Projet - Network Analysis for Information Retrieval

Lea Dahmani, Akouvi Tamatekou Projet M2 MIASHS

March 2024

Ce projet explore l'utilisation de techniques d'apprentissage automatique avancées pour le traitement et l'analyse de larges corpus documentaires, en l'occurrence ceux de Persee. En combinant des méthodes de NLP et d'apprentissage automatique avec une approche analytique méticuleuse, nous visons à doter la communauté académique d'un outil puissant qui facilite la découverte de connaissances et la compréhension des dynamiques de recherche. L'approche adoptée se concentre sur l'exploitation des embeddings de mots générés par BERT (Bidirectional Encoder Representations from Transformers), un modèle de langage pré-entraîné conçu pour capturer une large gamme de contextes et de nuances linguistiques.

Le projet a débuté par la collecte et la préparation d'un ensemble de données, consistant en les 100 premiers titres d'articles issus de neuf fichiers de données distincts, représentant divers domaines de recherche. Chaque titre a été transformé en un embedding dense via le modèle BERT, capturant ainsi les caractéristiques sémantiques profondes de chaque séquence de texte.

Avant de plonger dans les complexités de cette exploration, nous entreprendrons des analyses statistiques descriptives, y compris la génération de graphiques, pour saisir la distribution et les caractéristiques intrinsèques de notre corpus provenant de Persee.

Table des matières

1 Introduction **3**

2 Fonction 1 - Acquisition des données **4**
2.1 Analyse Exploratoire des données 4

3 Fonction 2 - Prise en compte de la structure du corpus **7**
3.1 Établissement de la Matrice d'Adjacence 7
3.2 Mesures de centralité 9
3.2.1 Distribution des degrés 9
3.2.2 La centralité eigenvector 10
3.2.3 Centralité de second ordre 11

4 Fonction 3 - Moteur de recherche **12**

5 Fonction 4 - Ajout de clustering **13**
5.1 Le clustering 13
5.2 Clustering Spectral 13
5.3 Clustering de Louvain 15

6 Fonction 5: Classification supervisée **15**

7 Conclusion **18**

1 Introduction

Le traitement et l'analyse de données textuelles représentent un domaine de recherche dynamique et en constante évolution, particulièrement dans le contexte académique où l'abondance des publications crée à la fois des opportunités et des défis. À travers ce projet, nous nous aventurons au cœur de cette problématique en mettant en avant une stratégie innovante pour l'organisation, la recherche, et l'analyse de documents académiques, s'appuyant sur un corpus substantiel provenant de Persee.fr, un portail de référence pour la numérisation et la diffusion du patrimoine scientifique francophone. En puisant dans cette ressource précieuse, notre initiative cherche à transcender les approches traditionnelles centrées sur les analyses des réseaux de citations, pour se focaliser sur l'exploration de la sémantique profonde des titres des documents.

En vue d'approfondir notre analyse et d'assurer une interprétation robuste des clusters formés, nous entamons notre exploration par l'agrégation de segments variés provenant de Persee, en mettant un point d'honneur à intégrer des échantillons thématiquement diversifiés.

Cette approche stratégique visant à introduire une richesse thématique dans notre corpus, renforce la solidité de notre méthode de clustering.

Au-delà du clustering, la classification des documents et le développement d'un moteur de recherche avancé, exploitant les séquences sémantiques des titres, se profilent comme des étapes cruciales de notre projet. Ces démarches visent à améliorer significativement l'accès aux documents pertinents et à faciliter la classification automatique des nouvelles entrées dans le corpus de Persee. Ces tâches, bien qu'exigeantes en termes de précision et d'analyse des données, promettent d'ajouter une valeur considérable à notre système.

Il est important de noter que, pour des raisons de contraintes de mémoire et d'efficacité de compilation, nous avons restreint notre analyse à seulement 30% de l'ensemble du corpus. Cette sélection a été faite de manière à préserver l'intégrité des tendances globales des données tout en garantissant la maniabilité du traitement informatique.

2 Fonction 1 - Acquisition des données

2.1 Analyse Exploratoire des données

Pour l'analyse exploratoire du corpus nous avons extraie des colonnes spécifiques qui incluent l'identifiant du document, le titre, l'abstract en français, les informations sur les auteurs et les citations.

