Network Analysis for IR, ensignement de Julien Velcin

Walid Ghalleb[1†] and Ugo Zennaro[2†]

[1]M2 MIASHS, Lyon2.
[2]M2 MIASHS, Lyon2.

Contributing authors: walid.ghalleb@univ-lyon2.fr;
u.zennaro@univ-lyon2.fr;
[†]Ces auteurs ont eu une contribution égale dans ce travail.

Keywords: réseaux, graphes, NLP, machine learning

1 Introduction

1.1 Objectif

Dans le cadre du projet *Network Analysis for IR*, nous devons développer un système de recherche d'information permettant une navigation efficace à travers un vaste corpus de données textuelles.

L'objectif principal de ce projet réalisé en binôme est de développer une solution d'analyse d'un corpus structuré qui comporte plusieurs fonctionnalités :

- Chargement rapide des données et affichage de quelques statistiques,
- Visualisation du corpus pour donner une idée de la structure des données,
- Inclusion d'un petit moteur de recherche permettant de faire des requêtes par mots clefs,
- Nouvelle structuration des données à l'aide de techniques de clustering,
- Classification supervisée des données en prenant en compte la structure et l'information textuelle.

Ce corpus est constitué des métadonnées descriptives des documents disponibles sur www.persee.fr, un portail dédié à la numérisation et à la diffusion du patrimoine scientifique.

1.2 Présentation du dataset

Les documents sont de tous types (articles, comptes-rendus etc.), majoritairement issus de revues en sciences humaines et sociales, francophones, et couvrent une production du dix-neuvième siècle à aujourd'hui. À l'origine, un document est contenu dans un fascicule qui appartient lui-même à une collection correspondant généralement à une revue scientifique. Pour permettre la navigation dans le portail Persée, les collections sont associées à une discipline principale. Le jeu de données décrit plus de 900 000 documents. Il a été produit à l'occasion du cas d'étude, à partir d'un sous-ensemble des fichiers de dumps de données liées du triplestore Persée à leur état d'octobre 2021. Dans notre cas d'étude, nous avons reçu les données sous format `.pickle`.

Pour chaque entrée de document, en plus du titre et sous-titre, des auteurs, de la date de publication, on peut trouver, quand ils existent, un résumé, des mots clés, une table des matières. Les données comprennent aussi, sans caractère exhaustif, des relations de citation entre documents de Persée. Pour les champs multi-valués, on trouvera une colonne par valeur, avec un indice entre crochets dans le nom des colonnes répétées. Par ailleurs, l'identifiant du document contient le code de collection qui permet de retrouver par correspondance la discipline associée.

2 Acquisition des données

Dans le cadre du projet, nous nous sommes tout d'abord consacré à l'acquisition des données, essentielle pour assurer la pertinence et la qualité de l'analyse ultérieure. Parmi les fichiers disponibles, nous avons choisi de nous concentrer sur un seul de ces fichiers pour initier notre analyse par soucis computationel. Pour chaque dataframe sélectionné, nous extrayons les colonnes pertinentes basées sur les identifiants, les créateurs, et les citations. Ces colonnes serviront à mettre en place le graphe. Enfin nous extrayons la colonne des titres des articles est extraite et sera utilisé comme méta-donnée associée aux noeuds.

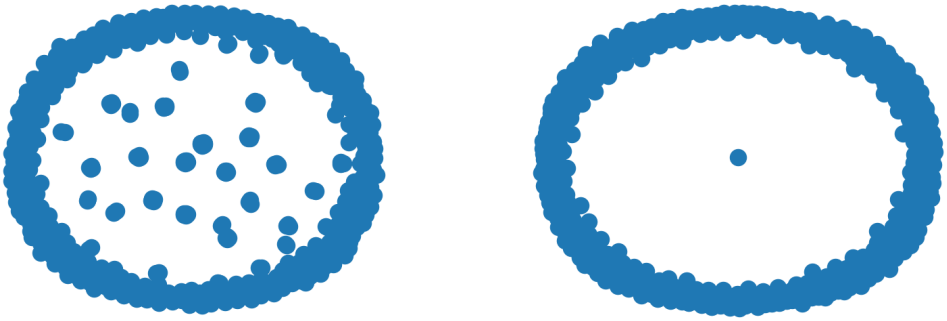
Une fonction supplémentaire a été mis en place pour extraire de l'indice du document le domaine associé. On utilisera ces domaines comme classes des documents dans la suite de ce travail.

3 Prise en compte de la structure du corpus

Par la suite, notre travail s'est concentré sur la prise en compte de la structure du corpus à travers la construction d'un graphe. Pour construire ce graphe, le jeu de données a été séparé en deux parties : une partie pour l'entraînement et une pour le test, avec une répartition stratifiée pour assurer une distribution équilibrée des domaines. Nous avons utilisé un échantillon de 500 données, dont 400 ont été allouées

à l'ensemble d'entraînement et 100 à l'ensemble de test.

