الاسم / اسلام السيد رمزى الغرباوى

سيكشن / 1

CHAPTER 9

Multiple Choice Questions (MCQs)

1. requests
2. requests.get()
3. soup.find_all("a")
4. The inner text of the tag
5. Selenium
6. Respecting site limits
7. None of the above

True / False Questions

1. True

2. False

3. True

4. True

5. False

Short Answer / Conceptual Questions

1. **Explain the difference between requests and Selenium in web scraping**

requests fetches static HTML content via HTTP requests. Selenium automates a browser and can handle dynamic JavaScript content

2. **What is the purpose of the robots.txt file on a website?**

It specifies which parts of the website web crawlers are allowed or disallowed to scrape

3. **What is the difference between find() and find_all() methods in BeautifulSoup?**

find() returns the first matching element, while find_all() returns a list of all matching elements.

4. **Why is it important to use headers like "User-Agent": "Mozilla/5.0" in requests.get()?**

Headers make the request appear as if it comes from a real browser,

helping avoid blocks from websites.

5. **List three possible formats to store scraped data**

   Answer: CSV, JSON, Excel (XLSX)

#Problem 1

```
from bs4 import BeautifulSoup

import requests


url = "https://example.com"

page = requests.get(url)
```

```
soup = BeautifulSoup(page.content , "lxml")

pageTitle = soup.find("title").text.strip()

print(f"Page title: {pageTitle}")
```

#Problem 2

```
from bs4 import BeautifulSoup

import requests


url = "https://example.com"

page = requests.get(url)


soup = BeautifulSoup(page.content , "lxml")

textLink = soup.find("a").get('href')

print(f"All links in the page: {textLink}"
```

#Problem 3

```
from bs4 import BeautifulSoup


page = """
    <table>
        <tr><th>Name</th><th>Age</th></tr>
        <tr><th>Alice</th><th>25</th></tr>
        <tr><th>Bob</th><th>30</th></tr>
    </table>
"""
soup = BeautifulSoup(page, "lxml")
rows = soup.find("table").find_all("tr")
listOfLists = []


for row in rows:
    cells = [cell.text for cell in row.find_all(["th", "td"])]
    listOfLists.append(cells)
```

```
print(listOfLists)
```

#Problem 4

```
from selenium import webdriver
from selenium.webdriver.common.keys import Keys
from selenium.webdriver.common.by import By
import time


driver = webdriver.Chrome()


driver.get("https://www.google.com")


search_box = driver.find_element(By.NAME, "q")


search_box.send_keys("Python Web Scraping")
search_box.send_keys(Keys.RETURN)
```

```
time.sleep(2)

print(driver.title)

driver.quit()
```

#Problem 5

```
import csv

from bs4 import BeautifulSoup


page = """

    <ul>

        <li>Apple</li>

        <li>Banana</li>

        <li>Cherry</li>

   </ul>
```

```
"""

soup = BeautifulSoup(page , "lxml")


tags = soup.find_all("li")

fruits = []

for tag in tags:

   fruits.append(tag.text)

print(fruits)

with open("fruits.csv", "w", newline="") as outputFile:

    writer = csv.writer(outputFile)

    writer.writerow(["Fruit"])

for fruit in fruits:

writer.writerow([fruit])
```