CEA List
Laboratoire d'Analyse Sémantique Texte et Image
Centre de Saclay 91191 Gif-sur-Yvette France
http://www.kalisteo.eu

Contact Julien Tourille
Tél +33 (0)1 69 08 49 02
E-mail julien.tourille@cea.fr

STAGE 2021 Réf : LASTI-2021-S2

Classification semi-supervisée de news sur des thèmes polarisants

Présentation du laboratoire d'accueil

Basé à Paris-Saclay, le CEA List est l'un des quatre instituts de recherche technologique de CEA Tech, direction de la recherche technologique du CEA. Dédié aux systèmes numériques intelligents, il contribue au développement de la compétitivité des entreprises par le développement et le transfert de technologies.

L'expertise et les compétences développées par les 800 ingénieurs-chercheurs et techniciens du CEA List permettent à l'Institut d'accompagner chaque année plus de 200 entreprises françaises et étrangères sur des projets de recherche appliquée s'appuyant sur 4 programmes et 9 plateformes technologiques. 21 start-ups ont été créées depuis 2003. Labellisé Institut Carnot depuis 2006, le CEA List est aujourd'hui l'institut Carnot Technologies Numériques.

Le Laboratoire d'Analyse Sémantique des Textes et des Images (LASTI) est une équipe de 25 personnes (chercheurs, ingénieurs, doctorants) menant des travaux de recherche sur les technologies de description et de compréhension du contenu multimédia (image, texte, parole) et des documents multilingues, en particulier à grande échelle. Les enjeux scientifiques sont :

- développer des algorithmes efficaces et robustes pour l'analyse et l'extraction de contenu multimédia, leur classification et analyse sémantique;
- reconstitution ou fusion de données hétérogènes pour interpréter des scènes ou documents;
- développer des méthodes et des outils pour la construction, la formalisation et l'organisation des ressources et connaissances nécessaires au fonctionnement de ces algorithmes;
- intégrer plusieurs de ces briques technologiques afin d'accéder à l'information et répondre à un besoin utilisateur (moteurs de recherche, chatbot, rapports synthétiques de veille).

Description du stage

Le fonctionnement des algorithmes de recommandation actuels induit l'apparition de bulles de filtres qui biaisent les informations proposées. Les utilisateurs reçoivent majoritairement des contenus en accord avec leurs convictions a priori menant ainsi à un appauvrissement des points de vue sur les sujets auxquels ils ont accès. En conséquence, une polarisation apparaît autour de sujets clivants (p.ex. l'efficacité de certains traitements pour la COVID-19 ou l'efficacité des politiques environnementales du gouvernement). Ce type de polarisation contribue à la dégradation du débat public et pousse certains citoyens vers des positions extrêmes qui contribuent à l'érosion du cadre démocratique de nos sociétés.

Le stage s'inscrit dans le cadre du projet ANR BOOM qui vise à ouvrir les bulles de filtres en utilisant des systèmes de recommandation qui proposent des informations diversifiées. Le stage sera centré sur le traitement automatique des langues et plus précisément sur le sujet de l'extraction d'information [1] qui est une composante essentielle de la future plateforme BOOM. Les objectifs principaux du stage sont :

- Définir une série de thèmes potentiellement polarisants et les modéliser en vue d'une leur détection automatique.
- Proposer une méthode semi-supervisée de collecte et filtrage de documents pertinents pour chaque thème.
- Proposer un outil de classification thématique basé sur des représentations textuelles profondes, telle BERT [2].
- Améliorer un système d'extraction et de normalisation d'entités nommées afin d'extraire les informations pertinentes relatives à un thème [3] [4].

CEA List
Laboratoire d'Analyse Sémantique Texte et Image
Centre de Saclay 91191 Gif-sur-Yvette France
http://www.kalisteo.eu

Contact Julien Tourille
Tél +33 (0)1 69 08 49 02
E-mail julien.tourille@cea.fr

Le poids donné à chaque objectif sera adapté en fonction des compétences et souhaits de la candidate ou du candidat sélectionné. La réalisation de ces objectifs sera facilitée par l'intégration des travaux de stage au sein d'outils déjà existants au laboratoire. Les résultats du stage seront complétés par l'exploitation d'une analyse d'opinions au niveau des entités afin de proposer des représentations sémantiques pertinentes en entrée du système de recommandation diversifiée.

Le stage nécessite des compétences en apprentissage profond qui seront développées pendant sa réalisation. Des connaissances en traitement automatique des langues ou au moins un fort intérêt pour le sujet sont également requis.

Selon les résultats obtenus, le stage pourra donner lieu à une publication scientifique.

Ce stage ouvre la possibilité de poursuite en thèse dans notre laboratoire.

Références

[1] Grishman, R. (2019). Twenty-five years of information extraction. Natural Language Engineering.
[2] Devlin, J., & al. (2018) Bert: Pre-training of deep bidirectional transformers for language understanding, NAACL 2019.
[3] Yadav, V. & al (2018). A Survey on Recent Advances in Named Entity Recognition from Deep Learning models, COLING 2018.
[4] Al-Moslmi, T. & al. (2020) Named Entity Extraction for Knowledge Graphs: A Literature Overview, IEEE Access.

Niveau demandé :	Ingénieur, Master 2
Durée :	6 mois
Rémunération :	entre 700 € et 1300 € suivant la formation.

Compétences requises :

- environnement de travail : linux
- maîtrise du langage de programmation Python
- expérience avec une bibliothèque de deep learning: Tensorflow, PyTorch ...
- connaissances en apprentissage automatique et en réseaux de neurones
- notions de base en traitement automatique des langues.