Network Analysis for Information Retrieval

TEKGOZ Sumeyye, DURIAUD Matéo M2 MIASHS

March 2024

Dans ce travail nous avons analysé un corpus de documents de recherche issus de PERSÉE. Lors de notre analyse nous avons été amenés à produire des graphes permettant la représentation des relations entre les auteurs, un moteur de recherche intégré ou encore la mise en évidence de certaines structures de données. Ce rapport contient la présentation des résultats que nous avons obtenus et doit être considéré en complément du notebook qui l'accompagne.

Table des matières

1 Acquisition des données

1.1 Découverte des données

Dans le cadre de ce travail universitaire, nous avons eu à travailler sur un corpus de plus de 900 000 documents provenant de la plateforme PERSÉE. Ce corpus contient majoritairement des articles scientifiques mais également des publications, des livres et d'autres sources littéraires notamment dans le domaine des sciences humaines et sociales francophones. Nous avons pu disposer d'un grand nombre d'informations concernant ces différents documents mais du fait de la complexité de la tâche nous avons été contraints de réduire le nombre de champs que nous avons analysés. En plus de cela, du fait de la quantité très importante de publications qui ont pu être ajoutés sur le site au fil des années, il existe beaucoup de données manquantes ou aberrantes qui pouvaient représenter une difficulté dans le traitement des informations.

Du fait de la taille conséquente du fichier contenant les informations (près de 5Go), nous avons pris la décision de travailler sur une partie réduite du jeu de données. En effet, nous souhaitions travailler sur un outil de programmation collaboratif (Google Colab) et nous nous sommes heurtés à l'impossibilité de traiter le fichier complet. Nous avons donc tout d'abord rassemblé les 9 documents pickle que nous avions à notre disposition sur l'un de nos ordinateur personnel et nous les avons cumulés dans un dataframe pandas. Nous avons ainsi remarqué que nous avions de nombreuses colonnes d'intérêts et pris la décision de nous concentrer sur le titre des articles, leur résumé, les 10 premiers auteurs ainsi que les 10 premières citations. Nous avons pris la décision de ne prendre que les 10 premiers résultats de ces colonnes pour chaque article car la très grande majorité ne dépassaient pas ce chiffre et que nous allions nous retrouver avec un jeu de donnée très déséquilibré et vide. Après avoir réalisé cette concaténation nous avons décidé de sélectionner de manière aléatoire sous python un échantillon représentant 40% de nos données de départ. Nous obtenons ainsi 363512 documents à analyser et grâce à notre sélection de champs uniquement 23 colonnes sur les 1041 d'origine. Cette sélection nous a paru adapté afin de permettre une réduction de la taille du fichier et un traitement plus rapide en terme de temps d'exécution tout en conservant une disparité très raisonnable correspondant aux données de départ.

1.2 Premières statistiques

Après avoir extrait le fichier au format qui nous intéressait nous l'avons mis à disposition sur un serveur GitHub afin qu'il soit simple à récupérer et à utiliser par la suite. Nous avons tout d'abord cherché à regarder les potentiels doublons et documents vides que nous pouvions avoir et avons pu voir qu'environ 60 000 documents portaient le même titre ou n'en avaient pas. Sur les environ 300 000 documents restants nous avons presque 90 000 auteurs différents ce qui nous ferait apparaître une moyenne théorique d'un peu plus de 3 articles par auteurs dans nos données. Malgré tout, nous avons supposé que certains indi-

vidus pouvaient être bien plus facilement considérés comme des co-auteurs dans des publications et que cela pouvait faussé nos résultats. En effet, nous avions dans l'idée qu'un membre éminent d'un laboratoire pourrait être amené à ne contribuer que de manière très partielle à certains articles mais tout de même apparaître dans la liste des co-auteurs du au fait que sa notoriété puisse participer à la crédibilité du document. Cette supposition s'est avérée juste puisque lorsque l'on regarde le nombre de publications de chacun on retrouve certains auteurs jusqu'à plusieurs centaines de fois. On peut alors supposer qu'il s'agisse soit d'entités qui participent donc nécessairement de manière plus courante ou des individus responsables de nombreuses publications et qui apparaissent donc très régulièrement.

