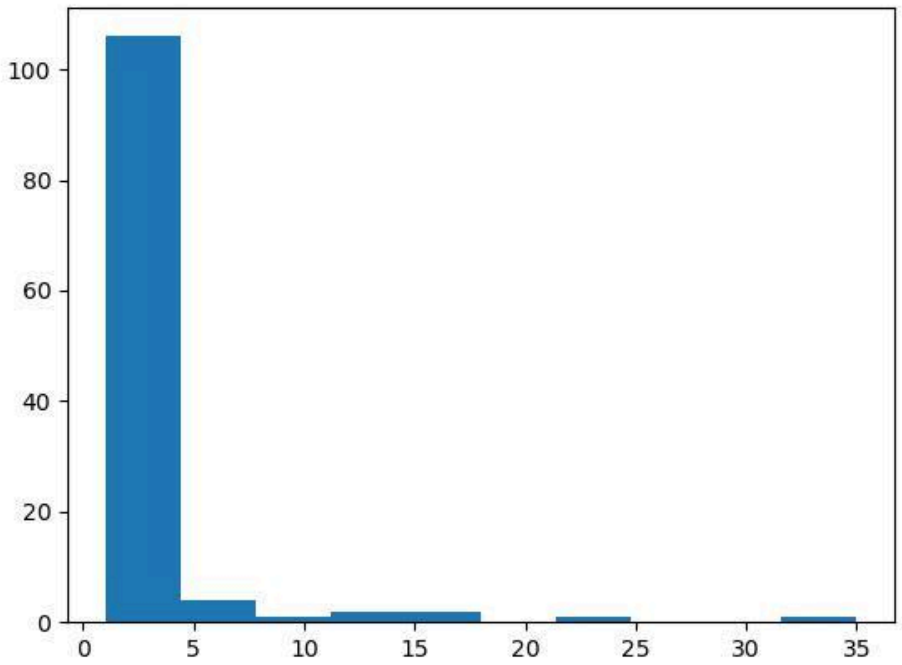
Notre graphe de réseaux a révélé plusieurs résultats intéressants. Tout d'abord, nous avons un nœud central qui en quelque sorte sert de liaison avec les autres nœuds. Nous pouvons apercevoir la construction de plusieurs groupes. Nous avons identifié des groupes d'auteurs

qui ne sont pas très loin du nœud central.On peut dire que ces derniers sont ceux qui agissent comme des facilitateurs de la collaboration, reliant diverses sous-communautés de chercheurs. Ces auteurs centraux ont tendance à publier avec un large éventail de collaborateurs et sont souvent cités dans de nombreux articles.
A la différence de ces derniers, les autres sont un peu plus inaccessibles ou ont probablement pas publié beaucoup d'articles.
Cependant,il faut reconnaître que les auteurs qui sont plus éloignés occupent la majorité des auteurs qui ont eu beaucoup d'arêtes.

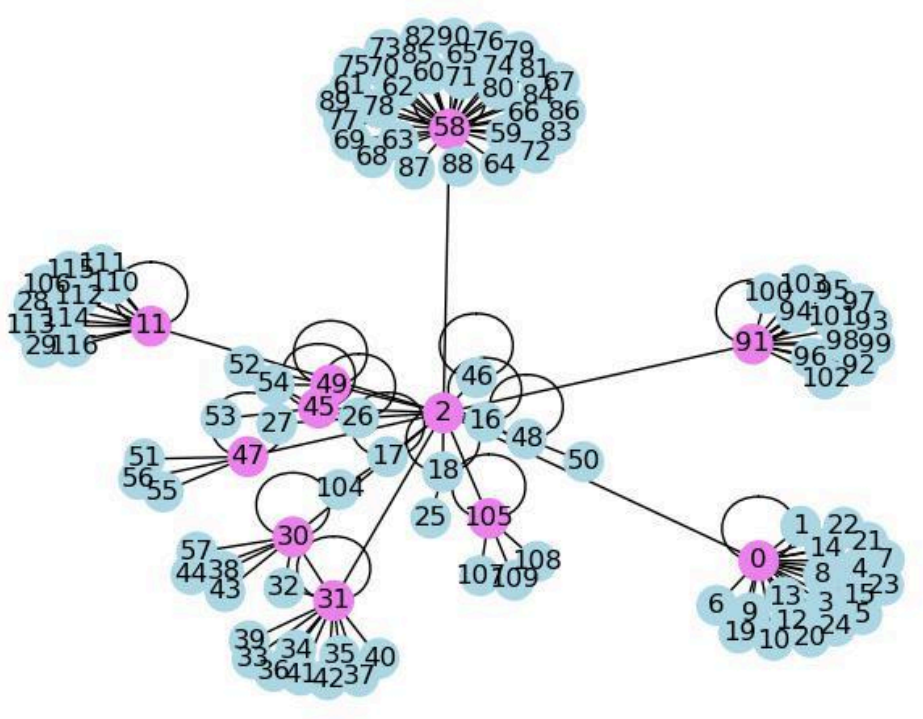
En outre, nous avons identifié plusieurs communautés de chercheurs qui se regroupent autour d'un thème de recherche spécifique.

