Sujet de stage : apprentissage de représentations pour la détection de nouveauté dans les flux textuels

Durée du stage : 5 à 6 mois

Rémunération mensuelle : gratification standard ~600 euros

Lieu du stage : laboratoire ERIC (plusieurs séjours à EDF prévus)

Orientation du stage : recherche *et* professionnelle

Encadrants : J. Cugliari et J. Velcin (ERIC), P. Suignard et M. Boumghar (EDF)

Problématique générale

Le stage se déroulera dans le contexte du projet DyNoFlu, financé par le programme Gaspard Monge et faisant intervenir des chercheurs du laboratoire ERIC et de l'équipe ICAME d'EDF Labs. L'objectif principal du projet consiste à modéliser conjointement l'évolution des thématiques abordées dans des flux textuels (fournis par le partenaire industriel) et l'apparition des documents nouveaux annonciateurs de changement. Un budget est ainsi prévu pour faire annoter certains textes comme présentant un caractère nouveau, soit parce qu'ils introduisent une nouvelle thématique, soit parce qu'ils changent la configuration des données (par ex. la fusion de deux thématiques déjà existantes). Ces indications supervisées doivent permettre d'apprendre un espace de représentation adapté à la tâche de détection de la nouveauté tout en la reliant à la dynamique de l'évolution des thématiques.

Dans ce contexte, la stage consiste tout d'abord à tester l'apport d'espaces de représentation des mots et/ou des documents déjà appris sur d'autres corpus (à l'instar de word2vec ou doc2vec) comme une alternative aux modèles thématiques déjà utilisé dans le cadre du projet. Une piste intéressante, et encore relativement nouvelle, consiste à combiner les deux approches afin d'améliorer les performances finales. Cette partie est encore non supervisée et l'évaluation pourra être réalisée sur des corpus dans lesquels la nouveauté est simulée de manière artificielle. Dans un second temps, il est demandé d'expérimenter des techniques d'apprentissage profond (*deep learning*) afin d'apprendre des espaces de représentation adaptés à la tâche de détection de la nouveauté en étant guidé, cette fois, par les étiquettes fournies par l'annotation manuelle. L'évaluation finale devra reposer sur au moins deux jeux de données : un jeu de données fourni par l'entreprise et un second jeu de données public.

Organisation du stage

Le stage se passera principalement dans les locaux du laboratoire ERIC avec plusieurs séjours prévus dans les locaux d'EDF Labs en région parisienne. Il profitera également d'une thèse CIFRE actuellement en cours entre les deux partenaires.

Quelques précisions sur le contexte industriel :

EDF surveille l'évolution des thématiques discutées dans différents types de corpus textuels : tweet, blogs,

réclamations, etc. Un plan de classement prédéfini permet de recourir à des algorithmes de classification supervisée performants, mais de nombreux documents se retrouvent mal ou même non classés. Cela peut être dû au fait que les catégories évoluent au fil du temps (principe de « dérive de concepts » ou concept drift), par exemple avec l'apparition de nouveaux termes dans le vocabulaire, ou au fait que de nouvelles catégories thématiques apparaissent. Dans ce contexte, l'analyse des signaux faibles sur la base des documents non classés est une piste envisagée sérieusement pour mieux appréhender ces évolutions. De manière plus générale, l'entreprise souhaite être en mesure de suivre les thématiques des textes dans le temps, que celles-ci soient récurrentes ou qu'elles apparaissent et disparaissent au fur et à mesure du temps.

Profil requis :

- connaissances avancées en fouille de données, fouille de textes / traitement automatique de la langue, modèles probabilistes d'apprentissage automatique
- compétences en programmation sous Python, si possible avec une première expérience avec les librairies de *deep learning* (Tensor Flow, Theano, Keras…)

Références bibliographiques

Das, R., Zaheer, M., & Dyer, C. (2015, July). Gaussian LDA for Topic Models with Word Embeddings. In ACL (1) (pp. 795-804).

Kusner, M., Y. Sun, N. Kolkin, et K. Weinberger (2015). From word embeddings to document distances. In International Conference on Machine Learning, pp. 957–966.

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In Proceedings of the 31st International Conference on Machine Learning (ICML-14) (pp. 1188-1196).