Projet : Network analysis

Réalisé par : ZIDAN Lama & ZIDAN Loubna

Master 2 MIASHS

Université Lumière Lyon2

22/mars/2024

Enseignants encadrants : VELCIN Julien

Table des matières

Introduction 2
1. Acquisition des données : 2
1.1. Décrire la source des données. 2
1.2. Téléchargement , analyse et prétraitement: 2
1.3. Statistiques du DataFrame Final 3
2. Prise en compte de la structure du corpus 4
2.1. Graphe de co-autorat 4
2.2. Caractéristiques Générales 5
2.2.1. Histogramme des Degrés 5
2.2.2. Matrice d'Adjacence 6
2.2.3. Analyse des Mesures de Centralité 6
2.2.4. Connectivité du Graphe 7
2.2.5. Densité du Réseau 7
2.3. Projeter les catégories sur les nouds du graphe : 7
3. classification supervisée : 7
3.1. Classification en fonction des caractéristiques structurelles : 7
3.2. Classification en fonction des caractéristiques structurelles : 8
3.3. Comparaison des deux modèles : 9
3.4. Interprétation et discussion : 10

Introduction

Notre projet consiste à exploiter les données fournies par la plateforme Persée, un riche répertoire de publications académiques. L'objectif est double : premièrement, cartographier et comprendre les réseaux de co-autorat pour révéler les collaborations et les communautés scientifiques, et deuxièmement, déployer des techniques de classification avancées pour prédire les catégories des documents à partir de leurs caractéristiques textuelles et structurelles.

En effet, notre démarche méthodologique intègre la classification supervisée des documents. Elle utilise non seulement le contenu textuel, tel que les titres et les résumés, mais aussi des mesures issues de l'analyse des réseaux, telles que le degré de centralité et l'intermédiarité des auteurs.

1. Acquisition des données :

Cette étape va fournir une base solide pour une compréhension approfondie du corpus étudié, ouvrant la voie à des analyses plus spécifiques, telles que l'examen des réseaux de co-auteurat ou l'application de méthodes de classification supervisée. La démarche méthodologique adoptée, alliant réduction ciblée des données et analyses statistiques détaillées, assure la pertinence et la richesse des informations extraites, cruciales pour le succès de notre étude.

1.1. Décrire la source des données.

La source des données pour notre projet provient de www.persee.fr, un portail dédié à la numérisation et à la diffusion du patrimoine scientifique, en particulier dans les domaines des sciences humaines et sociales. Le corpus se compose de métadonnées descriptives de plus de 900 000 documents, y compris des articles, des comptes-rendus, et d'autres types de documents principalement en langue française. Ces documents couvrent une période allant du dix-neuvième siècle à nos jours. Chaque document provient d'un fascicule qui fait partie d'une collection, généralement associée à une revue scientifique. Pour faciliter la navigation sur le portail Persée, les collections sont organisées par discipline principale, accessible via le lien.

Les données ont été générées à partir d'un sous-ensemble des fichiers de dumps de données liées (.rdf) du triplestore de Persée, disponibles à cette adresse, et représentent l'état des données en octobre 2021. Les données originales en format RDF-XML ont été converties en tableaux (dataframes pandas en Python), facilitant ainsi leur manipulation et analyse. Le jeu de données inclut également un tableau de correspondance entre les collections et leur discipline principale.

Pour chaque document, les métadonnées incluent, outre le titre et le sous-titre, les auteurs, la date de publication, et lorsque disponibles, un résumé, des mots-clés, une table des matières, ainsi que des informations sur les citations entre documents du portail Persée. De plus, l'identifiant de chaque document intègre le code de la collection, permettant l'exploitation des informations relatives à la discipline principale.

1.2. Téléchargement , analyse et prétraitement:

Dans le cadre de notre étude, nous avons procédé au téléchargement individuel de neuf fichiers au format Pickle, chacun correspondant à un ensemble distinct de données. Une analyse préliminaire des colonnes de chaque DataFrame a été effectuée afin d'identifier les variables les plus pertinentes pour notre recherche. Afin de systématiser la sélection des colonnes pertinentes à travers l'ensemble des neuf DataFrames, nous avons employé une boucle itérative. Cette méthode nous a permis de standardiser notre processus de sélection et d'assurer une cohérence dans l'ensemble des données traitées. Les colonnes sélectionnées pour leur importance dans notre analyse comprennent :

- 'dcterms:identifier', pour l'identification unique de chaque document,

- 'dcterms:title', pour le titre du document,
- 'dcterms:abstract{Literal}@fr', pour le résumé en langue française,
- 'marcrel:aut', pour l'identification des auteurs du document,
- 'bibo:numPages{Literal}', pour le nombre de pages du document, et
- 'persee:dateOfPrintPublication{Literal}(xsd:gYear)', pour l'année de publication imprimée du document.

