**WEB SCRAPING**

## **WORKSHEET – 1**

Answers:

MCQs:

1. B

2. C

3. A

4. D

5. B

6. C

7. C

8. C

9. D

10. A,B,C

11. Web scraping is the art of getting data from a website or downloading it.It can be done manually.Tool used is web scraper.It need not to visit all the pages of a website for information.It can be done both in small and large scale.eg ProWebScraper.It doesn’t obeys robots.txt file in most of the cases.

Web crawling is the art of looking up for data and indexing it.Tool used is spider bot or web crawler.It visits each and every page until the last line of information.It is mostly employed in large scale.eg.Google,Yahoo.It always obeys robots.txt.

12. It is file which is used as a mode of communication between the website developer and the web crawler.It contains standard information about which part of the website should be accessed or the parts which are probhited.Some web crawlers don’t follow robots.txt but there are some which follows like google,yahoo,etc.

It is very essential from the developer point of view as he can hide the files which are of no use or irrelevant and are of no use to the user.It helps in making the websites information accurate and add points to the website and enhances the work of Search engine optimistaion(SOE).It is also essential in fields where some private information are need to be hidden.

13. Static web pages are pages in which data do not change or remains static in nature until someone changes it manually.Static pages are built using HTML,CSS,Javascript.The site pages remains same for every user.eg.form validation page.

Dynamic web pages are lively in nature where the data changes from time to time or depending upon user to user.These pages are built using server side language such as PHP,AJAX.eg. sales stock pages,filpkart shopping pages.The site pages are different depending upon users choice.

14. <https://github.com/avinash75/web_page-_conatins_title_or_not.git>

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("http://www.example.com/")

if title == None:

print( "Title could not be found")

else:

print(title)