**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

Sol. B

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

A) selenium B) Beautiful soup

C) Requests C) scrapy

Sol. C

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

A) Browser based applications B) DOS applications

C) GUI applications D) All of the above

Sol. A

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

A) Scraper B) Crawler

C) Boat D) Spider

Sol. Crawler and spider

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’]

Sol. B

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

A) html.parser B) html5lib

C) lxml D) lxml-xml

Sol. A

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

Sol. D

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’

Sol. A

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

Sol. D

**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>

Sol. A and D

**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?

**Sol.** Though sometimes the two terms are used interchangeably the main difference is that webcrawlers usually focus on indexing the web while web scrapers extract or "scrape" data from webpages. 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?

**Sol.** A robots. txt file is a simple text file which webmasters can create to tell web crawlers which parts of a website should be crawled and which should not. The file is stored in the main directory (root) on the server. When a crawler arrives at a website, it first reads the robots.

1. What are static and dynamic web pages?

**Sol.** In staticweb pages, Pages will remain same until someone changes it manually. In dynamic **w**eb pages, Content of pages are different for different visitors

**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.

**Sol o/p**  None

<h1>Example Domain</h1>

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