UNIVERSITÉ LUMIÈRES LYON 2
Julien Velcin

M2 MIASHS
Coraly GERVASONI
Perrine MARTIN

NETWORK ANALYSIS FOR INFORMATION RETRIEVAL

PROJET

GRAPH,
CLUSTERING,
CLASSIFICATION,
MOTEUR DE RECHERCHE

16/03/2024

Introduction

Dans le cadre du projet en **Network Analysis for Information Retrieval**, notre objectif est de développer une solution d'analyse d'un corpus structuré, en mettant en pratique les concepts et les techniques vus en cours. Ce projet, réalisé en binôme, vise à explorer et à exploiter un ensemble de fonctionnalités permettant de comprendre la structure et le contenu d'un corpus de données.

Données

Les données que nous avons utilisées proviennent des métadonnées des documents disponibles sur le site *www.persee.fr*, un portail de numérisation et de diffusion du patrimoine scientifique. Ces documents, comprenant des articles, des comptes-rendus, etc., couvrent principalement les sciences humaines et sociales depuis le dix-neuvième siècle jusqu'à aujourd'hui.

Chaque document est associé à une collection correspondant généralement à une revue scientifique, avec une discipline principale pour la navigation. Notre jeu de données comprend plus de **900 000 documents**, avec des informations telles que le titre, les auteurs, la date de publication, un résumé, des mots-clés et des relations de citation.

Pour notre projet, nous avons sélectionné des documents dans diverses catégories telles que la religion, la littérature, l'art, l'archéologie, la sociologie, l'histoire, etc.

Analyses

Nous commençons par **présenter quelques statistiques** initiales sur les données, telles que le nombre de documents. Dans cette section initiale de notre rapport, nous offrons une vue d'ensemble du corpus, mettant en lumière sa structure et ses composants clés. Nous nous concentrerons sur la procédure d'**acquisition des données** et les techniques employées pour examiner la dynamique de **collaboration** parmi les auteurs.

En continuant notre exploration du projet, nous mettrons en œuvre plusieurs fonctionnalités supplémentaires pour enrichir notre analyse du corpus structuré. Premièrement, nous développerons des **moteurs de recherche** permettant à l'utilisateur de saisir un ou plusieurs mots-clés, afin de récupérer les titres des documents les plus pertinents correspondant à ces mots-clés. Cette fonctionnalité joue un rôle crucial dans la facilitation de l'exploration du corpus, en fournissant un accès rapide et efficace aux informations pertinentes. Nous aborderons ensuite le **clustering** des auteurs, une technique visant à regrouper les auteurs similaires en clusters distincts. Cette approche nous permettra de mieux appréhender les dynamiques de collaboration au sein du corpus, en identifiant des groupes d'auteurs partageant des intérêts ou des domaines d'expertise communs. Enfin, nous nous concentrerons sur la **prédiction de la catégorie principale** pour laquelle un auteur écrit.

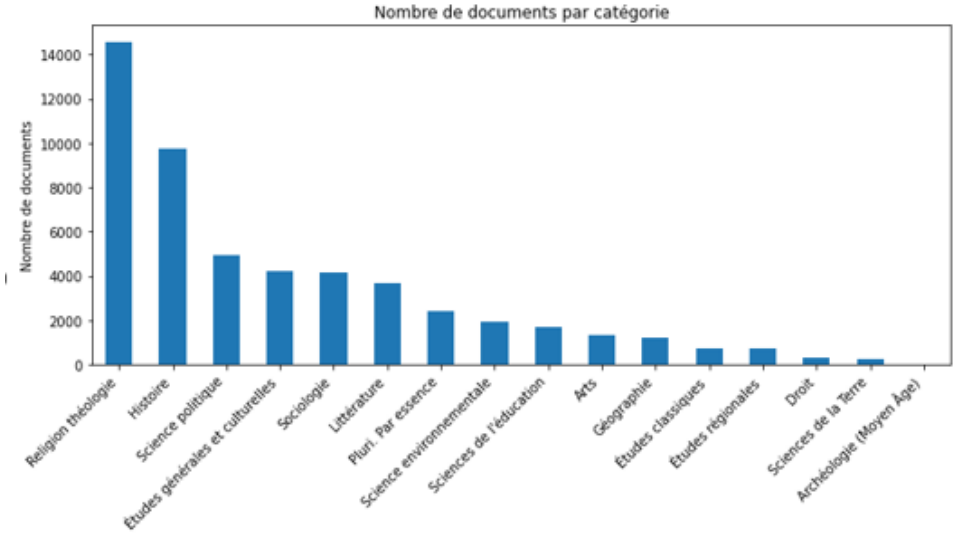
I) Acquisition des données

Nous avons sélectionné le jeu de données **"20240125_dataset_rscir_xxs.pickle"** pour notre étude, en raison de sa taille plus réduite facilitant les calculs. Ce jeu de données contient des documents classés dans plusieurs catégories (par ordre alphabétique), allant de "rscir" (Religion theologie) à "xxs" (Histoire).

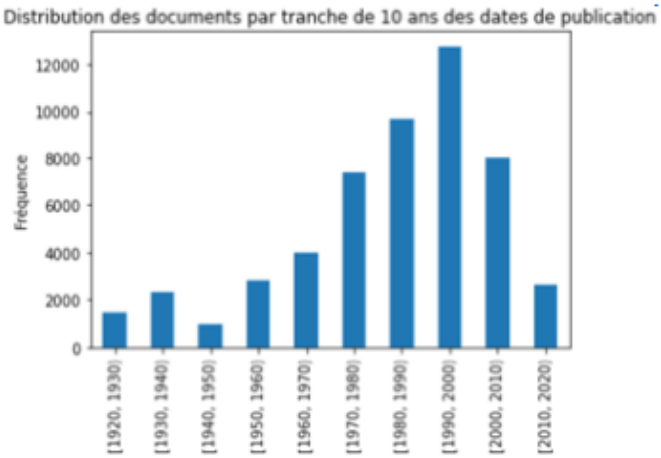
Initialement, le jeu de données comptait 59 529 lignes, mais nous avons restreint notre analyse aux documents rédigés en français, réduisant ainsi le nombre de lignes à 57 411. Suite à cela, nous avons constaté un nombre significatif de valeurs manquantes pour les résumés et les sujets des documents, notamment en

anglais (88%) et en français (plus de 95%). Pour garantir l'uniformité de notre analyse, nous avons donc choisi de nous concentrer exclusivement sur les titres des documents. Par ailleurs, nous avons supprimé 5 475 documents qui ne comportaient pas d'auteur renseigné, ce qui nous laisse avec un total de 51 936 documents pour notre étude.

