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 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. *computer vision* ou *theoretical computer science*). Le corpus que vous avez à votre disposition fournit le nom du journal ou de la conférence où ont été publiés les articles. Une catégorisation automatique est proposée dans la colonne *class*, à savoir :

- artificial intelligence (class 1): all the conference in machine learning, automatic decision systems, autonomous agents, natural language processing
- data science (class 2): information systems, databases, data mining, pretraitements, cleaning, business intelligence
- interface (class 3): visualization, human-machine interfaces, interaction
- computer vision (class 4): image processing, 2D and 3D, virtual reality
- network (class 5): networks, systems, security, mobile devices, iot, web
- theoretical computer science (class 6): theory of computer science, theorems, proofs, bounds, calculability, compilation, game theory
- specific applications (class 7): application to specific domains, such as humanities, biology, etc.
- other (class 8): all the other conferences

Ces classes ont été attribuées automatiquement à partir du nom du journal ou de la conférence (*venue*) en utilisant un LLM. L'objectif est alors de tester si vous pouvez apprendre automatiquement à classer les articles à partir du texte de leur titre et de leur résumé, mais aussi des liens qu'ils entretiennent entre eux. Il faut bien sûr utiliser ce que vous avez vu en cours et non des LLMs à cette fin.

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.

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