Il est à noter que nous avons délibérément choisi d'exclure les colonnes relatives aux citations, à savoir 'cito:cites{URIRef}[i]' et 'cito:isCitedBy{URIRef}[i]', de notre sélection. Cette décision a été guidée par notre focalisation sur l'étude de la co-autorat et des interactions entre auteurs, plutôt que sur les relations de citation entre les documents. De plus, parmi les différentes colonnes de résumé disponibles, nous avons privilégié la sélection du résumé en langue française ('dcterms:abstract{Literal}@fr'), afin de maintenir une homogénéité linguistique dans notre corpus d'étude.
Poursuivant nos analyses, nous avons décidé de réduire la taille des DataFrames pour cibler les informations les plus pertinentes, nous avons opté pour une stratégie sélective axée sur les contributions des auteurs les plus fréquents. Cette décision a été motivée par une analyse préliminaire des deux premiers DataFrames, qui a révélé une prédominance marquée de publications par un groupe restreint de six auteurs dans chaque DataFrame.

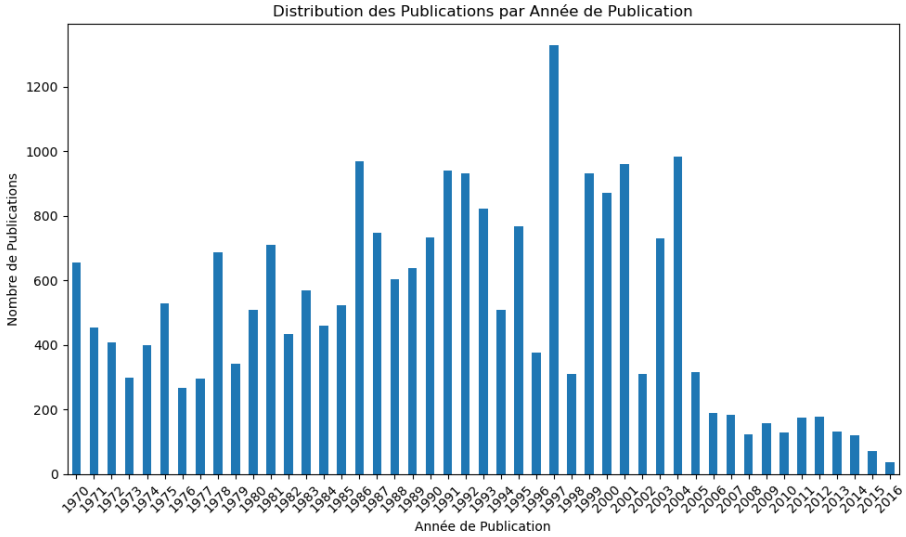
Nous avons institué une boucle itérative sur l'ensemble des neuf DataFrames afin d'extraire spécifiquement les entrées relatives aux 6 auteurs les plus fréquents dans chaque DataFrame, dans le but de réduire significativement le volume des données tout en conservant celles revêtant une importance cruciale pour notre analyse.

Cependant, étant donné la présence de multiples auteurs par document, cette approche a inévitablement abouti à un ensemble final contenant un nombre d'auteurs supérieur à six par DataFrame.

Toutefois, consciente de la taille encore imposante de cet ensemble, nous avons procédé à une nouvelle réduction basée sur des critères temporels. En sélectionnant exclusivement les documents publiés entre 1970 et 2016 (une plage plus restreinte par rapport à l'étendue initiale de 1897 à 2020) nous avons davantage affiné notre corpus, le ramenant à 23 823 lignes, optimisant ainsi notre capacité à explorer les relations de co-auteurat et les dynamiques au sein du corpus avec une plus grande précision.