L'objectif principal de notre étude est de **relier les auteurs ayant collaboré sur au moins un document**. Cela nécessite une compréhension approfondie de la structure et du contenu du corpus. Par ailleurs, le corpus compte au total 15 672 auteurs, soulignant ainsi l'ampleur et la diversité des données sur lesquelles repose notre analyse.



Le graphique à gauche présente le nombre de documents par catégorie. La très grande majorité des documents sont classés dans la catégorie "Religion théologie" (plus de 14 000), suivie de près par la catégorie "Histoire" (environ 10 000). Les catégories avec un faible nombre de documents, comme "Archéologie", ne sont pas visibles sur le graphique en raison de leur nombre réduit.



Le graphique à gauche montre une distribution du nombre de documents en fonction de la décennie de leur année de publication.

La décennie la moins représentée est celle des années 1940, avec seulement 941 documents. Les décennies les plus représentées sont celles des années 1990 et 1980, avec respectivement 12 772 et 9 637 documents. Les années 2010 ont moins de documents, avec respectivement 2 626.

Ces données suggèrent une augmentation progressive du nombre de documents publiés au fil des décennies, avec un pic particulier dans les années 1990 et 1980, suivies par les années 2000.

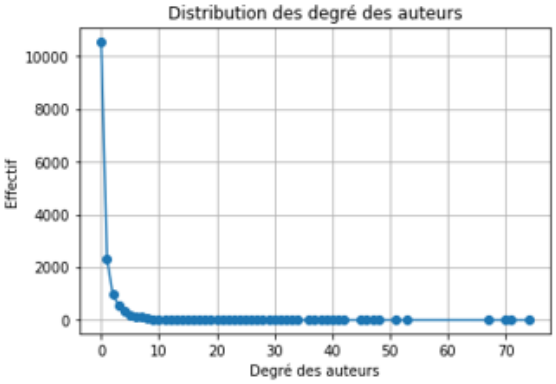
II) Prise en compte de la structure du corpus

Nous avons construit un **graphe de collaboration** entre les auteurs, qui nous a révélé un nombre impressionnant de 11 858 graphes connexes distincts. Cette multitude de graphes suggère une diversité importante dans les collaborations entre les auteurs du corpus, avec différents groupes travaillant sur des sujets variés. En examinant la taille des composantes de ces graphes, nous avons constaté que la majorité des composantes étaient composées d'un seul auteur, représentant **10 575 des 11 858 composantes**. Cela indique que la plupart des auteurs n'ont pas de collaborations fréquentes avec d'autres auteurs du corpus. Cependant, nous avons également observé un certain nombre de composantes plus importantes, allant

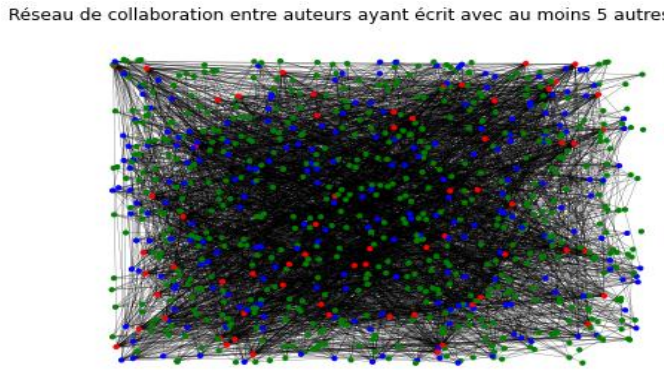
jusqu'à une taille maximale de 374 auteurs. Cela suggère l'existence de quelques communautés ou regroupements d'auteurs collaborant sur des sujets spécifiques.

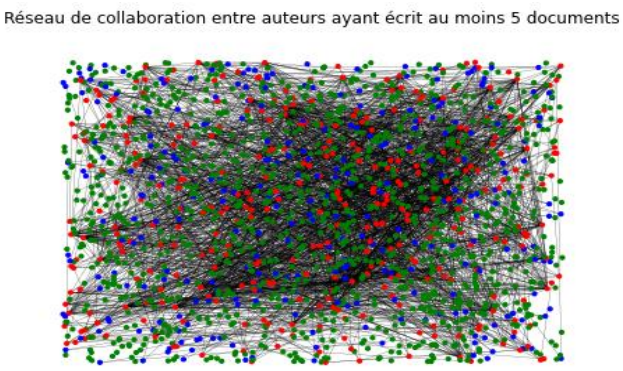
De plus, la largeur du graphe, qui mesure le nombre maximum de nœuds entre deux nœuds quelconques dans le graphe, est de 20. Cela signifie qu'il existe au moins un chemin reliant chaque paire de nœuds dans le graphe qui ne nécessite pas plus de 20 étapes. Cette mesure de la largeur du graphe nous donne un aperçu de la connectivité et de la densité globales du réseau de collaboration entre les auteurs.

La distribution des degrés des auteurs dans notre réseau de collaboration met en évidence la présence de "hubs" : quelques auteurs très connectés, tandis que la plupart ont peu ou pas de collaborations. Ces "hubs" jouent un rôle central dans le réseau de collaboration, ce qui souligne leur importance dans l'analyse des réseaux.

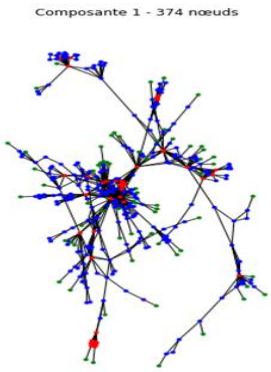


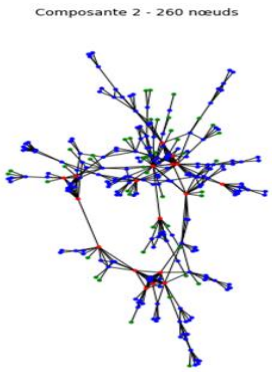
Pour rendre le graphe plus lisible, nous avons **visualisé uniquement les auteurs ayant un degré de collaboration d'au moins 5** dans le graphe ci-dessous à gauche et les auteurs ayant écrit au moins 5 documents à droite. Les nœuds sont respectivement colorés en rouge s'ils ont un degré d'au moins 30 et 20, en bleu s'ils ont un degré d'au moins 10, et en vert pour les nœuds avec moins de 10 collaborations. Ces visualisations nous permettent de mieux comprendre la structure du réseau de collaboration et de mettre en évidence les auteurs les plus influents :

