**WEB SCRAPING**

**WORKSHEET – 1**

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

* Which of the following extracts information from user generated content?

A) Java script tagging B) Web scraping

C) A/B testing D) MROCs

**Answer-**Web scraping

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

A) selenium B) Beautiful soup

C) Requests C) scrapy

**Answer -** C) scrapy

* Selenium tests \_\_\_\_\_\_\_\_\_\_?

A) Browser based applications B) DOS applications

C) GUI applications D) All of the above

**Answer**- Browser based applications

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

A) Scraper B) Crawler

C) Boat D) Spider

**Answer**-Scraper,crawler,spider

* 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’]

**Answer -**tag.name

* Which of the following is the default parser in Beautiful Soup?

A) html.parser B) html5lib

C) lxml D) lxml-xml

**Answer-**lxml

* 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

**Answer-**test a web application on firefox only

* 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’

**Answer-**all the attributes of the first webelement associated with the ‘given xpath’

* 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

**ANSWER- ‘a’** number of pixels horizontally

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

* Which of the following is(are) tags of HTML?

A) <a> B) <b>

C) <image> D) <href>

**Answer --**A**)**<a>B) <b> D) <href>

**-Q10 to Q13 are subjective answer type questions, Answer them briefly.**

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

**Answer**- **WEB CRAWLER** it is used for indexing the information on the page using bots also known as crawlers.The term crawling comes from the way a spider would crawl. That’s why a web crawler is also sometimes called a spider. It’s basically an internet bot that systematically browses (read crawls) the World Wide Web, usually for the purpose of web indexing.

**WEB SCRAPER** is basically extracting data from websites in an automated manner.Data scraping involves locating data and then extracting it. It does not copy and paste but directly fetches the data in a precise and accurate manner.

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

**Answer-**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.

**USE**-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.

* What are static and dynamic web pages?

**Answer -"Static" means unchanged or constant, while "dynamic" means changing or lively.**

**Static web pages** are generally simple HTML written pages which serve as response from browser to server in which all the information and data is static in nature and it does not get changed until someone changed it manually**.**

**Dynamic webpages** are the pages written in some more complex language such as ASP.NET in which data is rendered after some interpretation and capacity to produce distinctive content for different calls.

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

* Write a python program to check whether a webpage contains a title or not.
* Write a python program to access the search bar and search button on images.google.com.