AUTOMATISATION DE LA COLLECTE DE DONNÉES WEB POUR L'ENRICHISSEMENT DE LA BASE SIRENE

Furkan Narin et Rayan Ben Yacoub

Date : 10 / 03 / 2024

TABLE DES MATIÈRES

1. INTRODUCTION

Ce rapport présente le processus et les résultats de notre projet visant à enrichir les données de la base SIRENE avec des informations géographiques et supplémentaires via l'API Adresse et le web scraping sur Google Maps. Notre objectif était de compléter les données existantes avec les coordonnées géographiques des établissements et d'autres informations pertinentes pour faciliter l'analyse géospatiale et améliorer la qualité des données disponibles.

2. MÉTHODOLOGIE

2.1 EXTRACTION DES DONNÉES

Nous avons commencé par extraire les informations pertinentes de la base SIRENE, en particulier du fichier StockEtablissement, en utilisant la bibliothèque Pandas en Python. Étant donné la taille conséquente du fichier

(6,84 Go), nous avons opté pour une approche de lecture par blocs pour minimiser la consommation de mémoire et optimiser le traitement(d'abord par 100 établissements pour le fichier ‘StockEtablissement_utf8_1000.csv’ ensuite par 500 000 pour le fichier lourd).

2.2 CRÉATION ET UTILISATION DES API

Pour enrichir les données extraites, nous avons utilisé l'API Adresse pour obtenir les coordonnées géographiques des établissements à partir de leurs adresses postales. Nous avons également tenté d'utiliser le web scraping sur Google Maps pour récupérer des informations supplémentaires telles que le site web, le numéro de téléphone, et les avis des utilisateurs.

2.3 OBTENTION DES COORDONNÉES GÉOGRAPHIQUES

Grâce à l'API Adresse, nous avons réussi à associer à chaque établissement ses coordonnées géographiques (longitude et latitude). Cette étape a été cruciale pour permettre une analyse géospatiale ultérieure des données.

3. DIFFICULTÉS RENCONTRÉES

Nous avons rencontré plusieurs défis au cours de ce projet. Le principal a été la gestion du fichier de départ très lourd, ce qui a considérablement ralenti le processus d'extraction et de traitement des données. De plus, bien que nous ayons réussi à utiliser l'API Adresse pour obtenir les coordonnées géographiques, nous n'avons pas pu automatiser entièrement le processus de web scraping sur Google Maps en raison de restrictions et de la complexité du chargement dynamique des données.

4. CONCLUSION

Malgré les difficultés rencontrées, ce projet nous a permis de développer nos compétences en manipulation de données, en utilisation d'APIs et en web scraping. L'enrichissement de la base SIRENE avec des données géographiques et supplémentaires ouvre de nouvelles perspectives pour l'analyse et l'exploitation des informations des établissements. Pour les étapes futures, nous envisageons d'explorer des solutions pour optimiser le traitement des fichiers volumineux et pour surmonter les obstacles liés au web scraping automatisé.