Nous avons également cherché à mettre en évidence la discipline à laquelle se rattache chaque publication. Pour ce faire, nous avons utilisé le code situé au début de l'index de chaque article et qui représente une collection qui peut ensuite être rapproché de certaines discipline à partir d'un document qui nous a été fourni. Dans l'échantillon que nous avons utilisé nous avons ainsi 318 collections différentes par rapport aux 389 possibles dont nous avions connaissance. En faisant le rapprochement entre ces collections nous nous sommes rendu compte que seulement 95% de ces collections existaient dans le document des disciplines et que nous avions 13 collections dont nous n'avions aucune équivalence. Sur l'ensemble de nos 365 000 articles, ces 13 collections représentaient uniquement 729 articles donc nous avons pris la décision de continuer en attribuant les disciplines connues à chaque article et "Introuvable" dans le cas où nous ne savions pas.

2 Prise en compte de la structure du corpus

Dans la seconde partie de notre travail, nous avons cherché à représenter sous forme de graphes la structure de nos documents ainsi que les relations qui peuvent exister entre eux. Nous avons ainsi mis en évidence les liens entre chaque auteurs des différents l'article. Notre idée de départ étant premièrement que ces auteurs ont une forte probabilité de publier avec des co-auteurs différents à chaque fois mais que sur le long terme nous pourrions mettre en évidence l'existence de groupes avec une tendance à l'interaction bien plus élevée puisque appartenant à la même discipline. Cette théorie s'est avérée correcte puisque nous avons pu remarquer une séparation de nos auteurs qui n'interagissent donc pas avec tout les autres. Afin de permettre une meilleure clarté visuelle nous avons sélectionné uniquement les auteurs qui apparaissent au moins 100 fois dans l'ensemble des articles.

En faisant ce traitement nous avons mis en évidence l'existence de 29 composantes connexes dans notre graphe pour un total de 168 noeuds. En réalisant les mesures de centralités nous avons pu mettre en évidence que nous avions un noyau très important au centre de notre graphe et qui rassemble un grand nombre d'auteurs tandis que beaucoup d'autres sont complètement coupés des autres. Ces résultats s'expliquent en partie par le fait que nous ne faisons pas

apparaître les auteurs ayant peu de publications alors que ceux ci pourraient permettre l'apparition d'arêtes entre nos composantes connexes et donc modifier la centralité de l'ensemble du graphe. La distribution des degré semble très inégal dans le graphe que nos utilisons et nous sommes face à un cas ou notre noyau central est composé de noeuds extrêmement corrélés qui possèdent majoritairement un degré plus élevé que le reste des composantes connexes.

A partir de la matrice d'adjacence de notre graphe on peut obtenir une représentation des différents chemins entre les noeuds de notre graphe et nous pouvons tracer leur distribution. On remarque alors de manière très claire que nous sommes face à l'existence de plusieurs composantes connexes qui n'interagissent pas entre elles même si elles sont relativement proches. On peut supposer que le fait de considérer l'ensemble de nos données dans la construction de notre graphe pourrait avoir comme effet l'apparition d'arêtes entre ces composantes et ainsi le développement de relations entre les auteurs. Ce fonctionnement semble trahir l'aspect inter-disciplinaire qui peut exister dans un grand nombre de publications. Ces collaborations sont assez rares pour ne pas apparaître dans la liste des auteurs que nous avons sélectionnés mais des liens devraient apparaître entre pratiquement toutes les disciplines présentes dans nos données.

Lorsque l'on cherche à mesurer la centralité de notre graphe on remarque que toutes les méthodes permettent d'obtenir des résultats très similaires. Du fait de la connexité de notre graphe, cette remarque est assez cohérente puisque l'absence de relation entre la plupart des composantes de notre graphe empêche le partage de la centralité qui est donc restreinte à la composante la plus massive que nous avons.