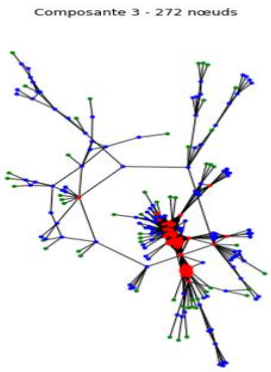


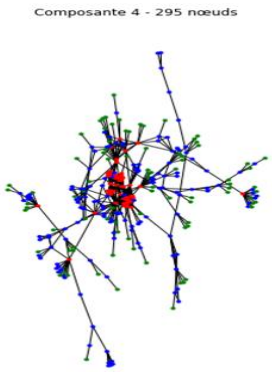


Pour présenter les **quatre plus grands graphes connexes** avec le plus de nœuds, où les nœuds rouges représentent les auteurs ayant au moins 10 collaborations, les nœuds bleus représentent ceux ayant au moins 2 collaborations, et les nœuds verts représentent ceux ayant une seule collaboration, nous avons utilisé une méthode de visualisation graphique. Ces graphes mettent en évidence les relations de collaboration les plus significatives entre les auteurs :









III) Moteur de recherche

a) Intégration des Données Textuelles

Nous explorons tout d'abord la mise en œuvre et les résultats d'un **moteur de recherche sémantique** basé sur les techniques de **TF-IDF** et la **similarité cosinus** pour traiter des requêtes de recherche dans un corpus de documents textuels. L'objectif principal est de développer une fonctionnalité permettant à l'utilisateur d'entrer un ou plusieurs mots-clés et de retourner un ensemble d'articles pertinents par rapport à cette requête.

Méthodologie

La première étape consiste à transformer les titres des documents en une liste, puis à nettoyer ces données en éliminant les mots vides à l'aide d'une liste préétablie de stop-words en français. Cette préparation est cruciale pour réduire le bruit dans les données et se concentrer sur le contenu sémantique pertinent. Ensuite, nous appliquons l'algorithme **TF-IDF** (Term Frequency-Inverse Document Frequency) pour indexer notre corpus. Cette technique met en évidence l'importance relative des mots dans les documents en fonction de leur fréquence, tout en prenant en compte leur rareté à travers le corpus, permettant ainsi une représentation riche et informative des documents.

Notre moteur de recherche utilise ensuite ces représentations TF-IDF pour calculer la **similarité cosinus** entre une requête de l'utilisateur et chaque document du corpus. Cette mesure de similarité permet de capturer l'orientation des mots-clés de la requête par rapport aux documents, indépendamment de leur longueur, offrant une appréciation quantitative de la pertinence des documents par rapport à la requête.

Résultats et Discussion

L'application de notre moteur de recherche à un ensemble de données a permis de générer un vocabulaire riche de **71 704 termes**, reflétant la diversité et la richesse du corpus. Une requête exemple composée des mots "juif" et "guerre" a illustré la capacité du moteur à identifier et à classer les documents en fonction de leur pertinence sémantique, mettant en avant des articles traitant de sujets historiques et sociétaux profonds comme la guerre d'Algérie, l'histoire juive pendant la seconde guerre mondiale et l'après-guerre.

Exemple des documents traitant le sujet de "juif" et "guerre" :

```
1 (23218): Une géographie de la guerre
2 (51823): La guerre d'Algérie
3 (50852): Poznanski Renée, Être juif en France pendant la seconde guerre mondiale
4 (50788): L'institut d'histoire juif de Varsovie
5 (49859): La guerre d'Algérie et les français
6 (32775): Maxime Rodinson, Peuple juif ou problème juif ? (coll. Petite collection Maspero, 249).
1981
7 (50064): Les français et la guerre d'Espagne
8 (9308): Villages d'après-guerre
9 (51621): Finir la guerre
10 (50580): Freud et la guerre
```

Les mots les plus fréquents dans le corpus, tels que *"histoire"*, *"siècle"*, et *"Jean",* montrent une inclination vers des thèmes historiques et biographiques, ce qui est cohérent avec la nature académique du corpus. La capacité à extraire aléatoirement des mots du vocabulaire, comme *"forsyth"* ou *"autographe"*, démontre également la diversité lexicale du corpus traité. Voir la liste entière des mots les plus fréquents en annexe.

b) Intégration des Relations Structurelles

Poursuivant notre exploration, nous avons mis en place l'intégration des relations structurelles entre documents et auteurs au-delà de la simple analyse textuelle. Ce défi nous conduit à explorer l'utilisation des **Graph Neural Networks (**GNNs), une avancée majeure permettant de calculer des représentations vectorielles de graphes qui prennent en compte non seulement les attributs des nœuds mais aussi leur connectivité.

Exploitation des Structures Relationnelles

Conscientes de la richesse sémantique et structurelle que représente la relation entre documents et auteurs, nous avons élaboré une méthode pour construire un graphe filtré, *G_filtered*, basé sur le degré des nœuds. Ce choix nous permet de concentrer notre analyse sur les composants les plus significatifs du réseau, en éliminant les nœuds de faible degré susceptibles de représenter des connexions moins pertinentes.

Intégration des Techniques TF-IDF dans les GNNs

Le cœur de notre approche réside dans l'application conjointe des techniques TF-IDF et des GNNs. D'une part, le TF-IDF nous permet de transformer les titres des documents en vecteurs de caractéristiques, capturant l'importance relative des mots au sein du corpus. D'autre part, en mappant ces vecteurs aux nœuds de notre graphe filtré et en les intégrant comme caractéristiques de nœuds dans un modèle **GNN**, nous sommes en mesure de calculer des embeddings enrichis qui reflètent à la fois le contenu textuel des documents et leur contexte structurel au sein du graphe.

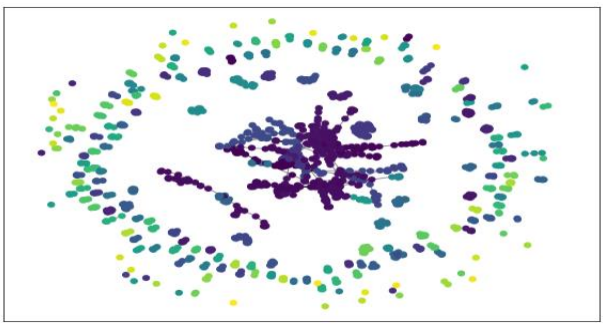
Défis et Limitations

La mise en œuvre de cette approche a cependant révélé des défis significatifs, notamment en termes de complexité computationnelle. Le traitement des grandes matrices de caractéristiques TF-IDF et l'entraînement des modèles GNN sur des graphes denses se sont avérés être particulièrement exigeants en ressources. Malgré la puissance théorique de notre méthode, les contraintes pratiques liées à la capacité de traitement de notre matériel informatique ont limité notre capacité à réaliser des expérimentations à grande échelle. Les tentatives d'entraînement du modèle ont régulièrement conduit à des redémarrages du noyau de calcul, reflétant les limitations de notre environnement expérimental.