Nous avons implémenté une fonction pour établir des liens entre les nœuds en fonction des colonnes cibles spécifiées dans notre dataframe. Deux types de graphes ont été générés : un graphe de citation où les liens sont établis entre les documents en fonction des citations qu'ils contiennent (illustration 1b) et un graphe d'auteurs où les liens sont créés entre les documents qui ont des auteurs en commun (illustration 1a).



(a) Par auteurs (b) Par citations

Fig. 1: Graphes proposés

Le graphe de citations ne montre aucun lien, on n'a aucune interconnexion entre les documents. En revanche, le graphe d'auteurs indique une plus grande interconnexion avec 204 liens sur 500 données. C'est ce graphe qu'on utilisera à partir de maintenant.

La faible quantité de liens observée s'explique par le fait que nous avons utilisé seulement 500 entrées et qui sont réparties sur de multiples disciplines telles que l'archéologie, les arts, le droit, etc. Il est intéressant d'observer dans l'illustration 2 que les liens du graphe se concentre particulièrement sur l'Histoire et la "Religion theologie". Cette diversité a été intentionnellement recherchée pour enrichir les résultats de la classification ou du clustering, permettant d'obtenir des groupes plus variés et interessants à analyser. En conséquence, les graphes obtenus sont très clairsemés. Il est à noter qu'avec des ressources informatiques plus avancées, il serait possible de traiter un volume de données plus important, ce qui augmenterait potentiellement le nombre de liens et améliorerait la qualité des résultats. La capacité de stockage limitée et la vitesse de traitement des scripts actuels, affectées négativement par l'emploi fréquent de boucles dans notre script, sont des limites actuelles qui soulignent la nécessité d'optimiser le code pour de futures emplois de notre travail.

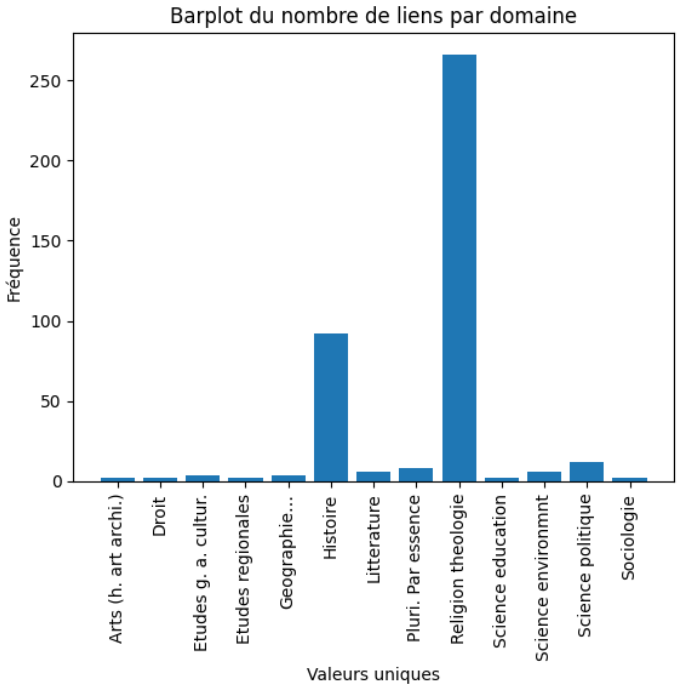


Fig. 2: Répartition des arcs dans les domaines

4 Moteur de recherche

Le moteur de recherche s'appuie sur l'encodage sémantique offert par BERT (Bidirectional Encoder Representations from Transformer) pour examiner les titres des documents. Ce modèle vise à classifier des séquences de mots. Dans le processus de classification, le modèle propose un encodage vectoriel des documents ensuite utilisé dans un modèle de classification : c'est dans cet encodage que réside l'intérêt du modèle. Ainsi, en partant du principe que cet encodage pouvait permettre d'encoder la sémantique de la séquence dans un vecteur de taille uniformisée quelque soit la taille de la séquence, que nous avons utilisé ce vecteur comme métadonnée descriptive des documents.

Dans le cadre du moteur de recherche, l'objectif était donc d'encoder par le même processus la séquence donnée en entrée du moteur et des proposer les articles qui, selon l'encodage proposée par BERT, étaient les plus proches du vecteur d'entrée. Pour cela nous avons calculé la distance euclidienne entre ce vecteur et chacun des vecteurs encodant les titres des documents. Le moteur de recherche suggère ainsi les 5 documents les plus similaires au vecteur d'entrée.

