**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

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

A) selenium B) Beautiful soup

C) Requests C) scrapy

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

A) Browser based applications B) DOS applications

C) GUI applications D) All of the above

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

A) Scraper B) Crawler

C) Boat D) Spider

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’]

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

A) html.parser B) html5lib

C) lxml D) lxml-xml

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

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’

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

**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>

**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 scrapers are used to scrape data from websites, whereas web crawler is used to index a website using bots. Once a website is indexed using web crawlers, we can scrape the data/information using a web scraping software.

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

Robots.txt file is always placed in the root directory of a website and tells a search engine, which pages in that website can be crawled. This text file will contain useragent, which defines the bot and list of webpages that are listed as “Allow” or “Disallow”. A crawler bot will not access those pages that are listed under “Disallow”.

1. What are static and dynamic web pages?

Static web pages are web pages that remain constant or the contents of such pages does not change frequently and are unchanged. Dynamic web pages, on the other hand, have contents that change frequently and are lively. Static web pages have the same content, each time a user visits that page, whereas, contents of dynamic web pages change, each time the user loads the same 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.

from bs4 import BeautifulSoup

import requests

def title(url):

res = requests.get(url)

bs = BeautifulSoup(res.content)

# bs

title = bs.find\_all("title")[0]

if title=="None":

print("No title available")

else: print(title.text)

title("https://www.github.com") # Write complete url including "http://"

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

from selenium import webdriver

def search\_image(Query):

# Open webdriver

driver = webdriver.Chrome('chromedriver.exe')

# Open URL

driver.get('https://images.google.com')

# Select the Seach bar and type Search Query

driver.find\_element\_by\_xpath("//input[@class='gLFyf gsfi']").send\_keys(Query)

# Click on Search Button

driver.find\_element\_by\_xpath("//button[@class='Tg7LZd']").click()

search\_image("Fun with python")