IV) Ajout de clustering

Cette section se concentre sur l'application de **trois méthodes de clustering principales**. À noter que le graphe `*G_filtered*` a été créé en ne gardant que les auteurs ayant plus de deux connexions (collaborations) avec d'autres auteurs. Tous nos modèles ont été utilisés avec le sous-ensemble filtré.

a. Clustering avec l'Algorithme de Louvain



L'**algorithme de Louvain** a été utilisé pour détecter les communautés. Les communautés sont des groupes d'auteurs qui collaborent plus fréquemment entre eux qu'avec des auteurs d'autres communautés.

La visualisation montre une répartition des auteurs en communautés, où chaque couleur représente une communauté différente. La position des nœuds est déterminée par l'algorithme

de positionnement de force *`spring_layout`*, qui essaie de positionner les nœuds de manière à minimiser les chevauchements et à placer les nœuds fortement connectés près les uns des autres.

Nous avons trouvé **211 clusters** pour un total de 1763 éléments (auteurs). Cela suggère une grande diversité de réseaux de collaboration dans notre corpus. Le nombre élevé de clusters par rapport au nombre d'auteurs indique également que beaucoup de clusters sont probablement très petits, ce qui peut indiquer soit des collaborations très spécialisées, soit des communautés isolées.

Nous avons analysé par catégorie, en produisant un tableau pivot qui résume le nombre de documents par catégorie au sein de chaque cluster.

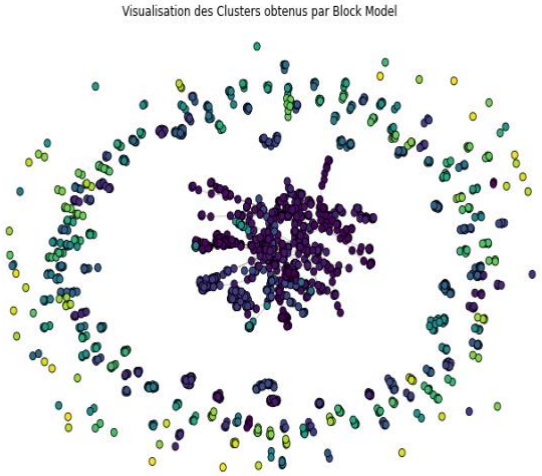
Cluster	rscir	rural	russe	rvart	salam	scrip	shmes	simon	slave	socco	sorci	sosan	sotra	spgeo	spira	sracf
19	0	0	0	12	0	0	0	0	0	0	0	0	0	0	0	0
210	0	0	0	11	0	0	0	0	0	0	0	0	0	0	0	0
3	10	151	0	9	0	5	0	0	5	0	15	0	0	562	0	1
130	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0
164	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0
17	0	0	130	1	0	1	1	0	494	0	0	0	0	0	0	0
105	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
5	0	1	0	0	2	0	24	0	1	0	0	0	0	0	0	0
103	0	0	0	0	0	1	12	0	0	0	0	0	0	0	0	0
147	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0	2
77	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0	2
94	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0
8	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	153
108	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	2
64	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	442	0	0	0	0	0	0	0	0	0	0
79	665	0	0	0	0	18	0	0	0	0	0	0	0	0	0	0

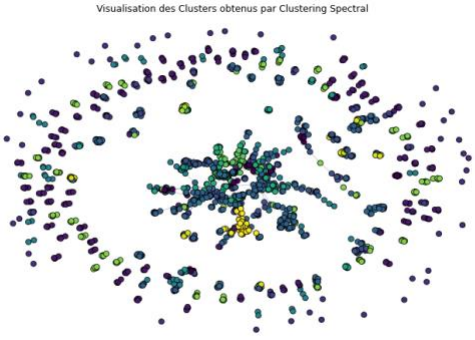
La présence d'un grand nombre de clusters avec une majorité de documents dans une catégorie particulière (par exemple, cluster 79 avec `rscir`) peut indiquer des communautés fortement spécialisées. Des clusters contenant un mélange de catégories peuvent indiquer des domaines interdisciplinaires ou des collaborations entre différents champs. De petits clusters ou des auteurs isolés peuvent représenter des chercheurs individuels ou des groupes qui n'ont pas de fortes collaborations externes ou qui travaillent dans des niches très spécialisées. Une catégorie représentée dans de nombreux clusters peut signifier un champ large et diversifié avec de multiples sous-domaines.

b. Clustering avec Black Modèle

La visualisation de ces clusters montre une répartition des couleurs qui indique une diversité dans la taille et la densité des groupes de recherche Nous avons ici **214 clusters** ici choisis automatiquement par notre algorithme (comme celui de Louvain).

En analysant le pivot table des clusters, on constate une répartition hétérogène des documents par catégorie au sein de chaque cluster, révélant une complexité dans les interactions et collaborations au sein du réseau. *Voir en annexe le pivot table.*





c. Spectral Clustering

Nous avons utilisé deux techniques distinctes pour estimer le nombre de clusters idéal, le spectre de la **matrice laplacienne** (voir annexe) et la **méthode Eigen Gap**, nous amenant à choisir 17 comme le nombre optimal de clusters. Notre méthode utilisait l'embedding pour affiner notre analyse des structures au sein du corpus de Persée.

