Sujet de TER année 2021 – 2022

Encadrants : N. Vincent – B. Bohet
nicole.vincent@u-paris.fr
baptiste.bohet@sorbonne-nouvelle.fr

Deep learning pour l'analyse de texte

Ce projet est proposé dans le cadre d'une collaboration avec le laboratoire THALIM (UMR7172) à l'Université Sorbonne Nouvelle. Il relève des humanités numériques mettant en œuvre des méthodes numériques pour analyser si certains éléments d'un texte peuvent être mis en évidence par apprentissage. De nombreux travaux s'intéressent au langage naturel, sa compréhension ou sa traduction. On étudiera ici des données textuelles. L'objectif du projet est de détecter si un texte est un texte original ou s'il s'agit d'une traduction. Les données, le texte, peuvent être considérées sous différentes formes comme une suite de caractères, de mots, de phrases. Il semble raisonnable ici de le considérer comme une suite de mots.
Le problème posé est un problème de classification à deux classes. La quantité de données qui seront disponibles étant suffisante, il est possible de considérer une approche statistique et on pense aux réseaux de convolution profonds qui permettent après une phase d'apprentissage, de construire un modèle discriminant les deux classes considérées ici, les textes traduits et les textes originaux.
Deux aspects sont à envisager au cours du travail, la représentation des données et l'apprentissage en lui-même.

Le travail comporte alors plusieurs étapes :

- Recenser les méthodes de codage d'un texte
- Recenser les méthodes de transformation d'un texte en image
- Réaliser l'apprentissage en fonction de la représentation du texte choisie
- Tester le système construit et étudier la quantité de texte nécessaire pour assurer une classification correcte
- On comparera l'efficacité de plusieurs méthodes.

Le logiciel sera codé en Python.