Pour tester la qualité de cette méthodologie nous avons choisi le premier article du dataset *La riziculture et la maîtrise de l'eau dans le Kampuchea démocratique*, l'avons tronqué et avons ainsi proposé comme séquence en entrée du moteur de recherche "La riziculture et la maîtrise de l'eau" en cherchant à voir si le moteur était déjà, en capacité de retrouver l'article original, et ensuite de proposer des articles similaires. Les titres des documents proposés sont les suivant :

- *L'avènement de la moralité et le rapport à la nature*
- *La crise des matières premières et les mesures internes d'organisation*
- *L'ajustement contre l'industrie*
- *La riziculture et la maîtrise de l'eau dans le Kampuchea démocratique*
- *L'équipement technique des campagnes*

Les résultats obtenus nous montrent plusieurs choses. La première est qu'on retrouve bien l'article original mais qu'il ne ressort pas en premier parmis les résultat ce qui est dommageable. Ensuite, il semblearait que la méthode comprend que la séquence en entrée porte sur la nature et de sa domestication par l'Homme car les autres titres portent sur ces thématiques.

Ainsi, il semblerait que la méthode soit efficace dans le cadre dans lequel nous avons voulu le mettre en place, néanmoins l'encodage proposée ayant initialement pour but la classification n'est surement pas optimisé pour l'interpretation sémantique et la méthode gagnerait surement à utiliser un modèle équivalent qui le soit. Aussi nous nous sommes limités à des encodages sur des titre et il aurait pû être bénéfique d'ajouter à cette démarche l'encodage de l'entiereté du document pour compléter ce que nous avons fait. Enfin une fonctionnalité qu'un encodage sémantique aurait du mal à prendre en compte seraient l'utilisation de noms propres qui aboutierait le moteur de recherche en permettant aux usagers de faire des recherches par auteur.

5 Classification supervisée

Ensuite, pour la classification, nous avons repris l'encodage de BERT et, malgré le fait que l'encodage soit optimisé pour d'autres classes que celle utilisée ici, on retrouve l'objet de ce pourquoi le modèle a été paramétré : la classification. Le classifieur originel étant capable de faire une classification sur un grand nombre de classes, nous avons décidé que le réutiliser l'encodage proposé par ce modèle pour un nombre plus restreint de classe était raisonable.

Ainsi, en utilisant ces métadonnées conjointement au graphe par auteur, nous avons décidé d'utiliser un GCN (*Graph Convolutional Network*), un réseau de neurone qui utilise à la fois la structure du graphe et le voisinage des noeuds et les métadonnées, ici l'encodage vectoriel des titres.

La classification nous a permis d'obtenir une accuracy de 62% sur les données tests. C'est un score assez faible mais sachant qu'on a un jeu de donnée d'entrainement

relativement petit par rapport au nombre de classes (400 individus pour 15 classes) et que le graphe utilisé est très creux on peut relativiser ce manque de performance et rationellement penser qu'avec le jeu de données ettendu et le graphe associé qui devrait multiplier les liens observés on aurait une accuracy bien meilleure.

Sur l'illustration 3 qui montre la confusion entre les classes faites par le modèle sur le jeu de données test on peut observer qu'on a bien une densité assez importante sur la diagonale et que les classes avec le plus de données sont globalement bien classées. On peut également observer des blocs de confusion entre certains domaines comme Histoire et Litterature, avec des confusions qui peuvent paraître cohérente (on peut penser que ces deux matière traitent d'écrit assez anciens par exemple).

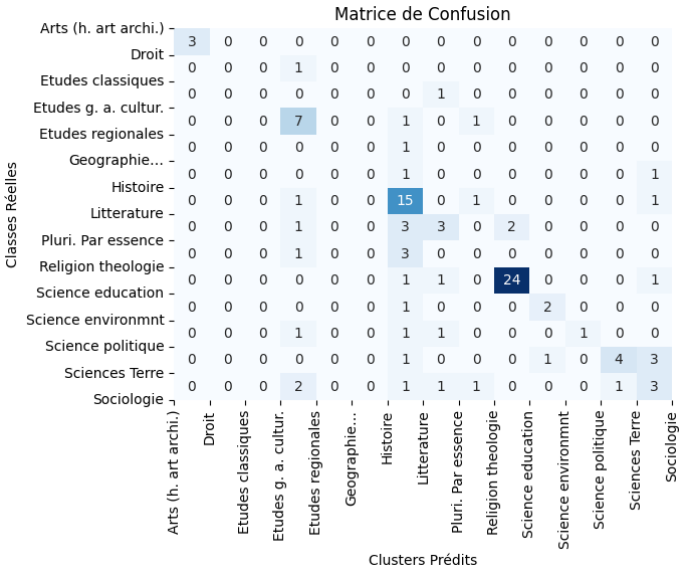


Fig. 3: Matrice de confusion de la cassification

6 Clustering

Pour finir, nos avons mis en place une méthode de clustering qui se base uniquement sur le graphe. Pour se faire nous avons utilisé un clustering dit spectral. Nous avons spécifié que la méthode propose autant de clusters que de domaines pour simili-superviser les résultats de la méthode. Pour comprendre la structure sous-jacente de notre ensemble de données et pour regrouper efficacement les données en clusters distincts, nous avons mis en œuvre un clustering spectral qui se base uniquement sur le graphe.

