INF34515 Projet en informatique I

Hiver 2024

Brève description de projet en collecte de données énergétique

Mise en garde

Ceci est un bref aperçu de ce que vous aurez à réaliser dans le cadre du projet, vous recevrez une version plus détaillée du « cahier de charge » en janvier.

Objectif

Concevoir un scrapper pour une province, ou un sous-ensemble des fournisseurs de services publics d'une province dans le but de collecter des données brutes de manière automatique et efficace.
Étant donné l'hétérogénéité des sources de données, l'architecture envisagée du scrapper sera modulaire - potentiellement un module par source de données.

- Le scrapper sera conçu pour naviguer à travers les différentes sources de données disponibles, telles que les sites web, les bases de données ou via des API afin de rassembler les données brutes pertinentes.
- Il est attendu que les métadonnées soient aussi colligées, telles que l'auteur, la date déclaration réglementaire pour les fournisseurs de services publics, etc.

Une fois ces données brutes collectées, l'objectif suivant est de les prétraiter en utilisant des techniques de nettoyage de données, de transformation et de normalisation. Ce processus de prétraitement sera soigneusement documenté pour assurer la reproductibilité, la qualité des données.

Une initiative similaire américaine intéressante et assez aboutie est le projet PUDL:
https://github.com/catalyst-cooperative/pudl

Du côté de la présentation des données et du type de données à collecter, les statistiques sur l'électricité du European Network of Transmission System Operators for Electricity (ENTSO-E) offrent un excellent exemple. Les données sont aisément accessibles à un public de "non-développeurs". https://www.entsoe.eu/data/power-stats/

Les données recueillies pourront ensuite être visualisées à l'aide d'une plateforme web quelconque.

L'équipe sera en mesure de consulter des experts dans le domaine.

Méthodologie

1. Identifier des cas d'utilisation pertinents pour la collecte de données en collaboration avec des experts du domaine. L'identification de ces cas spécifiques aidera à définir le type de données prioritaires à collecter dans les prochaines étapes du projet.
    1. Vous pouvez choisir une province ou un territoire canadien de votre choix, autre le Québec et le Nouveau-Brunswick.
2. Identifier les sources de documents pertinents.
    1. À valider
3. Collecter les données brutes: une fois que les documents pertinents ont été identifiés, l'équipe concevra un extracteur de données spécifique à chaque type/source de document et les stockera de façon structurée. Il est attendu que les métadonnées des documents soient aussi sauvegardées.
    a. Date de création, date de modification, source, auteurs, génération d'identifiant unique, etc.
4. Analyser les documents
5. Conception de structure des tables/dataframes
6. Extraction et transformation des données: standardisation/normalisation des données.

Une entente sur les droits d'auteurs devra aussi probablement être signée pour être éligible au financement éventuel de ce projet. Essentiellement, les outils seront forts probablement rendus disponibles sous une licence libre.