**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

**Answer: B) Web scraping**

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

A) selenium B) Beautiful soup

C) Requests C) scrappy

**Answer: C) Requests**

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

A) Browser based applications B) DOS applications

C) GUI applications D) All of the above

**Answer: 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

**Answer: 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’]

**Answer: 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

**Answer: C) lxml**

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

**Answer: 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’

**Answer: C) the list of all webelements 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

**Answer: 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>

Answer: A) <a> , 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?

**Answer - The main difference is web crawlers usually focus on indexing the web but web scrapers extract or "scrape" data from webpages.**

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

**Answer - 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.**

**Use of ‘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.**

1. What are static and dynamic web pages?

**Answer - "Static" means unchanged or constant, while "dynamic" means changing or lively. Therefore, static Web pages contain the same prebuilt content each time the page is loaded, while the content of dynamic Web pages can be generated on-the-fly. ... For example, the server may display the current time and date on the Web page.**

**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.

**Program code –**

**import selenium**

**from selenium import webdriver**

**DRIVER\_PATH='D:\software\chromedriver\_win32 (1)\chromedriver.exe'**

**driver = webdriver.Chrome(executable\_path= DRIVER\_PATH)**

**get\_title = driver.get("https://www.cognizant.com/")**

**# Printing the title of this URL**

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

**print("No Title")**

**else:**

**print(get\_title)**

**Output – “No Title”** but the website is highlighted **by Jupiter notebook**

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

**Program code –**

**import selenium**

**from selenium import webdriver**

**import os**

**DRIVER\_PATH='D:\software\chromedriver\_win32 (1)\chromedriver.exe'**

**driver = webdriver.Chrome(executable\_path= DRIVER\_PATH)**

**get\_title = driver.get("https://www.google.co.in/imghp?hl=en&tab=wi&authuser=0&ogbl")**

**def make\_directory(dirname):**

**current\_path= os.getcwd()**

**path= os.path.join(current\_path, dirname)**

**if not os.path.exists(path):**

**os.makedirs(path)**

**make\_directory("search")**

**search = driver.find\_elements\_by\_xpath("//input[@class='\_gLFyf gsfi']")**

**search = []**

**for search in search:**

**source= search.get\_attribute('src')**

**urls.append(source)**

**print(source)**

**Output – Jupiter notebook took control over google chrome images website with search bar as input class.**