Stage Master 2 / dernière année école d'ingénieur – 6 mois – 2024

Extraction de connaissances à partir de données textuelles : application à la découverte de règles de changements d'usage des sols

Contexte général et projets de recherche :
Les approches en modélisation spatio-temporelles nécessitent d'utiliser des connaissances qui sont ensuite formalisées en règles de type SI condition ALORS action (e.g. si l'altitude d'une parcelle cultivée est supérieure à 1700 m, la probabilité que la culture soit de type "pomme de terre" est très élevée). Dans ce contexte, une connaissance correspond à un état d'occupation ou d'utilisation des sols, un changement entre deux états selon des processus anthropiques ou non. Ces connaissances peuvent être qualitatives ou quantitatives.
L'un des moyens d'obtenir ces connaissances est la sollicitation d'experts par des entretiens, approche pouvant s'avérer coûteuse en temps et souffrant de biais (e.g. différence d'angle d'approche selon les disciplines). L'analyse de la littérature scientifique et technique représente une autre manière d'identifier des connaissances, complémentaires aux connaissances apportées par les experts, et de s'affranchir de certains biais (meilleure profondeur temporelle par exemple). La sélection de documents pertinents et leur analyse est une tâche chronophage pour laquelle des approches d'extraction automatique peuvent être utilisées. En particulier, les outils de traitement automatique du langage et de fouille de texte sont de bonnes solutions. Dans ce contexte, le traitement automatique du langage et l'analyse des résultats peuvent permettre : (1) l'identification et la sélection des connaissances d'intérêt, (2) la quantification de changements d'état et des processus qui lui sont liés, et (3) la transformation de ces connaissances en règles logiques.

Cette problématique s'inscrit dans le cadre et du projet CECC (Cycle de l'Eau et Changements Climatiques) en collaboration avec le projet TipHyc (Tipping points in the West African Hydrological Cycle), portant sur les changement de régimes hydrologiques en Afrique de l'Ouest. Il s'agit de mobiliser des méthodes informatiques afin d'extraire des connaissances à partir de données textuelles.

Objectifs du stage :
L'objectif est de développer une méthode d'extraction de connaissances transparente et reproductible permettant de mobiliser de larges volumes d'informations textuelles. Nous souhaitons plus particulièrement extraire les connaissances relatives à des transitions et évolutions d'usage et occupation des sols, ainsi qu'aux facteurs causaux de ces transitions (e.g. urbanisation, déforestation, mise en culture). Cet objectif se décline en deux sous-objectifs : (1) l'identification automatique de connaissances sur les changements d'occupation ou d'usage des sols et leurs processus en Afrique de l'Ouest, (2) l'analyse de la distribution statistique de ces connaissances dans un ensemble de documents techniques et scientifiques.

Le stage s'articulera en plusieurs étapes :

1. l'analyse des résultats d'une indexation automatique et d'expansion automatique de termes à partir d'un corpus de documents techniques et scientifiques

2. l'identification d'expressions "trigger" pour cibler les segments d'intérêt pouvant contenir des connaissances,

3. l'évaluation de méthodes d'extraction de connaissances à partir du corpus. Selon les enjeux méthodologiques identifiés, des approches supervisées ou non supervisées seront implémentées. Le point d'ancrage des connaissances à extraire sera les changements d'occupation et d'usage des sols et les processus qui leur sont associés,

4. l'analyse statistique des connaissances extraites afin d'identifier des connaissances majoritaires (très représentées dans le corpus) ainsi que les connaissances sous-représentées.

Les données d'étude (corpus de documents en anglais et en français) sont déjà constituées et ont préalablement été indexées à l'aide d'une nomenclature experte. Des outils d'extraction et de visualisation développés par l'UMR TETIS pourront librement être mobilisés au cours du stage.

Organisation du stage :
Le stage se déroulera sur une période de 6 mois, à compter de février 2024, dans les locaux d'HSM à Montpellier.
L'étudiant·e sera accueilli·e dans l'équipe HEC de l'UMR HSM (Hydrosciences Montpellier , 28 agents) et sera encadré·e par Arthur Crespin-Boucaud, géographe, post-doctorant à l'Institut de Recherche pour le Développement (IRD) au sein de l'UMR HSM, Sarah Valentin, chercheuse en fouille de données au Cirad à l'UMR TETIS et Christophe Peugeot, Hydrologue, chargé de recherche à l'IRD à HSM.

Le déroulement du stage se fera dans un contexte interdisciplinaire fort, en collaboration avec Maguelonne Teisseire (INRAE, UMR TETIS) et Nanée Chahinian (IRD, UMR HSM).

En fin de stage, les résultats du travail effectué seront présentés oralement aux membres des projets de recherche ainsi qu'aux équipes TETIS et HSM. En plus de la rédaction d'un mémoire de Master 2 ou de fin d'étude selon les attentes de sa formation, d'autres modalités de valorisation des résultats seront éventuellement envisagées avec les encadrants au cours du stage.

Compétences recherchées :
- Formation en informatique,
- Bonne maîtrise du langage de programmation Python,
- Connaissances en fouille de données et/ou apprentissage automatique voir traitement automatique du langage,
- Maîtrise de l'anglais écrit,
- Intérêt pour les applications socio-environnementales et le travail interdisciplinaire.

Candidature :
Envoyer CV, lettre de motivation et relevé de notes M1 (ou 4ème année) avant le 11/11/2023.
Sarah Valentin et Arthur Crespin-Boucaud
Mails: sarah.valentin@cirad.fr, arthur.crespin-boucaud@ird.fr