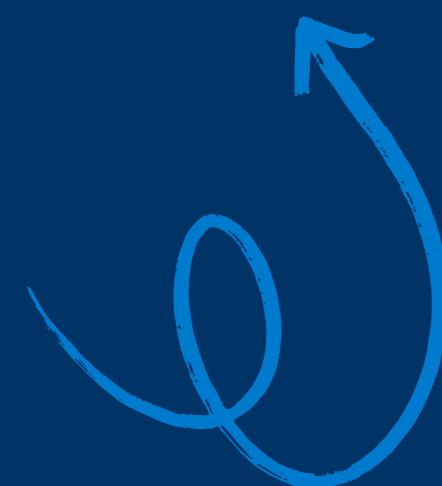


Web Scraping Leclerc Project

Yannis
Anya
Abdellatif





Introduction



Objectif

Collecter, centraliser et visualiser les promotions automatiquement.

Problématique

Les promotions changent régulièrement ce qui rend le suivi manuel difficile.

Données collectées

- Nom
- Produit
- Prix
- Remise
- Image
- Catégorie
- Lien
- Description
- Caractéristiques

Choix techniques

Selenium



Choisi pour sa capacité à interagir avec les sites web dynamiques (SPA) et gérer les interactions complexes.

SQLite



Base de données locale et légère, facilitant le stockage simple et rapide des données extraites.

Flask

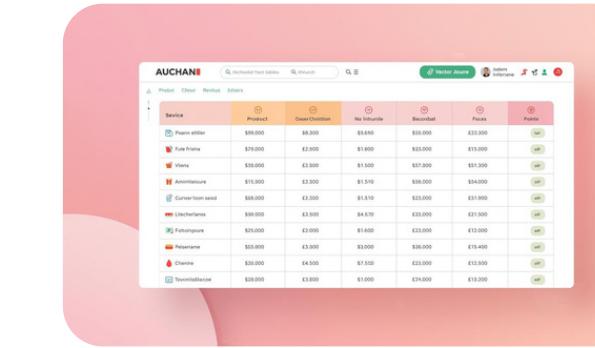


Flask

Framework pour la visualisation des données collectées via une interface web conviviale.

Sample Data Extrait

- Source E.Leclerc "Bons plans"



Nom des produits



Prix des produits



Lien des produits



Pourcentage / montant de la promo



Lien des images



Vendeur



Catégorie du produit

Profitez d'un sommeil réparateur avec le Matelas Express 140x190 cm. Conçu pour allier confort, soutien et fraîcheur, ce matelas hybride est équipé de ressorts ensachés Aeroflex™ et de 7 zones de couchage qui s'adaptent parfaitement aux courbes de votre corps. Grâce à sa structure innovante, il assure un maintien optimal de la colonne vertébrale tout en favorisant la circulation de l'air pour des nuits sans transpiration.

Les atouts du Matelas Express :

Technologie hybride : l'alliance parfaite entre mousse polyuréthane et ressorts ensachés pour un confort équilibré.
Ressorts ensachés Aeroflex™ : indépendance de couchage et excellente ventilation.
7 zones de confort : soutien ciblé pour la tête, les épaules, le dos, les hanches, les jambes, les genoux et les pieds.
Épaisseur de 18 cm : idéale pour un soutien efficace sans excès de hauteur, adaptée à tous types de sommiers.
Coutil tissé respirant : surface douce, fraîche et aérée pour un contact agréable avec la peau.
Mousse haute respirabilité : favorise une meilleure aération et réduit l'accumulation de chaleur corporelle.

Caractéristiques

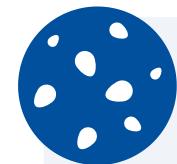
Hauteur produit emballé	110 cm
Largeur produit emballé	38 cm
Poids produit emballé	30 kg
Profondeur produit emballé	38 cm
Marque	Emma
Garantie légale de	2 ans

Voir plus

Description du produit

Caractéristiques du produit

Pipeline (scraping process)



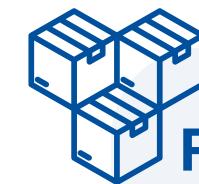
Accepter Cookies

Le scraper simule l'acceptation des cookies pour accéder à l'intégralité du contenu du site E.Leclerc.



Accéder au menu "Bons plans"

Le robot navigue vers le menu "Bons plans" pour concentrer l'extraction sur les offres spéciales et promotions.



Récupérer les cartes produits

Les informations préliminaires des produits sont extraites directement depuis les pages de liste, incluant prix et nom.



Parser les données

Des expressions XPath sont utilisées pour extraire la description, les caractéristiques tabulaires et la catégorie de chaque article.



Ouvrir les fiches détaillées

Chaque produit est visité individuellement pour récupérer des données plus riches et spécifiques à sa fiche technique.



