Network Analysis for Information Retrieval

JUNOY T – LEAL DE ALMEIDA G – M2MIASHS

Table des matières

Introduction

L'analyse des réseaux d'information est une branche cruciale de l'informatique et du machine learning qui se focalise sur la compréhension et l'exploitation des structures complexes et des données textuelles au sein des réseaux. Ces réseaux peuvent inclure des médias sociaux, des bases de données bibliographiques, et d'autres formes de réseaux informationnels où les données sont souvent représentées sous forme de texte et de relations entre les différents éléments ou nœuds du réseau. Le but principal est de modéliser ces réseaux de manière à extraire des connaissances utiles, ce qui peut inclure la classification de documents, la détection de communautés, la prédiction de liens, et la recherche d'informations pertinente.

Dans cette perspective, l'analyse des réseaux d'information englobe plusieurs défis techniques, notamment la représentation efficace des documents et la modélisation des relations entre eux. Les méthodes d'apprentissage de sujets et de représentation vectorielle des textes jouent un rôle central dans la résolution de ces défis, permettant de capturer l'essence sémantique des documents et la structure sous-jacente des réseaux.

La première section d'acquisition des données couvre le chargement des données, le calcul, l'affichage de quelques statistiques simple et graphique pour la visualisation de données.

La seconde partie de prise en compte de la structure du corpus, nous ferons quelques retraitement pour la préparation des données textuelles, nous utiliserons aussi des graphes pour visualiser et se faire idée de la topologie de vos données. Nous calculerons aussi quelques statistiques basées sur la structure des graphes.

Nous mettrons en œuvre un moteur de recherche d'article dans la troisième partie.

Dans la quatrième section, nous testerons un algorithmes de clustering classiques, le clustering spectral.

Enfin, en cinquième et dernière partie de ce travail, nous esseyrons de classifier les nos articles.

1. Acquisition des données

A. Introduction des données

Dans le cadre de ce projet, neuf fichiers de métadonnées bibliographiques au format .pickle ont été générés. Ces fichiers encapsulent les collections du programme Persée, connues pour leur vaste répertoire d'œuvres scientifiques numérisées et accessibles via data.persee.fr. En complément, un fichier .csv fait le lien entre les collections et les disciplines scientifiques correspondantes. Ce fichier de correspondance comprend deux colonnes essentielles : l'identifiant de la collection et le label de la discipline associée.

Persée, un portail dédié à la numérisation et à la diffusion du patrimoine scientifique francophone, se distingue par sa collection variée de documents. Ces derniers, allant d'articles à des comptes-rendus et bien d'autres types de publications, se concentrent principalement sur les sciences humaines et sociales. La période couverte s'étend du dix-neuvième siècle à nos jours, témoignant ainsi de l'évolution de la recherche dans ces domaines. Chaque document, initialement inclus dans un fascicule, fait partie d'une collection généralement associée à une revue scientifique.

Le jeu de données en question décrit plus de 900 000 documents. Toutefois, en raison de limitations liées à la capacité des ressources informatiques, l'analyse se limitera à un seul des neuf fichiers mentionnés. Ce fichier contient, pour chaque entrée documentaire, des informations telles que le titre, le sous-titre, les auteurs, et la date de publication. Lorsqu'elles sont disponibles, des informations supplémentaires comme un résumé, des mots-clés, ou une table des matières peuvent également être trouvées. En outre, le jeu de données intègre des relations de citation entre les documents, enrichissant ainsi l'analyse possible. La structure adoptée pour les champs à valeurs multiples sera conservée.

B. Chargement et description des données

L'analyse du premier fichier des disciplines révèle une certaine diversité avec près de 390 identifiants uniques correspondant à 26 disciplines différentes. Ce premier nuage de mot (Image 1) met en évidence l'histoire et la sociologie comme étant les disciplines les plus représentées, reflétant la richesse et la variété des sujets couverts par le corpus de données de Persée.

Image 1 - Nuage de mots des disciplines

Le jeu de données sélectionné pour cette étude est "20240125_dataset_recod_rqdi.pickle". Il contient pas moins de 154 484 documents répartis et possède 315 colonnes, qui ne sont pas toutes remplies. Parmi ces documents, 26 140 auteurs uniques ont été identifiés. Les cinq auteurs les plus prolifiques se distinguent par un nombre considérable de documents à leur actif, le plus actif ayant contribué à 1 242 documents, suivi de près par les autres dans une fourchette de 629 à 842 contributions.

Les documents sont datés de 1872 à 2018 et la distribution temporelle des documents permet de mettre en avant cinq années particulièrement productives, avec 1997 en tête, comptabilisant 2 406 documents, suivie de près par 1999 et 1994.

Quant à la nature des documents, les revues et articles représentent les types les plus fréquents, soulignant l'importance de ces formats dans la diffusion du savoir académique. Ce second nuage de mots (Image 2) illustre clairement la prédominance de ces supports dans le corpus.

