Lucas SUIZDAK

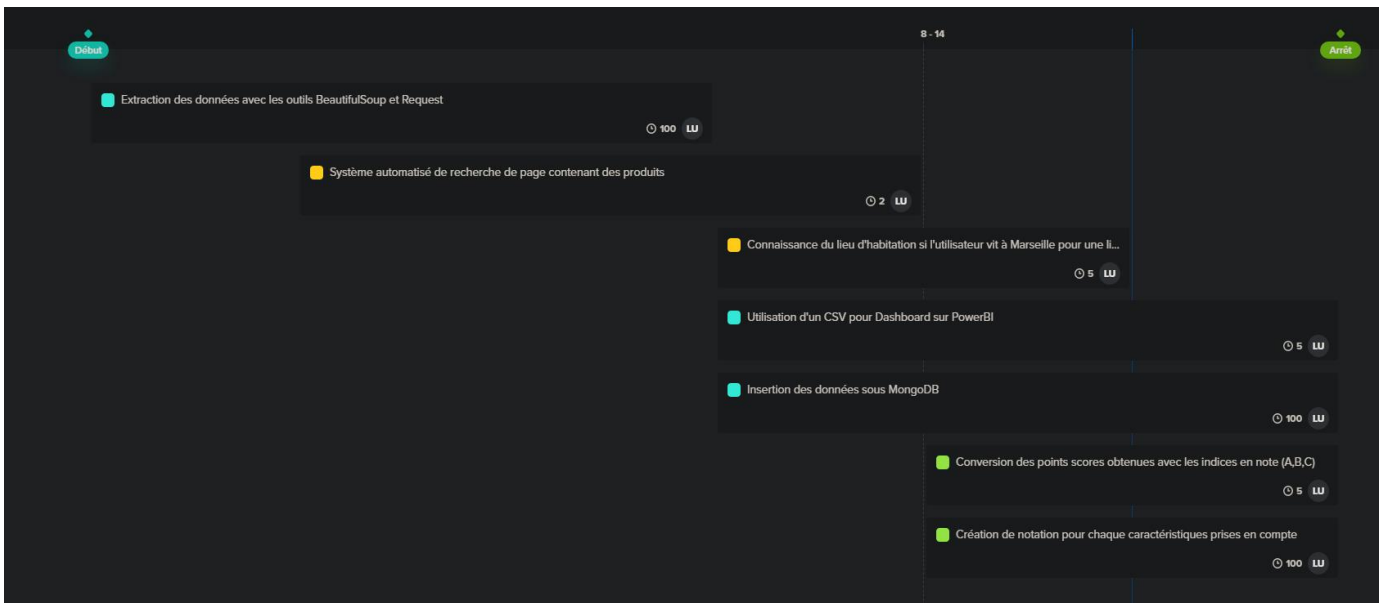
Web Scraping

Extraction de données des produits d'Hubsch

09/01/2023

Collection des données

Contexte :



L'équipe Eco Impact cherche à automatiser leur système de notation et travaille dans le développement d'un logiciel d'extraction de données les quelles permettent d'évaluer un produit.

Pour cela, l'équipe met sur la tâche 1 personne dont le travail consiste à rechercher un moyen de collecter les données de produits commercialisés sur un site afin d'y donner une note à chacun. Les exigences sont la création d'un programme automatisé permettant l'extraction d'informations, l'envoi de celles-ci sur une base de données MongoDB, la création d'un DashBoard et en finalité l'estimation de chaque produit avec l'utilisation des données collectées.

Pour cela, l'utilisation de Python, de la librairie Request et de BeautifulSoup est nécessaire pour achever cette tâche. La personne en charge du projet a partagé sa roadmap afin de pouvoir mieux se situer dans son travail et de s'organiser. Le voici à droite du texte.

La tâche demandée fait appel au webscraping sous Python. Cela consiste à chercher des clés dans le code d'une page afin de trouver des mots/phrases utiles pour l'estimation du produit. Les données à acquérir sont :

Le produit, le poids, la matière, le transport de distribution, le transport d'approvisionnement, le lieu de fabrication ainsi que les dispositifs supplémentaires concernant la fin et la durée de vie du produit concerné.

mongoDB

Requests

BeautifulSoup

Méthode utilisée

Voici le fonctionnement du programme. Celui-ci est découpé en 3 parties.

La première étape consiste à créer les listes que le programme va utiliser, comme le poids ou l'origine du produit. Suite à cela, il va parcourir plusieurs pages grâce à l'URL pour définir celles qui contiennent les données à extraire. Afin d'accéder au caractéristiques de chaque produit, il enregistre dans la liste « links_list » toutes les redirections de page de produit et avec une boucle for, il passe ensuite à l'étape 2, l'extraction de données.

En recherchant les mots clés comme « Poids », « Matériaux » ou autre, il extrait uniquement l'information recherchée dans la ligne. Certains paramètres sont appliqués selon la caractéristique à collecter.

Suite à cela, toutes les données sont présentes dans des listes.

Le programme a aussi des fonctionnalités pour permettre une recherche plus précise des données.

Afin de représenter ces données et de les sauvegarder, les listes sont envoyés sur une base de donnée MongoDB, système de gestion de base de données orienté documents.

Ces données pourront à la suite être utilisé pour une analyse afin d'être plus précis dans la notation. PowerBI permet ce travail en permettant la visualisation de données avec des systèmes de comparaison et de corrélation. Les données sont alors plus compréhensibles, ce qui facilite le travail de l'homme. Voici une analyse de notre extraction de données ; plusieurs caractéristiques y sont présentés comme le nombre de produits par matière ou le nombre de dispositifs supplémentaires augmentant la durée de vie.

La troisième partie de ce programme consiste ensuite à évaluer chaque produit selon ses caractéristiques. Pour cela, le produit va comparer ses données avec une liste d'indice de notation [voir le code].