M2 MALIA-MIASHS : projet Network Analysis for Information Retrieval (partie 3)

Julien Velcin, Université Lyon 2, Laboratoire ERIC

2024-2025

Exercice 7 : Prise en compte de la structure du corpus

Vous avez à votre disposition d'autres informations qui vous permettent de rapprocher deux documents lorsque : a) les articles partagent un ou plusieurs auteurs en commun, b) un article cite un autre article dans sa bibliographie, b) deux articles partagent un certain nombre de références bibiographiques en commun, d) les articles ont été publiés dans le même journal ou la même conférence. La similarité entre deux documents peut donc se baser sur leur similarité textuelle (ce que nous avons fait précédemment) *mais également* sur d'autres informations de proximité.

7.1 Commencez par construire la matrice d'adjacence des nœuds d'un graphe qui sont constitués par les articles scientifiques. A vous de choisir comment définir les liens qui unient ces œuds (références en commun, auteurs en commun...).

7.2 Vous pouvez à présent calculer un certain nombre de statistiques sur votre graphe, telles que la distribution des degrés, la largeur et la densité du graphe, les coefficients de clustering, etc. Cela peut vous aider à mieux prétraiter le graphe afin de garder un graphe de taille raisonnable pour le traitement ultérieur.

7.3 Visualisez le graphe en utilisant par ex. la librairie `ipysigma` ou `scikit-network`. Projeter certaines caractéristiques des nœuds, par exemple la thématique majoritaire de l'article (cf. partie 2).

7.4 Faites tourner un (ou plusieurs) algorithmes de clustering classiques, comme le clustering spectral ou Louvain. Il existe bien sûr d'autres solutions, tel que les *block models*. Ajoutez ces catégories calculées automatiquement dans vos données. Vous pouvez ainsi faire en sorte de les projeter sur la visualisation.

7.5 Calculez des représentations basées sur le graphe uniquement, par ex. en utilisant un algorithme comme Node2Vec ou DeepWalk. Projetez ces représentations afin de pouvoir les comparer avec la visualisation du graphe. Commentez les différences entre ces deux spatialisation du corpus.

7.6 Comparez la partition obtenue par clustering avec le résultat des catégorisations réalisées uniquement sur le texte (par ex. thématique). Une idée intéressante serait de combiner les deux représentations (textuelles et structurelles). Cette comparaison peut être uniquement réalisée de manière *qualitative* (vous donnez votre avis) ou vous pouvez aller jusqu'à comparer les partitions de manière quantitative, par ex. via des mesures comme l'ARI.