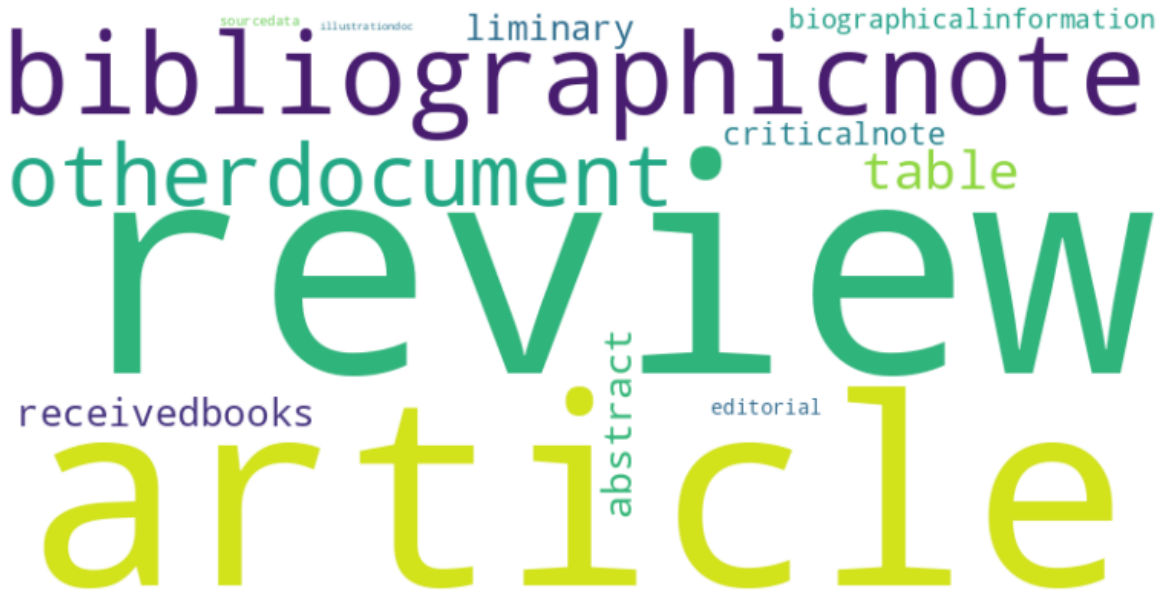

Image 2 - Nuage de mots des type de documents

Enfin, les thèmes abordés au sein de cette collection sont variés, avec une prédominance notable de la science, de l'éducation, de la religion et de la théologie. Ce troisième nuage de mots dévoile les préoccupations majeures des auteurs et des disciplines représentées, offrant un aperçu de la diversité thématique du corpus.

Image 3 - Nuage de mots des sujets

2. Prise en compte de la structure du corpus

Face aux limitations matérielles ne permettant pas de traiter l'intégralité des plus de 154 000 documents de notre corpus, nous avons opté pour un échantillon aléatoire représentant 0,05 % de l'ensemble. En supposant une répartition homogène des données, cet échantillon devrait, selon la loi des grands nombres, offrir une représentation fidèle de la diversité et des caractéristiques générales du dataset complet. Cependant, nous sommes conscients que des informations rares ou des cas atypiques, cruciaux pour notre analyse, pourraient être omis. Faute d'alternative, nous avons dû accepter cette contrainte. Désormais, notre dataset réduit se concentre sur des colonnes spécifiques telles que les auteurs, les références, les titres, le type de document, les sujets, et les identifiants des documents, résultant en un nouveau jeu de données comprenant 772 entrées et 68 colonnes.

Dans notre analyse, nous avons élaboré un graphe orienté (Image 4), où les nœuds représentent les documents et les arêtes symbolisent les citations entre eux. La distribution des degrés de ce graphe varie considérablement, suggérant une hétérogénéité dans la connectivité des documents. Nous observons une multitude de composantes connexes, indiquant que notre graphe est constitué de nombreux sous-ensembles interconnectés de manière autonome, sans liens directs avec d'autres groupes du graphe. Cette fragmentation se reflète également dans une densité de graphe extrêmement basse, mettant en lumière un réseau où la majorité des paires de nœuds ne sont pas connectées directement.

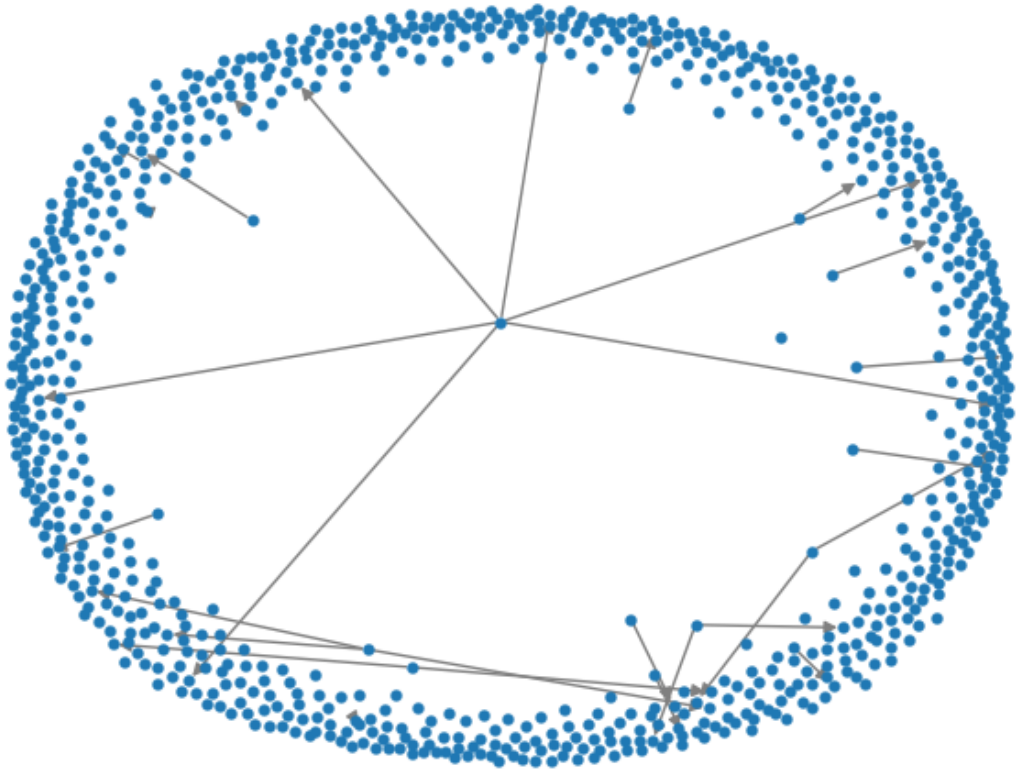

Image 4 - Graphe orienté Documents - Citations

Les résultats spécifiques de notre analyse montrent que bien que la majorité des articles ne soient pas directement connectés entre eux, il existe une petite proportion de documents qui jouent un rôle clé dans la connectivité du réseau, agissant comme des ponts entre différentes parties du graphe. Cette situation souligne l'existence de clusters isolés ainsi que de quelques nœuds hautement connectés.

Dans le prolongement de notre analyse sur les interactions documentaires, nous avons intégré la notion de type de document dans un graphe dédié, afin de déterminer si une corrélation existe entre le type de document et sa fréquence de citation. Cependant, l'analyse visuelle n'a pas révélé de lien direct, indiquant que le type de document n'influence pas significativement sa probabilité d'être cité.

Nous avons également exploré la structure de collaboration entre auteurs à travers un graphe orienté autour des notions d'auteur et de co-auteur. La distribution des degrés dans ce graphe montre une grande majorité de nœuds avec très peu de connexions, indiquant que nombreux sont les auteurs qui collaborent dans des cadres restreints. Néanmoins, quelques nœuds se détachent par un nombre significativement plus élevé de liens, reflétant des auteurs ayant une large étendue de collaborations.