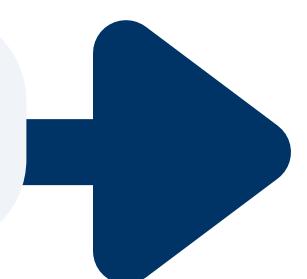
Gérer la pagination

Le système automatise le défilement entre les pages pour garantir que chaque produit d'une catégorie est identifié.



Insérer dans la base de données

Les données structurées et validées sont ensuite insérées dans une base de données SQLite pour une exploitation et une analyse ultérieures.



Données Collectées

162

Promotions scrapées

5

Pages de promo
scrapées

Exemples de Données Extraites

<u>id</u>	<u>sold_by</u>	<u>product_name</u>	<u>discount_text</u>	<u>price_eur</u>	<u>page_url</u>	<u>image_url</u>	<u>description</u>	<u>features</u>	<u>category</u>
Filtre	Filtre	Filtre	Filtre	Filtre	Filtre	Filtre	Filtre	Filtre	Filtre
10	E.Leclerc	Etagère 6 cases blanc	5 €	24.9	https://www.e.leclerc...	https://media.e.leclerc/...	Au design épuré, cette...	Profondeur du produit: ...	etagere
11	Techstarkshop	Smartphone Samsung ...	30 %	275.99	https://www.e.leclerc...	https://media.e.leclerc/...	Simplifiez-vous la vie...	Taille de l'écran (en ...	smartphones
12	E.Leclerc	Appareil à raclette ...	33 %	39.9	https://www.e.leclerc...	https://media.e.leclerc/...	Savourez de délicieux ...	Disponibilité des pièce...	NULL
13	Techstarkshop	Smartphone Samsung ...	30 %	339.79	https://www.e.leclerc...	https://media.e.leclerc/...	Le Samsung Galaxy A56 ...	Taille de l'écran (en ...	smartphones
14	tectake	TECTAKE Arbre à chat...	42 %	56.99	https://www.e.leclerc...	https://media.e.leclerc/...	Arbre à chat 141 cm ...	Libellé: TECTAKE Arbre ...	arbre_a_chat
15	E.Leclerc	Siège auto rotatif ...	40 €	139.9	https://www.e.leclerc...	https://media.e.leclerc/...	Le siège auto évolutif...	Positions: Dos à la ...	siege auto

XPaths Clés pour E.Leclerc

Ces XPaths sont essentiels pour l'extraction précise des données produits, depuis les listes jusqu'aux fiches détaillées, en assurant traçabilité et maintenabilité du code.

```
# 1) Cartes produits (listing)
XPATH_ALL_PRODUCT_CARDS = "/html/body/app-root/ng-sidebar-container/div/div/div[2]/app-template-details/div[2]/div[4]/div/div[2]/app-template-result-list/ul/li"

# 2) Lien vers la fiche produit (depuis la carte)
XPATH_PAGE_LINK_IN_CARD = "././a[@href][1]"

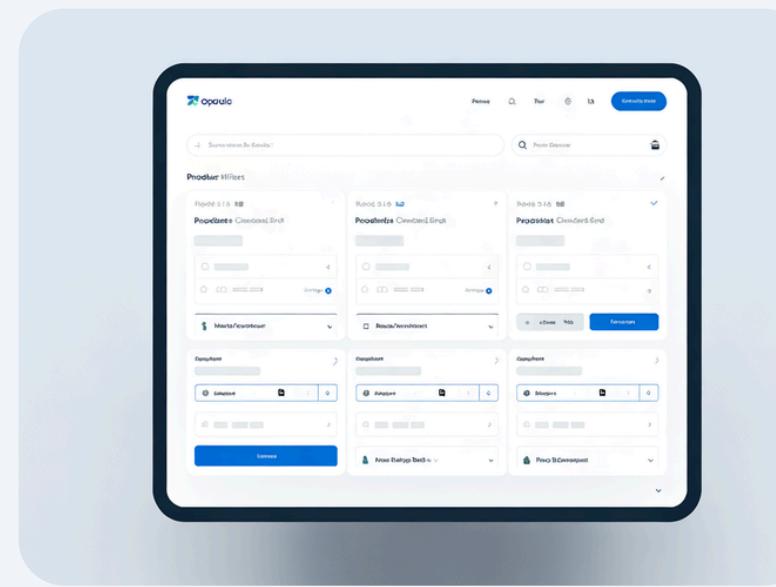
# 3) Description (sur la fiche)
XPATH_PRODUCT_DESCRIPTION = "/html/body/main/div/div/div[3]/section[1]/div"

# 4) Caractéristiques (tableau <tbody> complet)
XPATH_FEATURES_TBODY = "/html/body/main/div/div/div[3]/section[2]/div[1]/table/tbody"

# 5) Bouton "page suivante" (pagination robuste)
XPATH_NEXT_LI = "//li[contains(@class,'pagination-next')]")
```

