M2 MALIA-MIASHS : projet Network Analysis for Information Retrieval (partie 4)

Julien Velcin, Université Lyon 2, Laboratoire ERIC

2024-2025

Exercice 8 : Pour aller plus loin

Voilà plusieurs pistes qui vous permettront d'aller un peu plus loin dans la réalisation de cette application. Il n'est pas demandé de les explorer toutes : elles constituent des idées que vous pouvez plus ou moins développer.

8.1 Classification supervisée avec GNN Les exercices précédents ne vont pas jusqu'à introduire la supervision au niveau des nœuds du graphe pour résoudre des problèmes de classification. Par exemple, une tâche pourrait consister à prédire la catégorie de la publication en fonction de son domaine (par ex. *machine learning* ou traitement d'image) Le corpus que vous avez à votre disposition fournit le nom du journal ou de la conférence où ont été publiés les articles. Il suffit alors soit de chercher à prédire ce nom, soit à classer manuellement ces lieux de publication dans des catégories, puis prédire les catégories. Le corpus comportant de nombreux lieux, vous pouvez vous contenter d'apprendre à partir d'un sous-ensemble des articles publiés aux lieux les plus fréquents. Une expérimentation intéressante consiste alors à à faire varier la quantité de supervision et d'observer les différences.

8.2 Identification d'auteurs Une tâche intéressante consiste à essayer de trouver le nom des auteurs d'un article à partir de sa description textuelle. Cette tâche peut être définie comme un problème de recherche d'information dans laquelle on utilise un vecteur qui représente un auteur et on compare ce vecteur avec celui des documents. Une solution naïve consiste à placer l'auteur au barycentre de vecteurs des articles qu'il a publié. Une autre solution serait d'utiliser Doc2Vec en utilisant comme tag le nom de l'auteur, ce qui permet de calculer des représentations d'auteur. La difficulté peut être de trouver une bonne manière d'*évaluer* la solution proposée, par exemple en calculant le rang moyen du ou des véritables auteurs dans la liste retournée par le système.

8.3 Utilisation de techniques avancées de plongement L'idée ici est de remplacer la représentation vectorielle sur le vocabulaire des mots par une représentation plus avancée (par ex. : InferSent, USE, SBERT). L'objectif est clairement d'obtenir des espaces avec une meilleure estimation des similarités entre les documents.

Idée de barème pour le projet

<10 : le projet ne répond pas aux attentes car il ne permet pas (du tout) de naviguer dans le corpus, que ce soit par des requêtes ou des catégories (supervisées ou non)

10-12 : le projet répond très partiellement aux attentes avec une application fonctionnelle mais très limitée (par ex. un seul type de paramétrage, plusieurs fonctionnalités de base qui manquent (cf. exercices 1 à 7)

12-14 : le projet répond aux attentes mais se contentent d'implémenter quelques techniques sans chercher à aller plus loin

14-16 : bon projet qui répond aux attentes et explorent quelques pistes

16-18 : très bon projet qui répond particulièrement bien aux attentes, avec plusieurs pistes explorées en profondeur

18-20 : excellent projet