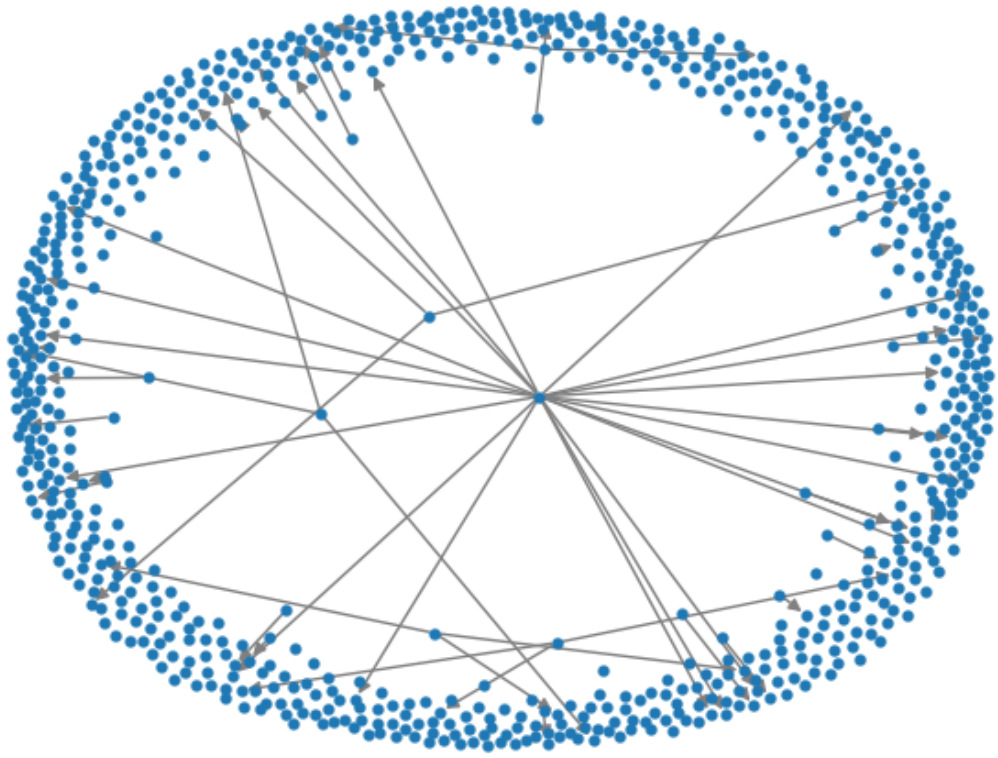

Image 5 - Graphe orienté Auteurs - Co-Auteurs

Les autres graphes, et visuels générés, sont disponibles en annexe. (,)

3. Moteur de recherche

Pour construire notre moteur de recherche, nous avons commencé par un nettoyage des données. Ce processus implique plusieurs étapes essentielles. Tout d'abord, nous sélectionnons uniquement la colonne contenant les titres des documents dans notre DataFrame. Ensuite, nous convertissons tous les titres en chaînes de caractères, puis en minuscules, pour uniformiser le format des données et réduire les variations dues à la casse. La prochaine étape consiste à éliminer les mots sans importance sémantique significative, ou "stopwords", du texte. Cela se fait en tokenisant chaque titre, c'est-à-dire en le découpant en mots individuels, puis en filtrant ceux qui sont considérés comme des stopwords en français. Nous avons décidé de ne pas supprimer la ponctuations, car certain titre couvre des périodes et sont notées comme suit : « XXXX-XXXX » en retirant la ponctuation nous supprimerions l'information date de certain documents. Ce qui pourrait réduire l'efficacité de la recherche.

Après la préparation des données, nous avons choisi d'utiliser TF-IDF (Term Frequency-Inverse Document Frequency) couplé à la similarité cosinus et la distance euclidienne pour notre moteur de recherche. TF-IDF est une technique statistique qui évalue l'importance d'un mot dans un document par rapport à une collection de documents. Elle combine la fréquence du terme (TF), qui mesure la fréquence d'apparition d'un mot dans un document, avec la fréquence inverse du document (IDF), qui diminue l'importance des mots apparaissant dans de nombreux documents du corpus. La similarité cosinus, quant à elle, est une mesure utilisée pour calculer à quel point deux documents sont similaires en traitant leurs textes comme des vecteurs dans un espace multidimensionnel, où l'angle entre les deux vecteurs reflète le degré de similarité. La distance euclidienne, en revanche, mesure la distance "directe" entre deux points (ou documents) dans cet espace vectoriel, offrant une alternative pour évaluer la dissimilarité entre les documents.

Une fois la transformation des titres en un espace vectoriel effectuée, nous pouvons rechercher un mot-clé en le transformant également selon le modèle TF-IDF. Ensuite, selon le choix de l'utilisateur entre la similarité cosinus et la distance euclidienne, nous calculons soit la similarité (cosinus) entre le vecteur de la requête et les vecteurs des titres, soit la distance (euclidienne) négative pour trier les résultats en ordre décroissant de pertinence. Les titres des documents les plus pertinents sont ensuite affichés, basés sur ces scores, permettant à l'utilisateur d'obtenir rapidement une vue d'ensemble des documents les plus proches de sa requête.

Image 6 - Résultats du moteur de recherche

4. Le Clustering

Pour le clustering, nous avons opté pour un clustering spectral, une technique qui se base sur les propriétés spectrales (valeurs et vecteurs propres) de la matrice de similarité des données. Le clustering spectral est particulièrement efficace pour identifier des structures complexes au sein des données en se fondant sur les relations de proximité entre points, ce qui le distingue des méthodes basées sur la distance euclidienne classique, comme le K-means. Cette méthode est souvent couplée avec des techniques comme le K-nearest neighbors (kNN) pour construire la matrice de similarité, permettant ainsi de capturer les relations locales entre les points.

Nous avons d'abord commencé par préparer nos données à l'aide du vectoriseur TF-IDF, configuré pour filtrer les mots très fréquents et ceux trop rares, et limiter l'espace de caractéristiques à 10 000. Ceci afin de réduire le bruit et de se concentrer sur les termes les plus significatifs pour l'analyse. Après avoir transformé les titres des documents en vecteurs TF-IDF, nous avons procédé à une visualisation naïve en utilisant UMAP (Uniform Manifold Approximation and Projection) pour une projection en deux dimensions. L'objectif était de se faire une idée visuelle de la distribution des données et d'identifier un nombre optimal de clusters.

