**WEB SCRAPING**

## **WORKSHEET – 1**

**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 D) scrapy

ANS C) Requests

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

A) Browser based applications B) DOS applications

C) GUI applications D) All of the above

ANS D) All of the above

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

A) Scraper B) Crawler

C) Boat D) Spider

ANS B) Crawler

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 A) tag.attrs

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

A) html.parser B) html5lib

C) lxml D) lxml-xml

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

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 B) <b>

**Q10 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 usually work on indexing of the webpage to improved ranking of websites.  
 Web scraper is used to extract a data from web pages

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

Ans – Robots.txt files tells information to web crawler to which page of website, you have to crawler or not.  
 This is used mainly to avoid overloading your site with requests.

Used of Robots.txt

User-agent: \*

Disallow: /

User-agent: \*

Disallow:

User-agent: Googlebot

Disallow: /example-subfolder/

1. What are static and dynamic web pages?

ANS

.

|  |  |  |
| --- | --- | --- |
| 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. |

**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.

**import selenium**

**from selenium import webdriver as wb**

**webD=wb.Chrome('chromedriver.exe')**

**driver = wb.Chrome()**

**url = "https://www.w3schools.com/"**

**driver.get(url)**

**# Get title of url**

**get\_title = driver.title**

**# show the title of URL**

**if get\_title =={}:**

**print("NO Title")**

**else:**

**print(get\_title)**

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

**import selenium**

**from selenium import webdriver as wb**

**webD=wb.Chrome('chromedriver.exe')**

**driver = wb.Chrome()**

**driver.get("https://images.google.com/")**

**que=driver.find\_element\_by\_xpath("//input[@name='q']")**

**que.send\_keys("Datatrainedacademy ")**

**que.send\_keys(Keys.RETURN)**