1.3. Statistiques du DataFrame Final

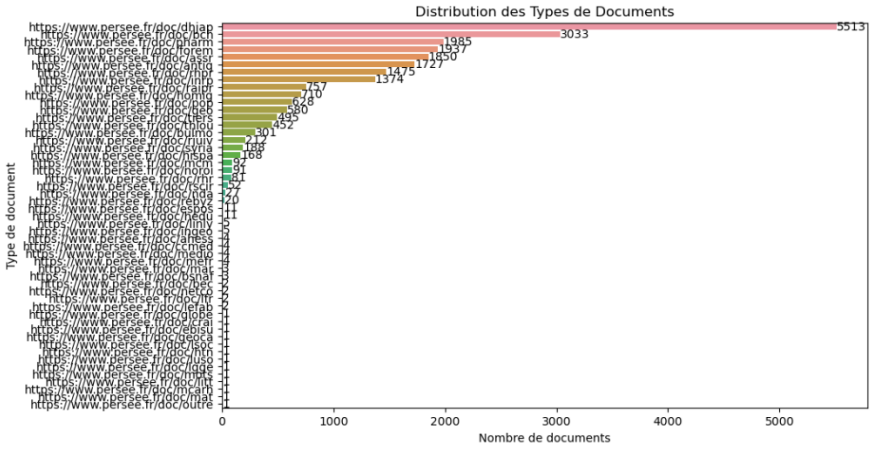
- Distribution Temporelle des Documents sur 47 années:



- Taille Moyenne des Documents : La taille moyenne des documents, mesurée en nombre de pages, s'est établie à environ 1.99 pages, reflétant la structure typique des contributions au sein

de notre sélection. Cette moyenne offre un aperçu de la densité et de la concision des travaux inclus.

- Nombre Total d'Auteurs Uniques : Avec un total de 278 auteurs uniques identifiés, notre corpus démontre une diversité significative en termes de contributions, soulignant l'étendue du réseau académique représenté.
- Distribution des Types de Documents : L'exploration des différents types de documents (articles, comptes-rendus, etc.) seront employé plus tard pour la prédiction des nouds de graphs.



- Nombre Moyen d'Auteurs par Document : Le calcul d'un nombre moyen de 3.33 auteurs par document met en avant la nature collaborative de la recherche dans les domaines couverts, indiquant une tendance à la co-auteurat.

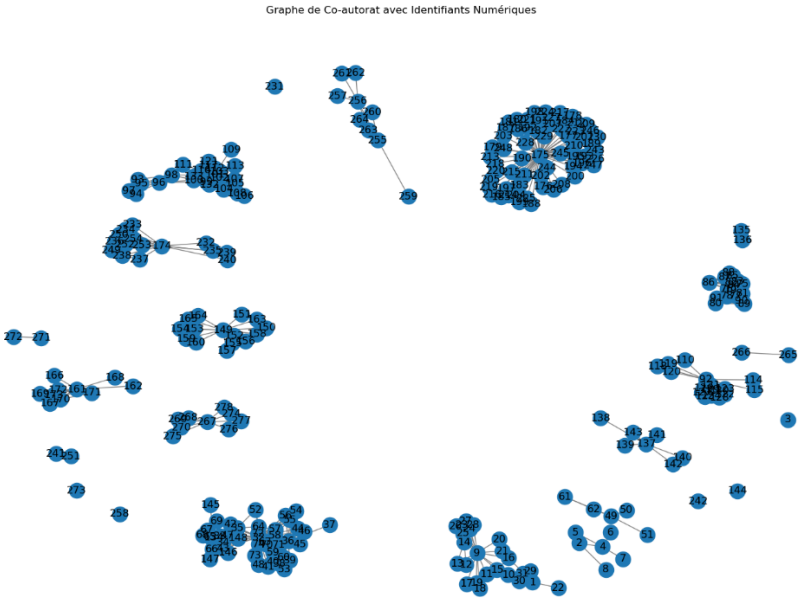
Suite à ces analyses, le DataFrame final a été sauvegardé sous format CSV pour faciliter le travail ultérieur et les manipulations suivantes.

2. Prise en compte de la structure du corpus

Dans la seconde partie de notre projet, nous avons développé une analyse de réseau pour explorer les relations de co-autorat au sein de notre corpus documentaire. Cette analyse s'appuie sur la construction d'un graphe où les nœuds représentent les auteurs uniques et les arêtes illustrent les liens de collaboration entre ces auteurs.