Sur ce premier graphe (Image 7), nous avons observé une projection homogène des données, rendant la tâche de sélection de clusters compliquée. Cependant, nous avons opté pour la formation de 3 clusters, basée sur cette visualisation. Pour affiner notre clustering, nous avons ensuite appliqué le clustering spectral en utilisant kNN pour déterminer les relations de voisinage. La performance des clusters formés a été évaluée à l'aide du silhouette score, qui mesure la qualité de l'assignation des points à leurs clusters en fonction de la proximité et de la séparation entre clusters. Nous avons choisi de former 2 clusters kNN en fonction des résultats graphiques obtenus.

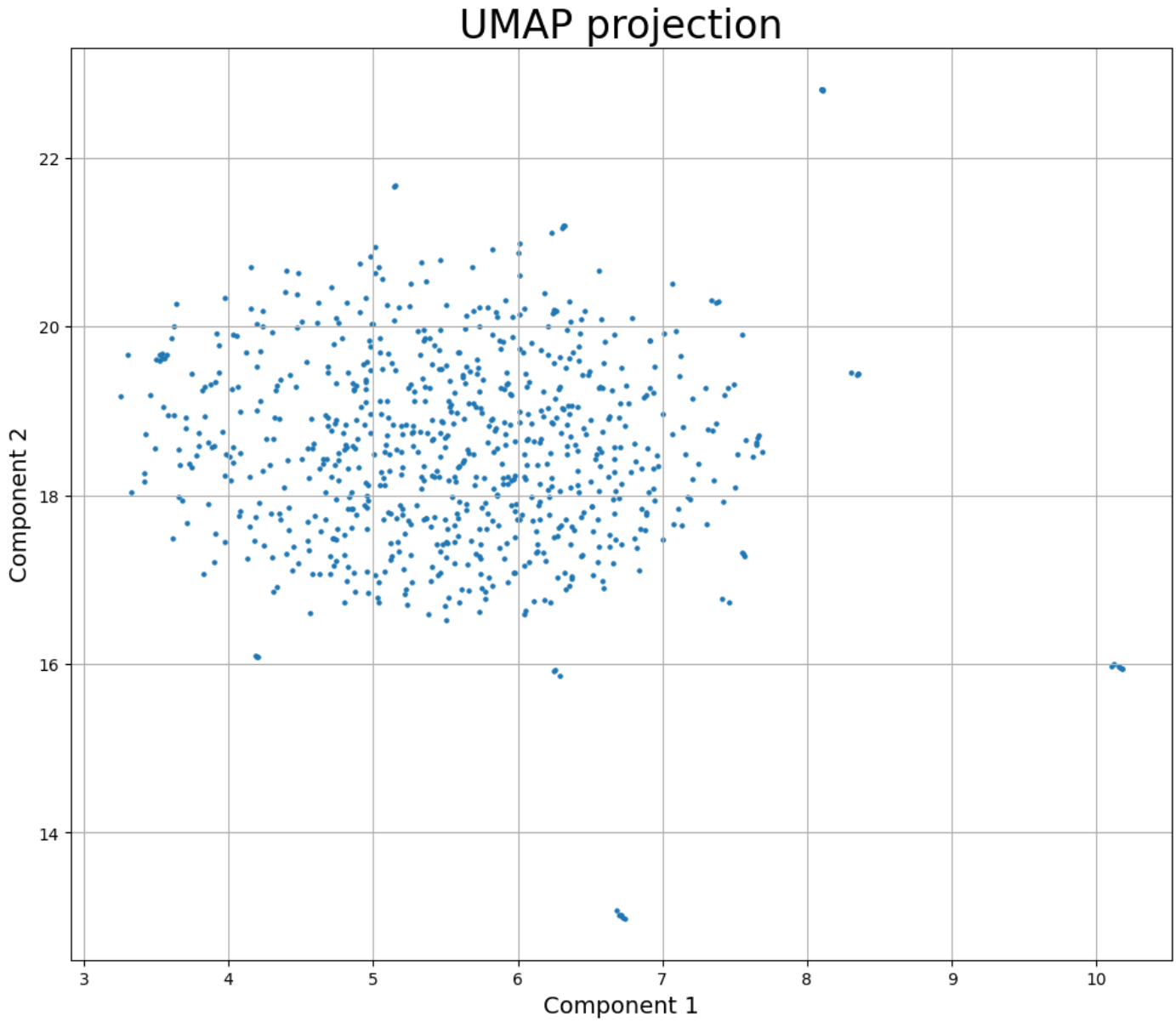

Image 7 - Projection en 2 dimensions de la visualisation naïve UMAP

Nous avons ensuite mis en application du clustering spectral et la projection des catégories calculées sur une nouvelle visualisation TSNE (t-distributed Stochastic Neighbor Embedding), une autre technique de réduction de dimension pour la visualisation de données à haute dimension. Les points sont colorés différemment selon leur appartenance à l'un des clusters identifiés, offrant ainsi une représentation visuelle claire de la segmentation des données. Cette visualisation (Image 8) montre un cluster central (orange) et deux autres populations de clusters (gris et rouge) gravitant autour, indiquant la présence de sous-groupes distincts au sein des données, chacun avec ses propres caractéristiques.