La visualisation des clusters a révélé des groupes bien définis au sein du réseau. mais légèrement moins bien que les deux techniques précédentes. En particulier, le cluster 0 s'est distingué par une concentration notable de documents dans plusieurs catégories. Voir en annexe le pivot table. Il est essentiel de noter que, malgré l'importance des clusters ayant un grand nombre de documents, la dispersion équilibrée dans d'autres clusters suggère une intégration et une collaboration scientifique étendue. Ce modèle de dispersion souligne l'importance de ne pas se fier uniquement aux nombres élevés, mais de considérer également la valeur des interactions moins fréquentes, mais potentiellement plus innovantes et intersectorielles.

d. Analyse comparative des résultats

Nous avons mené une analyse comparative des résultats obtenus par les trois méthodes de clustering. Cette comparaison a été réalisée à l'aide des indices **ARI** (Adjusted Rand Index) et **AMI** (Adjusted Mutual Information), deux mesures statistiques qui évaluent la similitude entre deux assignations de clusters en tenant compte du hasard.

```
ARI Louvain vs Spectral (clean): 0.4955705662364283
AMI Louvain vs Spectral (clean): 0.5798726927252068

ARI Louvain vs Block Model (clean): 0.9918540244405937
AMI Louvain vs Block Model (clean): 0.9952636009434522

ARI Spectral vs Block Model (clean): 0.4974246730358297
AMI Spectral vs Block Model (clean): 0.5804748751393967
```

Les scores ARI et AMI entre les **méthodes de Louvain et Spectral** sont relativement modérés (ARI: 0.496, AMI: 0.580), suggérant que, bien que ces méthodes partagent une certaine cohérence en termes de regroupement, il existe des différences significatives dans la manière dont elles partitionnent les données. Cela pourrait être dû aux principes sous-jacents distincts de ces méthodes, Louvain se basant sur la modularité tandis que le clustering spectral utilise la structure du graphe et ses propriétés spectrales.

En revanche, les scores ARI et AMI entre les **méthodes de Louvain et Block Model** sont exceptionnellement élevés (ARI: 0.992, AMI: 0.995), indiquant une concordance quasi-parfaite dans les partitions de cluster. Cela suggère que les communautés détectées par Louvain sont très similaires à celles identifiées par le Block Model, ce qui renforce la validité de ces clusters.

La comparaison entre **Spectral et Block Model** montre des résultats similaires à ceux de Louvain vs Spectral, avec des scores ARI et AMI modérés (ARI: 0.497, AMI: 0.580). Cela réitère que le clustering spectral fournit une perspective unique sur la structuration des données qui se distingue des deux autres méthodes.

Nous avons également modifié le Spectral Clustering Embedding avec 214 clusters au lieu de 17. Voir les résultats en annexe.

V)Classification supervisée

Pour la classification, nous nous sommes concentrés sur les auteurs ayant rédigé **au moins 5 documents** pour garantir une représentation significative. Dans chaque modèle, nous avons utilisé la concaténation des labels de chaque document pour chaque auteur qui les a écrits, en plus des autres caractéristiques textuelles et structurelles. Voici un aperçu de chaque méthode utilisée :

a. Random Forest :

Il s'agit d'un algorithme d'apprentissage supervisé pour la classification. Il construit plusieurs arbres de décision lors de l'entraînement et combine leurs prédictions pour obtenir une prédiction globale plus robuste. Nous avons utilisé les caractéristiques textuelles des documents en plus des labels pour prédire les catégories.

b. GAT (Gated Graph Neural Network) :

Le GAT est un réseau de neurones récurrent adapté à l'apprentissage sur des graphes. Il utilise une attention pondérée pour saisir les relations entre les nœuds du graphe, ce qui le rend efficace pour analyser des données structurées telles que les réseaux de collaboration d'auteurs. Dans notre cas, nous avons utilisé deux couches de neurones pour intégrer à la fois les caractéristiques textuelles et structurelles des documents, en mettant l'accent sur les interactions entre les auteurs.

c. GCN (Graph Convolutional Network) :

Similaire au GAT, le GCN est également un réseau de neurones récurrent conçu pour l'apprentissage sur des graphes. Il utilise des couches de convolution pour agréger les informations des nœuds voisins dans le graphe, permettant de capturer les motifs et les structures complexes des données. Nous avons également utilisé deux couches de neurones pour intégrer à la fois les caractéristiques textuelles et structurelles des documents, en exploitant les relations entre les auteurs.

Il est à noter que le graphe utilisé présente une limitation où plus de 1000 nœuds ont un degré 0 sur 1740, ce qui rend le graphe moins informatif. C'est pourquoi nous avons concentré notre analyse sur les deux aspects combinés.

d. Résultats

Après avoir présenté les différentes méthodes utilisées pour la classification, nous passons à l'exposition des résultats obtenus. Nous avons évalué chaque modèle en termes de temps d'exécution, de précision (accuracy), de matrice de confusion et d'autres métriques pertinentes.

Pour évaluer les performances de chaque modèle, nous avons divisé nos données en deux ensembles distincts : une **base d'entraînement** comprenant 80 % des données et une **base de test** comprenant les 20 % restants. Nous avons entraîné chaque modèle sur la base d'entraînement et évalué sa performance sur la base de test.

	Accuracy (en %)	Execution Time (sec)
Random Forest	95.11	11.43
GAT	79.89	134.73
GCN	76.15	174.64

Les résultats montrent que **Random Forest** obtient la meilleure précision parmi les trois méthodes, avec une accuracy de 95.11 %, et présente également le temps d'exécution le plus rapide, avec seulement 11.43 secondes. Cela en fait un choix attrayant pour des tâches où la précision et le temps d'exécution sont critiques.

En comparaison, le modèle **GAT** affiche une accuracy légèrement inférieure à celle de Random Forest, avec un score de 79.89 %, et nécessite un temps d'exécution intermédiaire de 134.73 secondes. Il convient mieux aux tâches où la structure du réseau est importante.

Quant au modèle **GCN**, il présente l'accuracy la plus basse parmi les trois méthodes, avec un score de 76.15 %, et affiche le temps d'exécution le plus long, avec 174.64 secondes. Cependant, il reste une option pour des tâches nécessitant une exploitation spécifique de la structure du réseau.

Il est crucial de calculer des métriques supplémentaires telles que la précision, le rappel et le score F1 en plus de l'accuracy qui correspond au taux de prédictions positives, notamment lorsque les catégories sont réparties de manière déséquilibrée dans le jeu de données.

Modèle	Précision (en %)	Recall (en %)	F1-score (en %)
Random Forest	96.59	93.84	94.85
GAT	82.05	81.01	81.16
GCN	83.99	73.49	76.76

Les résultats dans le tableau ci-dessus montrent les moyennes non pondérées des métriques de performance (précision, rappel et score F1) pour chaque modèle dans l'ensemble des 16 catégories :

- La **précision** mesure la proportion de prédictions positives qui étaient correctes parmi toutes les prédictions positives faites par le modèle.
- Le **rappel** (recall) mesure la proportion de véritables positifs qui ont été correctement identifiés parmi tous les vrais positifs dans les données.
- Le **score F1** est la moyenne harmonique de la précision et du rappel. Il combine ces deux métriques en un seul nombre, offrant ainsi une mesure globale de la performance du modèle qui tient compte à la fois des faux positifs et des faux négatifs.

