**WEB SCRAPING – ANSWER SHEET**

**Name- SAURABH UPADHYAY**

**In Q1 to Q9, only one option is correct, Choose the correct option:**

1. Which of the following extracts information from user generated content?

A) Java script tagging B) Web scraping

C) A/B testing D) MROCs

**Ans) (B) Web scraping**

1. Which of the following is not a web scraping library in python?

A) selenium B) Beautiful soup

C) Requests C) scrapy

**Ans) (C) scrapy**

1. Selenium tests \_\_\_\_\_\_\_\_\_\_?

A) Browser based applications B) DOS applications

C) GUI applications D) All of the above

**Ans) (A) Browser based applications**

1. Task of crawling is performed by a complex software which is known as:

A) Scraper B) Crawler

C) Boat D) Spider

**Ans) (A) Scraper**

1. Which of the following commands is used to access name of a tag in Beautiful Soup?

A) tag.attrs B) tag.name

C) tag,id C) tag[‘id’]

**Ans) (B)tag.name**

1. Which of the following is the default parser in Beautiful Soup?

A) html.parser B) html5lib

C) lxml D) lxml-xml

1. In selenium the webdriver is used to?

A) design a test using selenese

B) test a web application on firefox only

C) execute tests on HtmlUnit browser

D) to download any content from a webpage

**Ans) D) to download any content from a webpage**

1. In selenium, driver**.find\_elements\_by\_xpath(‘given xpath’)** returns:

A) the first webelement associated with the ‘given xpath’

B) the url of first webelement associated with the ‘given xpath’

C) the list of all webelements associated with the ‘given xpath’

D) all the attributes of the first webelement associated with the ‘given xpath’

**Ans) (C) lxml**

1. The script **‘window.scrollBy(0,a)** scrolls the webpage by?

A) **‘a’** number of horizontal spaces

B) **‘a’** number of lines

C) **‘a’** number of pixels horizontally

D) **‘a’** number of pixels vertically

**Ans) D) ‘a’ number of pixels vertically**

**In Q10, more than one options are correct, Choose all the correct options:**

1. Which of the following is(are) tags of HTML?

A) <a> B) <b>

C) <image> D) <href>

**Ans:**

A) <a> B) <b>

D) <href>

**Q11 to Q13 are subjective answer type questions, Answer them briefly.**

1. What is the main difference between a web scraper and a web crawler?

**Ans:**

Web crawlers work by browsing to a series of webpages and analyzing their contents for links to other webpages. The links to the other webpages are then followed and searched for more links. The process of following and recording these links is referred to as “crawling.” While crawling through various web pages can reveal useful information about the structure of the web, extracting data from those sites, or “web scraping”, captures the content of those pages which can then be analyzed to reveal more information about the crawled pages. Many web crawlers utilize web scraping to contextualize the pages that they have crawled.

A web scraper's main purpose is to extract data from webpages. Web scrapers often have the ability to browse to different pages and follow links. Though web scrapers can crawl to different pages their primary purpose is scraping the data on those pages, not indexing the web.

1. What is **‘robots.txt’** file? What is the use of **‘robots.txt’** file?

**Ans:**

**What is a robots.txt file?**

A robots.txt file tells search engine crawlers which pages or files the crawler can or can't request from your site. This is used mainly to avoid overloading your site with requests; it is not a mechanism for keeping a web page out of Google. To keep a web page out of Google, you should use noindex directives, or password-protect your page.

**What is robots.txt used for?**

robots.txt is used primarily to manage crawler traffic to your site, and usually to keep a page off Google, depending on the file type:

1. What are static and dynamic web pages?

**Ans:**

**Static Web pages:**  
Static Web pages are very simple. It is written in languages such as HTML, JavaScript, CSS, etc. For static web pages when a server receives a request for a web page, then the server sends the response to the client without doing any additional process. And these web pages are seen through a web browser. In static web pages, Pages will remain the same until someone changes it manually.

**Dynamic Web Pages:**  
Dynamic Web Pages are written in languages such as CGI, AJAX, ASP, ASP.NET, etc. In dynamic web pages, the Content of pages is different for different visitors. It takes more time to load than the static web page. Dynamic web pages are used where the information is changed frequently, for example, stock prices, weather information, etc.

**Difference between Static and Dynamic Web Pages:**

|  |  |  |
| --- | --- | --- |
| **SL.NO** | **STATIC WEB PAGE** | **DYNAMIC WEB PAGE** |
| 1. | In static web pages, Pages will remain same until someone changes it manually. | In dynamic web pages, Content of pages are different for different visitors. |
| 2. | Static Web Pages are simple in terms of complexity. | Dynamic web pages are complicated. |
| 3. | In static web pages, Information are change rarely. | In dynamic web page, Information are change frequently. |
| 4. | Static Web Page takes less time for loading than dynamic web page. | Dynamic web page takes more time for loading. |
| 5. | In Static Web Pages, database is not used. | In dynamic web pages, database is used. |
| 6. | Static web pages are written in languages such as: HTML, JavaScript, CSS, etc. | Dynamic web pages are written in languages such as: CGI, AJAX, ASP, ASP.NET, etc. |
| 7. | Static web pages does not contain any application program . | Dynamic web pages contains application program for different services. |
| 8. | Static web pages require less work and cost in designing them. | Dynamic web pages require comparatively more work and cost in designing them. |

**Q14 and Q15 are programming practice questions. Solve it using JUPYTER NOTEBOOK and paste the solution in your answer sheets.**

1. Write a python program to check whether a webpage contains a title or not.

**Ans.**

# import libreries

import requests

from bs4 import BeautifulSoup

# Define a variable assign the url

url='https://www.w3schools.com/'

res = requests.get(url=url)

res

<Response [200]>

#scarp the all content and store in soup variable

soup = BeautifulSoup(res.content, 'html.parser')

#find the title in scrap content and store in variable t if is avaiable.

t=soup.find('title')

# To check that title present is the content or not.

if t in soup() :

print("Title of the webpage is :-", t.text)

else:

print("Web page has no title!")

Out put :-**Title of the webpage is :- W3Schools Online Web Tutorials**

1. Write a python program to access the search bar and search button on images.google.com.

**Ans**.

# import libreries

import time

from selenium import webdriver

from selenium.webdriver.common.keys import Keys

from selenium.webdriver.common.by import By

driver=webdriver.Chrome(executable\_path=r"D:\webscraping\chromedriver.exe")

url="https://images.google.com/"

driver.get(url)

#Accessing the search bar using find\_element\_by\_name method

search\_bar=driver.find\_element(By.NAME,"q")

#clicking on search button

search\_button=driver.find\_element\_by\_xpath("//div[@class='FAuhyb']")

#input which you want search

search\_word= 'corona'

search\_bar.send\_keys(search\_word)

#clicking on search button

search\_button.click()

time.sleep(7)

**driver.quit()**