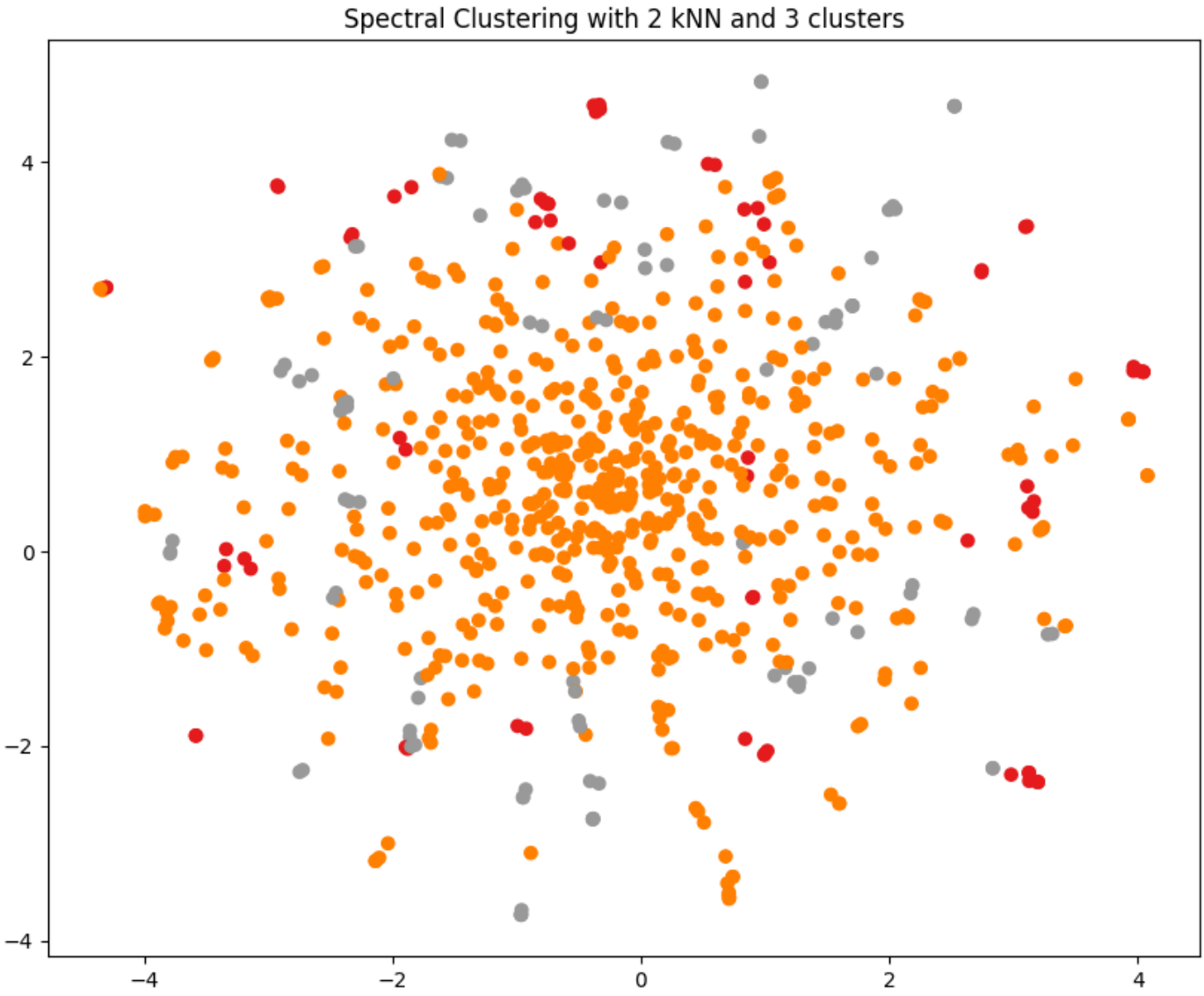

Image 8 - TSNE Visualisation des clusters

5. Tache de Classification supervisée

En préparation à la tâche de classification, où l'objectif était de prédire le sujet en fonction du titre, nous avons d'abord augmenté la taille de l'échantillon pour permettre une division en ensembles d'entraînement et de test. Le nouveau dataset contient 0,5% du dataset originel soit 7724 lignes et 68 colonnes. Pour extraire des caractéristiques significatives des titres, nous avons utilisé BERT (Bidirectional Encoder Representations from Transformers), un modèle de traitement du langage naturel pré-entraîné capable de générer des représentations vectorielles (embeddings) riches en contexte pour le texte. Nous commençons par initialiser un tokenizer et un modèle BERT pour convertir chaque titre en un ensemble d'embeddings correspondant au token initial ([CLS]), réputé pour capturer l'essence globale de la phrase.

En plus des embeddings BERT, nous avons également vectorisé les titres en utilisant TF-IDF, mais avons finalement choisi d'utiliser uniquement les embeddings BERT pour la classification. Pour la classification, un modèle de forêt aléatoire (Random Forest Classifier) a été entraîné sur les embeddings BERT des titres. Ce choix de modèle est motivé par sa capacité à gérer efficacement les données de haute dimension et à fournir des résultats robustes même en présence de bruit dans les données. Les données ont été divisées en ensembles d'entraînement et de test, permettant ainsi d'évaluer la performance du modèle sur des données non vues.

Les résultats montrent une précision globale de 52.168%, ce qui indique une performance modérée du modèle. Les rapports de classification détaillés par sujet révèlent des variations significatives en termes de précision, de rappel et de score F1, avec certains sujets comme "Hist. sc. techniques" et "Linguistique" obtenant des scores particulièrement bas en rappel, signifiant que le modèle a du mal à identifier correctement ces catégories. En revanche, des sujets comme "Droit" et "Science education" affichent des performances relativement meilleures, soulignant la variabilité de la capacité du modèle à prédire différents sujets basés sur les titres.

Ces variations de performance peuvent être attribuées à plusieurs facteurs, notamment la distribution inégale des sujets dans le dataset, la qualité des embeddings BERT pour représenter le contenu sémantique des titres, et les limitations inhérentes du modèle de forêt aléatoire face à des données textuelles complexes. La représentation des résultats sur un graphe a visé à fournir une visualisation des clusters de sujets prédits, mais il semble que cette approche n'ait pas produit des insights probants, soulignant les défis associés à la visualisation de données de haute dimension et à l'interprétation des modèles de classification complexes.