- **Les degrés**

Voyons voir le nombre de relations , nous avons pour chaque nœud. Elles sont assez déséquilibrées. Il y a de ces auteurs qui collaborent assez souvent et y'en a d'autres très rarement. Ce qui fait qu'ils apparaissent très peu dans des articles . En moyenne , les auteurs ont au plus 2 deux co-auteurs pour leur article.



Awa Syr DIAGNE
Université Lumière Lyon 2 ,5 Av. Pierre Mendès France, 69500 Bron France
Professeur Julien Velcin ,Professeur d'informatique
Université Lumière Lyon 2
Cours de Network Analysis



Nous avons voulu voir une autre représentation de graphes mais cette fois-ci selon le degré et avec des couleurs où les nœuds bleues représentent les auteurs avec des collaborations inférieures à 3 auteurs et en violet l'inverse . Il y a une domination des bleues ce qui est tout à fait logique .Ces auteurs sont en dessous de la moyenne de co-autorat.
Il faut savoir aussi que les auteurs ayant moins de 3 co-auteurs sont ceux qui se trouvent à l'extrémité des graphes:ce sont des nœuds isolés.Il peuvent être des cibles en cas de reconsolidation du réseau.

Les auteurs avec un degré élevé peuvent être considérés comme centraux, car ils ont potentiellement plus d'influence et de connexions directes avec d'autres auteurs .

Ces auteurs peuvent être appelés des hubs et jouent un rôle crucial dans la connectivité globale du réseau. Ils sont des points de passage clés pour l'information dans le réseau.

Cependant, notre réseau n'est pas très stable du fait de sa distribution déséquilibrée en termes de degré. Ce qui veut dire qu'ils sont "fragiles" et sont plus susceptibles d'avoir des noeuds critiques.

3. Moteur de recherche

Nous voulons à présent créer un moteur de recherche qui ,en précisant à la fonction des mots clés ,nous retourne le texte ou l'article dont nous avons besoin.
Tout d'abord , il était nécessaire de faire quelques traitements préliminaires :

- Pour pouvoir constituer un corpus , il a fallu utiliser une colonne contenant que des valeurs textuelles.Donc on s'est concentré sur la variable *dcterms:title{Literal}'*cette variable stock le titre de chaque article. Donc l'objectif était de pouvoir les regrouper afin d'avoir un grand corpus.
- les valeurs de cette colonnes ont été par la suite mis en format UTF8
- Les valeurs de la colonne ont été regroupées en un seul corpus où par la suite, nous avons procédé à une tokenisation.

Ensuite , il sera nécessaire d'indexer le corpus.Elle consiste à créer une structure de données qui associe chaque terme ou mot des articles à une liste d'articles dans lesquels il apparaît.
Pour cette partie , il sera plus question de trouver une similarité entre les mots du corpus. Le bag of words est très utile pour représenter le texte sous forme de vecteurs numériques. L'idée principale derrière est de simplifier le texte en le traitant comme un "sac de mots" où l'ordre des mots n'a pas d'importance, seules les occurrences de mots importent.
Dans cette étude , la représentation TF-IDF va nous permettre d'ajouter un poids à chaque mot. Elles sont notamment importantes pour trouver une similarité entre deux mots. On pourrait utiliser la fonction cosinus qui calcule l'angle entre deux vecteurs pour mesurer la similarité entre deux auteurs du graphe ou aussi entre deux articles.

Le moteur de recherche se concentrera principalement sur les similitudes entre les titres des articles pour les regrouper lorsque l'un de ces mots clés est parmi les mots les plus importants du corpus.

