Récolte et Analyse des connaissances partagées sur Youtube sur la thématique des ravageurs en production légumière et arboricole

Stage master 2 – 2025 (6 mois)

Contexte général et projet de recherche

Ce stage s'inscrit dans les activités interdisciplinaires de l'UMR TETIS du projet STAY - Savoirs Techniques pour l'Autosuffisance, sur YouTube (financement CNRS) - en partenariat avec le LISIS (Laboratoire Interdisciplinaire Sciences Innovations Sociétés). Des pratiques agricoles sont aujourd'hui partagées et commentées sur YouTube, plateforme d'hébergement de vidéos dont la popularité n'est plus à démontrer. En effet, en février 2023, les données Médiamétrie indiquaient 48 millions d'utilisateurs uniques en France. Disponible à tout le monde, la plateforme permet à quiconque – professionnels de l'agriculture ou pas - de devenir créateur de contenu, les caractéristiques et la qualité des informations ainsi partagées faisant l'objet d'une littérature déjà abondante. Cette littérature montre entre autres que YouTube constitue pour ses utilisateurs une source d'informations qui contribue aux appréciations qu'ils se font d'une situation, et qui peut influencer leur jugement et leur action parfois de manière significative.

Qu'il s'agisse d'utilisateurs ou de producteurs de contenu, ils peuvent être à la fois des professionnels (exploitants agricoles, Chambres d'Agriculture...) et des amateurs (des jardiniers engagés dans l'autoproduction alimentaire à l'échelle d'un potager ou petit verger, militants…). Nous nous intéressons tout particulièrement au sujet des ravageurs en production légumière et arboricole.

L'objectif du stage est double :

(1) dresser un inventaire le plus exhaustif possible des chaînes YouTube pouvant être consultées afin d'obtenir des informations concernant les techniques de production légumière et arboricole – avec une attention particulière aux chaînes faisant référence aux techniques de lutte contre les ravageurs - en distinguant les chaînes produites par des professionnels de l'agriculture et les chaînes alimentés par des amateurs. Il s'agira dans un premier temps d'identifier les mots-clés pertinents et d'une liste de thèmes susceptibles de faire l'objet de recherches sur YouTube

(2) réaliser de façon automatique une catégorisation des contenus, en s'appuyant sur les statistiques et métadonnées, en termes:

- d'année d'apparition
- de nombre d'abonnés, de nombre de commentaires, de nombre de vues et de nombre de likes, avec une analyse de l'évolution temporelle de ces indicateurs d'identification des repères temporels marquants pour l'apparition et l'évolution en termes de succès de ces chaînes (épidémie de Covid, des évènements climatiques significatifs, etc.)
- de production de contenu, en termes quantitatifs
- de catégories des producteurs de contenu (classification à construire) de types de contenu proposés et de thèmes abordées – relatifs aux techniques agricoles et plus particulièrement aux techniques de lutte contre les ravageurs
- de type de stratégie économique employée par les créateurs de contenu – en termes de nombre de publicités et d'autres sources de revenu (contrats, cagnotte Tipeee..).

Le/la stagiaire pourra s'appuyer sur une production académique récente (Bruhl 2023) concernant un sujet similaire, à savoir la thèse de Guillaume Bruhl intitulée « État des lieux de la vulgarisation scientifique vétérinaire francophone sur Youtube ». Les implémentations s'intégreront dans la plateforme en cours de développement du projet.

De façon plus précise, les activités à réaliser dans le cadre de ce stage consisteront à :

- Adapter une méthode de webscrapping afin de faire la récolte des vidéos, en adéquation avec les thématiques définies (mots clefs thématiques sélectionnés avec les experts)
- Mettre en oeuvre une étape de classification automatique à l'aide des modèles de langues et grands modèles de langues (LM et LLM) selon une catégorisation définie en collaboration avec les experts du domaine
- Proposer un tableau de bord permettant de naviguer entre les différentes données et métadonnées, mettre en œuvre des mesures de discréminences spécifiques.
- Structurer le jeu de données constitué afin de le déposer sur le dataverse d'INRAe.

Référence

Guillaume Bruhl. État des lieux de la vulgarisation scientifique vétérinaire francophone sur Youtube. Sciences du Vivant [q-bio]. 2023. ⟨https://dumas.ccsd.cnrs.fr/dumas-04344450⟩

Lotfi, Chaimaa & Srinivasan, Swetha & Ertz, Myriam & Latrous, Imen. (2021). Web Scraping Techniques and Applications: A Literature Review. 10.52458/978-93-91842-08-6-38

Qi, Danrui and Jiannan Wang. CleanAgent: Automating Data Standardization with LLM-based Agents ArXiv abs/2403.08291 (2024)

Ramy Baly, Georgi Karadzhov, Jisun An, Haewoon Kwak, Yoan Dinkov, Ahmed Ali, James Glass, and Preslav Nakov. 2020. What Was Written vs. Who Read It: News Media Profiling Using Text Analysis and Social Media Context. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 3364–3374, Online. Association for Computational Linguistics

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. « The FAIR Guiding Principles for Scientific Data Management and Stewardship ». Scientific Data 3 (1): 160018. https://doi.org/10.1038/sdata.2016.18

Organisation

Le stage rémunéré se déroulera sur une période de 6 mois, à compter de février 2024. L'étudiant·e sera accueilli·e au sein de l'UMR TETIS, à la Maison de la Télédétection (Montpellier) et sera encadré·e par Laura Maxim, chercheuse au LISIS (Laboratoire Interdisciplinaire Sciences Innovations Sociétés) et Maguelonne Teisseire, directrice de recherche à l'UMR TETIS (INRAE). Des réunions hebdomadaires sont prévues conjointement aux échanges informels en continu avec les encadrants du stage afin de discuter de l'avancée du travail et des éventuelles difficultés rencontrées.

Profil recherché

Le/la stagiaire aura un profil en informatique avec des connaissances en traitement automatique de la langue et/ou apprentissage automatique, avec un intérêt pour le travail interdisciplinaire. Une expérience dans le langage de programmation Python est un plus.

Candidature

Les candidatures (CV, lettre de motivation et relevé de notes M1 – ou 4ème année) sont à envoyer à maguelonne.teisseire@inrae.fr et laura.maxim@cnrs.fr avec pour Objet « Candidature Stage1 – Projet STAY »

Date limite des candidatures : 20/11/2024.