

Offre de stage : Identification et classification automatique de style de textes littéraires francophones nativement numérique

Contact : julien.velcin@univ-lyon2.fr

Période envisagée : février/mars - juillet 2021 (4 à 6 mois)

Gratification : indemnités de stage légales (3,60 euros par heure pour 35 heures de travail par semaine, soit environ 550 € par mois)

Localisation : Laboratoire ERIC, Université Lyon 2 (campus de Bron)

Mots-clefs : réseaux de neurones profonds, modèles génératifs, traitement automatique du langage, fouille de données textuelles, Big Data

Contexte du stage

Le projet LIFRANUM vise à constituer et analyser le corpus des productions littéraires francophones nativement numériques. Il s'agit d'un projet financé par l'Agence Nationale de la Recherche (ANR) qui regroupe un laboratoire de sciences humaines (MARGE), un laboratoire d'informatique (ERIC) et la Bibliothèque Nationale de France (BnF). Dans le cadre de ce projet, le laboratoire ERIC recherche un.e candidat.e pour un stage qui débutera en février ou mars 2021.

Mission :

La personne recrutée aura pour mission d'engager un traitement du contenu littéraire produit et diffusé 100 % sur le web (par exemple, twittérature, forme de littérature numérique pratiquée sur Twitter). A cette fin, un corpus de données textuelles collecté par les chercheurs du projet sera mis à disposition afin d'alimenter les modèles d'apprentissage profond. Les modèles devront intégrer la représentation des auteurs dans un espace latent afin de répondre à des questions pratiques comme calculer des similarités entre auteurs ou découvrir des communautés dont la production partage des similitudes.

En particulier, une question intéressante (voir par exemple Tikhonov and Yamshchikov, 2018) est de savoir s'il est possible de résoudre des tâches de "transfert de style" (*style transfer*) sur le texte comme cela a été fait sur les images (Gatys et al., 2016). Pour contribuer à répondre à cette question, le/la stagiaire aura pour mission de tester des approches récentes de génération automatique de texte employant des réseaux de neurones profonds (*deep learning*).

Afin de mettre en place une expérimentation, la personne recrutée aura l'opportunité d'utiliser des solutions innovantes de traitement de données à large échelle (*big data*) pour les humanités numériques (Ruest et al, 2020). Enfin, le/la stagiaire pourra aussi être amené à participer à la proposition de nouvelles solutions adaptées aux données du projet LIFRANUM.

Références

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. "Image style transfer using convolutional neural network". Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference, pages 2414–2423 (2016).

Tikhonov, Alexey, and Ivan P. Yamshchikov. "What is wrong with style transfer for texts?." arXiv preprint arXiv:1808.04365 (2018).

Nick Ruest, Jimmy Lin, Ian Milligan, and Samantha Fritz. "The Archives Unleashed Project: Technology, Process, and Community to Improve Scholarly Access to Web Archives". In Proc. of the ACM/IEEE JCDL '20, New York, NY, USA, 157–166 (2020). DOI: <https://doi.org/10.1145/3383583.3398513>

Profil recherché

Nous recherchons un·e candidat·e ayant des compétences solides en analyse/fouille de données, en programmation (Python de préférence) et si possible des notions de traitement automatique des langues (*natural language processing*) / fouille de données textuelles (*text mining*) et d'apprentissage automatique (*machine learning*). Un intérêt pour le travail pluridisciplinaire serait un plus.

Déroulement du stage

Le stage se déroulera dans les locaux du laboratoire ERIC, sur le campus Porte des Alpes (à environ 30 min. du centre-ville de Lyon en tram). Il sera encadré par un enseignant-chercheur permanent accompagné par un post-doctorant et un doctorant travaillant sur les thématiques du stage. Une réunion hebdomadaire est prévue, en plus des réunions organisées pour le projet LIFRANUM.

Procédure de candidature

Les candidats doivent envoyer les documents suivants à l'adresse : julien.velcin@univ-lyon2.fr avant le 10 décembre 2020 :

- CV
- lettre de motivation
- relevés de notes des deux dernières années

Les candidat·es retenu·es seront convoqué·es pour un entretien durant la semaine du 14 décembre. Les résultats devraient être communiqués avant les congés de fin d'année.