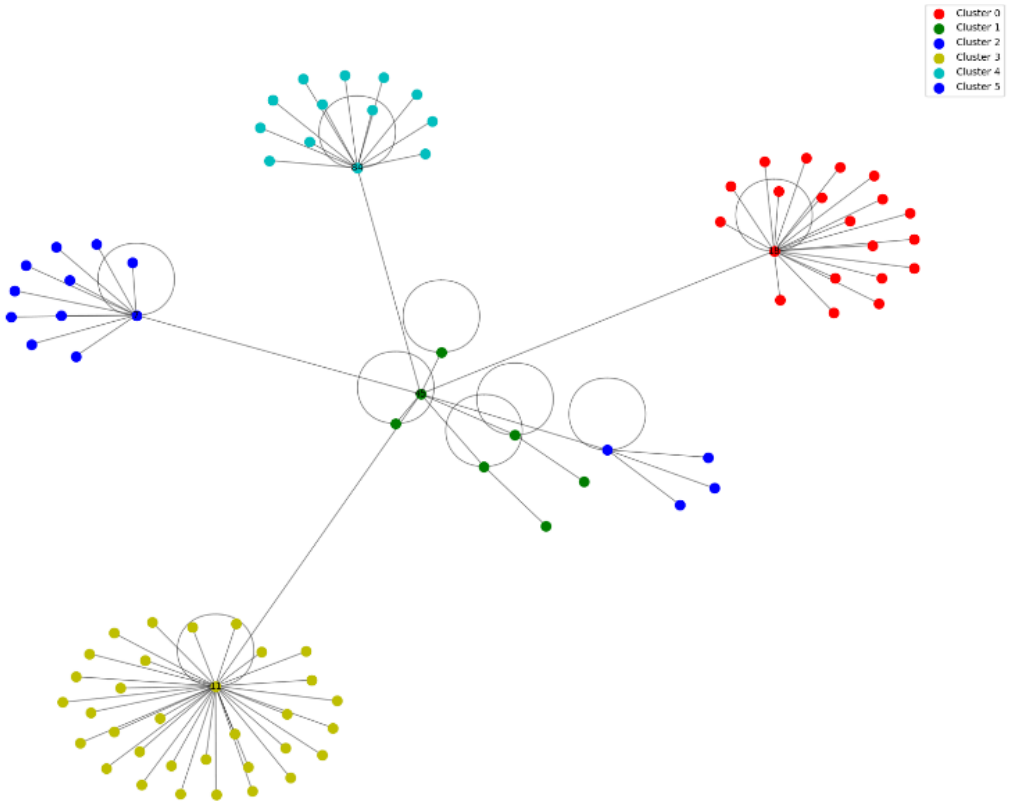
4. Ajout de clustering

Le clustering vise à regrouper les nœuds similaires ou fortement connectés dans des ensembles cohérents, révélant ainsi des sous-groupes dans le réseau.
Deux méthodes de clustering ont été utilisées.

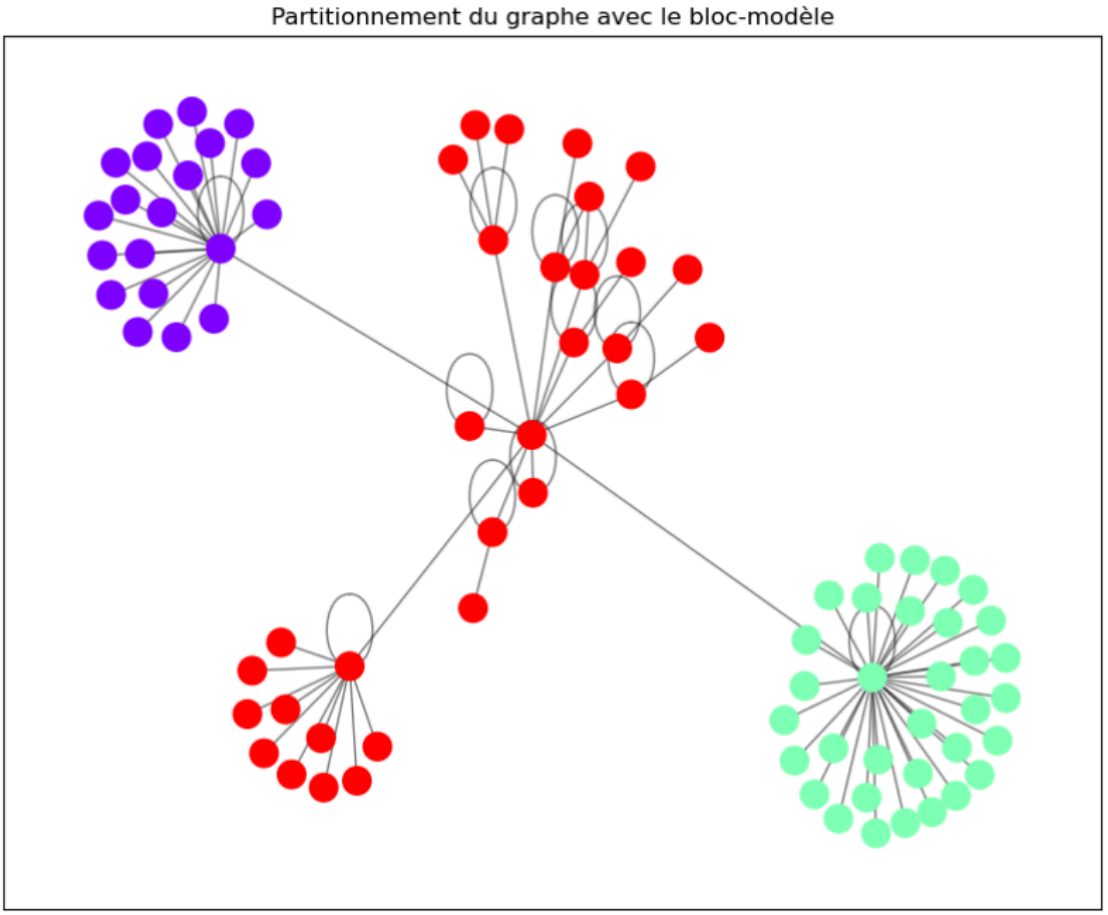
- La méthode Louvain :