3 Moteur de recherche

Par la suite notre objectif a été de mettre en place un système de recherche permettant à partir d'un ou plusieurs mots-clés de faire ressortir les articles correspondants. Afin de réaliser cela nous avons choisi d'utiliser un modèle de langage pré-entraîné à partir de transformers et nous avons fait apparaître les documents qui étaient les plus proches de la requête. Nous avons préféré travailler à partir d'un modèle de transformers car cela permet une plus grande flexibilité vis à vis de la sémantique contenu dans chaque terme. Notre choix de modèle s'est porté sur *all-MiniLM-L6-v2* qui a l'avantage d'être un modèle déjà réduit qui est donc très rapide à exécuter et intéressant dans notre cas sans avoir de puissance de calcul importante. De plus, il s'agit d'un modèle qui n'est pas uniquement spécialisé en français mais qui présente l'avantage d'avoir été entraîné sur d'autres langues également ce qui est un avantage car nous savons qu'une part de nos articles ne sont pas en langue française.

Pour le moteur de recherche, notre première idée a été de nous baser uniquement sur le titre car c'est généralement un assez bon résumé du sujet abordé dans l'article. Pour cela notre méthode a consisté à extraire la liste des titres de tout les documents et la transformer en vecteurs à partir de notre modèle de

traitement du langage (embedding). Nous enregistrons nos tenseurs dans des variables afin de ne pas avoir à régénérer tout les résultats à chaque fois. De cette manière nous pouvons sauvegarder nos embeddings et les réutiliser plus tard ce qui est nettement plus rapide. Nous avons tenté de les rendre disponible en ligne afin d'accélérer encore plus le temps d'exécution mais la taille des tenseurs était trop importante pour être mise en ligne sur github donc nous avons pris la décision de conserver la génération à l'intérieur du code. Après avoir testé notre code avec les titres nous avons décidé d'ajouter également les résumés dans l'embedding et de permettre à l'utilisateur de choisir entre faire une recherche uniquement dans le titre, dans le résumé ou dans l'ensemble.

Nous avons ensuite mis en place un petit interface graphique en python qui nous permet d'avoir une barre de recherche dans laquelle il est possible de rentrer les mots clés qui nous intéressent. Ces mots clés sont ensuite transformés sous forme de tenseurs à partir du même modèle que nos textes et nous faisons ensuite un calcul de symétrie avec chacun des textes du corpus. Une fois que nous avons notre score nous faisons alors un classement des textes les plus ressemblant et nous affichons leurs titres pour que l'utilisateur puisse les retrouver. Puisque nous avions mis en évidence l'appartenance de chaque texte à une catégorie nous avons également rajouter une case à cocher qui permet d'activer un second champ qui peut être rempli afin de filtrer nos textes selon la catégorie qui nous intéresse. Nous avons construit deux interface avec tout d'abord un champ libre dans lequel l'utilisateur peut préciser la catégorie qui l'intéresse s'il la connaît et un second dans lequel le choix se fait à partir d'une liste déroulante.

4 Ajout de clustering

Dans cette section nous avons tenté de mettre en place une méthode de clustering de Louvain ainsi que le clustering spectral avec le graphe que nous avions obtenu dans la section 2. L'objectif était d'explorer la structure des communautés dans le réseau d'auteur que nous avons. L'algorithme de Louvain est souvent utilisé car il repose sur l'optimisation de la modularité des graphes. Il permet en général d'obtenir une assez bonne séparation des communautés tout en ayant un temps de traitement assez réduit qui est intéressant dans le cas de notre jeu de données conséquent. Le clustering spectral permet une projection de notre graphe dans un espace de dimension réduit.

Nous avons construit un graphe d'auteurs à partir de nos données, où chaque nœud représente un auteur et chaque arête représente une collaboration entre deux auteurs. Ensuite, nous avons appliqué l'algorithme Louvain pour détecter les communautés dans le graphe. Nous avons également utilisé le clustering spectral pour comparer les résultats avec ceux de Louvain. Les communautés détectées ont été projetées sur la visualisation du graphe pour une analyse visuelle.

