

Bonjour Sam,

Je viens de terminer la version bêta du programme qui permet de surveiller les prix du revendeur Books to Scrape. Voici le lien vers mon repository Github : https://github.com/mcourte/OC_Projet2.git.

Vous trouverez dans le README.md toute la démarche pour exécuter le programme.

Vous trouverez en pièce jointe le dossier .zip des données générées, il possède un dossier par catégorie de livre, dans lequel il y a un fichier .csv avec l'ensemble des datas extraites, et un dossier Images comprenant les couvertures de tous les livres.

Concernant le processus du pipeline ETL, voici comment notre programme y répond :

-Extraction (Extract)

Utilisation de la bibliothèque Python "requests" pour extraire des données à partir de sites web. Dans notre cas, nous utilisons cette fonctionnalité pour récupérer les données de livres à partir de "<http://books.toscrape.com/>".

-Transformation (Transform)

Une fois les données extraites, plusieurs opérations de transformation sont faites pour les préparer à l'analyse. Par exemple, les noms de catégories sont nettoyés (suppression des chiffres et caractères inutiles), De plus, les données extraites sont organisées en les associant à une catégorie et en créant des fichiers CSV pour chaque catégorie. Ces fichiers contiennent des informations sur les livres, telles que défini dans le cahier des charges

-Chargement (Load)

Une fois les données transformées, nous les chargeons dans des fichiers CSV correspondant à chaque catégorie. Le programme permet la création de dossiers par catégories, dans lequel vous trouverez un fichier « book_data.csv » ainsi qu'un dossier Images contenant l'ensemble des couvertures de livre avec comme nom le titre du livre.

N'hésitez pas à revenir vers moi si vous avez des questions,

Cordialement,

Magali Courté

Développeur d'application Python