Лабораторная 2. Часть 2. Работа с Selenium.
Кузьменко Сергей БД-231м

1) Шаг 1. Установка Selenium

wget https://dl.google.com/linux/direct/google-chrome-stable_current_amd64.deb
sudo apt-get install -y ./google-chrome-stable_current_amd64.deb
sudo apt-get -y install python3-pip
pip3 install selenium
pip3 install webdriver-manager

2) Шаг 2. Создать тестовый файл для проверки работоспособности Selenium и запустить

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

3) Шаг 3. Определим путь драйвера webdriver.Chrome.
``bash
find /home/dba -name "chromedriver"
В поиске находим путь для виртуальной машины dba :
/home/dba/.wdm/drivers/chromedriver/linux64/129.0.6668.58/chromedriver
linux64/chromedriver

4) Шаг 4. Создать скрипт sel.py для парсинга веб-сайта и выполнить
обновленный код

```python
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/chromedriver"
# Шаг 1: Настройка драйвера Selenium
service = Service(chrome_driver_path)
driver = webdriver.Chrome(service=service)
# Шаг 2: Открытие страницы
url = "https://www.insurancejournal.com/top-100-insurance-agencies/"
driver.get(url)
# Шаг 3: Ждем несколько секунд для загрузки страницы
time.sleep(5)

# Шаг 4: Находим все элементы списка агентств
ranks = []
agency_names = []
pc_revenues = []
other_revenues = []
offices = []

agency_list = driver.find_elements(By.CLASS_NAME, 'rankings-agency-list-item')

# Шаг 5: В цикле извлекает Ранг, Название, P\C Revenue, Other Revenue и Офис
for agency in agency_list:
    rank = agency.find_element(By.CLASS_NAME, 'agency-rank').text.strip()
    ranks.append(rank)

    name = agency.find_element(By.CLASS_NAME, 'agency-name').text.strip()
    agency_names.append(name)

    pc_revenue = agency.find_element(By.CLASS_NAME, 'agency-revenue').text.strip().replace('P\C REVENUE\n', '').strip()
    pc_revenues.append(pc_revenue)

    # Извлечение Other Revenue (если отсутствует, добавляем None)
    try:
        other_revenue = agency.find_element(By.CLASS_NAME, 'agency-other-revenue').text.strip().replace('OTHER REVENUE\n', '').strip()
    except:
        other_revenue = None
    other_revenues.append(other_revenue)

    office = agency.find_element(By.CLASS_NAME, 'agency-office').text.strip().replace('OFFICE', '').strip()
    offices.append(office)
# Шаг 6. Создание DataFrame
data = {
    'Rank': ranks,
    'Agency Name': agency_names,
    'P/C Revenue': pc_revenues,
    'Other Revenue': other_revenues,
    'Office': offices
}
df = pd.DataFrame(data)

print(df.head())
# Шаг 7. Закрытие драйвера
driver.quit()
```

Результат работы

```
dba@dba-vm:~$ python3 sel.py
   Rank                     Agency Name   P/C Revenue   Other Revenue               Office
0     1  Alliant Insurance Services/Confie  $2,791,081,357  $1,127,643,532     Irvine, California
1     2                    Acrisure LLC  $2,731,289,283    $305,608,697  Grand Rapids, Michigan
2     3             HUB International Ltd.  $2,602,873,000  $1,701,966,000      Chicago, Illinois
3     4                         Lockton  $2,143,306,000    $625,054,000   Kansas City, Missouri
4     5                 AssuredPartners  $1,813,550,941    $859,893,000        Orlando, Florida
dba@dba-vm:~$
```

ПРИЛОЖЕНИЕ (КОД 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/chromedriver"
# Шаг 1: Настройка драйвера Selenium
service = Service(chrome_driver_path)
driver = webdriver.Chrome(service=service)
# Шаг 2: Открытие страницы
url = "https://www.insurancejournal.com/top-100-insurance-agencies/"
driver.get(url)
# Шаг 3: Ждем несколько секунд для загрузки страницы
time.sleep(5)

# Шаг 4: Находим все элементы списка агентств
ranks = []
agency_names = []
pc_revenues = []
other_revenues = []
offices = []

agency_list = driver.find_elements(By.CLASS_NAME, 'rankings-agency-list-item')

# Шаг 5: В цикле извлекает Ранг, Название, P\C Revenue, Other Revenue и Офис
for agency in agency_list:
    rank = agency.find_element(By.CLASS_NAME, 'agency-rank').text.strip()
    ranks.append(rank)

    name = agency.find_element(By.CLASS_NAME, 'agency-name').text.strip()
    agency_names.append(name)

    pc_revenue = agency.find_element(By.CLASS_NAME, 'agency-
revenue').text.strip().replace('P\C REVENUE\n', '').strip()
    pc_revenues.append(pc_revenue)

    # Извлечение Other Revenue (если отсутствует, добавляем None)
    try:
     other_revenue = agency.find_element(By.CLASS_NAME, 'agency-other-
revenue').text.strip().replace('OTHER REVENUE\n', '').strip()
    except:
     other_revenue = None
```

```python
        other_revenues.append(other_revenue)

        office = agency.find_element(By.CLASS_NAME, 'agency-
office').text.strip().replace('OFFICE', '').strip()
        offices.append(office)
# Шаг 6. Создание DataFrame
data = {
    'Rank': ranks,
    'Agency Name': agency_names,
    'P/C Revenue': pc_revenues,
    'Other Revenue': other_revenues,
    'Office': offices
}
df = pd.DataFrame(data)

print(df.head())
# Шаг 7. Закрытие драйвера
driver.quit()
```