esprit
Se former autrement

Data science project

Scrapping des données avec selenium

Fatma mrabti
Dorra Bouzidi
Oussema khalifa ben mimouna
Mohamed ali lamouchi
Fatma mhadhbi
Aymen kefi

Sommaire

1-Objectif

Le script a pour objectif de recueillir des données sur les accidents de véhicules en france à partir du site https://www.data.gouv.fr/fr/datasets/bases-de-donnees-annuelles-des-accidents-corporels-de-la-circulation-routiere-annees-de-2005-a-2022/?fbclid=IwAR0fPLr6O9g32O3GV6J4vcYXEnkLjuhm_xshJ3OipPy4PA1ulF16G7UT6X4
Pour automatiser la navigation sur le site, le script utilise la bibliothèque Selenium by python.

2- Description des données et technologies utilisées

2.1 Description des données

Les données utilisées proviennent du fichier national des accidents corporels de la circulation, géré par l'ONISR, recueillant des informations sur les accidents en France de 2005 à 2022. Chaque entrée décrit un accident impliquant au moins un véhicule avec des victimes nécessitant des soins, enregistré par les forces de l'ordre. Les bases de données annuelles, composées de fichiers CSV (Caractéristiques, Lieux, Véhicules, Usagers), fournissent des détails sur les caractéristiques de l'accident, son lieu, les véhicules et les victimes.

Spécifications de la base

La base Etalab de données des accidents corporels de la circulation d'une année donnée, est répartie en 4 rubriques sous la forme pour chacune d'elles d'un fichier au format csv.

1. La rubrique CARACTERISTIQUES qui décrit les circonstances générales de l'accident
2. La rubrique LIEUX qui décrit le lieu principal de l'accident même si celui-ci s'est déroulé à une intersection
3. La rubrique VEHICULES impliqués
4. La rubrique USAGERS impliqués

*Etalab est une administration française qui a pour mission de promouvoir l'ouverture des données publiques en France