2.1. Graphe de co-autorat

Dans la première tentative de visualisation, nous avons constaté que les URI des auteurs en tant qu'étiquettes rendaient le graphe surchargé et difficile à interpréter. Ce problème est inhérent aux graphes denses où le chevauchement des étiquettes peut masquer la structure sous-jacente du réseau. Pour remédier à cela, nous avons assigné à chaque auteur un identifiant numérique unique. Cette étape de conversion a permis de réduire la complexité visuelle du graphe et d'accentuer les arêtes et les nœuds.



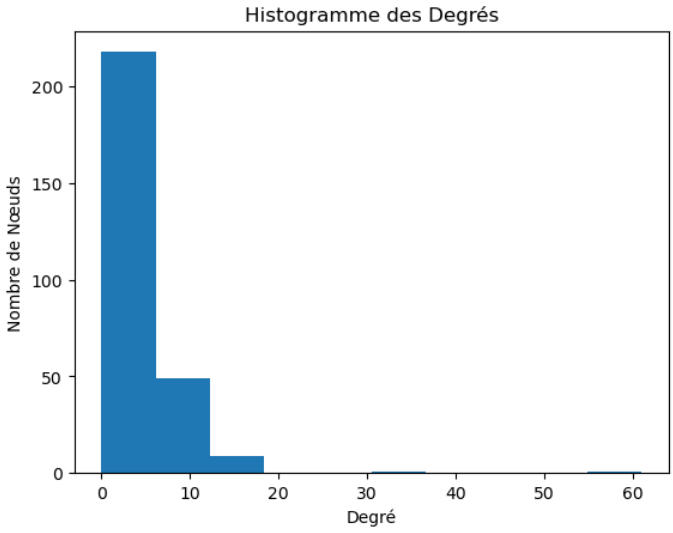
La visualisation optimisée, présentée ci-dessus, offre désormais une meilleure lisibilité et permet d'observer clairement les différentes composantes du graphe. On peut distinguer des groupes d'auteurs fortement interconnectés ainsi que des auteurs isolés ou des petits clusters, reflétant la diversité des collaborations au sein du corpus.

Nous avons ensuite procédé à une série d'analyses pour extraire des informations quantitatives et structurelles à partir de notre réseau.

2.2. Caractéristiques Générales

Le graphe est composé de 278 nœuds, représentant chaque auteur unique, et de 595 arêtes, illustrant les collaborations entre ces auteurs.

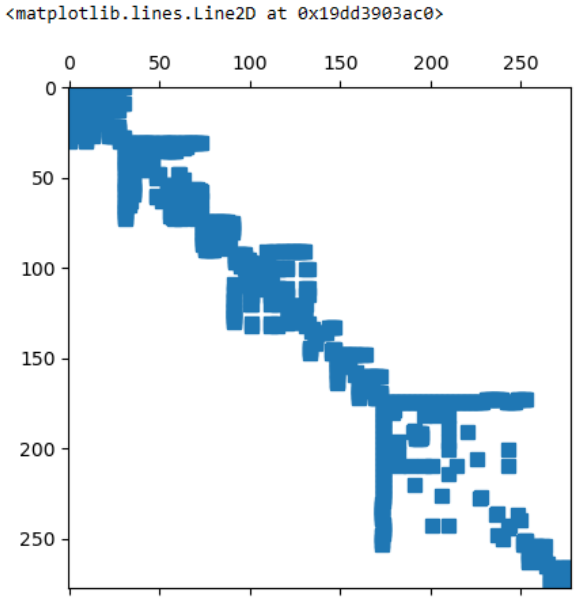
2.2.1. Histogramme des Degrés



L'histogramme des degrés montre la distribution de la connectivité des auteurs au sein du réseau (première image). La majorité des auteurs ont un nombre de connexions relativement bas, comme le démontre la décroissance rapide des fréquences à mesure que le degré augmente. Cela suggère que

notre réseau possède une structure de type "longue traîne" où quelques auteurs sont très connectés, tandis que la plupart ont peu de co-auteurs.