Ces éléments ont été sélectionnés pour leur pertinence dans l'analyse des documents académiques. Chaque jeu de données extrait a été ajouté à une liste qui, une fois le processus de parcours terminé, a été combinée en un DataFrame unique. Cette consolidation a permis de créer une vue intégrée et exhaustive du corpus pour faciliter les opérations d'analyse.

La deuxième étape de l'analyse exploratoire a consisté à renommer les colonnes du DataFrame consolidé pour améliorer la lisibilité et simplifier l'accès aux données.

Un mapping des noms de colonnes a été établi, remplaçant les intitulés techniques issus des standards de métadonnées bibliographiques par des noms plus courts et intuitifs. Les identifiants RDF ont été transformés en désignations simplifiées, telles que 'titre', 'abstract_fr','auteur_0' , 'cite_par_0' rendant le DataFrame plus accessible à l'utilisation.

Ces deux étapes fondamentales de préparation des données ont jeté les bases nécessaires pour l'analyse exploratoire détaillée du corpus. Elles nous ont permis d'établir une structure de données propre et bien organisée, cruciale pour la suite du projet qui impliquera des visualisations de données, des analyses statistiques, et potentiellement l'application de modèles de machine learning pour extraire des insights plus profonds du corpus.

Dans la continuation de notre analyse exploratoire sur le corpus académique extrait de Persee.fr, nous avons procédé à une évaluation quantitative approfondie des données dont la synthèse a révélé une variété dans les données.

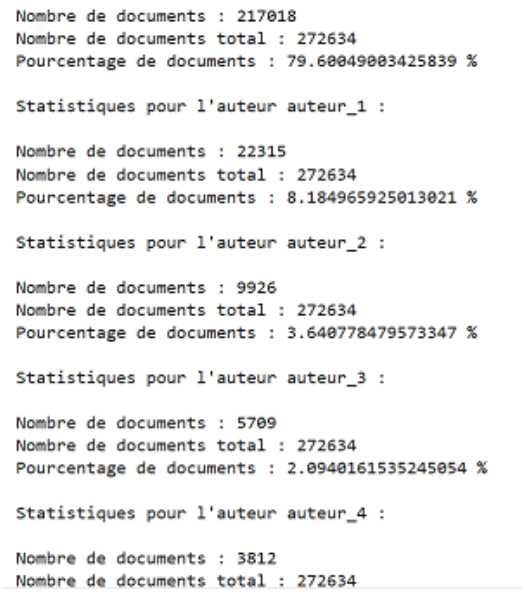


Figure 1: Nombre de document par auteur

L'analyse a démontré une riche diversité au sein des colonnes, avec un nombre élevé d'identifiants uniques et de titres.

Un autre aspect crucial de l'analyse a été l'identification des valeurs manquantes, révélant une absence notable d'abstracts pour une grande partie du corpus et une déperdition d'informations concernant les auteurs et les citations. Cette observation indique que, bien que le dataset soit riche et varié, il existe des lacunes significatives dans les métadonnées disponibles, en particulier pour les abstracts et les auteurs secondaires.

Pour continuer nous avons entrepris une démarche systématique pour quantifier et examiner la contribution de chaque auteur au sein de notre corpus académique consolidé. L'objectif est de dégager une compréhension plus fine de la répartition des documents entre les différents auteurs et d'apprécier l'étendue de leur participation.

Les 2 graphiques obtenus illustrent de manière frappante la distribution des contributions des auteurs au sein d'un vaste corpus académique. Dans le premier graphique, nous observons une barre très élevée pour l'auteur 0, indiquant qu'il est de loin l'auteur le plus prolifique du corpus, avec un nombre de documents s'élevant à 217 018. La prédominance de cet auteur est telle qu'elle éclipse considérablement celle des autres auteurs, dont les contributions, bien que non négligeables, sont nettement moins volumineuses.

Le deuxième graphique, représentant le pourcentage de documents par auteur, réaffirme cette observation. L'auteur 0 apparaît comme un cas atypique, représentant environ 79.6% du corpus, une proportion massivement plus grande comparée à celle des autres auteurs. La courbe chute brutalement après l'auteur 0 et se stabilise à des niveaux beaucoup plus bas pour les auteurs suivants, ce qui suggère que le corpus est dominé par un nombre limité d'auteurs principaux.

