Sujet de stage recherche
Master 2 AFD 2017-2018

Titre : Apprentissage automatique et TAL pour l'indexation de la documentation SNCF

Encadrant(s) : Nathalie Camelin nathalie.camelin@univ-lemans.fr, Nicolas Dugué nicolas.dugue@univ-lemans.fr

Contexte du stage *:* Au Mans, financement sur projet en collaboration avec la SNCF

Mots-clés : Apprentissage supervisé, traitement automatique de la langue naturelle, clustering

Sujet du stage

Le groupe SNCF connaît actuellement une transformation digitale et se tourne de plus en plus vers des technologies susceptibles de faire appel à de l'intelligence artificielle appliquée au traitement d'informations écrites ou orales. La documentation métier est aujourd'hui en pleine mutation, avec des métiers qui se digitalisent, plus mobiles et de nouveaux modes de consommation de l'information. Divers projets internes ont permis d'enclencher une transition vers le numérique, pour trouver la juste information au bon moment. Au-delà de la numérisation des documents se pose la question de nouveaux systèmes intelligents d'accès aux contenus, d'aide à l'interprétation et à la saisie. L'objectif du projet dans lequel s'inscrit ce stage est ainsi d'identifier, de maquetter et d'évaluer des solutions capables d'enrichir les initiatives actuelles de documentation numérique. Les champs d'application sont l'aide à la rédaction, la recherche d'information et la navigation dans les contenus textuels.

Dans ce contexte, le stage a pour objectif d'étudier des méthodes pour l'indexation fine de cette documentation numérique. En particulier, il s'agit d'étudier les méthodes existantes (tf idf, zipf law, glove, lsa) et de fournir des outils capables de décrire ces documents au contenu technique via des mots-clés issus du vocabulaire métier. Cette description doit pouvoir ensuite être utilisé pour réaliser des traitements tels que des regroupements thématiques de documents.

Il s'agira en particulier pour ces mots-clés du vocabulaire métier, d'apprendre des représentation vectorielles pertinentes, et ainsi d'étudier des approches de l'analyse distributionnelle sur petit corpus en langue spécialisée. Pour réaliser ce travail, nous disposerons d'un lexique incomplet qui pourra notamment être utilisé comme ressource pour le vocabulaire technique, ainsi que de corpus pré-traités dont les documents sont également catégorisés en thématique.

Bibliographie

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543)

Ferrer i Cancho, R., & Solé, R. V. (2001). Two Regimes in the Frequency of Words and the Origins of Complex Lexicons: Zipf's Law Revisited. *Journal of Quantitative Linguistics*, *8*(3), 165-173.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), 391.

Leaman, R., & Gonzalez, G. (2008, January). BANNER: an executable survey of advances in biomedical named entity recognition. In *Pacific symposium on biocomputing* (Vol. 13, pp. 652-663).