**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**

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

A) selenium B) Beautiful soup

C) Requests C) scrapy

**Answer :** as per my knowledge all four are libraries for web scraping

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: 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

**Answer : C**

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 : D**

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**

**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 and B**

**Q11 to Q13 are subjective answer type questions, Answer them briefly.**

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

**Ans11. Web Scraper** extracts or “scrape” data from webpages.

**Web Crawler** usually focus on indexing the web.

There is a fair bit of overlap between the web crawlers and web scrapers.

Web crawlers work by browsing to a series of webpages and analyzing their contents for links to

other webpages. The links to the other webpages are then followed and searched for more links.

The process of following and recording these links is referred to as “crawling.” While crawling

through various web pages can reveal useful information about the structure of the web, extracting

data from those sites, or “web scraping”, captures the content of those pages which can then be analyzed

to reveal more information about the crawled pages. Many web crawlers utilize web scraping to

contextualize the pages that they have crawled.

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

**Answer 12 robots.txt** file :- The robots exclusion standard, also known as the robots exclusion protocol

or simply robots.txt, is a standard used by websites to communicate with web crawlers and other web robots.

The standard specifies how to inform the web robot about which areas of the website should not be

processed or scanned. Robots are often used by search engines to categorize websites.

1. What are static and dynamic web pages?

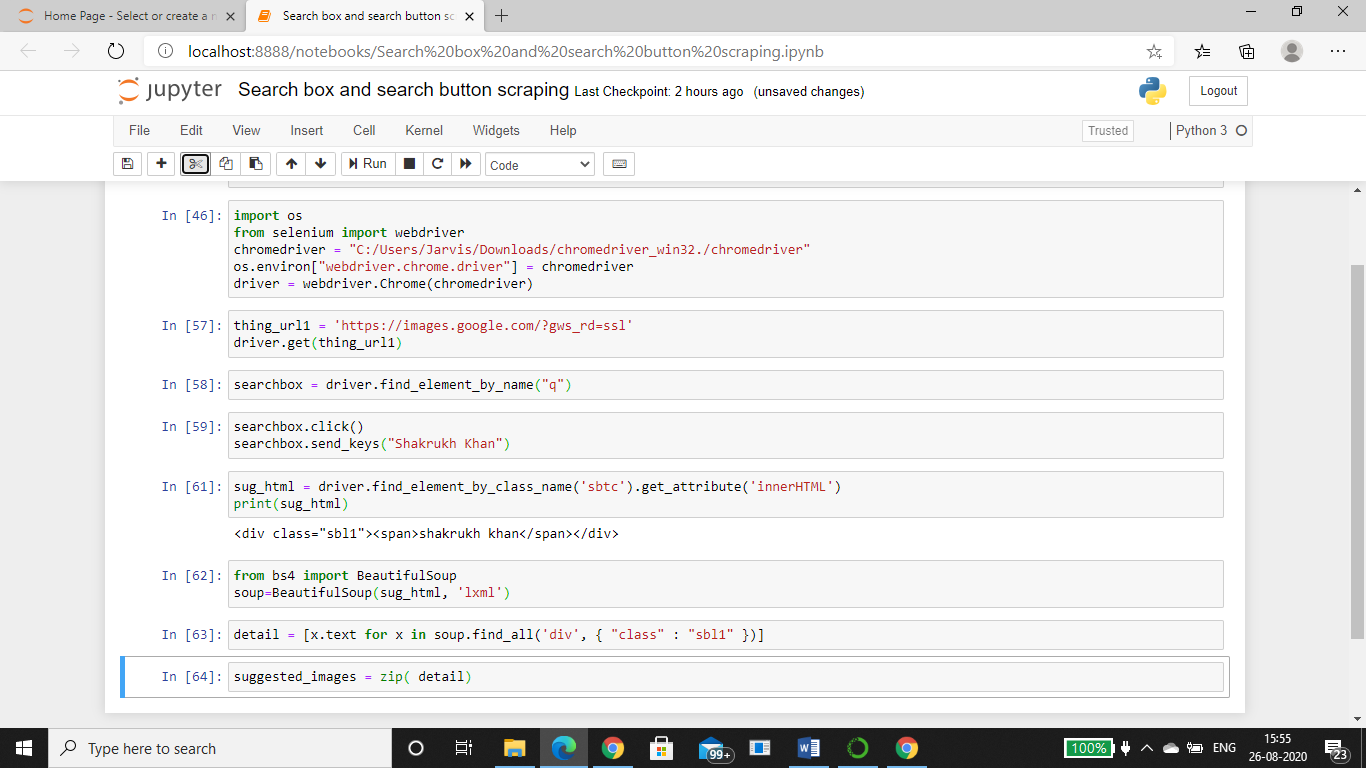
**Answer 13.** Web pages can be either static or dynamic. "Static" means unchanged or constant, while

