Cahier des charges PFE

1. Intitulé de projet :

Analysis of twitter data on sustainable development goals

Développement d'un outil d'analyse de données pour extraire de la connaissance à partir des tweets collectés. IL y aura 300000 tweets à analyser.

1. Contexte et objectifs :

L’objectif du projet est de donner une idée sur les interactions en ligne et la révélation des perceptions des intervenants au sujet du SDG (objectifs de développement durable) liés aux villes. Nous vous fournirons des villes avec de nouveaux outils d’information, soutenant leur stratégie afin d’atteindre le niveau SDG 11 et de ses objectifs. Ces outils d’information permettront de combler le fossé entre la vision des intervenants divers, de l’Organisation des Nations Unies, organismes gouvernementaux et non gouvernementaux, à la société civile et le secteur privé. Plus précisément, la proposition va analyser les données postées sur la plate-forme Twitter pour fournir des organismes et des citoyens avec rétroaction opérationnelle sur comment SDG doit être perçus et mis en œuvre.

1. Description de travail à faire :

Le sujet comporte deux volets à savoir :

1. Analyse des hashtags

Différents Hashtags ont été utilisés dans les tweets comme #Goal11, #SDG11 ou #SDGs. par conséquent, la première étape de l’analyse est de construire un graphique détaillé avec les hashtags différents utilisés dans les tweets et leurs liens. Un (cooccurrence) lien entre deux hashtags exprime le fait que les deux hashtags ont été utilisé dans la même Tweet. L’équipe de recherche va calculer le nombre de fois chaque hashtag a été utilisé dans les tweets et les nombre de fois où deux hashtags ont été utilisés dans les tweets mêmes. Ce faisant, les hashtags identifiés peuvent être utilisés comme un vocabulaire étendu pour chercher d’autres tweets qui peuvent être associés à SDG 11. Par conséquent, le corpus de données sera enrichie, offrant ainsi une vue plus complète des données.

1. Identification des sujets

La technologie qui sera utilisée dans la fouille de texte sera élaborée dans la partie état de l'art à savoir le traitement automatique du langage naturel (PNL) et l'apprentissage automatique (Machine Learning ML), basé sur la modélisation du langage continu linéaire et non linéaire, par exemple l’analyse sémantique latente (Dumais, 2004) et l'apprentissage profond (Deep Learning Bengio al 2013). L'apprentissage automatique (ML) va être utilisé pour analyser des tweets par thèmes et pour analyser les sites Web analysés par thèmes. Par conséquent, l’équipe de recherche verra les sujets qui sont les plus abordés et ceux qui ne le sont pas. Les approches les plus connus d'apprentissage automatique(ML) assureront des résultats valables, y compris l’apprentissage supervisé avec formation et jeux de test, ainsi qu’inter-validation, et les jugements de l’intérêt pour l’utilisateur.

1. Les outils à utiliser :

Le logiciel R sera utilisé avec d'autres outils à identifier (dans la partie état de l’art).

Clarifies le point concernant le fait que tu vas développer un nouveau outil de A jusqu'à Z et si oui quel sera sa valeur ajoutée? Ou bien tu vas adapter un outil existant?

Tu vas te baser sur des algorithmes existants de traitement de langage naturel que nous allons améliorer par rapport aux tweets collectes. Comment allons-nous les améliorer? C'est l'objectif de la recherche. On va regarder des dizaines de milliers de tweets et on verra par après.

Et dans quel objectif?

- L'objectif est d'analyser des tweets relies aux objectifs du développement durables

- Tu vas améliorer des algorithmes existants pour traiter des tweets relies aux objectifs de développement durables. Améliorer des algorithmes veut dire également le développement d'un outil (que je ne peux décrire car il fait partie de la recherche) pour traiter les tweets.