

Sujet de stage École ingénieur / M2

CIRAD - UMR TETIS, Montpellier, France

Titre

Adapter les modèles de type ChatGPT pour interroger en langage naturel une base de données mondiales sur les acquisitions foncières

Keywords

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, d'extraction de bois, de finance du carbone, d'activités industrielles, de production d'énergie renouvelable, de conservation de la nature et de tourisme, dans les pays à revenus faibles ou intermédiaires.

Référence mondiale sur les phénomènes d'acquisitions de terres dans le monde académique, les données de la Land Matrix restent encore peu utilisées pour accompagner l'action publique malgré d'importants efforts pour que les données soient en accès libre. Nous faisons l'hypothèse que l'accès aux données et leur exploitation restent des freins importants à l'utilisation des informations contenues dans la base de la Land Matrix. En effet, une certaine expertise en informatique et une bonne connaissance des champs de variables sont nécessaires pour que l'utilisateur puisse accéder aux informations nécessaires.

L'objectif du stage est de simplifier l'accès aux données de la base Land Matrix en permettant son interrogation en langage naturel. Les récentes avancées en intelligence artificielle, dont bénéficient les modèles de langues de type OpenAI/ChatGPT ou meta/Llama-2, offrent de nouvelles opportunités pour convertir des questions, formulées en langage naturelle, en langage informatique (génération de code python et/ou SQL). Ces modèles généralistes nécessitent, cependant, d'être ajustés à la structure de données de la base Land Matrix afin de n'extraire que les données réellement attendues par les utilisateurs. Les stratégies d'ajustement sont un attendu de ce stage dont l'objectif final est de proposer un agent conversationnel (Chat bot) capable de générer les requêtes informatiques à partir d'une formulation en langage naturel. Une requête potentielle serait un croisement de filtres géographiques, filtres de types d'investisseurs et de filtres d'intentions d'investissements, par exemple : « j'aimerais savoir quelles banques soutiennent des investissements miniers (lithium) en Argentine ».

Compétences du candidat/e :

- Analyse des données (collecte, exploration, mise en lien)
- Programmation (préférentiellement Python)

- Capacité d'analyses, rédactionnelles et de synthèse
- Travail d'équipe

Informations complémentaires :

Durée de 6 mois, à partir de février 2024.

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), Jérémy Bourgoïn, Roberto Interdonato (CIRAD, UMR TETIS) et Marie Gradeler (ILC).

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 jeremy.bourgoin@cirad.fr

Date limite pour l'envoi du dossier : 30 Novembre, 2023