Stage de Master 2 #DigitAg: Application de méthodes de text-mining et Natural Language Processing (NPL) pour l'extraction automatisée de données web pour la création d'une base de données des traits phénotypiques des adventices tropicales

Période : Février-Juillet 2024

Résumé:

L'objectif de ce stage est de définir, extraire et organiser des données de traits phénotypiques et de distribution géographique des adventices tropicales (plantes qui s'établissent spontanément dans les systèmes tropicaux cultivés) afin de construire une base de données complète et exploitable pour décrire et comparer leur caractéristiques et leur diversité. Les plantes adventices des cultures tropicales ont un impact significatif sur la productivité agricole mais aussi sur de multiples services écosystémiques comme le maintien de la fertilité des sols, la réduction de l'érosion, la régulation des bioagresseurs etc. Une meilleure connaissance des traits phénotypiques des adventices tropicales permettrait d'améliorer notre compréhension de leur fonctionnement et de leur écologie, qui déterminent leur développement aux sein des cultures, leur réponse aux facteurs environnementaux et aux pratiques agronomiques, ainsi que leur distribution géographique. C'est donc une étape cruciale pour développer des stratégies de gestion efficaces.

Des approches inspirées par l'écologie fonctionnelle sont de plus en plus utilisées en agroécologie, car elles permettent de comprendre, quantifier et piloter les mécanismes physiologiques et écologiques et qui déterminent l'influence des pratiques agricoles et la manière dont les espèces interagissent entre elles et avec leur environnement au sein des systèmes cultivés. On peut considérer comme "trait" toute caractéristique phénologique, physiologique, morphologique, ou anatomique (ex: date de floraison, surface des feuilles, hauteur de la plante, densité des tissus…) associée à des fonctions écophysiologiques et écologiques (ex: reproduction, tolérance à la sécheresse, vitesse de croissance, compétitivité face à d'autres espèces…).

Aujourd'hui, différentes sources de données de traits de plantes sont accessibles en ligne et contiennent de nombreuses informations concernant les adventices tropicales. Les articles scientifiques et publications académiques fournissent parfois des données détaillées issues de recherches spécifiques et sont souvent disponibles dans des revues spécialisées et lors de conférences. Les bases de données publiques et spécialisées, telles que le GBIF (Global Biodiversity Information Facility), TRY (Plant Trait Database) ou GIFT (Global Inventory of Floras and Traits), compilent de nombreuses informations sur la distribution et les traits des plantes. Les sites des herbiers, tels que l'Herbier de Paris, l'herbier numérique de Kew Gardens ou l'herbier tropical du Cirad, offrent également des données botaniques utiles qui peuvent renseigner sur les traits des plantes . En plus de ces sources, des plateformes généralistes comme Wikipedia peuvent parfois fournir des informations générales sur les espèces de plantes adventices. Ces sources variées peuvent permettre de constituer une base de données riche et étendue sur les traits des adventices tropicales. L'exploration et l'extraction de données de ces sources web représentent néanmoins un défi de taille par la quantité, l'hétérogénéité et la qualité des informations qu'elles contiennent qui nécessite la mise en place de méthodes numériques automatisées jusque-là peu utilisées en écologie fonctionnelle et en agroécologie.