En somme, ces visualisations offrent une perspective quantitative sur l'influence et la présence des auteurs dans le corpus , et elles mettent en évidence l'importance de tenir compte de la contribution individuelle des auteurs lors de l'analyse des tendances et des schémas de publication dans les sciences humaines et sociales.

Pour plus d'analyse approfondie , nous avons fait le graphique du nombre de documents par date de publication

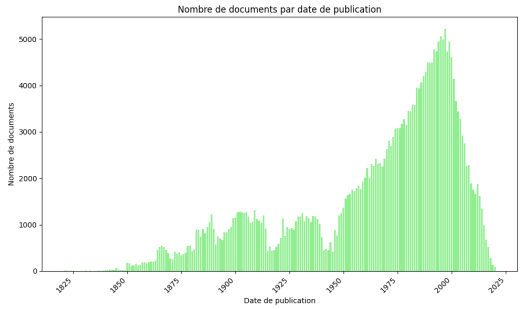


Figure 2: Nombre de documents par date de publication

L' histogramme en vert trace le nombre de documents publiés entre 1825 et 2025. La fréquence de publication s'accroît au fil du temps, avec le nombre le plus élevé de documents apparaissant juste avant 2025. Le graphique indique des périodes de croissance et de baisse, avec une tendance nettement ascendante à mesure que l'on se rapproche de l'année 2025.

Nous avons ensuite réalisé un code nous fournissant des données représentant le nombre d'auteurs par année de publication pour une série d'années sélectionnées,

s'étendant du 19ème siècle à la première moitié du 21ème siècle. Ces données illustrent une tendance variable du nombre d'auteurs au cours du temps, avec des pics significatifs en certaines années comme en 1996 avec 3883 auteurs, et des valeurs plus faibles dans d'autres, comme en 1917 avec 210 auteurs.

Sur la base de ces informations, nous avons développé un code pour visualiser le nombre d'auteurs en fonction de leur année de publication qui nous renvoie ce graphique

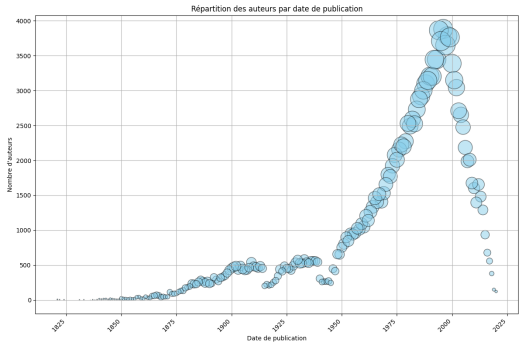


Figure 3: Répartition des auteurs par date de publication

Le graphique ci-dessus est un nuage de points où chaque point représente le nombre d'auteurs pour une année donnée, avec la taille de chaque point proportionnelle au nombre d'auteurs. L'axe horizontal montre les années de 1825 à 2025 et l'axe vertical indique le nombre d'auteurs, allant de 0 à environ 4000. On observe que la densité des points et leur taille augmentent significativement vers les années récentes, indiquant une hausse du nombre d'auteurs par année de publication au fil du temps. Il y a quelques années avec des pics particulièrement élevés, correspondant probablement aux années où le nombre d'auteurs était le plus grand.

Suite à l'analyse exploratoire détaillée du corpus de données. Nous nous focaliserons dès à présent sur l'analyse des titres des documents.

3 Fonction 2 - Prise en compte de la structure du corpus

3.1 Établissement de la Matrice d'Adjacence

Dans notre cas la matrice d'adjacence est un tableau bidimensionnel qui capture la proximité sémantique entre les documents basée sur leurs titres. Chaque élément de la matrice représente le degré de similitude entre les titres

de deux ou de plusieurs documents : une valeur proche de 1 indique une forte similité sémantique, tandis qu'une valeur proche de 0 indique peu ou pas de relation.