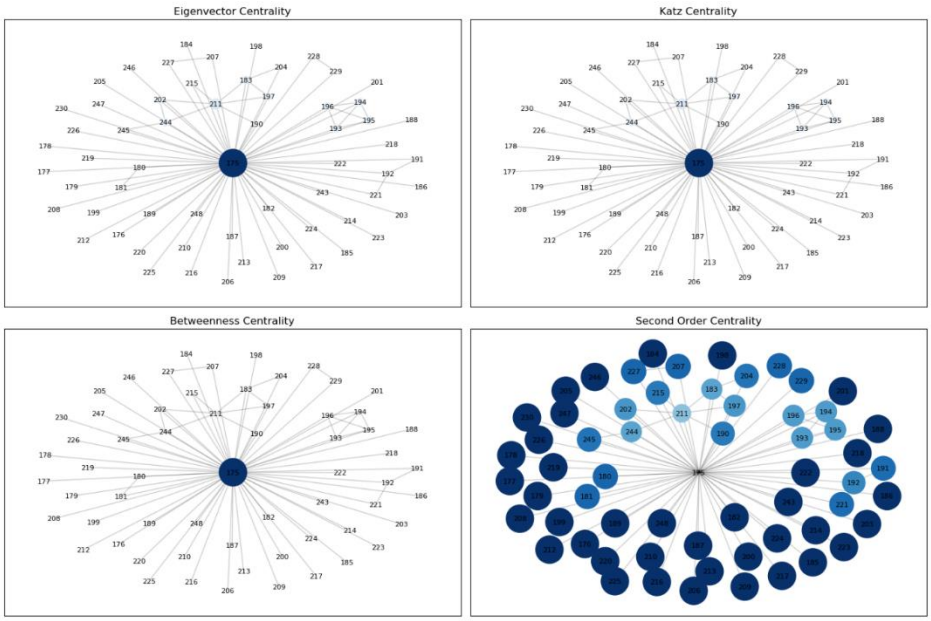
2.2.2. Matrice d'Adjacence



En utilisant plt.spy(A), nous avons visualisé la matrice d'adjacence du graphe, offrant une représentation binaire des connexions entre les auteurs. Les points bleus indiquent la présence d'une arête entre deux nœuds. Cette visualisation confirme la sparsité de notre graphe, avec un nombre d'arêtes bien inférieur au nombre total de paires de nœuds possibles.

2.2.3. Analyse des Mesures de Centralité

L'analyse des mesures de centralité est cruciale pour comprendre l'influence et l'importance des nœuds individuels dans notre réseau de co-autorat. Nous avons calculé et visualisé quatre mesures de centralité différentes :



- Eigenvector Centrality : Dans notre graphe, les nœuds avec une haute centralité de vecteur propre sont affichés plus larges et plus colorés, reflétant leur score de centralité.

- Katz Centrality : Les nœuds avec une haute centralité de Katz sont représentés de manière plus prononcée, indiquant leur nombreuses connexions directes et indirectes.
- Betweenness Centrality : Les nœuds qui apparaissent fréquemment sur les chemins les plus courts sont visualisés avec une taille ou une teinte plus marquée, ce qui souligne leur rôle de "pont" au sein du réseau.
- Second Order Centrality : Les nœuds centraux en termes de centralité de second ordre sont représentés avec une taille ou une couleur adaptée pour mettre en avant leur position stratégique dans le réseau.

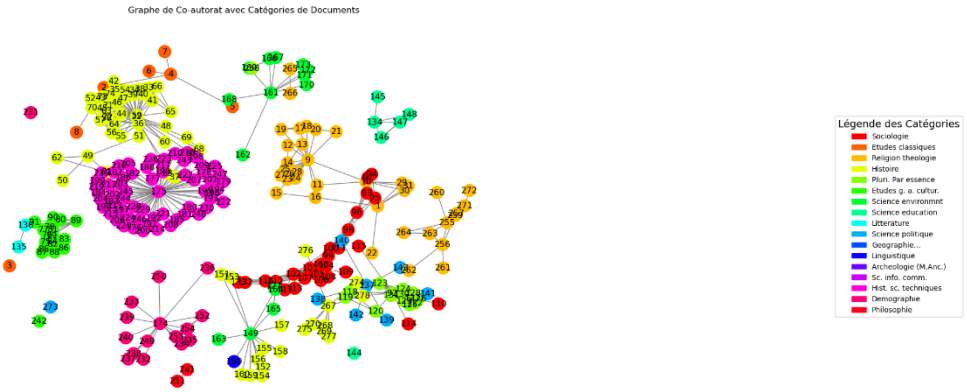
2.2.4. Connectivité du Graphe

Notre analyse de la structure globale du réseau a révélé que le graphe est constitué de 25 composantes connexes. Chaque composante connexe représente un sous-ensemble d'auteurs qui sont interconnectés entre eux par des co-autorats, mais qui ne sont pas directement liés aux auteurs des autres composantes connexes.