Le **random Forest** obtient les performances les plus élevées en termes de précision, rappel et score F1, avec des moyennes respectives de 96.59%, 93.84%, et 94.85%. Cela suggère que le modèle Random Forest est le plus robuste pour la classification des documents dans les différentes catégories étudiées.

En revanche, les **modèles GAT et GCN** présentent des performances légèrement inférieures. Le modèle GAT affiche des moyennes de précision, de rappel et de score F1 de 82.05%, 81.01%, et 81.16% respectivement, tandis que le modèle GCN obtient des moyennes de 83.99%, 73.49%, et 76.76%.

Dans l'ensemble, les résultats des trois modèles montrent des performances satisfaisantes, voire excellentes, avec de nombreuses catégories atteignant des scores de 100 % pour plusieurs métriques. Cependant, il y a des exceptions notables, notamment pour les catégories de géographie et de droit.

Pour la catégorie de géographie, le **modèle GAT** obtient des scores de 0.60 pour la précision, 0.50 pour le rappel et 0.55 pour le score F1. Ces scores inférieurs pourraient indiquer que le modèle a du mal à généraliser ou à distinguer efficacement les documents de cette catégorie par rapport aux autres. Des améliorations pourraient être nécessaires dans la manière dont les informations géographiques sont traitées ou intégrées dans le modèle pour obtenir de meilleurs résultats.

De même, pour la catégorie de droit, le **modèle GCN** présente des scores plus bas, avec un rappel de 0.50 et un score F1 de 0.33. Cela suggère que le modèle a des difficultés à rappeler correctement les documents pertinents de cette catégorie, ce qui peut indiquer des lacunes dans la façon dont les aspects juridiques sont appréhendés ou représentés dans le modèle.

Il convient de noter que dans le graphe, plus de 1000 nœuds sur 1740 ont un degré de 0, limitant ainsi l'information disponible dans le graphe. Cela pourrait expliquer en partie des résultats moins performants pour les modèles de réseaux de neurones, qui peuvent avoir du mal à capturer des informations significatives à partir d'un graphe avec de nombreux nœuds de degré nul.

Conclusion

Dans le cadre de notre cursus de M2 MIASHS, nous avons pu développer un système de recherche d'information et d'analyse approfondie d'un corpus textuel conséquent, issu de Persée.

À travers ce travail, nous avons d'abord entrepris la **collecte**, le **nettoyage** et la **structuration de données** extrêmement diverses, provenant de plus de 900 000 documents, afin de les rendre exploitables pour notre analyse. Cela a constitué un premier défi majeur, compte tenu de la variété des documents et des informations qu'ils contenaient, allant du titre et des auteurs à des données plus complexes comme les résumés, les mots-clés, et même les relations de citations entre documents.

Ensuite, en nous appuyant sur ces données préparées, nous avons développé un **moteur de recherche avancé**, permettant une exploration efficace du corpus par mots-clés. L'aspect le plus innovant de notre projet réside dans l'application de **méthodes de clustering** et de **classification supervisée**, telles que le Random Forest, le Gated Graph Neural Network (GAT), et le Graph Convolutional Network (GCN). Ces techniques nous ont permis d'explorer de nouvelles structurations des données, révélant des dynamiques de collaboration entre auteurs et des regroupements thématiques au sein du corpus.

En conclusion, ce projet a non seulement confirmé notre capacité à appliquer de manière concrète les enseignements de notre formation, mais a également ouvert la voie à de futures recherches. Les méthodologies développées et les résultats obtenus offrent un solide fondement pour des études ultérieures, que ce soit dans le cadre académique ou professionnel.

Annexes

Moteur de recherche

Liste des mots les plus fréquents :

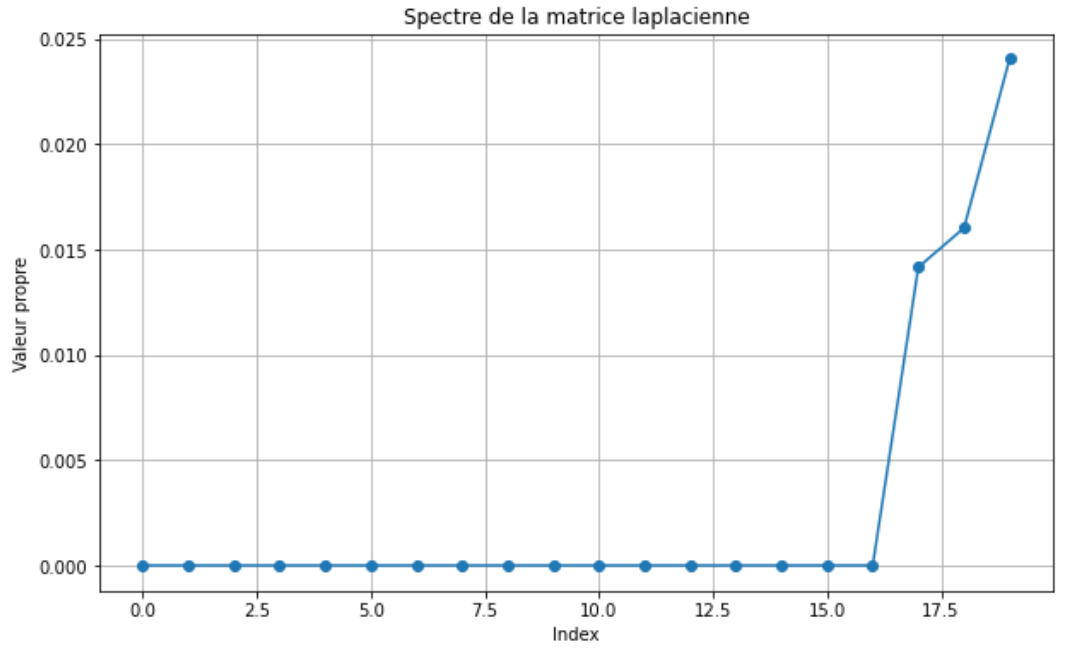
	word	value
0	the	603.112738
1	of	440.357898
2	in	421.144442
3	histoire	409.774889
4	coll	396.967178
5	propos	380.302738
6	jean	370.631220
7	and	363.921052
8	siècle	313.558215
9	saint	294.215433
10	développement	282.615617
11	introduction	277.352633
12	france	267.766142
13	pierre	235.145273
14	théologie	223.982534
15	der	206.874880
16	monde	196.235977
17	politique	194.858621
18	und	189.052815
19	géographie	186.907145
20	présentation	181.775234
21	ii	180.407287
22	paris	179.918754
23	chronique	178.948255
24	vie	174.868920
25	église	169.409380
26	espace	166.649789
27	paul	165.627892
28	éd	162.373591
29	dieu	159.469105