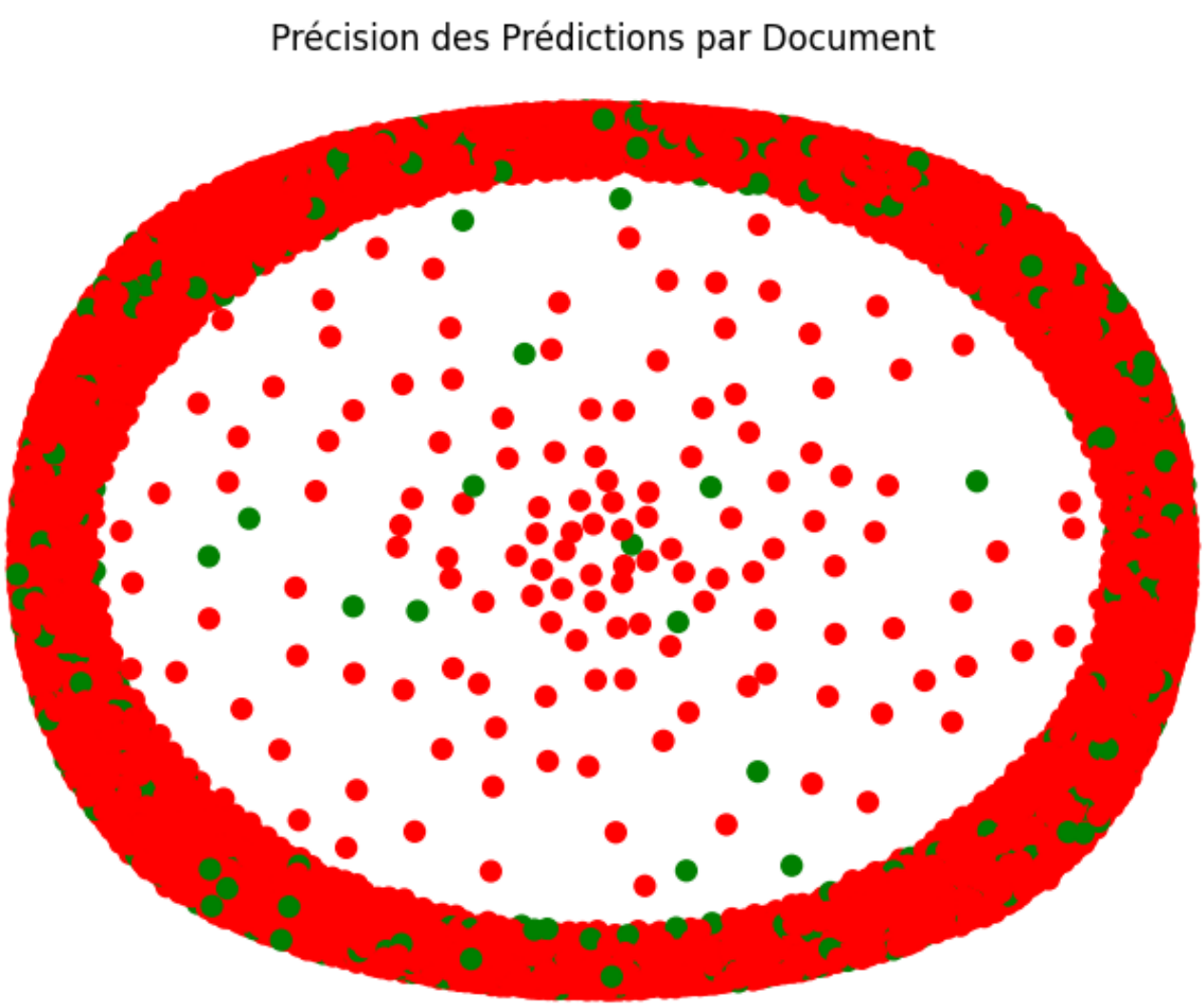

Image 9 - Prédictions des sujets

Conclusion

L'objectif de ce projet était de déchiffrer les structures complexes et les données textuelles au sein d'un corpus volumineux, ici, issu du programme Persée, englobant une richesse de documents scientifiques. Mais aussi de découvrir et de tester de nouvelles méthodes d'analyses. L'analyse initiale a fourni une vue d'ensemble des documents, certaines caractéristiques. Cette phase a également mis en lumière les défis posés par la taille du corpus, conduisant à l'adoption d'un échantillon pour les analyses ultérieures.Nous avons pu visualiser les connections complexes et identifier des clusters isolés ainsi que des nœuds de connexion clés, pour une compréhension plus profonde de la dynamique du réseau d'information.
Un moteur de recherche basé sur TF-IDF, la similarité cosinus, et la distance euclidienne a ensuite été réalisée. L'application du clustering spectral a offert une méthode pour regrouper les documents en fonction de leur contenu sémantique, malgré les défis liés à la sélection du nombre optimal de clusters. Enfin, la tâche de classification supervisée, bien que confrontée à des performances modérées, a illustré le potentiel des embeddings BERT et des modèles de forêt aléatoire pour prédire le sujet des documents basé sur leurs titres.

La démarche exploratoire et analytique vu dans notre projet souligne l'importance et la complexité de traiter et d'analyser de vastes ensembles de données textuelles. Malgré les avancées technologiques et les méthodes de pointe employées, nous pouvons réfléchir à quelques axes d'amélioration et d'exploration pour enrichir notre compréhension et notre capacité à extraire des connaissances utiles de ces données. Tout d'abord, l'extension de la capacité de traitement pourrait permettre d'inclure une plus grande partie du corpus, offrant une vue d'ensemble plus complète et potentiellement révélant des patterns et des relations jusqu'alors inaperçus. Ensuite, l'extension des analyses à d'autres aspects des données, tels que l'évolution temporelle des thèmes de recherche, pourrait offrir de nouvelles perspectives sur la dynamique des champs représentés dans le corpus. Cela pourrait inclure des études longitudinales ou des analyses de séries temporelles pour suivre l'évolution des sujets de recherche au fil du temps.