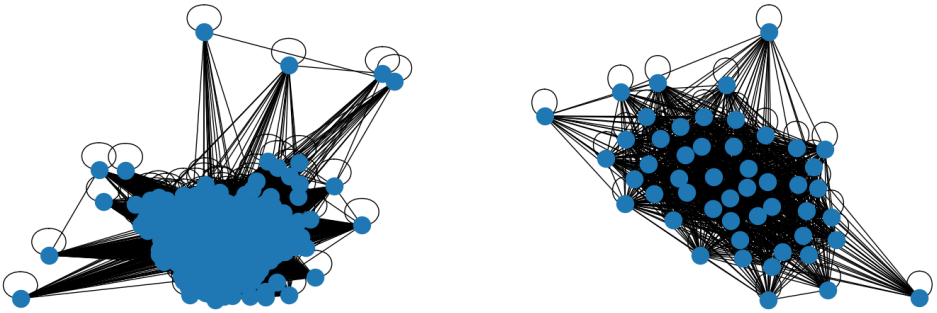
Pour créer cette matrice nous avions commencé par l'application du modèle BERT pour convertir les titres textuels des documents en vecteurs numériques, qui sont ensuite utilisés pour calculer une matrice de corrélation, reflétant la similarité sémantique entre les paires de documents.

```
# construction d'une matrice de correlation entre -1 et 1
corr = (encoded_sequences @ encoded_sequences.t())
max = torch.max(corr)
min = torch.min(corr)
adjacence = torch.relu(((corr - min) / (max - min)) * 2 - 1)
```

Les valeurs obtenues sont normalisées pour tomber dans un intervalle allant de -1 à 1, offrant ainsi une échelle standardisée pour mesurer la similarité. Ensuite, toute similitude négative, qui ne serait pas pertinente pour notre analyse, est filtrée à l'aide de la fonction ReLU, résultant en une matrice d'adjacence où seules les similarités positives sont maintenues, indiquant des liens sémantiques significatifs entre les documents.

```
G = nx.from_numpy_array(np.matrix(adjacence))
nx.draw(G)
```

Nous avons ensuite fait la transformation de la matrice d'adjacence calculée à partir des similarités sémantiques entre documents en un graphe réseau. La bibliothèque NetworkX est utilisée ici pour créer le graphe à partir de cette matrice, permettant ainsi de visualiser les relations entre les titres des documents comme un réseau où les titres des documents sont les nœuds et les similarités sémantiques sont les arêtes. La fonction draw est ensuite utilisée pour produire une visualisation de ce réseau.



(a) Graphe matrice d'adjacence (b) Le graphe des 50 premiers noeuds

Figure 4: Graphiques de matrice d'adjacence et des premiers noeuds

Le graphe est une visualisation d'un réseau, où les nœuds représentent des documents individuels, et les arêtes entre eux dénotent la similarité sémantique de leurs titres. Le graphe montre une concentration dense de liens ou d'arêtes au centre, cela indique que de nombreux documents ont des titres sémantiquement similaires, formant un cluster fortement interconnecté.

3.2 Mesures de centralité

La mesure de centralité permet d'évaluer l'importance ou l'influence de chaque document dans le réseau.

Il existe plusieurs types de mesures de centralité :

3.2.1 Distribution des degrés

- La Centralité de Degré : Elle mesure le nombre de connexions (arêtes) qu'un document (nœud) a avec d'autres documents.

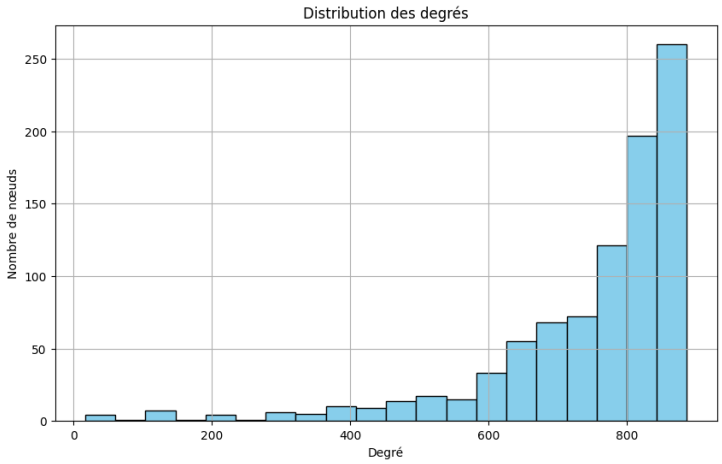


Figure 5: Histogramme de distribution des degrés

Le graphique ci dessus est un histogramme qui représente la "Distribution des degrés". L'axe horizontal (abscisses) est étiqueté "Degré" et l'axe vertical (ordonnées) est étiqueté "Nombre de nœuds". L'axe des degrés semble aller de 0 à plus de 800, en incréments de 100 ou 200, tandis que l'axe du nombre de nœuds commence à 0 et monte à plus de 250.

