**WEB SCRAPING**

## **WORKSHEET – 1 Solution**

**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: D**

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

A) selenium B) Beautiful soup

C) Requests C) scrapy

**Answer: A**

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

A) Browser based applications B) DOS applications

C) GUI applications D) All of the above

**Answer: A**

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

A) Scraper B) Crawler

C) Boat D) Spider

**Answer: B**

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**

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

A) html.parser B) html5lib

C) lxml D) lxml-xml

**Answer: 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

**Answer: A**

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**

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**

1. Which of the following is(are) tags of HTML?

A) <a> B) <b>

C) <image> D) <href>

**Answer: A and C**

**Answer 11:**

**Web crawling** is a process performed by web crawlers which includes systematic browsing of the Web in order to analyse web pages and add them to the database if needed. Web crawler starts with a page or a list of pages and follows all links on them according to certain patterns and instructions.

For example, web crawling is performed 24/7 by all search engines (e.g. Google, Bing, Yahoo etc.) to extend and update their index databases.

**Web scraping** is a technique used for extracting data from websites. It’s based on web crawling because before you can extract any data from a webpage you have to crawl it. We can create a web scraping script ourselves allowing us to find and extract data according to custom parameters you set.

Web scraping can be used for retrieving such info as phone numbers,images, prices, pieces of code/text etc. There are various techniques used to find certain pieces of data on webpages like regular expressions, XPath, CSS selector etc.

In simple terms, I’d say that scraping is an addon of crawling which allows extracting data from crawled pages.

**Answer 12 :**

Robots.txt file is a small text file but it plays an important role in SEO. It provides instructions to Search engines like what to crawl and what to ignore. This file is really very useful. Through this file you can instruct SEs to ignore any duplicate pages, internal search results pages, any “not in use” folder, tags, images, PDFs, etc. on your website. But if you do not have any robots.tx file on your server then search engines will have a free run to crawl and index anything they find on your website.

**Answer 13:**

A static web pages is a page that are manually created and contains only HTML, JavaScript and CSS and no a interaction with database and it does need to have a content manager.

A dynamic web pages has HTML, JavaScript and CSS, and also a backend that can be made in PHP, Java, Asp.net and the web app has a connection with database that stores the posts, images, videos etc and the app has also a content manager and through it the owner of the page can add/update/delete content/images/videos and so on.

**Answer 14 and Answer 15 is in the Jupyter Notebook**