Annexes

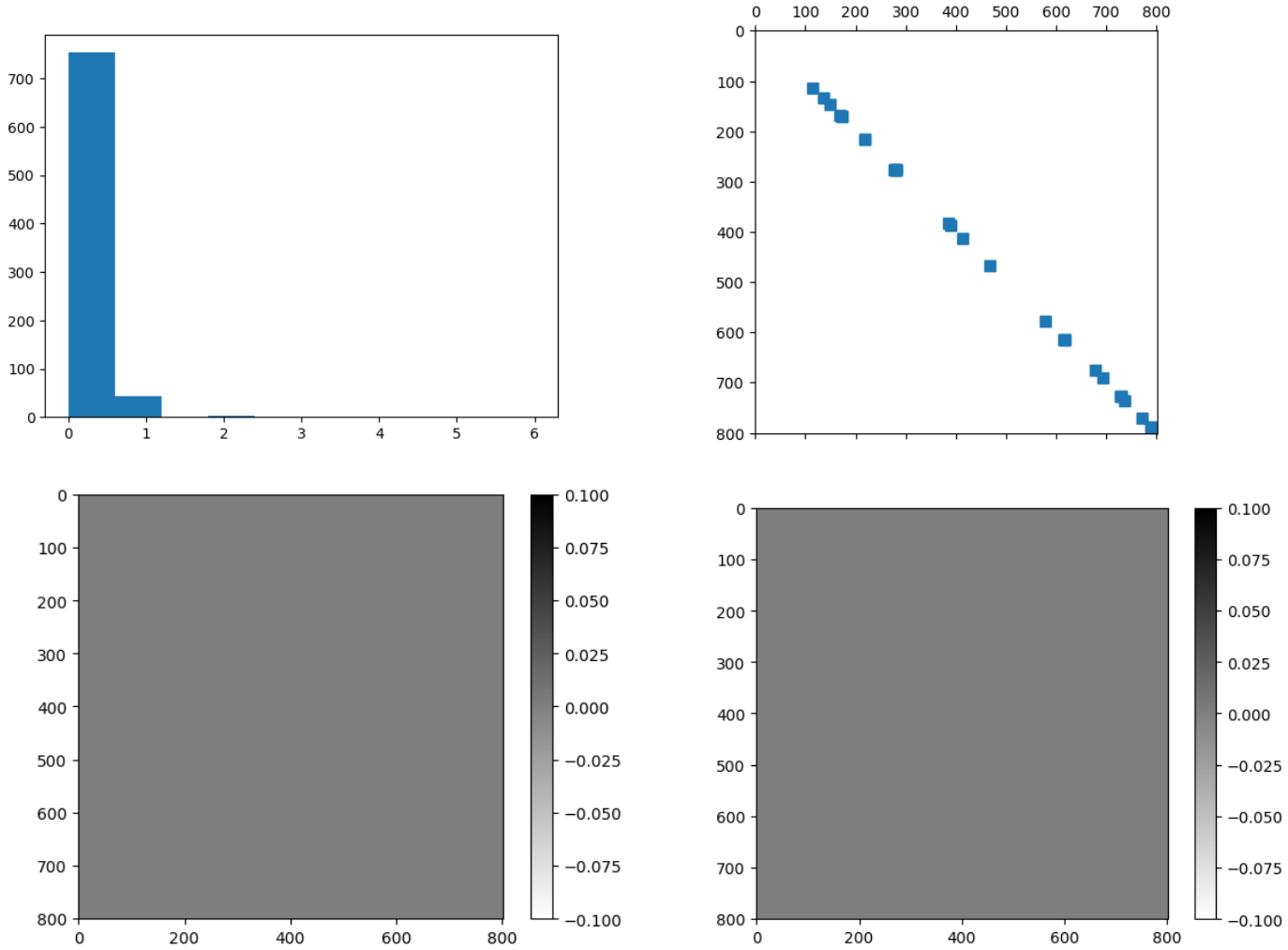

Image 10 - Visualisation des propriétés structurelles du graphe Document - Citations

