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?

C) Requests

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

A) Browser based application

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

D) Spider

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

C) tag[‘id’]

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

C) lxml

1. In selenium the webdriver is used to?

D) to download any content from a webpage

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?

C) ‘a’ number of pixels horizontally

In Q10, more than one options are correct, Choose all the correct options:

1. Which of the following is(are) tags of HTML

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?

ANS11-web crawlers 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?

ANS12-A robots.txt file is a set of instructions for bots This file is included in the source files of most websites. A web crawler bot will follow the most specific set of instructions in the robots.txt file. Dynamic webpages are the pages written in some more complex language such as ASP.NET in which data is rendered after some interpretation and capacity to produce distinctive content for different calls.

1. What are static and dynamic web pages?

ANS13-Static web pages are generally simple HTML written pages which serve as response from browser to server in which all the information and data is static in nature and it does not get changed until someone changed it manually.

Dynamic webpages are the pages written in some more complex language such as ASP.NET in which data is rendered after some interpretation and capacity to produce distinctive content for different calls.

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.

Ans14-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://www.linkedin.com/"))

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

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

Ans15-from selenium import webdriver

from selenium.common.exceptions import StaleElementReferenceException

driver=webdriver.Chrome('C:\\Users\\hp\\Downloads\\chromedriver\_win32\\chromedriver.exe')

my\_page=driver.get('https://images.google.com/?gws\_rd=ssl')

search\_bar=driver.find\_elements\_by\_xpath("//div[@class='RNNXgb']")

search\_button=driver.find\_elements\_by\_xpath("//span[@class='z1asCe MZy1Rb']//svg")