**WEBSCRAPING**

## **WORKSHEET – 1 (Solutions)**

**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

**B**

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

A) selenium B) Beautiful soup

**C) Requests** C) scrappy

**C**

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

**A) Browser based applications** B) DOS applications

C) GUI applications D) All of the above

**A**

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

A) Scraper **B) Crawler**

**C) Boat**  **D) Spider**

**B,C,D**

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’]

**B**

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

A) html.parser B) html5lib

**C) lxml**  D)lxml-xml

**C**

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**

**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’**

**C**

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**

**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>

**A,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?

**Web scrapper :**

* It is used for **downloading** information.
* It **need not visit all the pages** of website for information.
* A Web Scraper **doesn’t obey robots.txt** in most of the cases.
* Data de-duplication is not necessarily a part of Web Scraping.
* From here we will get specific information which type of information we are looking to scrap only those data only we will get ,like while scrapping data from amazon if we need price and rating ,brand in that case we will get only that data with web scrapper.

**Web crawler :**

* It is used for **indexing** of Web pages .
* It **visits each and every page**, until the last line for information.
* It always **obeys robots.txt**.
* Data de-duplication and integral part of Web Scraping.
* Web crawler are normally used by major search engines such as **Google ,Bing ,Yahoo** and some other .

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

**robots.txt** file tells search engine crawlers which pages or files the crawler can or can’t request from your site. A search engine bot like Google bot will read the robots.txt file prior to crawling your site to learn what pages it should deal with.

**Uses:**

* The use of **‘robots.txt’** file is to block bots, spiders, and other crawlers from accessing any site’s pages.
* When bots, spiders and other crawlers hit a site, they can utilize extensive resource (memory and CPU) usage and can also lead to high load on the server and slow down a site.
* Robots.txt files are used to limit server load from bots, spiders, and other crawlers, and they also tell search engines what content on your site they should and should not index.

1. **What are static and dynamic web pages?**

**Static web pages :**

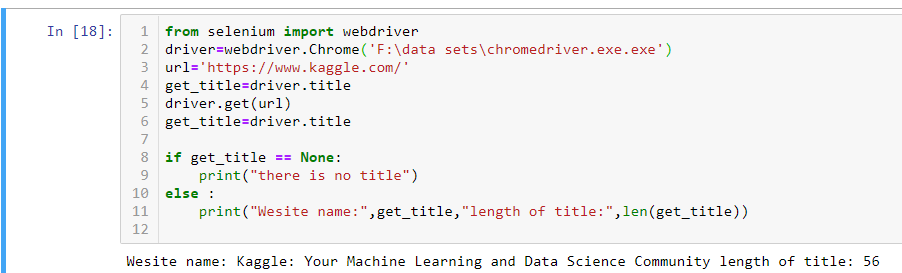
* Static website is the basic type of website that is easy to create. You don't need the knowledge of web programming and database design to create a static website. Its web pages are coded in HTML.
* Contents don't change unless manually edited by someone . Uses file extensions .html, for its pages.

**Dynamic Web pages:**

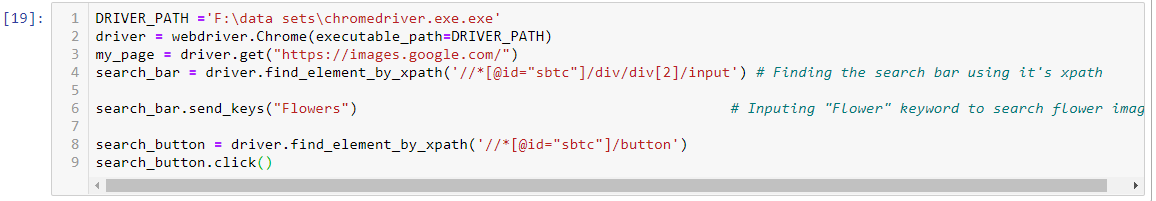
* Contents update and change on their own without human intervention.Uses file extensions .php, .jsp .asp etc.
* Dynamic website is a collection of dynamic web pages whose content changes dynamically. It accesses content from a database or Content Management System. Therefore, when you alter or update the content of the database, the content of the website is also altered or updated.

**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.

****

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

****