Ajout de clustering

Pivot Table Block Model :

BlockModelCluste	rscir	rural	russe	rvart	salam	scrip	shmes	simon	slave	socco	sorci	sosan	sotra	spgeo	spira	sracf	stice	syria	thlou	tiers	tigr	tlgpa	topoi
0	13	0	0	0	0	0	0	0	537	0	0	0	0	0	0	0	0	2	0	1	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	38	0	0	0	0	0	0	0
2	0	110	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
3	10	151	0	9	0	5	0	0	5	0	15	0	0	562	0	1	3	1	0	187	36	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12	0	0
5	0	1	0	0	2	0	1	0	0	0	0	0	0	0	0	0	0	2947	0	0	0	0	37
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	45	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	4	0	69	83	0	0	0	0	0	0	5	0	0	0
8	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	153	0	181	0	0	0	0	117
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	191	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	12	0	0	0	0	0	0
13	0	35	0	0	0	0	0	0	0	0	0	0	0	138	0	0	0	0	0	386	148	1	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	4	1	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
18	0	0	0	0	0	442	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0
19	0	0	130	1	0	1	1	0	494	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
21	0	4	0	0	0	2	0	0	0	11	0	19	447	0	0	0	0	0	0	3	0	0	1
22	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	481	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Technique Nombre Cluster Spectral Clustering :



Pivot Table Spectral Clustering :

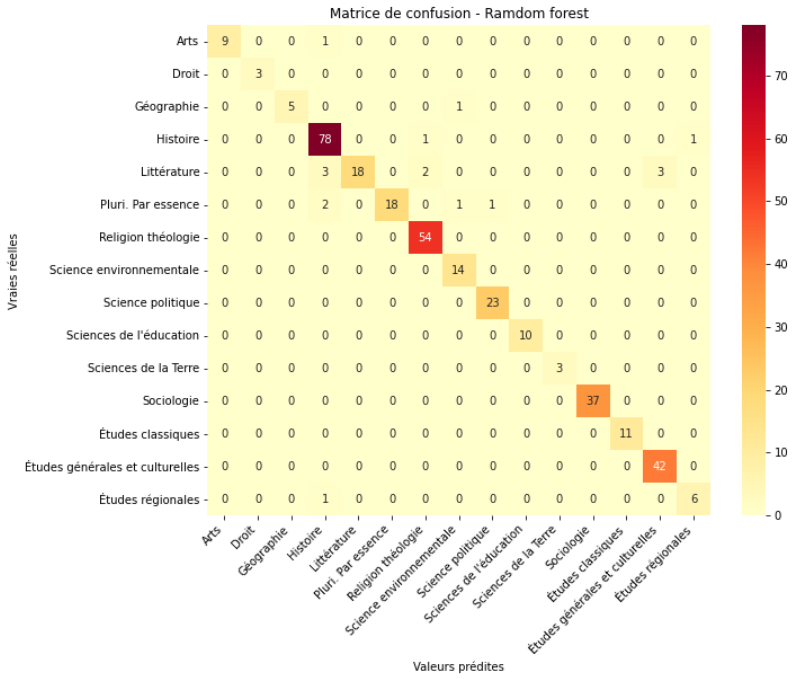
Cluster	rscir	rural	russe	rvart	salam	scrip	shmes	simon	slave	socco	sorci	sosan	sotra	spgeo	spira	sracf	stice	syria	thlou	tiers	tigr	tlgpa	topoi	versa	vibra
0	678	65	126	1	0	40	36	6	1008	6	21	49	243	232	1	72	43	249	2	19	40	88	109	0	64
1	0	9	0	0	0	124	3	0	2	0	10	14	71	97	1	13	9	1	3	83	38	5	0	0	0
2	10	18	0	10	0	17	7	0	5	5	15	19	22	20	14	3	8	12	0	102	2	0	5	0	1
3	1	6	23	23	0	3	0	0	39	0	2	5	11	22	10	40	5	0	206	125	4	4	8	0	0
4	0	5	0	7	0	5	9	0	10	5	1	10	55	230	10	40	18	67	0	32	14	0	22	0	0
5	0	0	0	0	0	0	1	198	0	0	0	0	0	28	2	29	0	0	0	1	120	1	0	3	0
6	0	234	0	0	0	10	1	0	0	5	0	2	41	62	111	17	24	41	376	178	18	2	15	0	3
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	42	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	184	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	13	0	0	0	2	0	0	0	0	0	38	60	49	0	30	0	0	0	25	4	0	0	9	0
10	0	1	0	0	0	0	0	0	0	0	0	0	0	15	0	0	0	0	0	0	4	0	0	0	0
11	0	0	0	0	0	16	1	0	0	0	0	0	0	0	2	93	162	45	0	0	0	0	30	0	0
12	0	1	0	0	2	0	1	0	0	0	0	0	0	0	0	0	0	2738	0	0	0	0	9	0	0
13	0	0	0	0	0	299	0	277	0	0	0	4	0	3	0	0	4	0	0	0	1	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	21	0	0	0	0	0
15	0	7	0	7	1	0	0	0	1	1	0	11	64	73	2	13	3	5	0	241	0	58	5	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0

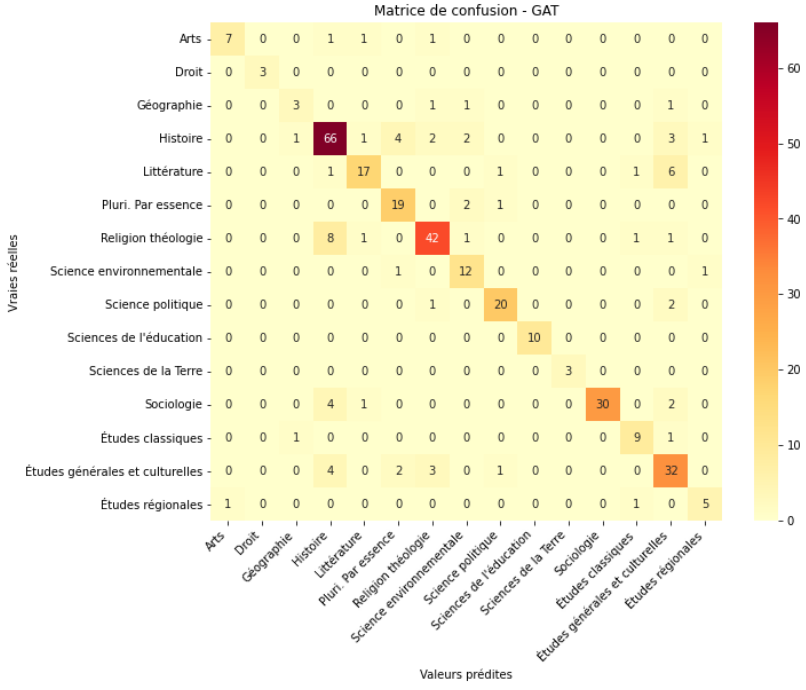
Modification cluster du clustering spectral à 214 au lieu de 17 :

