Justine - Gabriel - Helena

Rapport collectif - SAE : Collecte de données web

Répartition des missions :

Nous nous sommes mis d'accord pour répartir le travail de la manière suivante :

- Gabriel s'occupe de la partie web scraping, ce qui lui laisse la liberté de choisir son sujet.

- Justine et Héléna travaillent sur la partie API.

PARTIE API :

Au départ, chacune développait son code séparément. Nous avons ensuite mis en commun nos travaux afin de comparer les avantages et les inconvénients de chaque version.

Le code de Justine était plus simple et s'appuyait davantage sur le contenu du cours, tandis que celui d'Héléna était plus complet, comportait des améliorations pertinentes, mais était également plus complexe et faisait davantage appel à l'IA.

Après réflexion sur les attentes du projet, nous avons choisi de partir du programme de Justine, que nous avons ensuite amélioré afin d'obtenir un résultat plus ergonomique et visuellement plus sympathique pour l'utilisateur.

Pour ce faire, nous avons utilisé l'IA, qui a modifié certaines parties du code, parfois de manière non souhaitée. Toutefois, le rendu final étant plus intéressant, nous avons décidé de conserver certaines de ces modifications.

Nous avons aussi passé du temps à chercher un indicateur pertinent pour la dernière partie du travail, que nous avons réalisée ensemble. Nous en avons trouvé deux indicateurs intéressants que nous avons répartit sur les deux codes.

Ce qui fonctionne :

- Graphique interactif
- Affichage automatique de la carte ou de la page web
- Différentes couleurs en fonction des conditions
- Indicateur (belle version) les stations nécessitant un rééquilibrage

Ce qui ne fonctionne pas :

Afficher les stations les plus importantes par commune sur la carte. (Indicateur supplémentaire)

Streamlit, c'est pourquoi nous avons 2 codes : une "belle" version (qui ne comprend pas de carte folium) et une version beaucoup plus basique.

PARTIE WEB SCRAPING :

Le choix du sujet a été long et surtout très problématique au niveau technique : j'ai dans un premier temps décidé de partir sur un projet que je ne pensais pas si ambitieux, le sujet était "Toutes les routes mènent-elles bien à Rome ?" Il s'agissait de faire du web scraping sur Google Maps pour récupérer des données sur les grandes routes et les croisements de routes dans 15 grandes villes de pays Européens, puis élaborer tout un script qui évaluait si oui ou non, toutes les routes menaient bien à Rome. Cependant, un premier problème est survenu très rapidement : bien que Google maps soit le site le plus scrapé dans le monde en 2025, le faire était trop complexe (et illégal), donc j'ai décidé de faire la même chose mais cette fois-ci sur OpenStreetMap.
C'est alors qu'un second problème est survenu lui aussi assez rapidement : télécharger des données sur l'ensemble des routes à 50km autour des centres des grosses villes de 15 pays nécessitait non seulement une puissance de calcul et de ram impressionnante, mais cela représentait surtout bien plus de 500 fichiers JSON de plusieurs milliers de lignes chacuns, donc techniquement ce n'était pas tenable.
Suite à cela, j'ai décidé de changer de sujet et de choisir quelque chose de moins ambitieux et techniquement faisable. J'ai alors essayé au moins 5 ou 6 autres sujets mais à chaque fois quelque chose bloquait : trop de données, scraping illégal/infaisable, mauvais format de données ect…

Je me suis alors enfin résolu à faire une variante du sujet du cours, c'est-à-dire scraper Wikipédia afin d'obtenir des données sur les sites de l'UNESCO en France, ce que j'ai finalement réalisé avec succès.

J'ai aussi passé beaucoup de temps à la confection du site web dans lequel on retrouve la carte du scraping, ainsi que le code du WS et ceux de l'API. L'usage de nouvelles technologies (react.js; serveurs "Vite"; .jsx ect..) pour faire l'arrière-plan a été très long mais est à mon point de vue une réussite.

Ce qui fonctionne :

- Les graphiques pertinents
- la carte en elle-même

Ce qui ne fonctionne pas :

- Sur mon pc, la carte ne se lance pas automatiquement dans le navigateur quand je lance le script pourtant tout est sensé fonctionner.