**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?

B) Web scraping

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

All are.

A) selenium B) Beautiful soup

C) Requests C) scrapy

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

A) Browser based applications

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

B) Crawler

C) Boat D) Spider

1. Which of the following commands is used to access name of a tag in Beautiful Soup?

B) tag.name

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

C) lxml

1. In selenium the webdriver is used to?

A) design a test using selenese

1. In selenium, driver**.find\_elements\_by\_xpath(‘given xpath’)** returns:

C) the list of all webelements associated with the ‘given xpath’

1. The script **‘window.scrollBy(0,a)** scrolls the webpage by?

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>

**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 that the scraper can extract data from a given page however, it takes a crawler to extract data from the given page and also the references given on that page.

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

**Answer:** This is a file also known as robots exclusion protocol which includes instructions for web robots telling parts of websites should be and the parts which should not be processed. However, this is just a suggestion and it is not mandatory for web robots to follow such instructions. This file is mainly used to avoid overloading the site.

1. What are static and dynamic web pages?

**Answer:** Static web pages do not change unless the content is manually changed. However, dynamic web pages use database and the content is change using this database during the calls. Dynamic sites or pages are easy to manipulate and user friendly.

**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.

**Answer:** file name: [CheckWebPageTitle.ipynb](https://github.com/DipeshChothwe/Worksheets/blob/master/CheckWebPageTitle.ipynb)

#python program to check whether a webpage contains a title or not

from urllib.request import urlopen

from urllib.error import HTTPError

from bs4 import BeautifulSoup

def CheckTitle(url):

    try:

        html = urlopen(url)

    except HTTPError as e:

        return None

    try:

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

        title = soup.title.name

        if title == None:

            return "Title could not be found"

        else:

            return "This site has a title"

    except AttributeError as e:

        return None

print(CheckTitle("https://www.crummy.com/software/BeautifulSoup/bs4/doc/#find-parents-and-find-parent"))

print(CheckTitle("https://www.google.co.in/"))

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

**Answer:** file name: [AccessGoogleImage.py](https://github.com/DipeshChothwe/Worksheets/blob/master/AccessGoogleImage.py)

from selenium import webdriver

from selenium.webdriver.common.keys import Keys

DRIVER\_PATH = 'G:\FliproboAssgn\chromedriver\_win32\Chromedriver.exe'

browser = webdriver.Chrome(executable\_path = DRIVER\_PATH)

search = ['nature']

for book in search:

  browser.get('https://www.google.com/imghp')

  e = browser.find\_element\_by\_name('q')

  e.send\_keys(search)

  e.send\_keys(Keys.ENTER)