Amélioration des modules d'une plateforme de textmining

Contexte général

Ce stage s'inscrit dans le cadre des projet ANR CROQUIS "Collecte, représentation, complétion, fusion et interrogation de données de réseaux d'eaux urbains hétérogènes et incertaines" financé par l'Agence Nationale de la Recherche (ANR-21-CE23-0004) et STARWARS "STormwAteR and WastewAteR networkS heterogeneous data AI-driven management" financé par le l'action Marie Skłodowska-Curie du programme d'innovation Horizon Europe, (HORIZON-MSCA-2021-SE-01). Dans ces projets des chercheurs en Sciences de l'Eau et en Intelligence Artificielle unissent leurs efforts pour proposer de nouvelles méthodes pour la représentation, la complétion, la fusion, l'archivage, la réparation et l'interrogation des données hétérogènes décrivant les réseaux d'eau. Les informations textuelles utilisées dans le projet proviennent de la plateforme de textmining WEIR-P qui a été développée lors du projet "MeDo" (Megadonnées, données liées et fouille de données pour les réseaux d'assainissement-2018- 2020), financé par la Région Occitanie-Pyrénées-Méditerranée à travers le dispositif "Rechercheet Société(s) 2017[1].

La plateforme WEIR-P combine des techniques de recherche d'information (RI) et d'extraction d'information (IE) adaptées à la langue française et au domaine des eaux usées. Le traitement en lui-même comporte 5 étapes : collecte de documents ; reconnaissance d'entités nommées ; extraction de relations sémantiques ; cartographie et visualisation de données ; exportation vers une base de données relationnelle et intégration dans un Système d'Informations Géographique.

Objectif

Dans le cadre des projets CROQUIS et STARWARS, un nouveau modèle de données et une ontologie pour les réseaux d'assainissement unitaires et séparatifs a été élaboré. L'objectif du stage est de faire évoluer la plateforme WEIR-P, notamment son module de reconnaissance d'Entités Nommées (EN) sur la base de cette ontologie. Pour cela, un processus de benchmarking (série d'évaluations et protocoles de test) doit être défini et appliqué selon les critères et besoins des experts du projet.

Il s'agira plus particulièrement de :

- Définir et mettre en œuvre un protocole d'évaluation en accord avec l'équipe du projet.
- Ré-entraîner les modèles de reconnaissance d'EN intégrés dans la plateforme existante sur la base de la nouvelle ontologie.
- Tester les nouveaux algorithmes de REN inclus dans ces librairies selon le protocolepré-établi.
- Restituer les résultats et mettre en œuvre les améliorations en modifiant au besoin l'architecture existante.

Profil recherché

- Master 2 en informatique.
- Outils et langages : Python (JAVA et OWL seront un plus)
- SGBD MongoDB et outils NLP (souhaité)
- Capacité de travail dans une équipe pluri-disciplinaire.

Encadrement et équipe projet

Encadrement : Serge Conrad et Nanée Chahinian, UMR HydroSciences Montpellier.
Équipe projet : Carole Delenne, Stéphane Debard et Batoul Haydar, UMR HydroSciences Montpellier.
Francesca Frontini et Franco Alberto Cardillo, CNR-ILC Pise.

[1] http://webmedo.msem.univ-montp2.fr/

Divers

- Durée : 5 mois avec possibilité d'extension.
- Gratification : Taux légal en vigueur
- Localisation : UMR HSM Montpellier
- Candidature : Envoyer un CV + relevés de notes des deux dernières années à : nanee.chahinian@ird.fr et serge.conrad@umontpellier.fr