Objet : Première rencontre avec nos encadrants

CAROT Axel

24 septembre 2023

1 Présentation

En premier lieu, les commanditaires se sont présentés. Nous avons fait l'entretien avec Roberto INTERDONATO et Sarah VALENTIN, deux chercheurs du CIRAD, qui est un organisme français de recherche agronomique et de coopération internationale pour le développement durable des régions tropicales et méditerranéennes.

2 Les données

Les données textuelles sont collectées depuis environ deux ans grâce au web scraping. Elles proviennent de journaux locaux (plus de 20 000 articles), de transcriptions de vidéos YouTube, ou encore de podcasts abordant les actualités des différents pays de l'Afrique de l'Ouest (avec un focus sur le Burkina Faso, le Bénin et le Sénégal). Les données comportent aussi des sources d'informations plus anciennes. Elles nous ont été transmises dans un dossier partagé via WeTransfer.

3 La documentation

Les commanditaires nous ont fourni une documentation composée de :
- Différents articles sur la sécurité alimentaire
- Un notebook d'analyse descriptive et des lexiques.
- Une liste de bibliothèques Python à utiliser (NLTK, Spacy, Transformers).

4 Réponses à nos questions

Quels types de visualisations sont privilégiés pour représenter les résultats de l'analyse exploratoire ?

Les commanditaires nous ont donné quelques idées et quelques conseils : Wordcloud, radar plot, cartographie, s'inspirer de la bibliographie, ne pas s'attarder sur la prédiction car les données sont bruyantes, faire attention au contexte d'utilisation des mots.

Pourquoi le problème est important pour le laboratoire ? Quels sont les enjeux ?

Le CIRAD a une approche pluridisciplinaire et n'est pas spécialisé en statistique. Notre travail permettrait de compléter différentes visions.

Quels sont les défis spécifiques liés à l'utilisation de données textuelles en français dans ce projet, et comment envisage-t-on de surmonter ces défis ?

Le principal défi est le fait que les données soient non structurées et sans ponctuations (à cause de la transcription des vidéos et podcasts), ce qui augmente le bruit des données. Ensuite, il va nous falloir identifier les articles corrélés avec la sécurité alimentaire, puis savoir parmi eux, ceux qui sont pertinents.

Est-ce que l'utilisation de D3js est obligatoire ou d'autres logiciels de data visualisation peuvent être utilisés ?

Les commanditaires n'ont pas de préférences.

Quelles seraient les variables clés et les métriques permettant de juger de la qualité et de la précision des informations extraites et enrichies ?

— **Variables :** catégories et entités nommées (date, lieu, événement).
— **Qualité :** rappel et précision (des termes que nous devront approfondir grâce à la bibliographie).

Comment reformulerions-nous le sujet de manière non technique ? (Validée par les commanditaires)

Nous analysons les données textuelles qui ne sont pas forcément utilisées par les gouvernements sur le thème de la sécurité alimentaire en Afrique de l'Ouest dans le but de s'adapter à une forte croissance démographique, une agriculture pluviale très dépendante des conditions pluviométriques, des risques sécuritaires et sanitaires.

Quelles seront les grandes étapes de notre projet ?

— Étiquetage : annotation.
— Analyse descriptives.
— Extraction d'entités nommées.
— Analyse structurelle.

Est-ce qu'il y a des erreurs à ne pas commettre qui pourraient nous faire perdre du temps ?

Les erreurs à ne pas faire selon les commanditaires sont :
— Une mauvaise organisation.
— Négliger la partie préliminaire et le nettoyage.
— Ignorer le bruit.

À quelle fréquence doit-on vous faire part de nos avancées et comment ?

Nous avons convenu d'effectuer un rendez-vous environ toutes les deux semaines en fonction de nos avancées. Les commanditaires ont tout de même insisté sur l'importance de les solliciter par mail si nous rencontrons des difficultés.

Quelles sont les principales tendances ou informations clés que l'on espère découvrir en effectuant une analyse exploratoire des données textuelles ? Y a-t-il des attentes concernant son impact ?

Les attentes au niveau de l'information : trouver des événements ponctuels (sécheresse, inondations) avec un fort impact sur la durée et la montée des prix. Un exemple d'impact : modèle prédictif ciblé (ex : prix du blé).

5 À faire avant la prochaine réunion

En premier lieu, nous avons prévu de mettre en place les outils collaboratifs : Discord, mailing list (avec les commanditaires).

Par la suite, nous allons procéder à la planification du projet, les rôles et les tâches des membres du groupe. Il va notamment falloir se répartir la documentation et en faire des comptes rendus pour les autres membres.

Certains membres du groupe n'ont pas eu l'occasion d'utiliser GitHub. Ils devront donc apprendre à effectuer les actions de base (pull, commit, push, etc.) à l'aide d'un TP sur GitHub que nous avons effectué en L3 MIASHS en cours de Gestion de projet.