Les barres de l'histogramme montrent la fréquence des nœuds pour chaque degré spécifié. La plupart des barres sont basses, indiquant un petit nombre de nœuds pour la plupart des degrés. Cependant, il y a une augmentation notable dans les barres correspondant aux degrés élevés. Par exemple, la plus haute barre, située à l'extrême droite du graphique, indique qu'il y a un nombre élevé de nœuds avec un degré d'environ 800, suggérant une distribution de degrés qui est lourde dans la queue ou qui suit une loi de puissance.

3.2.2 La centralité eigenvector

- Centralité d'Eigenvector : Cette approche évalue non seulement le nombre de connexions d'un document, mais aussi la qualité de ces connexions. Si un document est connecté à d'autres documents qui sont eux-mêmes bien connectés, cela augmente son score de centralité d'Eigenvector. Cela pourrait signaler un document fondamental ou un

titre qui est central en raison de son association avec d'autres documents importants.

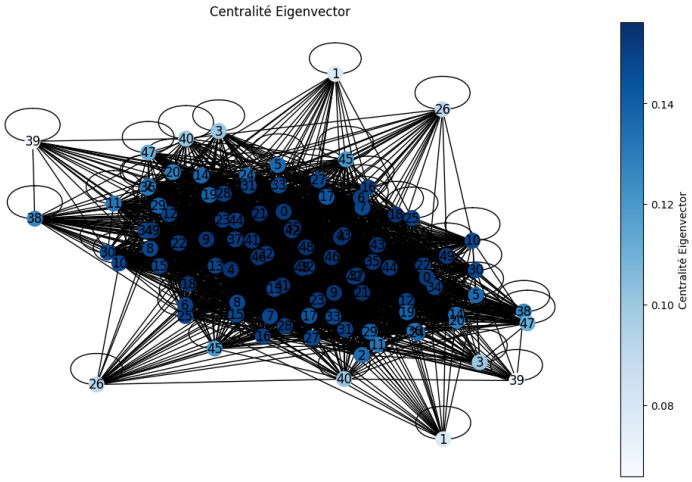


Figure 6: Graphique de la centralité eigenvector

Le graphique illustre la centralité eigenvector des nœuds au sein d'un réseau, une mesure de l'influence de chaque nœud basée sur la quantité et la qualité de ses connexions. Les nœuds, qui sont des entités comme des individus ou des articles, sont liés par des arêtes qui représentent leurs relations. La taille et la couleur des nœuds varient en fonction de leur score de centralité eigenvector, avec des nœuds plus grands et plus foncés indiquant une plus grande influence. La barre de couleur sert d'échelle pour cette métrique, avec des tons de bleu reflétant l'éventail de centralité au sein du réseau.

3.2.3 Centralité de second ordre

- La centralité de second ordre indiquée fait référence à une mesure de centralité qui prend en compte non seulement les connexions directes d'un nœud (comme le ferait la centralité de degré), mais aussi les connexions de ses voisins, offrant une vision plus globale de l'influence d'un nœd dans le réseau.

  La barre de couleur sur le côté indique l'intensité de la centralité de second ordre pour chaque nœud. Les nœuds avec une couleur plus foncée

