Sujet de stage École ingénieur / M2

CIRAD - UMR TETIS, Montpellier, France

Titre

Intégration des bases de données sur l'extraction de ressources minières avec NLP et modèles de langage

Keywords

Traitement Automatique du Langage (TAL), Grand Modèle de Langage, Ressources Minières, Acquisitions de Terres à Grande Échelle, Transition Énergétique

Contexte

L'initiative Land Matrix (https://landmatrix.org) et son observatoire mondial des acquisitions de terres à grande échelle ont pour objectifs de créer une source fiable de données permettant d'alimenter les débats et de mettre en œuvre des actions éclairées sur les transactions foncières à grande échelle. La Land Matrix collecte des données sur les tentatives prévues, conclues et échouées visant à acquérir des terres au moyen d'achat, de location ou de concession à des fins de production agricole, extraction de bois, extraction minière, finance du carbone, activités industrielles, production d'énergie renouvelable, conservation de la nature et tourisme, dans les pays à revenus faibles ou intermédiaires.

Bien que les données de la Land Matrix restent la référence mondiale sur les phénomènes d'acquisitions de terres dans le monde académique, la couverture des sites d'exploitation minière dans la base n'est toujours pas optimale, pour plusieurs raisons historiques et liées à des soucis d'accès aux données. D'autre part, le suivi des activités minières (et des investisseurs associés) sur ces sites est au cœur des études concernant la transition énergétique, qui est à son tour une dimension importante d'une stratégie globale de lutte contre le changement climatique. Ces études peuvent également contribuer à sensibiliser aux injustices distributives et à la répartition inéquitable des coûts, les pays cibles supportant la plupart des coûts sociaux et environnementaux de l'extraction des ressources dans des régions marquées par l'insécurité foncière et alimentaire et l'instabilité en termes de gouvernance.

L'objectif du stage est de construire une base de données la plus complète possible sur les sites d'exploitation minière dans le monde, en incluant les informations sur les investisseurs derrière ces sites et les informations géospatiales associées (coordonnées GPS et/ou polygones). Pour y parvenir, il faudra intégrer les données Land Matrix (https://landmatrix.org/) avec celles d'une autre plateforme, ResourceContracts (https://www.resourcecontracts.org/). ResourceContracts est un référentiel en ligne de contrats pétroliers, gaziers et miniers. Le site comprend la version PDF et le texte intégral de chaque contrat, des étiquettes des principaux termes financiers, sociaux, environnementaux, opérationnels et juridiques et des outils de recherche et de comparaison des contrats. Des techniques de Traitement Automatique du Langage (TAL), possiblement avec l'utilisation des modèles de type LLM (Large Language Models - Grand Modèle de Langage) seront nécessaires pour compléter cette tâche d'intégration. Finalement, la base obtenue sera mise en lien avec une autre base qui fournit l'étendue géographique des sites minières sous forme des polygones (i.e., shapefiles). Des méthodes de télédétection pourront être mobilisés à côté des techniques de TAL pour compléter cette deuxième étape. La base finale sera enfin utilisée pour mener deux cas d'études analytiques sur des pays spécifiques, qui sont des hotspots de l'extraction des ressources minière nécessaires à la transition énergétique : la République démocratique du Congo et l'Argentine.

Compétences du candidat/e :

- Analyse des données (collecte, exploration, mise en lien)
- Programmation (préférablement Python)
- Capacités d'analyse, rédactionnelle et de synthèse
- Travail d'équipe
- Des expériences en TAL et/ou Télédétection seront considérées comme un plus.

Informations complémentaires :

Durée de 6 mois, à partir de février 2025.
Le stage se déroulera au CIRAD, dans l'UMR TETIS (Territoire, Environnement, Télédétection et Information Spatiale), située dans les locaux de la Maison de la Télédétection à Montpellier.
Le stagiaire sera encadré par Rémi Decoupes (INRAE, UMR TETIS), Roberto Interdonato et Simon Madec (CIRAD, UMR TETIS), Jérémy Bourgoin (CIRAD/ILC) et Marie Gradeler (ILC).
Si des résultats significatifs sont obtenus, le stage peut conduire à une publication scientifique.

Financement :

Le stage est soutenu par le projet Land Matrix. La rémunération du stagiaire sera de l'ordre de 600 euros par mois.

Modalité de candidature :

Attention : cette proposition ne concerne que les stages d'étudiants sous convention avec une école ou une université : il ne s'agit pas d'une offre d'emploi.
Envoyer une lettre de motivation d'une page, précisant en outre la durée et période possible du stage, un CV détaillé, et un relevé des notes à : remy.decoupes@inrae.fr et roberto.interdonato@cirad.fr , en précisant dans l'objet « CANDIDATURE STAGE LAND MATRIX 2025 ».

Date limite pour l'envoi du dossier : 06 Decembre, 2024