**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

ANS - B

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

A) selenium B) Beautiful soup

C) Requests C) scrapy

ANS- C-Requests.

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

A) Browser based applications B) DOS applications

C) GUI applications D) All of the above

ANS-D

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

A) Scraper B) Crawler

C) Boat D) Spider

ANS – 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’]

ANS – 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

ANS -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

ANS - 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’

ANS – 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

ANS-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>

ANS – 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?

ANS- The main difference is that web crawlers usually focus on indexing the web while web scrapers extract or scrape data from webpages.

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

ANS**-** Robots.txt is a text file webmasters create to instruct web robots (typically search engine robots) how to crawl pages on their website. The robots.txt file is part of the the robots exclusion protocol (REP), a group of web standards that regulate how robots crawl the web, access and index content, and serve that content up to users.

This is used mainly to avoid overloading your site with requests; it is not a mechanism for keeping a web page out of Google.

1. What are static and dynamic web pages?

Ans**- 1.** In static web pages, Pages will remain same until someone changes it manually and while in dynamic web pages, Content of pages are different for different visitors.

2. Static Web Pages are simple in terms of complexity while Dynamic web pages are complicated.

3. In static web pages, Information are change rarely and in dynamic web page, Information are change frequently.

**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.

**ANS -** from urllib.request import urlopen

from urllib.error import HTTPError

from bs4 import BeautifulSoup

def getTitle(url):

try:

html = urlopen(url)

except HTTPError as e:

return None

try:

bsObj = BeautifulSoup(html.read(), "lxml")

title = bsObj.body.h1

except AttributeError as e:

return None

return title

print(getTitle("http://www.example.com/"))

<h1>Example Domain</h1>

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

**ANS-**

pip install icrawler

from icrawler.builtin import GoogleImageCrawler

google\_Crawler = GoogleImageCrawler(storage = {'root\_dir': r'path'})

google\_Crawler.crawl(keyword = 'query', max\_num = 10)

OUPUT - url