**WEB SCRAPING**

## **WORKSHEET – 1**

1. B
2. C
3. A
4. D,B,C
5. B
6. A
7. D
8. D
9. C
10. A,B
11. Web scraping is basically extracting data from websites in an automated manner. It is automated because it uses bots to scrape the information or content from websites. Data scraping involves locating data and then extracting it. It does not copy and paste but directly fetches the data in a precise and accurate manner. It does not limit itself to the web; data can be scraped virtually from anywhere it is stored. It does not have to be from the Internet. It is about data and not where it is stored.

The term crawling comes from the way a spider would crawl. That’s why a web crawler is also sometimes called a spider. It’s basically an internet bot that systematically browses (read crawls) the World Wide Web, usually for the purpose of web indexing. It is used for indexing the information on the page using bots also known as crawlers. Crawling through every nook and crevice of the World Wide Web, the spider locates and retrieves the information lying in the deeper layers. Web crawlers or bots navigate through heaps of data and information and procure whatever is relevant for your project.

1. A robots.txt file is a simple text file which webmasters can create to tell web crawlers which parts of a website should be crawled and which should not. The file is stored in the main directory (root) on the server. When a crawler arrives at a website, it first reads the robots.txt file to determine which parts of the website it should crawl and which parts it should ignore, according to the so-called Robots Exclusion Standard Protocol. You don’t have to create a robots.txt file but it’s often advisable to do so.
2. In static web pages, Pages will remain same until someone changes it manually. In dynamic web pages, Content of pages are different for different visitors. Static Web Pages are simple in terms of complexity. Dynamic web pages are complicated. Static Web Page takes less time for loading than dynamic web page. Dynamic web page takes more time for loading.

14) 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("https://www.keltonglobal.com/method/omnibus-surveys/"))

print(getTitle("https://in.pinterest.com/sbudtamba/wensite-design/"))

15) import requests

from bs4 import BeautifulSoup

import shutil

import re

searchs = ['Dogs','GTX','Virus']

url = 'https://images.google.com/imghp?hl=en&gl=ar&gws\_rd=ssl'

for search in enumerate(searchs):

payload = {'k': search}

headers = {'user-agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/80.0.3987.149 Safari/537.36'}

response = requests.get(url, headers=headers, params=payload)

soup = BeautifulSoup(response.text, 'html.parser')

print(soup)