Dans ce contexte, le stage a pour objectif de collecter, de traiter et d'organiser des données pertinentes des traits des plantes en utilisant des techniques de web scraping et le text mining. La finalité du stage est de construire une base de données relationnelle structurée regroupant les traits des adventices tropicales qui sera un outil précieux pour les chercheurs, permettant d'effectuer différentes analyses, pour une meilleure compréhension de la réponse des espèces adventices aux facteurs environnementaux et aux pratiques agricoles, ainsi que de leur impact sur le fonctionnement des cultures tropicales. Le/la stagiaire sera responsable du développement des scripts de web scraping en utilisant principalement le langage de programmation Python via des bibliothèques spécialisées. Une première étape sera d'évaluer et de comparer les performances spécifiques de différentes méthodes de text mining (Spacy, GLiNER, UniversalNER) pour l'extraction des entités visées dans le cadre de ce travail. Les méthodes retenues permettront d'extraire des informations spécifiques sur les plantes adventices et leurs traits phénotypiques et éventuellement leur répartition géographique. Les données seront ensuite normalisées et structurées de manière cohérente pour leur intégration dans une base de données relationnelle (PostgreSQL). Enfin, des analyses descriptives des données recueillies pourront être réalisées pour évaluer l'état et le volume des connaissances accumulées dans la base de données, et la distribution des valeurs de traits des espèces en fonction de leur distribution au travers des gradients environnementaux, géographiques et agronomiques.

Les compétences requises pour ce stage incluent une bonne maîtrise du langage de programmation Python. Une expérience avec des concepts et des outils de web scraping et de text mining serait fortement appréciée. Des connaissances en gestion de base de données relationnelles et en analyses statistiques descriptives appliquées à de gros jeux de données seraient également appréciées. Une compréhension des concepts de base en écologie végétale et en agroécologie serait un plus. Enfin, le/la candidat(e) devra avoir une bonne capacité à travailler en autonomie et de manière rigoureuse, ainsi qu'une capacité à synthétiser et présenter de manière claire et informative à une équipe multidisciplinaire l'approche numérique développée et les données récoltées au cours du stage.

Ce stage offre une opportunité pour un/une étudiant(e) de Master 2 d'appliquer ses compétences en programmation à un projet dans les domaines des sciences de l'information de la biodiversité (éco-informatique) et de l'écologie des adventices tropicales. Le/la stagiaire contribuera directement à la construction d'une base de données pour la recherche et le développement dans le secteur agricole, tout en acquérant une expérience pratique et approfondie en gestion de de gros jeux de données complexes et hétérogènes. De plus, ce projet de stage bénéficiant d'un financement DigitAg (Institut Convergences Agriculture Numérique), pourra permettre à l'étudiant(e) de s'intégrer dans la communauté scientifique #DigitAg (participation aux séminaires, contact avec les doctorants, post-docs, entreprises…).

Le/la stagiaire sera encadré(e) par une équipe pluridisciplinaire de chercheurs du CIRAD: Grégoire Blanchard (écologie fonctionnelle des plantes tropicales, Guadeloupe) et Sandrine Auzoux (systèmes d'information, gestion des données, la Réunion), avec un appui d'autres chercheurs spécialisés en agronomie, en écologie fonctionnelle et en fouille de texte automatisée (Aude Ripoche, agronome modélisatrice, étude des interactions Cultures-adventices; Karim Barkaoui, écologie fonctionnelle/ éco-informatique; Mathieu

Roche, Data Mining, Intelligence Artificielle). Il/elle participera aux événements DigitAg, en particulier la DigitAgora (3 jours) qui aura lieu en Avril 2024 à Montpellier.

Profil souhaité:

Formation de niveau Master 2 en cours dans le domaine de l'intelligence artificielle ou de la science des données.

Compétences et savoir-être attendus :

- Aisance relationnelle, facilité d'écoute et d'animation
- Capacités d'initiative
- Aisance rédactionnelle
- Recherche bibliographique
- Anglais scientifique
- Intelligence Artificielle et science des données
- Un intérêt pour l'agronomie

Conditions

Période de stage : 5 à 6 mois, dates flexibles à partir du 1er trimestre 2025

Gratification de stage : 670€ nets / mois + logement sur le site de travail + tickets restaurants + billets d'avion aller-retour

Stage basé en Guadeloupe (station de Neufchateau, Capesterre-Belle-Eau) avec une mission sur Montpellier pour Digitagora.

Renseignements pour le stage :

pour candidater, merci d'adresser un CV et une lettre de motivation à Grégoire BLANCHARD, gregoire.blanchard@cirad.fr

Date limité de candidature : 5 Décembre 2024