2.2.5. Densité du Réseau

La densité du graphe, calculée à 0.0155, est une mesure qui reflète la proportion des liens effectifs par rapport au nombre maximal possible de liens dans un réseau. Dans notre cas, la faible densité indique que le réseau est relativement clairsemé, suggérant que bien que des collaborations existent, elles sont loin de former un maillage serré de co-autorat. Cela peut être interprété comme une indication que notre corpus est composé d'une large gamme de documents couvrant divers sujets, où les auteurs ont tendance à former des groupes de collaboration plutôt restreints.

2.3. Projeter les catégories sur les nouds du graphe :



Pour approfondir l'analyse, nous avons projeté les catégories de documents sur les nœuds du graphe, colorant chaque nœud selon la catégorie du document auquel l'auteur a contribué. Cette représentation chromatique enrichit notre compréhension de la structure du réseau, révélant comment les différentes disciplines ou domaines d'étude se regroupent ou s'interpénètrent.

3. classification supervisée :

3.1. Classification en fonction des caractéristiques structurelles :

Dans la troisième partie de notre étude, nous avons entrepris la tâche de classifier les documents de notre dataset en 17 catégories distinctes en utilisant des informations textuelles provenant de leurs titres et résumés. L'objectif était d'évaluer la capacité d'un modèle d'apprentissage automatique à prédire correctement la catégorie d'un document sur la base de son contenu textuel. Pour chaque document, nous avons créé une colonne unique en concaténant le titre et le résumé. La vectorisation TF-IDF a

ensuite été utilisée pour convertir les textes concaténés en un ensemble de caractéristiques numériques. Ce processus permet de réduire l'influence des mots fréquents dans le corpus tout en mettant en exergue les termes clés spécifiques à chaque document. Nous avons choisi le RandomForestClassifier comme modèle de classification. Les résultats de précision du modèle sont illustrés dans l'image ci-dessous.

```
Accuracy: 0.6690451206715635

                      precision    recall  f1-score   support

Archeologie (M.Anc.)       0.00      0.00      0.00         3
         Demographie       0.63      0.29      0.40       131
   Etudes classiques       0.73      0.56      0.64       326
Etudes g. a. cultur.       0.62      0.99      0.76      1080
Hist. sc. techniques       0.78      0.73      0.76       388
            Histoire       0.56      0.09      0.16       723
        Linguistique       0.00      0.00      0.00         1
         Litterature       0.45      0.24      0.31        42
         Philosophie       0.45      0.47      0.46       147
  Pluri. Par essence       0.69      0.28      0.39        40
  Religion theologie       0.61      0.82      0.70       807
   Science education       0.80      0.75      0.77       300
  Science environmnt       0.50      0.29      0.36        21
   Science politique       0.86      0.59      0.70       244
         Sociologie       0.82      0.82      0.82       512

            accuracy                           0.67      4765
           macro avg       0.57      0.46      0.48      4765
        weighted avg       0.67      0.67      0.63      4765
```

La précision globale a atteint 0.669, indiquant que le modèle a correctement classé environ 66,9% des documents dans leur catégorie appropriée.

Une analyse plus détaillée des métriques de performance révèle une variance significative à travers différentes catégories. Il est important de noter que les catégories avec un faible nombre de documents (support faible) tels que "Archéologie (M.Anc.)" ont montré des scores de précision, de rappel et de score F1 de 0.00, indiquant une performance de classification faible pour ces catégories. Cela pourrait suggérer que le manque de données suffisantes pour ces catégories a eu une influence négative sur l'accuracy du modèle. En revanche, des catégories avec plus de données telles que "Sociologie" et "Science politique" ont présenté de meilleurs résultats, ce qui souligne l'importance de la quantité de données dans l'entraînement des modèles de machine learning.