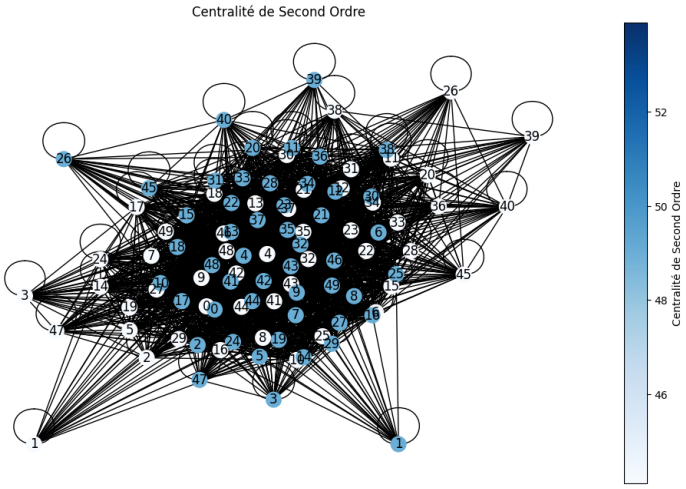


Figure 7: Graphique de la centralité de second ordre

sur l'échelle ont une centralité plus élevée, signifiant qu'ils sont non seulement bien connectés directement, mais qu'ils sont aussi centraux dans la structure globale du réseau à travers leurs voisins.

4 Fonction 3 - Moteur de recherche

Le moteur de recherche sert de porte d'entrée avancée pour des corpus académiques. Utilisant les dernières avancées en traitement du langage naturel et en analyse sémantique, il permet aux utilisateurs d'accéder efficacement à une vaste base de données documentaire. En saisissant des requêtes liées à des sujets spécifiques, les utilisateurs peuvent rapidement retrouver des documents pertinents, grâce à un algorithme qui analyse la similarité sémantique entre la requête et les titres des documents.

Nous avons créé un moteur de recherche interactif, à l'aide chatGPT pour naviguer dans le corpus des documents en utilisant le modèle BERT pour traiter la sémantique des requêtes. Lorsqu'un utilisateur entre des mots-clés, le moteur calcule la similarité cosinus entre la requête et les titres des documents, triant et affichant les résultats par pertinence. Avec une interface utilisateur intuitive basée sur un widget de texte, cet outil permet une découverte approfondie et contextuelle des documents dans le corpus.

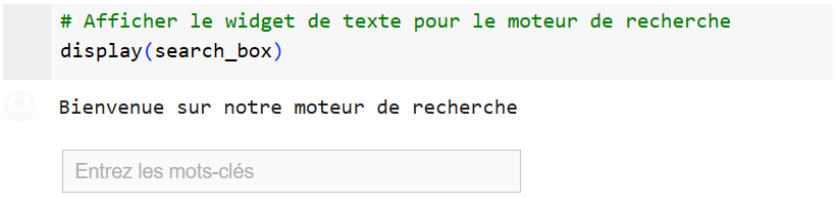


Figure 8: Image du moteur de recherche

5 Fonction 4 - Ajout de clustering

5.1 Le clustering

Le clustering, ou segmentation, est une technique d'apprentissage automatique non supervisée qui vise à regrouper un ensemble d'objets (dans ce cas, des documents) de telle sorte que les objets dans le même groupe, appelé cluster, sont plus similaires entre eux qu'avec ceux d'autres groupes. Cette similarité est souvent mesurée à l'aide de distances ou de mesures de similarité spécifiques.

Dans le contexte de notre projet, le clustering des documents basé sur les titres nous permet de :

- Découvrir des Structures Cachées : Identifier des groupes naturels de documents qui partagent des caractéristiques communes, souvent révélant des structures et des relations qui ne sont pas immédiatement évidentes.

- Faciliter l'Organisation des Données: Regrouper les documents similaires aide à organiser le corpus de manière logique, facilitant la gestion des données et l'accès à l'information.

5.2 Clustering Spectral

Nous avons donc établi un code qui implémente le clustering spectral sur la matrice d'adjacence dérivée de la similarité sémantique entre les titres des documents. Le clustering spectral est une méthode avancée d'analyse de clusters qui se base sur la théorie des graphes et est particulièrement efficace pour identifier les structures de communauté dans les données connectées comme dans notre cas, le graphe sémantique de documents.

```
from sklearn.cluster import SpectralClustering
import matplotlib.pyplot as plt
```

```
# Effectuer le clustering spectral
n_clusters = 5  # Nombre de clusters a  trouver
spectral = SpectralClustering(n_clusters=n_clusters,

affinity='precomputed', random_state=0)

clusters = spectral.fit_predict(adjacence)

# Afficher le graphe avec les clusters colores
pos = nx.spring_layout(G)
nx.draw(G, pos, node_color=clusters, cmap=plt.cm.Set1,

with_labels=False)
plt.show()
```

Le code utilise la méthode de clustering spectral pour regrouper les documents en cinq clusters distincts basés sur la similarité sémantique de leurs titres. Après avoir défini le nombre de clusters souhaités, la fonction SpectralClustering est donc appliquée à la matrice d'adjacence. Ensuite, le graphe est visualisé avec une couleur différente attribuée à chaque cluster pour différencier les groupes de documents.