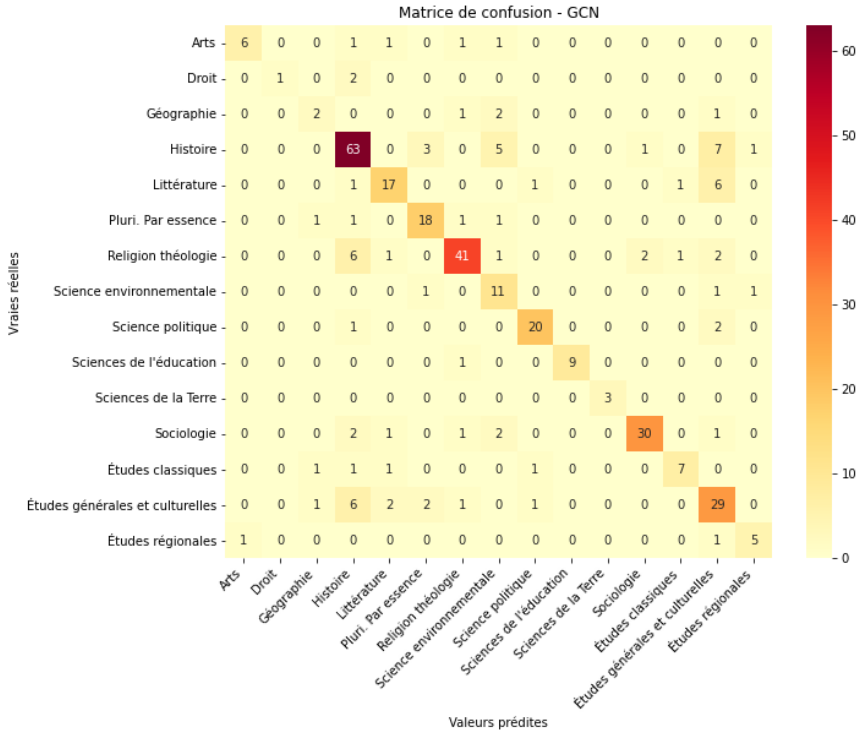
```
ARI Louvain vs Spectral (clean): 0.3421259417795706
AMI Louvain vs Spectral (clean): 0.7371889413622188

ARI Louvain vs Block Model (clean): 0.9918540244405937
AMI Louvain vs Block Model (clean): 0.9952636009434522

ARI Spectral vs Block Model (clean): 0.3458241487886765
AMI Spectral vs Block Model (clean): 0.740617803438982
```

Classification supervisée

Matrice de confusion







Métriques par catégorie

```
In [1371]: metric_rd
Out[1371]:
                          Catégorie  Précision    Recall  F1-score
0                              Arts   1.000000  0.900000  0.947368
1                             Droit   1.000000  1.000000  1.000000
2                        Géographie   1.000000  0.833333  0.909091
3                          Histoire   0.917647  0.975000  0.945455
4                       Littérature   1.000000  0.692308  0.818182
5                 Pluri. Par essence   1.000000  0.818182  0.900000
6                 Religion théologie   0.947368  1.000000  0.972973
7           Science environnementale   0.875000  1.000000  0.933333
8                  Science politique   0.958333  1.000000  0.978723
9            Sciences de l'éducation   1.000000  1.000000  1.000000
10              Sciences de la Terre   1.000000  1.000000  1.000000
11                         Sociologie   1.000000  1.000000  1.000000
12                  Études classiques   1.000000  1.000000  1.000000
13  Études générales et culturelles   0.933333  1.000000  0.965517
14                  Études régionales   0.857143  0.857143  0.857143
```

```
In [1372]: metric_GAT
Out[1372]:
                          Catégorie  Précision    Recall  F1-score
0                              Arts   0.875000  0.700000  0.777778
1                             Droit   1.000000  1.000000  1.000000
2                        Géographie   0.600000  0.500000  0.545455
3                          Histoire   0.785714  0.825000  0.804878
4                       Littérature   0.809524  0.653846  0.723404
5                 Pluri. Par essence   0.730769  0.863636  0.791667
6                 Religion théologie   0.840000  0.777778  0.807692
7           Science environnementale   0.666667  0.857143  0.750000
8                  Science politique   0.869565  0.869565  0.869565
9            Sciences de l'éducation   1.000000  1.000000  1.000000
10              Sciences de la Terre   1.000000  1.000000  1.000000
11                         Sociologie   1.000000  0.810811  0.895522
12                  Études classiques   0.750000  0.818182  0.782609
13  Études générales et culturelles   0.666667  0.761905  0.711111
14                  Études régionales   0.714286  0.714286  0.714286
```

```
In [1373]: metric_GCN
Out[1373]:
                          Catégorie  Précision    Recall  F1-score
0                              Arts   1.000000  0.600000  0.750000
1                             Droit   1.000000  0.333333  0.500000
2                        Géographie   1.000000  0.666667  0.800000
3                          Histoire   0.702128  0.825000  0.758621
4                       Littérature   0.764706  0.500000  0.604651
5                 Pluri. Par essence   0.750000  0.818182  0.782609
6                 Religion théologie   0.781818  0.796296  0.788991
7           Science environnementale   0.733333  0.785714  0.758621
8                  Science politique   0.869565  0.869565  0.869565
9            Sciences de l'éducation   1.000000  1.000000  1.000000
10              Sciences de la Terre   1.000000  1.000000  1.000000
11                         Sociologie   0.909091  0.810811  0.857143
12                  Études classiques   0.777778  0.636364  0.700000
13  Études générales et culturelles   0.595745  0.666667  0.629213
14                  Études régionales   0.714286  0.714286  0.714286
```