Stage Master 2 Pro ou école ingénieur

Développement et benchmarking d'une plateforme de Text Mining
dans le cadre du Projet **MeDO**

Encadrement : Thierry Bonnabaud La Bruyère (UM-UMR HSM), Laurent Deruelle (Berger-Levrault)
Equipe Projet : Mathieu Roche (TETIS, Cirad), Maguelonne Teisseire (TETIS, Irstea), Nanée Chahinian (HSM/IRD)

Contexte général :

Le stage se déroulera dans le cadre du projet MeDO (Megadonnées données liées et fouille de données pour les réseaux d'assainissement) porté par le laboratoire HydroSciences Montpellier. Ce projet a pour objectif de tirer profit des mégadonnées disponibles sur le web pour renseigner la géométrie et l'historique d'un réseau d'assainissement. L'originalité de l'approche réside à la fois dans la combinaison des différentes techniques de fouille de données, dans la multiplicité des sources analysées (forums de lecteurs, sites web municipaux, appels d'offres...) et dans leur adaptation à l'étude d'un réseau d'eau urbain. Le consortium scientifique du projet, qui bénéficie de la subvention de la Région Occitanie-Pyrénées-Méditerranée à travers le dispositif « Recherche et Société(s) 2017 », regroupe les UMRs HydroSciences Montpellier, Praxiling et TETIS ainsi que l'entreprise Berger-Levrault.

Objectif :

L'objectif de ce stage est de participer au développement de la chaîne de traitement de documents textuels du projet. Plus particulièrement, il concerne l'intégration d'un outil de NLP (Natural Language Processing) à choisir parmi ceux de référence (Stanford, Polyglot, OpenNLP, etc.). 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.

De façon plus précise, le stage sera décomposé en plusieurs étapes :

1. Prise en main de la chaîne de traitement actuelle du projet MeDO et des différents outils NLP à évaluer ;
2. Formatage des différentes informations propres aux documents textuels de la base de données (Sous MongoDB) selon les spécificités des différents outils choisis ;
3. Définition et mise en œuvre du protocole d'évaluation en accord avec l'équipe du projet selon des critères standards (Précision, Rappel, …) ;
4. Réalisation de séries de tests sur le Gold Standard (annoté à l'aide de l'outil Brat) selon des extractions spécifiques ;
5. Restitution des résultats et proposition d'une architecture adaptée aux besoins spécifiques du projet MeDo avec la mise en œuvre correspondante.

Compétences requises :

Langages Python et Java,

SGBD MongoDB et outils NLP (souhaité)

Capacité de travail en équipe pluridisciplinaire.

Divers :

Durée : 6 mois

Gratification : taux légal en vigueur

Localisations : HSM – TETIS - Berger Levrault – Montpellier

Candidature : Envoyer un CV + relevés de notes des deux dernières années à
nanee.chahinian@umontpellier.fr et maguelonne.teisseire@irstea.fr