2.2 Technologies utilisées

Editeur de code

VsCode

Langage de programmation

Python

Bibliothéque

Selenium

Navigateur

Chrome

2.3 Script

index.py

```python
from selenium import webdriver
from selenium.webdriver.chrome.options import Options
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.common.by import By
from selenium.webdriver.support import expected_conditions as EC
from selenium.webdriver.common.keys import Keys
import time

# Set up Chrome options
chrome_options = Options()
chrome_options.add_argument("--start-maximized")  # This opens Chrome in fullscreen mode
chrome_options.add_experimental_option('detach', True)

# Set up the webdriver with the configured options
driver = webdriver.Chrome(options=chrome_options)
# Navigate to the website
url = 'https://www.data.gouv.fr/fr/'
driver.get(url)

# Wait for up to 3 seconds
time.sleep(3)
# Locate and click on the 'Données' link
try:
    donnees_container = WebDriverWait(driver, 10).until(
        EC.presence_of_element_located((By.XPATH, '//li[contains(.,"Données")]'))
    )
    donnees_container.click()
except Exception as e:
    print(f"An error occurred: {e}")

# Wait for up to 3 seconds
time.sleep(3)
# Find the input field and type the text
try:
    search_input = driver.find_element(By.ID, 'search-input-1')
    search_input.send_keys("accidents corporels", Keys.RETURN)
except Exception as e:
    print(f"An error occurred: {e}")

# Locate and click on the 'Bases de données annuelles des accidents' link
try:
    donnees_container = WebDriverWait(driver, 10).until(
        EC.presence_of_element_located((By.XPATH, '//li[contains(.,"Bases de données annuelles des accidents corporels de la circulation routière - Années de 2005 à 2022")]'))
    )
    donnees_container.click()
except Exception as e:
    print(f"An error occurred: {e}")

element = driver.find_element(By.ID, "resources-panel")
driver.execute_script("arguments[0].scrollIntoView();", element)

# Find usagers-2022.csv and click the download link
try:
    download_link = driver.find_element(By.XPATH, '//a[@href="https://www.data.gouv.fr/fr/datasets/r/62c20524-d442-46f5-bfd8-982c59763ec8"]')
    download_link.click()
except Exception as e:
    print(f"An error occurred: {e}")
# Wait for up to 5 seconds
time.sleep(5)
# Find vehicules-2022.csv and click the download link
try:
    download_link = driver.find_element(By.XPATH, '//a[@href="https://www.data.gouv.fr/fr/datasets/r/c9742921-4427-41e5-81bc-f13af8bc31a0"]')
    download_link.click()
except Exception as e:
    print(f"An error occurred: {e}")
# Wait for up to 5 seconds
time.sleep(5)
# Find lieux-2022.csv and click the download link
try:
    download_link = driver.find_element(By.XPATH, '//a[@href="https://www.data.gouv.fr/fr/datasets/r/a6ef711a-1f03-44cb-921a-0ce8ec975995"]')
    download_link.click()
except Exception as e:
    print(f"An error occurred: {e}")
# Wait for up to 5 seconds
time.sleep(5)
# Find carcteristiques-2022.csv and click the download link
try:
    download_link = driver.find_element(By.XPATH, '//a[@href="https://www.data.gouv.fr/fr/datasets/r/5fc299c0-4598-4c29-b74c-6a67b0cc27e7"]')
    download_link.click()
except Exception as e:
    print(f"An error occurred: {e}")
# Wait for up to 5 seconds
time.sleep(5)
# Find description-des-bases-de-donnees-annuelles-2022.pdf and click the download link
try:
    download_link = driver.find_element(By.XPATH, '//a[@href="https://www.data.gouv.fr/fr/datasets/r/8ef4c2a3-91a0-4d98-ae3a-989bde87b62a"]')
    download_link.click()
except Exception as e:
    print(f"An error occurred: {e}")
```

2.4 Details du script

1-Importation des bibliothèques

-Importez les bibliothèques nécessaires pour le WebDriver de Selenium, les options Chrome, l'attente, et le temps.

2-Configuration des options Chrome

-Configurez les options de Chrome, y compris le démarrage du navigateur en mode plein écran, et le détachement du WebDriver du navigateur Chrome.

3-Navigation sur le Site

- Accès à l'URL de Data.gouv.fr. Attente de 3 secondes pour laisser la page se charger. Localisation et clic sur le lien "Données". Attente de 3 secondes. Recherche du terme "accidents corporels" dans le champ de recherche.

4-Téléchargement des Données

- Localisation et clic sur le lien "Bases de données annuelles des accidents corporels de la circulation routière - Années de 2005 à 2022". Défilement vers la section des ressources. Téléchargement des fichiers CSV (usagers-2022.csv, vehicules-2022.csv, lieux-2022.csv, caracteristiques-2022.csv) en utilisant les liens spécifiques.

5-Gestion des Erreurs

- Capture des erreurs éventuelles lors de la localisation des éléments ou du téléchargement des fichiers, avec affichage des messages d'erreur.

5-Attentes Temporaires

- Pause de 5 secondes entre chaque téléchargement pour assurer le chargement correct des pages.

3- Résultat et conclusion

3.1 Résultat

Num_Acc	id_usager	id_vehicule	num_veh	place	catu	grav	sexe	an_nais	trajet
202200000001	1 099 700	813 952	A01	1	1	3	1	2008	5
202200000001	1 099 701	813 953	B01	1	1	1	1	1948	5
202200000002	1 099 698	813 950	B01	1	1	4	1	1988	9
202200000002	1 099 699	813 951	A01	1	1	1	1	1970	4
202200000003	1 099 696	813 948	A01	1	1	1	1	2002	0

3.2 Conclusion

En conclusion, le script de scraping de données mis en place avec Selenium a permis de collecter efficacement des informations sur les collisions de véhicules. L'utilisation de Selenium avec python a permis d'extraire efficacement les données depuis le page web donnée.