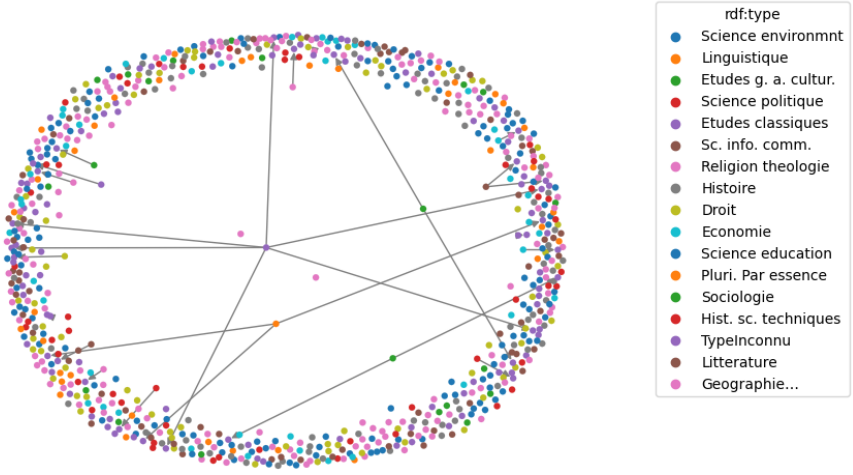

Image 11 - Image 12 - Graphe orienté Document – Citations par type de doucments

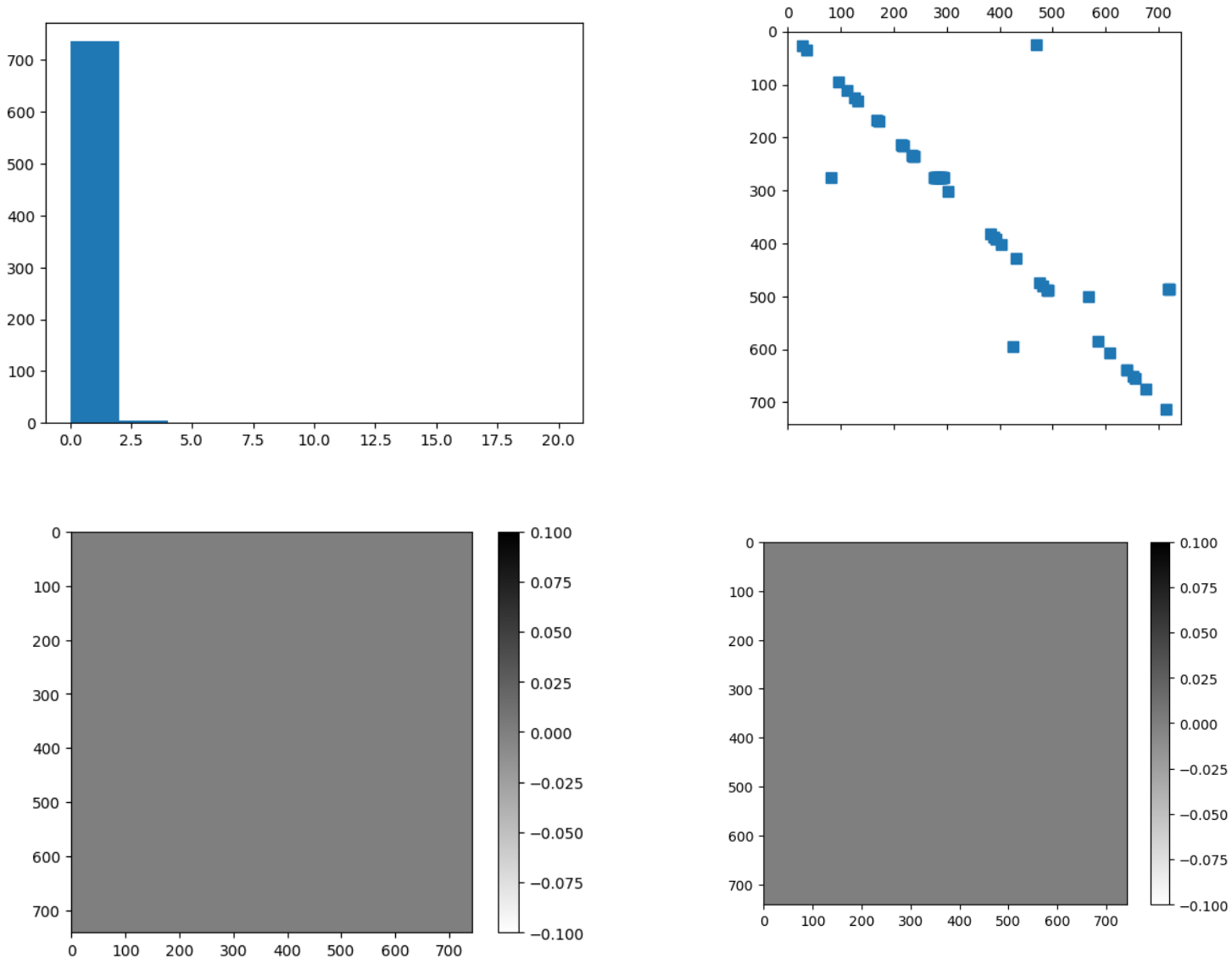

Image 13 – Visualisation des propriétés structurelles du graphe Auteurs - Co-Auteurs

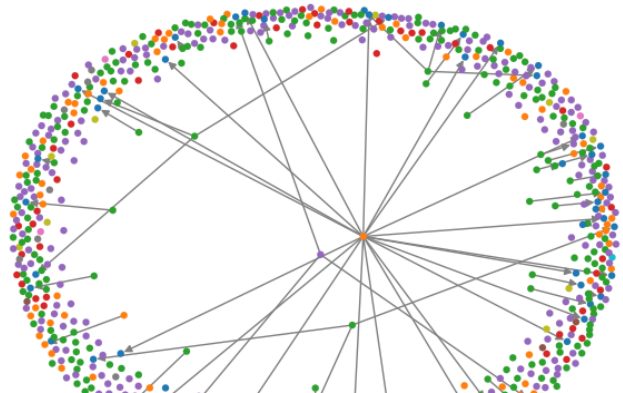

Image 14 - Graphe orienté Auteurs - Co-Auteurs par type de documents

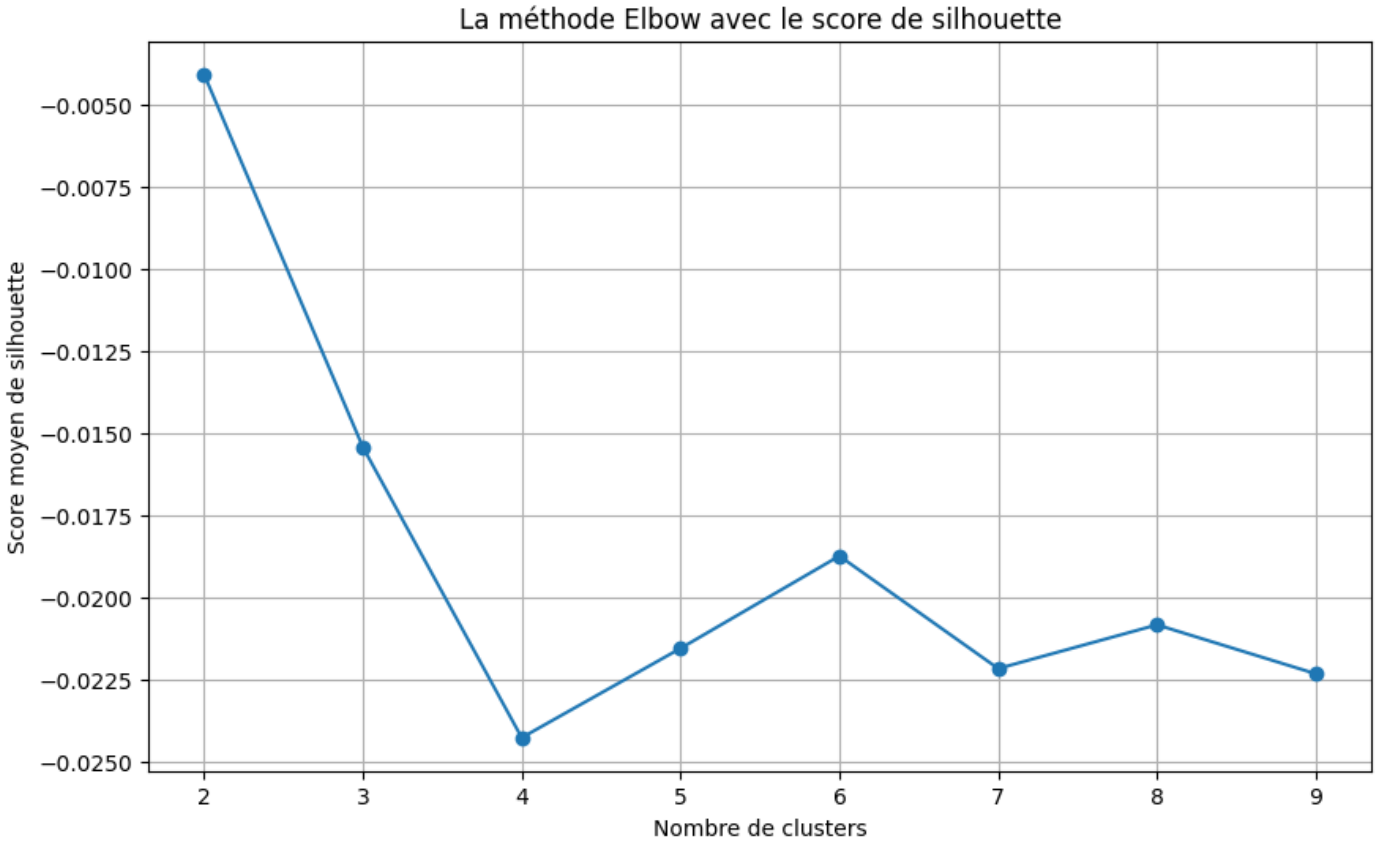

Image 15 – Score de silhouette