L'algorithme Louvain a identifié plusieurs communautés distinctes dans le réseau d'auteurs. La visualisation du graphe avec les communautés colorées a révélé des regroupements significatifs d'auteurs, suggérant des domaines de

recherche apparentés ou des collaborations fréquentes. En revanche, le clustering spectral a produit des résultats légèrement différents, avec certaines communautés fusionnant ou se séparant par rapport à Louvain. On remarque notamment ces différences au niveau de la centralité de notre graphe puisqu'il y a une forte différence de classification dans le cas du clustering spectral tandis que nous avons une représentation relativement homogène à l'aide de l'algorithme de Louvain. Certains auteurs étant regroupés différemment selon les méthodes utilisées, la question de la sensibilité des algorithmes de clustering aux caractéristiques des données devient importante. Les communautés détectées par chaque méthode présentent des tendances disciplinaires similaires, mais avec des nuances dans la composition des groupes.

Cette étude a démontré l'importance de comparer différentes méthodes de clustering pour l'analyse des réseaux d'auteurs. Bien que Louvain et le clustering spectral aient révélé des structures communautaires similaires dans notre réseau, ils ont également révélé des différences significatives dans certains cas. Cette analyse fournit des informations précieuses sur les dynamiques de collaboration entre les auteurs et on remarque certains échanges qui apparaissent dans des composantes connexes qui n'auraient à priori pas de liens.

5 Classification supervisée

Dans cette dernière section nous avons tenté de prédire l'appartenance de chaque texte à une catégorie à partir du titre et du résumé des articles. Cette tâche s'est révélé assez complexe car l'information qui est contenue dans les titres est bien souvent assez peu indicatrice de l'exacte discipline du document. De plus, le résumé de chaque article est souvent assez général et ne permet donc pas de discriminer efficacement les disciplines auxquels peuvent appartenir les textes.

Nous avons utilisé deux méthodes afin de réaliser cette classification. Nous avons tout d'abord pris la vectorisation à partir de Tfidf en complément avec un modèle linéaire et nous avons ensuite réalisé le même traitement en utilisant une vectorisation à partir de Word2Vec. Nous avons réussi à obtenir une performance légèrement meilleure dans le cadre de Tfidf mais nous n'avons pas réussi à obtenir une précision meilleure que 43%. Une piste d'amélioration pour ce travail aurait consisté à vérifier les types d'erreurs les plus fréquentes et essayer de voir si une cohérence pouvait apparaître. On peut en effet imaginer que certaines disciplines sont assez proches entre elles et que leur distinction est difficile à modéliser. Un travail de rassemblement des disciplines se ressemblant pourrait ainsi permettre d'obtenir de meilleurs résultats au niveau de notre classification.

6 Remarques et conclusion

Grâce à notre traitement des données PERSÉE nous avons pu mettre en évidence certaines limites du dataset et de notre approche. Pour commencer,

nous avons été confronté au fait que certaines des collections qui apparaissent dans nos données n'avaient aucune discipline de correspondance. Ce sujet peut être compris comme le fait que nous n'avions pas la liste complète à notre disposition mais cela peut aussi venir du fait que certaines collections ne sont plus utilisées et ne devraient donc peut être pas exister. Ce sujet nous semble donc important à faire remonter car si la quantité d'articles concernés actuellement est assez restreinte, nous pouvons craindre une dégradation de cette problématique si nous ignorons sa provenance.

La masse importante d'informations et le nombre très élevé de publications ont également été un frein à notre travail car cela nécessitait la mise en plus de ressources plus importantes, de temps de calcul plus élevé et c'est pour ces raisons que nous avons travaillé sur un échantillon réduit des données d'origine. Malgré cela nous sommes parvenu à mettre en évidence certains comportements parmi les auteurs faisant des publications sur PERSÉE et l'existence de relations entre les disciplines de sciences humaines et sociales qui peuvent se compléter sur certains sujets. Un travail plus approfondi sur ces données pourrait nous permettre d'en apprendre plus sur les habitudes de publications et peut être nous permettre de mettre en place un système permettant la classification automatique des nouvelles publications dans la catégorie qui leur correspond le mieux.