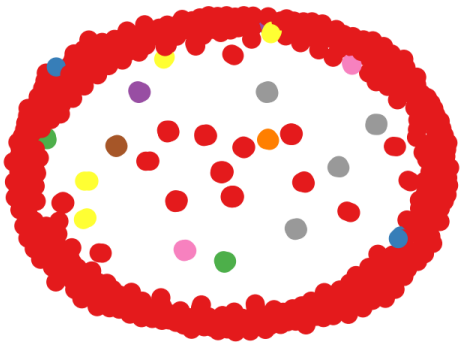


Fig. 4: Caption

Dans l'illustration 4, on observe une répartition et des composition des clusters assez hétérogène et qui manquent de cohérence. Ici encore on déplorera le manque d'arcs que le graphe propose.

La qualité du clustering a été quantifiée à l'aide du rand score, qui est une mesure de la similitude entre deux affectations de clusters. Pour ce faire, nous avons dû mapper les étiquettes de classes réelles aux indices de clusters pour permettre la comparaison. L'indice de Rand obtenu était de 27,4%, ce qui indique un faible accord entre les clusters prédits et les vrais labels.

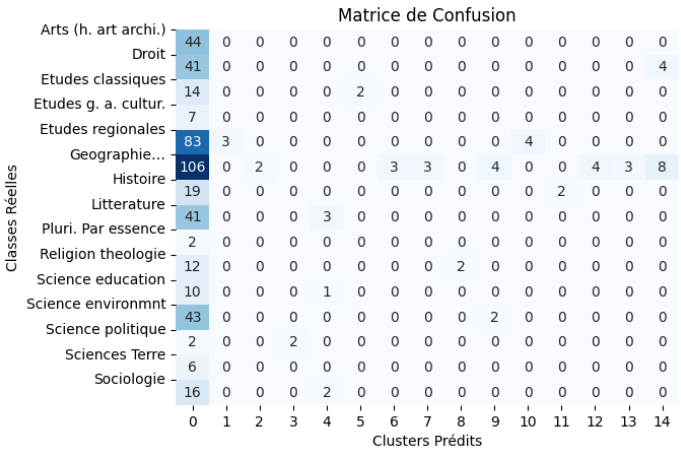


Fig. 5: Visualisation des clusters dans le graphe

Enfin dans l'illustration, 5 nous remarquons que la méthode ne permet pas vraiment de différencier les différents domaines et que le clusters proposés ne dégagent

pas de thématiques interprétables. Peu importe les thématiques, la majorité des documents se retrouvent dans la classe 0, et très peu se retrouvent en dehors. Notre méthode a impliqué la définition du nombre de clusters en fonction du nombre de classes uniques identifiées, une démarche qui, en théorie, paraît logique mais qui, en pratique, ne garantit pas l'alignement avec la structure sous-jacente des données.

7 Conclusion

Dans le cadre de notre projet *Network Analysis for IR*, nous avons entrepris une exploration approfondie d'un vaste corpus de données textuelles issues du portail Persee.fr. Notre objectif était de développer un système de recherche d'information sophistiqué, en utilisant des techniques de NLP, notamment le modèle BERT pour l'encodage sémantique des titres des documents, et en intégrant des techniques de clustering avancées ainsi qu'une classification supervisée.

Toutefois, notre travail a été confronté à des contraintes techniques significatives, notamment des limitations en termes de capacité de mémoire, ce qui nous a contraints à limiter la taille de notre jeu de données. Ces contraintes ont eu pour conséquence que le graphe produit était porteur d'assez peu d'informations. Notre classification et le moteur de recherche se sont montré assez performant ce qui valide l'interet d'analyse par NLP des titres, en contraste marqué avec le clustering qui s'est avéré nettement moins performant. Nos analyses ont révélé des confusions récurrentes entre certains domaines, comme l'histoire, les études générales et culturelles, et les sciences de la Terre, soulignant les défis posés par l'utilisation de métriques de similarité textuelle pour capturer des distinctions conceptuelles fines entre les disciplines.