**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?

**Answer**: B) Web scraping

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

**Answer**: C) Requests

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

**Answer**: A) Browser based applications

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

**Answer**: D) Spider

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

**Answer**: B) tag.name

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

**Answer**: A) html.parser

1. In selenium the webdriver is used to?

**Answer**: A) design a test using selenese

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

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

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

**Answer**: 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?

**Answer**: 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 web crawlers usually focus on indexing the web while web scrapers extract or "scrape" data from webpages.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.

A web scraper's main purpose is to extract data from webpages. Web scrapers often have the ability to browse to different pages and follow links. Though web scrapers can crawl to different pages their primary purpose is scraping the data on those pages, not indexing the web.

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

**Answer**: 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. The REP also includes directives like meta robots, as well as page-, subdirectory-, or site-wide instructions for how search engines should treat links (such as “follow” or “nofollow”).

If you don’t want certain pages or files to be listed by Google and other search engines, you need to block them using your robots.txt file.

Robots.txt files are useful if you want search engines not to index:

\* Duplicate or broken pages on your website

\* Internal search results pages

\* Certain areas of your website or an entire domain

\* Certain files on your website such as images and PDFs

\* Login pages

\* Staging websites for developers

\* Your XML sitemap

1. What are static and dynamic web pages?

**Answer**: Static Web pages:

Static web pages are made of “fixed code,” and unless the site developer makes changes, nothing will change on the page. Think of it like a brochure for a business. Static Web pages are very simple. It is written in languages such as HTML, JavaScript, CSS, etc. For static web pages when a server receives a request for a web page, then the server sends the response to the client without doing any additional process. And these web pages are seen through a web browser. In static web pages, Pages will remain the same until someone changes it manually.

Dynamic Web Pages:

Dynamic Web Pages are written in languages such as CGI, AJAX, ASP, ASP.NET, etc. In dynamic web pages, the Content of pages is different for different visitors. It takes more time to load than the static web page. Dynamic web pages are used where the information is changed frequently, for example, stock prices, weather information, etc. There’s a simple way to determine if a site is dynamic. If you can interact with it, it’s a dynamic site. For example, dynamic sites allow you to create a user profile, comment on a post, or make a reservation.

**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**:

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

title = getTitle(url)

if title == None:

return "Title could not be found"

else:

return title

print(getTitle("https://amazon.in/"))

print(getTitle("http://flipkart.com/"))

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

**Answer**:

import selenium

from selenium import webdriver as wb

from selenium.webdriver.common.keys import Keys

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

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

#Accesing the search bar

element = driver.find\_element\_by\_name("q")

element.send\_keys('fliprobo')

#Accesing the search button

element1 = driver.find\_element\_by\_class\_name('Tg7LZd').click()