OFFRE DE STAGE EN INFORMATIQUE (niveau master ou ingénieur)

Pour le projet **CLexIC** (Création Lexique Innovation Crowdfunding)
Porteuse du projet : **Anna Pappa** (ap@up8.edu)
Durée du stage : 5/6 mois (début : Mars 2023)
Gratification **:** montant légal en vigueur (environ 600€ / mois).
Laboratoire d'accueil : Equipe PASTIS du **LIASD** (EA 4383), Université Paris 8, 2 rue de la liberté, 93526 Saint Denis

Sujet de stage

- Création automatique d'un lexique multilingue dont les termes désignent l'*innovation non technologique* (initialement en anglais et en français et par la suite en d'autres langues présentes sur les sites de plus de 4000 entreprises).
- Étude comparative contextuelle des termes désignant l'*innovation non technologique*, issus de deux corpus différents : un corpus composé de textes extraits de sites web des entreprises participantes et un autre corpus composé de textes issus des sites de crowdfunding (projets collaboratifs innovants).

Problématique

L'écosystème du crowdfunding présente un intérêt particulier de par la multitude des projets proposés et le financement participatif des 'foules'. Les descriptifs des activités et projets proposés sur les plateformes de crowdfunding sont différents des descriptions des produits et des services sur les sites des entreprises.
Nous voulons identifier et analyser le contexte dans lequel l'*innovation non technologique* est représentée dans ces descriptifs.

Étapes et réalisation

Le stage comportera les étapes suivantes (la réutilisabilité des ressources et les codes produits feront l'objet d'une documentation tout au long du stage) :
1. Création corpus (Boudabous & Pappa, 2021) à partir des textes issus des plateformes du crowdfunding.
2. Annotation automatique du nouveau corpus avec apprentissage actif (Ren & al., 2021, Li et al., 2020) et apprentissage par transfert (Weiss & al., 2016).
3. Création d'une version de détection des caractéristiques lexicales portant sur l'innovation non technologique à partir d'un modèle BERT (Koufakou & al., 2020).
4. Rédaction du rapport de stage, et mise en forme des ressources et codes produits.

Compétences particulières et formation requise

Ce stage s'adresse aux étudiant.e.s de master 1 ou 2 ou étudiants ingénieurs en informatique.
- Connaissances des : techniques de scraping, méthodes d'apprentissage profond et modèles de type BERT.
- Très bon niveau en programmation Python (avec les framework Pytorch, Tensorflow).
- Connaissances d'outils TAL appréciées (outils fondés sur l'apprentissage, modèles de langue, classifieurs, si possible outils statistiques de lexicométrie).
- Curiosité et volonté de tester de nouvelles méthodes.

Candidature

L'étudiant-e sera accueilli-e dans les locaux de l'Université Paris 8 au laboratoire LIASD.

Contacts :.

Revekka Kyriakoglou, LIASD, kyriakoglou@up8.edu
Alice Millour, LIASD, am@up8.edu

References

Maroua BOUDABOUS and Anna PAPPA. 2021. WebT-IDC: A web tool for intelligent dataset creation a use case for forums and blogs. *In Procedia Computer Science*, volume 192, pages 1051–1060. Elsevier.
Pengzhen REN, Yun XIAO, Xiaojun CHANG, Po-Yao HUANG, Zhihui LI, Brij B GUPTA, Xiaojiang CHEN, and Xin WANG. 2021. A survey of deep active learning. ACM Computing Surveys (CSUR), 54(9):1–40.
Kun LI, Chengbo CHEN, Xiaojun QUAN, Qing LING, and Yan SONG. 2020. Conditional Augmentation for Aspect Term Extraction via Masked Sequence-to-Sequence Generation. *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7056–7066. Association for Computational Linguistics.
Karl WEISS, Taghi M. KHOSHGOFTAAR, & DingDing WANG. A survey of transfer learning. *J Big Data* **3**, 9 (2016).
Anna KOUFAKOU, Endang Wahyu PAMUNGKAS, Valerio BASILE, and Viviana PATTI. HurtBERT: Incorporating Lexical Features with BERT for the Detection of Abusive Language. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 34–43, 2020, Association for Computational Linguistics.