Sujet de stage

Equipe d'accueil :

Crédit Mutuel Arkéa / Pôle Innovation et Opération / DAEI / service IA-Factory

Encadrement :

Riwal LEFORT (riwal.lefort@arkea.com)

Titre du stage :

Résumé de documents textuels

Sujet détaillé :

Au sein du groupe Arkéa, de nombreux sujets sont liés à l'analyse des documents textuels. Les cas d'usage sont nombreux, allant de l'analyse des données du web, jusqu'aux traitements des demandes de financement, en passant par l'analyse des notes descriptives des projets internes. Bien souvent, le grand volume de données et la complexité des sujets traités rendent l'analyse manuelle particulièrement complexe et chronophage. Le stage porte sur le développement d'outils d'intelligence artificielle pour l'aide à l'analyse automatique de ces bases de données.

L'objectif final du stage est de proposer un prototype qui peut contenir les éléments suivants :

- identification automatique des thèmes abordés dans les documents,
- extraction automatique des éléments d'intérêts,
- résumé automatique des documents,
- classification automatique des documents (routage, scoring, indicateurs, etc),
- interprétation des scores de classification.

Les méthodes proposées vont s'appuyer sur les dernières contributions en *Machine Learning*, et plus particulièrement en *Natural Language Processing* (NLP) (*transformers* [1], BERT [2], GPT3 [3]). Nous effectuerons à la fois de la classification et de la génération [4] de textes. Dans ce contexte, une part importante du stage pourrait être consacrée au développement d'outils pour la récupération des données en provenance du web (*web scrapping*).

Les compétences attendues pour ce stage s'articulent autour des domaines suivants :

- l'informatique (capacité à développer en python (torch, tensor-flow), capacité à développer des codes de *web scrapping*),
- les mathématiques (capacité à comprendre des articles scientifiques et des algorithmes, capacité à analyser des résultats),
- la communication (capacité à communiquer avec le métier qui exprime son besoin et ses attentes, être force de proposition).

Mots clés :

Machine Learning, traitement du langage naturel (NLP), python, résumé automatique, classification automatique.

Références :

[1] "*Attention is all you need*", A. Vaswani *et. al.*, NIPS, 2017

[2] "*BERT* : *Pre-Trained of deep bidirectional transformers for language understanding*", NAACL-HLT, 2019

[3] "*Language models are few-shot learners*", T. B. Brown *et. al.*, 2020

[4] *Generative Adversial Networks*