Génération de graphes de citation pour la recherche d'information

Encadrement : Florian Boudin et Maël Houbre (prénom.nom@univ-nantes.fr)
Lieu du TER : LS2N (Site de la FST)

Contexte

La quantité d'articles scientifiques publiées sur les bibliothèques numériques a augmenté ces dernières années de manière exponentielle. De par le nombre croissant d'articles, les relations entre articles comme les citations augmente également. [1] A notamment montré que l'utilisation des liens entre documents (similarité sémantique, hyperliens...) permet d'améliorer les capacités de raisonnement d'un modèle sur des tâches complexes telles que le Question Answering et le NLU. L'utilisation particulière des réseaux de citations peut permettre d'obtenir d'autres informations sur un texte donné en analysant les articles qu'il cite ou par lesquels il est cité. Ces informations complémentaires peuvent ensuite à l'instar de [2], être utilisée pour l'extraction de mots-clés.

Objectifs

L'objectif de ce Travail d'Etude et de Recherche (TER) est de développer un outil permettant de facilement générer et visualiser le graphe de citation d'un corpus donné. Plus précisément, il s'agit de fournir au laboratoire, un outil permettant de facilement analyser un corpus et/ou extraire des informations à partir du réseau de citation. Un effort particulier sera mis sur l'optimisation du code, la taille du corpus pouvant varier de l'ordre du millier à plusieurs millions de documents. S'il est satisfaisant, l'outil final sera mis à disposition de la communauté scientifique sous licence libre.

Profil des candidat·e·s

Deux personnes sont attendues. Ces dernières doivent avoir des appétences pour le TAL et la fouille de donnée. Des compétences en programmation Python et des connaissances sur les graphes sont appréciées

Bibliographie

1. Yasunaga Michihiro, Leskovec Jure,Liang Percy. LinkBERT : Pretraining Language Models with Document Links. 2022. http://arxiv.org/abs/2203.15827

2. Caragea Cornelia, Bulgarov Florian Adrian, Godea Andreea, Das Gollapalli Sujatha. Citation-Enhanced Keyphrase Extraction from Research Papers: A Supervised Approach. 2014. http://aclweb.org/anthology/D14-1150