Вариант 16. Гадильшина Валентина Евгеньевна, БД-231м

Лабораторная работа 2. Парсинг HTML. XPath+Selenium

SELENIUM

Извлечь данные через BeautifulSoup таблицы на сайте https://fred.stlouisfed.org/data/PRIME.

Задача выполняется в виртуальной машине selenium_dba_bmstu (логин: dba, пароль: 1).

Ход работы:

Шаг 1. Установка Selenium.

```
dba@dba-vm:~$ wget https://dl.google.com/linux/direct/google-chrome-stable_current_amd64.deb
--2024-09-29 20:58:49--  https://dl.google.com/linux/direct/google-chrome-stable_current_amd64.deb
Resolving dl.google.com (dl.google.com)... 142.250.179.78, 2a00:1450:4007:813::200e
Connecting to dl.google.com (dl.google.com)|142.250.179.78|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 111860972 (107M) [application/x-debian-package]
Saving to: ‘google-chrome-stable_current_amd64.deb.1’

google-chrome-stabl 100%[===================>] 106,68M  6,37MB/s    in 25s

2024-09-29 20:59:15 (4,22 MB/s) - ‘google-chrome-stable_current_amd64.deb.1’ saved [111860972/111860972]

dba@dba-vm:~$ sudo apt-get install -y ./google-chrome-stable_current_amd64.deb
[sudo] password for dba:
Reading package lists... Done
dba@dba-vm:~$ sudo apt-get -y install python3-pip
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
dba@dba-vm:~$ pip3 install selenium
Defaulting to user installation because normal site-packages is not writeable
Requirement already satisfied: selenium in ./.local/lib/python3.10/site-packages (4.25.0)
Requirement already satisfied: trio-websocket~=0.9 in ./.local/lib/python3.10/site-packages (from selenium)
(0.11.1)
dba@dba-vm:~$ pip3 install webdriver-manager
Defaulting to user installation because normal site-packages is not writeable
Requirement already satisfied: webdriver-manager in ./.local/lib/python3.10/site-packages (4.0.2)
Requirement already satisfied: packaging in ./.local/lib/python3.10/site-packages (from webdriver-manager) (
```

Шаг 2. Создать тестовый файл для проверки работоспособности Selenium.

```
dba@dba-vm:~$ nano test.py
```

```
  GNU nano 6.2                                    test.py
from selenium import webdriver
from selenium.webdriver.chrome.options import Options
from selenium.webdriver.chrome.service import Service
from webdriver_manager.chrome import ChromeDriverManager

options = Options()
# options.add_argument('--headless')
# options.add_argument('--no-sandbox')
options.add_argument('--disable-dev-shm-usage')
driver = webdriver.Chrome(service=Service(ChromeDriverManager().install()), options=options)

driver.get("https://python.org")
print(driver.title)
driver.close()
```

Шаг 3. Проверка наличия файла и запуск тестового файла.

```
dba@dba-vm:~$ ls
Desktop                                 Music      test.py
Documents                               Pictures   thinclient_drives
Downloads                               Public     Videos
google-chrome-stable_current_amd64.deb  snap
google-chrome-stable_current_amd64.deb.1  Templates
dba@dba-vm:~$ python3 test.py
Welcome to Python.org
```

Шаг 4. Поиск пути к chromedriver.

```
dba@dba-vm:~$ find /home/dba -name "chromedriver"
/home/dba/.wdm/drivers/chromedriver
/home/dba/.wdm/drivers/chromedriver/linux64/129.0.6668.70/chromedriver-linux64/chromedriver
/home/dba/.wdm/drivers/chromedriver/linux64/129.0.6668.58/chromedriver-linux64/chromedriver
```

Шаг 5. Создание файла sel.py.

```
dba@dba-vm:~$ nano sel.py
```

Шаг 6. Обновленный код с использованием Selenium в sel.py.

```
  GNU nano 6.2                                    sel.py
from selenium import webdriver
from selenium.webdriver.chrome.service import Service
from selenium.webdriver.common.by import By
import pandas as pd
import time

# Укажите путь к chromedriver
chrome_driver_path = "/home/dba/.wdm/drivers/chromedriver/linux64/129.0.6668.58/chromedriver-linux64/chrome>

# Шаг 1: Настройка драйвера Selenium
service = Service(chrome_driver_path)
driver = webdriver.Chrome(service=service)

# Шаг 2: Открытие страницы
url = "https://fred.stlouisfed.org/data/PRIME"
driver.get(url)

# Шаг 3: Ждем несколько секунд для загрузки страницы и данных через JavaScript
time.sleep(5)

# Шаг 4: Поиск таблицы по ID
table = driver.find_element(By.ID, 'data-table-observations')

# Шаг 5: Извлечение заголовков таблицы (thead)
headers = []
thead = table.find_element(By.TAG_NAME, 'thead')
for th in thead.find_elements(By.TAG_NAME, 'th'):
    headers.append(th.text.strip())

# Шаг 6: Извлечение строк данных (tbody)
rows = []
tbody = table.find_element(By.TAG_NAME, 'tbody')
for tr in tbody.find_elements(By.TAG_NAME, 'tr'):
    cells = tr.find_elements(By.TAG_NAME, 'td')
```

Шаг 7. Результат работы скрипта.

```
dba@dba-vm:~$ nano sel.py
dba@dba-vm:~$ python3 sel.py
           DATE VALUE
0    1955-08-04  3.25
1    1955-10-14  3.50
2    1956-04-13  3.75
3    1956-08-21  4.00
4    1957-08-06  4.50
..          ...   ...
354  2023-02-02  7.75
355  2023-03-23  8.00
356  2023-05-04  8.25
357  2023-07-27  8.50
358  2024-09-19  8.00

[359 rows x 2 columns]
dba@dba-vm:~$
```