Rapport sur le Scraping de Données des Annonces de Voitures sur Wandaloo.

Projet PFA - EMSI

Réalisé par:
Anas Filali
Youssef Khouili
Zakaria Riady

Contents

1 Introduction

Ce rapport complet fournit une analyse détaillée du processus de scraping des données à partir des annonces de voitures d'occasion sur le site Web de Wandaloo. Le script Python utilise une combinaison des bibliothèques `Requests`, `BeautifulSoup` et `CSV` pour extraire, organiser et stocker des informations sur chaque annonce de voiture. Le rapport couvre l'architecture, le flux de travail, les objectifs et les considérations pré/post-exécution, ainsi que des détails sur les informations sources et la structure des données.

2 Objectif

L'objectif principal du script de scraping de données est de rassembler et d'organiser systématiquement les données du site Web de Wandaloo, en se concentrant spécifiquement sur les annonces de voitures d'occasion. Le script vise à collecter des informations détaillées pour chaque voiture, y compris des détails essentiels tels que le titre, le prix, l'URL de l'image, l'URL des détails, le type de carburant, l'année de fabrication, la puissance, le kilométrage, la date de l'annonce et la ville de l'annonce.

3 Flux de travail et Architecture

3.1 Flux de travail

Le script suit une architecture modulaire utilisant trois bibliothèques clés :

- **Bibliothèque Requests :** Pour effectuer des requêtes HTTP afin de récupérer le contenu HTML.
- **Bibliothèque BeautifulSoup :** Pour l'analyse HTML et la navigation.

- **Bibliothèque CSV :** Pour lire et écrire des fichiers CSV pour le stockage des données.

3.2 Étapes du flux de travail

1. **Importer les bibliothèques :** Importer les bibliothèques Python nécessaires (`requests`, `BeautifulSoup`, `csv`).
2. **Définir la fonction `scrape_page` :** Créer une fonction (`scrape_page`) pour scraper les pages individuelles et extraire les annonces de voitures.
3. **Spécifier l'URL de base et le nombre de pages :** Définir l'URL de base pour les annonces de voitures d'occasion sur le site Wandaloo. Spécifier le nombre de pages à scraper.
4. **Boucle principale de scraping :** Itérer à travers le nombre spécifié de pages. Appeler la fonction `scrape_page` pour chaque page. Accumuler les résultats dans la liste `all_car_listings`.
5. **Spécifier le nom du fichier CSV :** Définir le nom du fichier CSV pour stocker les données extraites.
6. **Écrire les données dans le fichier CSV :** Ouvrir le fichier CSV en mode écriture. Utiliser la bibliothèque CSV pour écrire les données collectées dans le fichier CSV.

3.3 Objectifs atteints

- Extraction réussie d'informations détaillées sur les annonces de voitures d'occasion.
- Organisation des données extraites dans un format structuré adapté à l'analyse.

4 Informations pré-exécution

4.1 Dépendances

S'assurer que les bibliothèques Python requises (`requests`, `beautifulsoup4`) sont installées.

```
!pip install requests
!pip install beautifulsoup4
```

4.2 Considérations éthiques

- Respecter les pratiques éthiques de scraping, en respectant les conditions d'utilisation et le fichier robots.txt du site Web.
- Éviter de surcharger le site Web avec un trop grand nombre de requêtes pour prévenir d'éventuels problèmes.

4.3 Exigences d'entrée

- Le script nécessite l'URL de base des annonces de voitures d'occasion sur le site Web de Wandaloo.
- Spécifier le nombre de pages à scraper.

5 Informations post-exécution

5.1 Sortie

- Les données extraites sont stockées dans un fichier CSV nommé "car_listings-data.csv."
- Le fichier inclut des informations détaillées sur chaque annonce de voiture.

5.2 Maintenance

- Vérifier périodiquement le script pour les mises à jour ou les changements dans la structure du site Web.
- S'assurer de la conformité aux éventuels changements dans les conditions d'utilisation ou les politiques de scraping.

5.3 Analyse des données

- Le fichier CSV généré peut être utilisé pour diverses tâches d'analyse des données, telles que les tendances du marché, l'analyse des prix ou les préférences des clients.

5.4 Journalisation

- Mettre en œuvre des mécanismes de journalisation pour capturer d'éventuelles erreurs ou problèmes pendant l'exécution du script.

continue]Surveillance
continue

- Surveiller régulièrement la structure du site Web pour d'éventuelles modifications qui pourraient affecter la fonctionnalité du script.

6 Informations sur la source et structure des données

6.1 Informations sur la source

Le script cible le site Web de Wandaloo (URL : `https://www.wandaloo.com/occasion/`) pour extraire des données sur les annonces de voitures d'occasion.

6.2 Structure des données

Les données extraites pour chaque annonce de voiture comprennent le titre, le prix, l'URL de l'image, l'URL des détails, le type de carburant, l'année de fabrication, la puissance, le kilométrage, la date de l'annonce et la ville de l'annonce.

6.3 Structure HTML

Les éléments HTML clés et les classes utilisées pour l'extraction des données incluent les éléments de liste (`<li>`), le titre (`<p class="titre">`), le prix (`<p class="prix">`), l'URL de l'image (`<img>`), l'URL des détails (`<a class="btn">`), les détails (`<ul class="detail">`) et la date et la ville (`<p class="infos">`).

6.4 Processus d'extraction des données

Le script utilise des requêtes HTTP, une analyse HTML et des techniques d'organisation des données pour extraire et structurer les informations.

6.5 Recommandations

- Surveiller périodiquement la structure du site Web pour des changements.
- Respecter les pratiques éthiques de scraping et la validation des entrées.

7 Conclusion

Le script atteint avec succès son objectif de scraper des données à partir des annonces de voitures d'occasion sur Wandaloo, fournissant un ensemble de données précieux pour une analyse ultérieure. L'architecture détaillée, le flux de travail et les considérations garantissent un processus d'extraction de données robuste et fiable. La surveillance continue, le respect des meilleures pratiques et la prise en compte des informations sources et de la structure des données contribuent à l'efficacité du script pour fournir des informations sur le marché des voitures d'occasion.