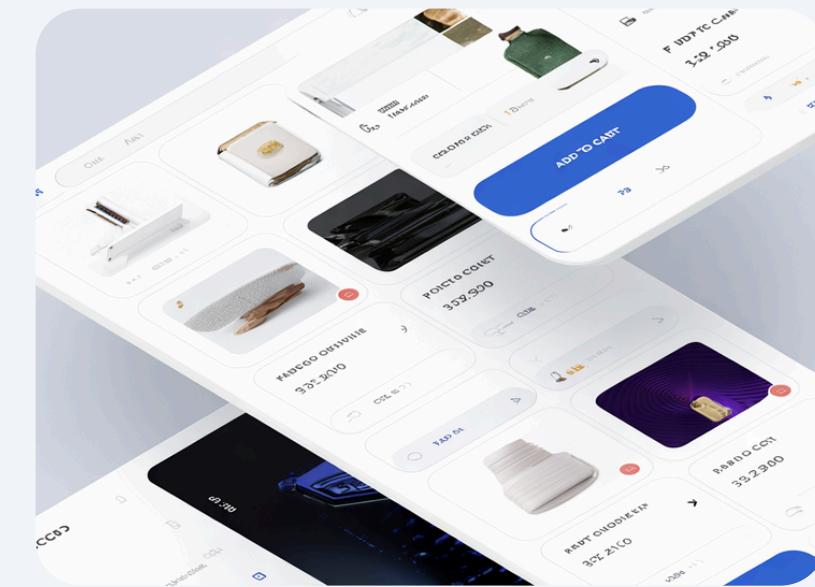
Démonstration Front - Interface Flask

Découvrez la facilité d'utilisation et la richesse fonctionnelle de notre interface web Flask, conçue pour une exploration intuitive des données produits.



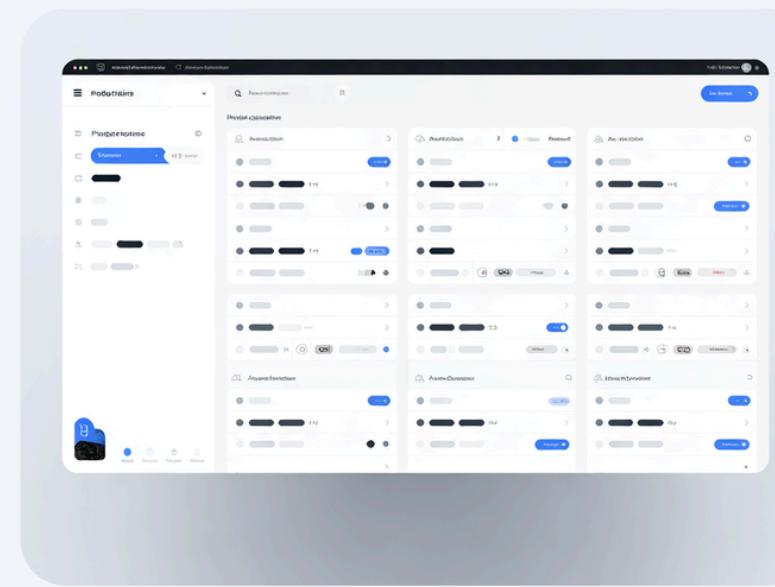
Filtres de recherche avancés

Explorez les produits avec des filtres intuitifs par prix, catégorie ou disponibilité.



Affichage en cartes élégantes

Visualisation claire des produits sous forme de cartes élégantes, affichant images, noms, prix et descriptions en un coup d'œil.



Système de catégorisation intuitif

Naviguez sans effort grâce à une catégorisation logique et détaillée, permettant de trouver rapidement les articles souhaités.



Limites & Améliorations Futures



Performance & Temps

L'ouverture séquentielle des pages (un onglet par fiche) ralentit l'extraction. L'utilisation de modes headless ou de requêtes HTTP directes (requests) peut optimiser ce processus.



A/B Tests XPaths

Les XPaths peuvent changer avec les mises à jour du site. Implémenter des tests A/B sur les sélecteurs ou un système de mapping adaptatif pour maintenir la robustesse du scraping.



Parallélisation

Explorer des méthodes d'extraction simultanée pour réduire significativement les temps de scraping et augmenter le volume de données traitées.



Exports & Tableau de Bord

Mettre en place des exports vers des formats standardisés (CSV, Parquet) et développer un tableau de bord pour une visualisation immédiate des données.



Normalisation des Catégories

Définir et implémenter une structure de catégories uniforme pour les produits, facilitant l'analyse comparative et l'intégration aux systèmes internes.

Conclusion

Bilan du projet

- Automatiser la collecte et la centralisation des promotions
- Pipeline complet intégrant Selenium, SQLite et Flask.
- Extrait, nettoie et stocke les données de manière fiable
- Offre une interface simple pour la visualisation

Perspectives

Les prochaines étapes visent à améliorer :

- la performance
- la résilience
- la valorisation des données