3.2. Classification en fonction des caractéristiques structurelles :

Après avoir exploré la classification textuelle, notre étude s'est poursuivie avec une classification supervisée basée sur les caractéristiques structurelles des nœuds dans le graphe. Cette approche a pour but de prédire la catégorie d'un document en se basant sur la manière dont il est intégré dans le réseau de co-autorat, plutôt que sur le contenu textuel du document lui-même. Pour ce faire, nous avons extrait des mesures de centralité structurelle - degré, proximité et intermédiation - pour chaque nœud du graphe. Ces mesures reflètent divers aspects de la position d'un nœud dans le réseau. Ces caractéristiques ont été compilées dans un dataframe features_df, qui a ensuite été divisé en ensembles d'entraînement et de test à l'aide de la fonction train_test_split de la bibliothèque sklearn.model_selection, avec 20% des données réservées pour le test. Le modèle RandomForestClassifier a été choisi pour effectuer la classification en raison de sa capacité à gérer les dépendances non linéaires et son efficacité sur les données de grande dimension. Après entraînement et test sur les ensembles de données pertinents, le modèle a produit des résultats de classification que nous avons évalués à l'aide de métriques standard, telles que la précision, le rappel, le score F1 et le support.

Les résultats obtenus par le modèle sont illustrés ci-dessous :

```
Accuracy: 0.8363636363636363

                     precision    recall  f1-score   support

         Demographie       1.00      0.67      0.80         3
Etudes g. a. cultur.       1.00      0.67      0.80         3
Hist. sc. techniques       0.92      1.00      0.96        12
            Histoire       0.90      0.82      0.86        11
          Litterature       0.00      0.00      0.00         1
  Pluri. Par essence        1.00      1.00      1.00         5
  Religion theologie        1.00      0.67      0.80         9
   Science education        1.00      1.00      1.00         1
  Science environmnt        1.00      0.50      0.67         2
   Science politique        0.33      1.00      0.50         1
         Sociologie         0.58      1.00      0.74         7

           accuracy                             0.84        55
          macro avg        0.79      0.76      0.74        55
       weighted avg        0.88      0.84      0.84        55
```

L'accuracy globale atteint 0.836, indiquant que le modèle a correctement classé environ 83.6% des instances dans l'ensemble de test. Les métriques détaillées par catégorie révèlent des performances variables :

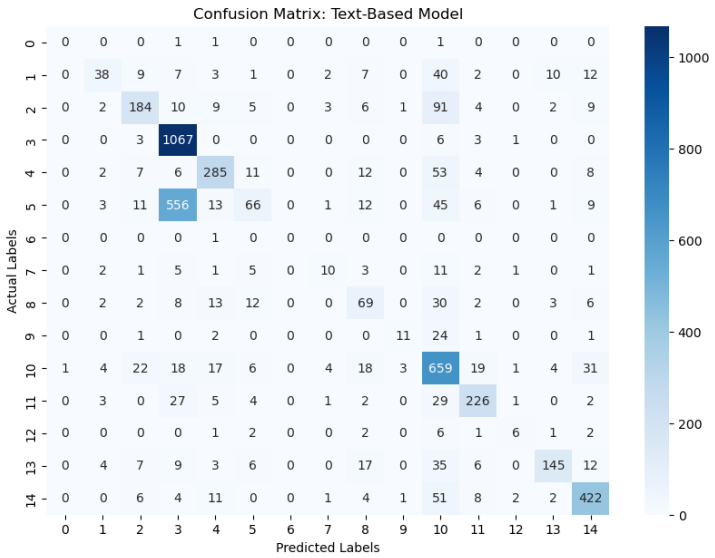
Des catégories comme "Histoire technique" et "Pluri. Par essence" montrent d'excellentes performances avec des précisions et des rappels élevés, signifiant que le modèle était capable de les identifier correctement avec une grande fiabilité. Cependant, d'autres catégories comme "Science politique" et "Sociologie" présentent une précision relativement faible, ce qui suggère que le modèle a classé à tort d'autres catégories dans ces classes. Les métriques macro et pondérées moyennes, respectivement 0.79 et 0.88 en précision et 0.76 et 0.84 en rappel, fournissent une vue d'ensemble des performances du modèle. La moyenne pondérée, en particulier, reflète une performance favorable en tenant compte de la taille de l'échantillon pour chaque catégorie (support). Il est important de noter que les catégories avec un support faible (comme "Science politique" avec seulement un échantillon dans l'ensemble de test) peuvent produire des statistiques de performance trompeuses. Une classification correcte ou incorrecte peut drastiquement changer les métriques, résultant en une précision et un rappel extrêmes. Ces catégories nécessitent donc une interprétation prudente et mettent en lumière l'importance d'avoir un ensemble de données équilibré pour une évaluation précise du modèle.