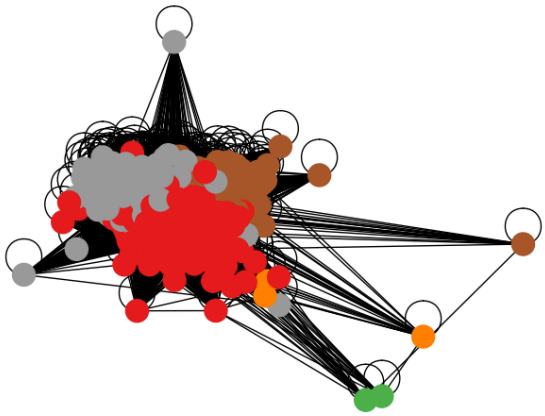


Figure 9: Graphe du clustering spectral

Le graphe du clustering spectral est une visualisation de la manière dont

les documents sont groupés en fonction de leur similarité sémantique. Les nœuds représentent les documents et les couleurs différentes indiquent des clusters distincts identifiés par l'algorithme de clustering spectral.

Les nœuds fortement connectés suggèrent des groupes de documents étroitement liés, tandis que des liens plus faibles pourraient indiquer des relations thématiques moins directes.

5.3 Clustering de Louvain

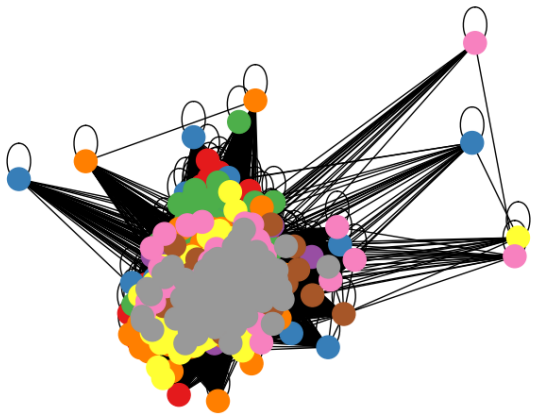


Figure 10: Graphe du clustering de Louvain

6 Fonction 5: Classification supervisée

En apprentissage automatique, la classification est une tâche supervisée qui consiste à prédire la catégorie à laquelle appartient une entrée donnée. On utilise des données historiques étiquetées pour entraîner un modèle qui, après avoir appris les caractéristiques associées à chaque catégorie, peut classer de nouvelles entrées inconnues avec précision.

Dans le cas de notre projet la classification utilise les données sémantiques extraites des titres des documents pour assigner chaque document à un domaine thématique spécifique. En apprenant des patterns à partir d'un ensemble de données étiqueté, le modèle peut ensuite prédire le domaine des nouveaux documents, ce qui enrichit l'organisation du corpus et soutient des recherches ciblées.

Pour cette partie, nous avons divisées les données en deux ensembles, un pour l'entraînement et un pour le test, avec un ratio de 80/20. Cela permet au modèle d'apprendre sur un ensemble de données et de valider ses performances sur un autre ensemble indépendant. Nous avons utilisé le modèle de régression logistique qui est entraîné sur les données encodées, où encoded_sequences sont les caractéristiques extraites des titres et y_domaine sont les étiquettes de domaine correspondantes. Le modèle entraîné est ensuite utilisé pour prédire le domaine des documents de l'ensemble de test. La précision (accuracy) du modèle est calculée pour évaluer sa performance, c'est-à-dire sa capacité à prédire correctement le domaine d'un document.

L'accuracy de 0.88 signifie que le modèle de classification de régression logistique a correctement prédit le domaine de 88.89% des documents dans l'ensemble de test. C'est un résultat assez élevé, suggérant que le modèle est assez précis dans la classification des documents selon leur sémantique, basée sur les données apprises lors de l'entraînement.

