OFFRE de STAGE M1 – ANSES – SPAAD

Durée : minimum de 2 mois (stage rémunéré)

Localisation : ANSES, service SPAAD, Maisons-Alfort (94)

Contact : virginie.chesnais@anses.fr

Evaluation des outils de classification de texte pour le nettoyage de métadonnées issues de bases de données publiques

Ces dernières décennies, la fréquence d'émergences de maladies infectieuse a augmenté dans le monde selon des mécanismes et des voies de circulation variés. De nombreux facteurs ont pu être impliqués dans cette évolution, notamment l'environnement, le climat, les pratiques agro-industrielles, les comportements humains. La démocratisation du séquençage d'ADN des agents pathogènes a permis de constituer des bases de données publiques regroupant plusieurs milliers de génomes bactériens. Une limite à l'utilisation de ces données reste l'absence d'harmonisation des métadonnées contextuelles décrivant les échantillons disponibles dans ces bases de données.

Dans le cadre d'un projet de recherche, nous avons récupéré et annoté manuellement plusieurs milliers de données en utilisant un référentiel Européen permettant de décrire des échantillons similaires de la même façon. Nous souhaitons évaluer la possibilité d'automatiser ces taches de nettoyage de données par des approches de texte miniing (classification de texte) afin de faciliter l'inclusion de données nouvellement déposées sur les bases de données publiques. Des premiers tests réalisés avec l'algorithmes fastext ont permis d'obtenir une accuracy > 95% sur la classification des échantillons en trois groupes : aliments, environnement, prélèvement humain. L'étudiant aura pour objectif de i. confirmer ces premiers résultats en incluant les nouveaux échantillons récupérés plus récemment et encore non-annoté, ii. tester d'autres outils de classification de texte et iii. évaluer la possibilité de classifier plus en détail les échantillons (ex : annoter tous les fromages vs les légumes)