En résumé, les résultats montrent que les caractéristiques structurelles du réseau peuvent servir de prédicteurs significatifs pour la classification des catégories de documents, bien que la performance varie selon la distribution des catégories dans les données.

3.3. Comparaison des deux modèles :

Afin d'évaluer et de comparer l'efficacité de nos modèles de classification basés sur des caractéristiques textuelles et structurelles, nous avons analysé les erreurs de prédiction à l'aide de matrices de confusion. Ces matrices permettent de visualiser les performances de chaque modèle en montrant la fréquence des vraies classes par rapport aux prédictions des classes.

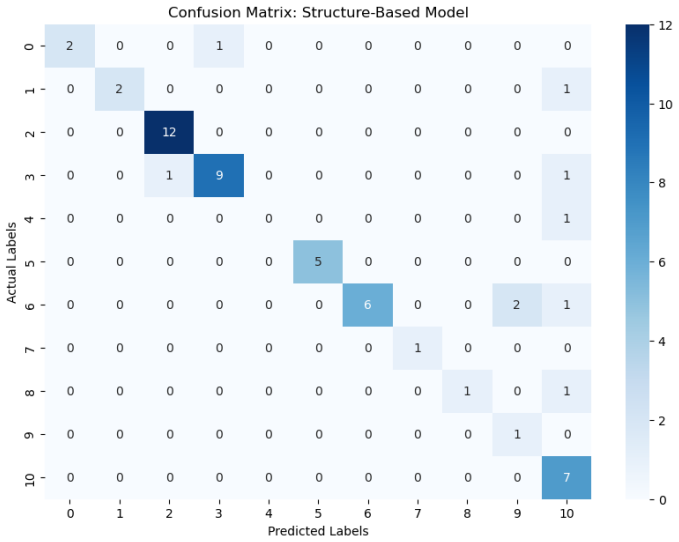
Matrice de Confusion pour le Modèle Basé sur le Texte :



La matrice de confusion pour le modèle basé sur les caractéristiques textuelles révèle des tendances intéressantes :

- Des valeurs élevées sur la diagonale principale, telles que pour les catégories correspondant aux indices 3 et 10, indiquent un taux élevé de vrais positifs, c'est-à-dire de bonnes prédictions.
- Certaines catégories montrent des confusions avec d'autres, comme indiqué par les valeurs non nulles hors de la diagonale. Par exemple, pour la catégorie correspondant à l'indice 5, il y a eu des confusions avec les catégories 3 et 10.
- Le modèle semble avoir une difficulté particulière avec les catégories ayant un petit nombre d'échantillons (faible support), comme illustré par les plus petites valeurs ou les zéros dans certaines lignes et colonnes de la matrice.

Matrice de Confusion pour le Modèle Basé sur la Structure :



3.4. Interprétation et discussion :

- La comparaison des matrices suggère que le modèle textuel, tout en ayant tendance à plus d'erreurs de classification globales, est capable de prédire un éventail plus large de catégories. D'autre part, le modèle structurel, bien Il est évident que le modèle structurel a un ensemble plus petit de catégories prédites correctement, mais ces prédictions tendent à être plus concentrées,

avec moins de dispersion des erreurs. Par exemple, les catégories 2 et 6 montrent des concentrations élevées de prédictions correctes.

- Cependant, la tendance à ne pas prédire de nombreux cas, illustrée par les zéros dans plusieurs cellules de la matrice, indique que le modèle peut être trop restrictif ou pas assez général pour couvrir toutes les classes efficacement.

qu'il semble plus précis dans les catégories qu'il peut prédire, a tendance à rater ou à ignorer un plus grand nombre de catégories.

Cette analyse révèle l'importance d'une stratégie combinée qui pourrait tirer parti des forces de chaque approche. La classification textuelle est peut-être mieux adaptée lorsque la diversité des catégories est grande et que les caractéristiques textuelles sont distinctives, tandis que la classification structurelle pourrait être privilégiée pour prédire des catégories où la position dans le réseau est un indicateur fort de la catégorie du document.

En conclusion, les résultats soulignent que l'équilibre et la complémentarité des modèles textuels et structurels sont essentiels pour améliorer la performance globale de classification dans des systèmes de recommandation, de recherche ou de filtrage d'informations.