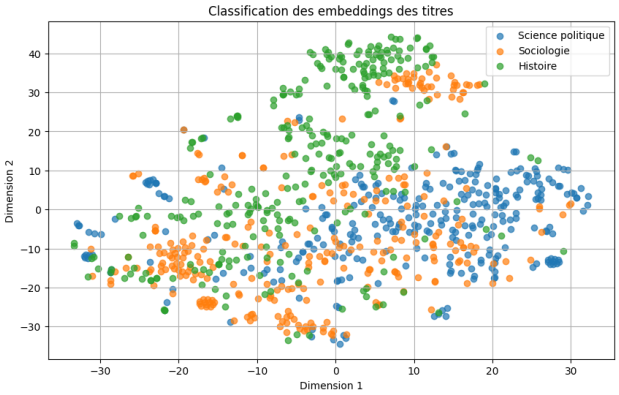


Figure 11: Classification des embeddings des titres

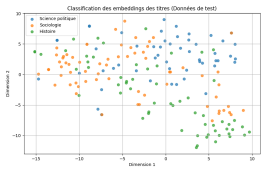
Ce graphe semble être un nuage de points bidimensionnel représentant la classification des embeddings des titres des documents dans un espace réduit. Chaque point représente l'embedding d'un titre de document, et les couleurs indiquent différentes catégories, telles que la science politique, la sociologie et l'histoire. La dispersion des points dans cet espace réduit suggère des regroupements par domaine, permettant de visualiser la séparation ou l'overlap entre les catégories basées sur le contenu sémantique des titres.

Science Politique : Les points bleus représente les documents liés à la science politique. Ils sont groupés ensemble, cela suggère que les titres de ces

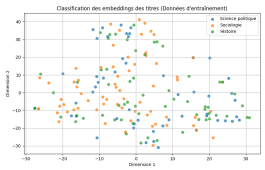
documents partagent une sémantique commune qui les distingue des autres domaines.

Histoire : Les points verts symbolisent le domaine de l'histoire les documents associés à la sociologie. Un regroupement distinct de ces points indique que le modèle reconnaît et sépare efficacement les titres en rapport avec des thématiques de l'Histoire.

Sociologie : Les points oranges sont ceux des documents associés à la sociologie. Si ces points forment un cluster ou sont dispersés de manière spécifique, cela peut révéler comment les sujets sociologiques sont représentés dans l'espace des embeddings des titres et leur relation avec les autres domaines.



(a) Classification (Données de tests)



(b) Classification(Données de train)

Figure 12: Graphique de test et de train

Ces graphiques de classification des embeddings des titres reflètent la dispersion des documents sur deux dimensions principales après réduction dimensionnelle. La distribution des points dans l'espace bidimensionnel montre comment les modèles distinguent les documents de science politique, de sociologie et d'histoire dans les ensembles de test et d'entraînement. Une dispersion large indique une variabilité significative au sein de chaque catégorie, tandis que des groupes plus concentrés suggéreraient une cohésion thématique plus forte. Ces visualisations servent à évaluer si les modèles conservent une distinction claire entre les domaines lorsqu'ils sont confrontés à de nouvelles données (test) et pendant l'apprentissage (entraînement).

Si les points ne sont pas trop dispersés dans le modèle de classification des embeddings des titres, cela suggère que les titres de documents appartenant au même domaine thématique (science politique, sociologie, histoire) ont tendance à être regroupés plus étroitement. Cela peut indiquer que les titres au sein d'un même domaine partagent des caractéristiques sémantiques similaires et que le modèle peut capter ces similitudes pour effectuer une classification cohérente.

Une faible dispersion suggère également que le modèle de classification a réussi à établir des frontières distinctes entre les différents domaines, ce qui est un bon signe pour la capacité du modèle à généraliser et à classifier correctement de nouveaux documents.

7 Conclusion

Notre projet a démontré comment l'application de techniques avancées d'apprentissage automatiques et de classification peut révéler des structures cachées dans un corpus de textes académiques, offrant ainsi des perspectives précieuses pour l'organisation et la récupération efficace de l'information.