Elle se distingue par sa capacité à détecter efficacement les communautés dans des graphes de grande taille tout en étant relativement rapide et adaptable à différents types de

réseaux.Nous voyons qu'il a quand même pu les séparer en clusters.Nous avons choisis arbitrairement le nombre de clusters.



- Les blocs models :
  Elles fournissent une approche pour modéliser la structure d'un graphe en identifiant des sous-groupes de nœuds qui sont fortement connectés entre eux, mais faiblement connectés avec les nœuds en dehors de leur groupe.



L'algorithme de clustering s'est basé sur la proximité des nœuds . Les nœuds les plus proches forment un groupe.

Bien que les deux approches visent à détecter les structures communautaires dans les réseaux, les méthodes de block models adoptent une approche probabiliste pour modéliser la structure du réseau, tandis que la méthode Louvain est basée sur une heuristique de maximisation de la modularité pour identifier les communautés. La méthode Louvain est souvent préférée dans les cas où la simplicité, la rapidité et l'évolutivité sont des préoccupations importantes, tandis que les modèles de blocs peuvent être plus adaptés lorsque des modèles probabilistes plus sophistiqués sont nécessaires pour capturer la structure du réseau de manière plus précise.

(NB : ce dernier paragraphe a été pris de CHATGPT!)

5. Fonction 5 : Classification Supervisée

Pour cette partie, nous nous focalisons sur les auteurs ,les co-auteurs et le titre. L'idée serait de pouvoir prédire les auteurs qui sont susceptibles de publier un type de document ,ceci se fera par l'intermédiaire du titre du document.Si l'auteur ou le co-auteur ont l'habitude de publier des documents sur le domaine de la géographie par exemple, nous pouvons prédire en fonction du nom de l'auteur ou du co-auteur le domaine d'étude du document.

(nous n'avons pas pu y aller jusqu'au bout)

Conclusion

Dans cette analyse de réseau consacrée aux relations entre auteurs et coauteurs dans un graphe, nous avons examiné attentivement la structure complexe et les dynamiques subtiles qui animent le réseau de collaborations scientifiques..Avec cette analyse, nous n'avons pas pu faire apparaître des interactions entre les auteurs et leurs collaborations, nous avons pu mettre en évidence également des motifs de coopération, d'influence et de centralité au sein du réseau des co-auteurs. Ces résultats auraient pu permettre de révéler la richesse des relations interpersonnelles qui se manifestent dans le domaine de la recherche scientifique, mettant en lumière l'importance de la collaboration et de l'échange d'idées dans la production et la diffusion des connaissances.

Mais néanmoins, l'analyse des mesures de centralité a mis en lumière les acteurs clés qui jouent un rôle crucial dans le réseau des co-auteurs, agissant comme des facilitateurs de collaboration, des médiateurs d'idées et des vecteurs d'influence au sein de leur communauté scientifique.

En conclusion, l'analyse des relations entre auteurs et coauteurs dans un réseau peut fournir des perspectives précieuses sur la dynamique sociale et académique qui influence la production scientifique. Elles peuvent guider les décisions relatives aux collaborations, aux stratégies de recherche et aux initiatives professionnelles visant à favoriser une collaboration efficace et innovante dans le domaine scientifique.