"dynamic" means changing or lively. Therefore, static Web pages contain the same prebuilt content

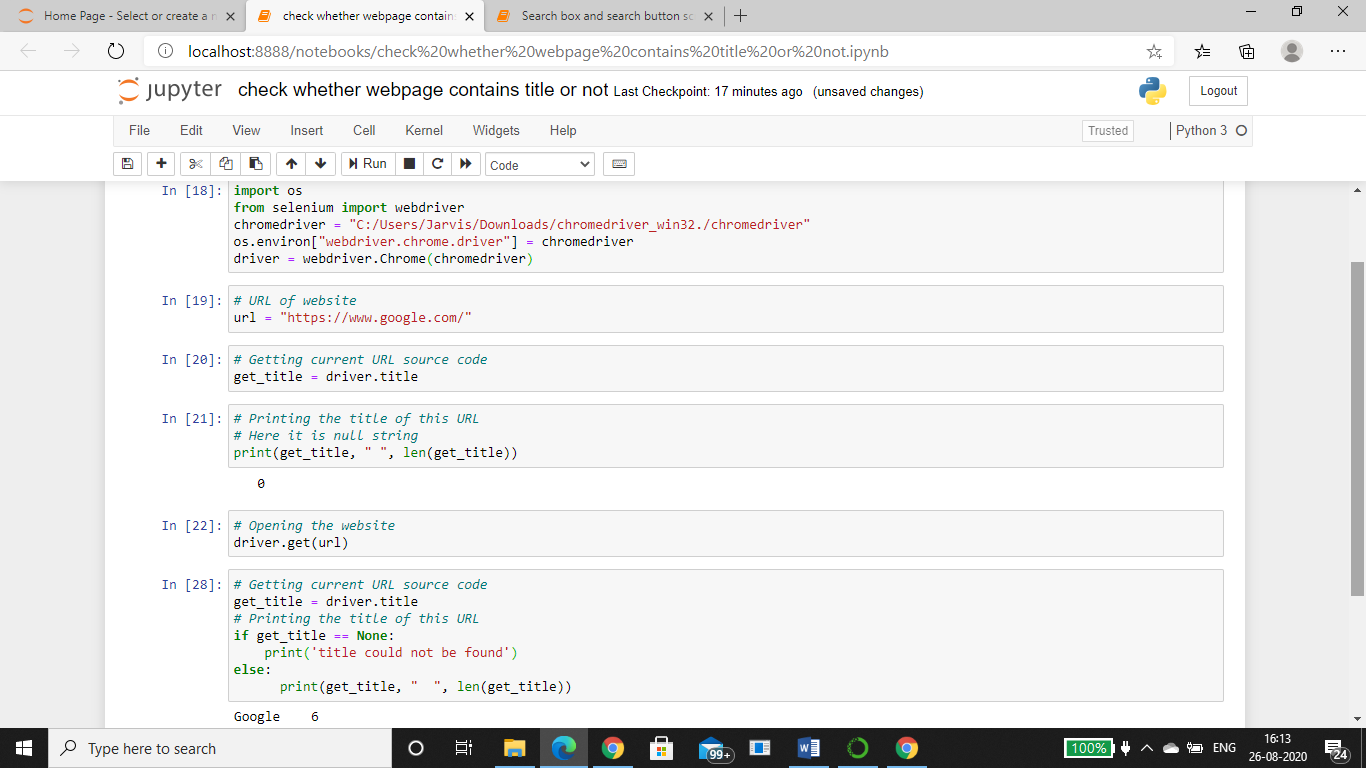
each time the page is loaded, while the content of dynamic Web pages can be generated on-the-fly.

**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.

****