Stage Master 2

Biodiversité et pratique de recherche : extraction automatique de mots-clés caractérisant les thématiques saillantes issues de données textuelles

Durée : 5 à 6 mois à partir de février 2020

Encadrement : Mathieu Roche (TETIS) et Christian Leclerc (AGAP)

Contact : mathieu.roche@cirad et christian.leclerc@cirad.fr

Description :

De nombreux travaux de fouille de textes permettent (i) de faire émerger les descripteurs linguistiques les plus significatifs (mots, syntagmes) à partir d'un corpus puis (ii) de les regrouper. Ceci permet de mettre en relief, de manière automatique, les thématiques abordées dans les textes facilitant l'organisation et l'indexation des documents, la recherche d'information, la compréhension et l'analyse des textes. Il permet aussi de comparer, pour une période donnée, les approches privilégiées par différentes unités de recherche, ou encore de décrire l'évolution de ces approches au cours du temps. Cette analyse portera sur Biodiversité et pratique de recherche au Cirad, avec l'objectif d'appliquer la méthode à d'autres thématiques, notamment le territoire et la mobilité.

La réalisation du premier point (identification des descripteurs linguistiques significatifs) s'appuie, en grande partie, sur l'utilisation de méthodes d'extraction de la terminologie à partir de textes, en combinant méthodes linguistiques et statistiques pour constituer une liste de descripteurs linguistiques. La deuxième étape du processus consiste à utiliser ces descripteurs afin de mettre en lumière les différentes thématiques abordées dans les textes. Pour découvrir des structures thématiques "cachées" dans les corpus de textes, les méthodes appelées "topic models" seront utilisées, notamment, le modèle probabiliste génératif LDA, i.e. Latent Dirichlet Allocation.

Dans ce contexte, les objectifs du stage sont déclinés selon 4 sous-tâches :

(1) Intégrer des outils de la littérature d'extraction de la terminologie (en particulier BioTex - http://tubo.lirmm.fr:8080/biotex) et des approches LDA dans le cadre du développement d'un système générique et utilisable par des non informaticiens.
(2) Intégrer et combiner des ressources sémantiques (vocabulaire contrôlé) fournies par les utilisateurs aux méthodes d'extraction de la terminologie.
(3) Étudier la valeur structurante des termes rares (queue de distribution) associées aux fonctions de rangs propres aux systèmes d'extraction de la terminologie. De nouvelles fonctions de rangs pourront alors être proposées, pour mettre en valeur les termes rares et pertinents.

Lieu et gratification :

Ce stage basé au Cirad à Montpellier (https://www.cirad.fr/) bénéficie d'une gratification mensuelle de 580 euros.

Profil :

Master 2 ou École d'Ingénieur en Informatique / Science des Données