M2 MALIA-MIASHS : projet Network Analysis for Information Retrieval (partie 2)

Julien Velcin, Université Lyon 2, Laboratoire ERIC

2024-2025

Exercice 4 : Structurer le corpus

Jusqu'à présent, le corpus n'est qu'un ensemble de documents. Afin de faciliter l'exploitation (interrogation, visualisation, navigation) de ce corpus, il est important de le structurer par exemple en regroupant les documents dans des catégories. Une solution consiste à une utiliser un algorithme de clustering de votre choix (k-means, modèle de mélange, etc.). Cet algorithme peut être utilisé à partir des différentes représentations vectorielles que nous avons vues en cours :

- espace des mots (avec les différents systèmes de pondération),
- espace de plongement (naïf, Doc2Vec, autres)
- espace thématique (par ex. avec LDA).

Vous pouvez essayez plusieurs solutions afin de comparer les résultats.

Exercice 5 : Visualisation du corpus

Cette étape consiste à proposer une ou plusieurs visualisation du corpus. Il s'agit par exemple de montrer :

- Les termes les plus employés dans le corpus, par exemple via des nuages de mots.
- Les co-occurrences de mots les plus observés.
- Les thématiques extraites par un algorithme comme LDA.
- Les catégories extraites à la section précédente (via les espaces de plongement et/ou les thématiques).

Une fonctionnalité intéressante serait de pouvoir sélectionner un ou plusieurs mots et de voir dans quelle partie du corpus (document, thématique, cluster) il(s) se situe(nt).

Exercice 6 : Etiqueter les catégories construites

Les catégories et/ou les thématiques ne sont pas toujours simples à interpréter. Pour aider à l'interprétation, on peut calculer pour chaque catégorie :

- les mots les plus fréquemment employés (que l'on pourrait représenter sous forme de nuage),
- les termes fréquents les plus intéressants, en utilisant par ex. des collocations,

- les termes les plus discriminants en pénalisant les termes employés dans trop de catégories ou thématiques.
- les documents les plus centraux à la catégories (par ex. proches du centre d'inertie ou de la moyenne).