**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?

C) Requests

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

A) Browser based applications

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?

C) execute tests on HtmlUnit browser

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

B) the url of first webelement 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?

Though sometimes the two terms are used interchangeably the main difference is that web crawlers usually focus on indexing the web while web scrapers extract or "scrape" data from webpages.

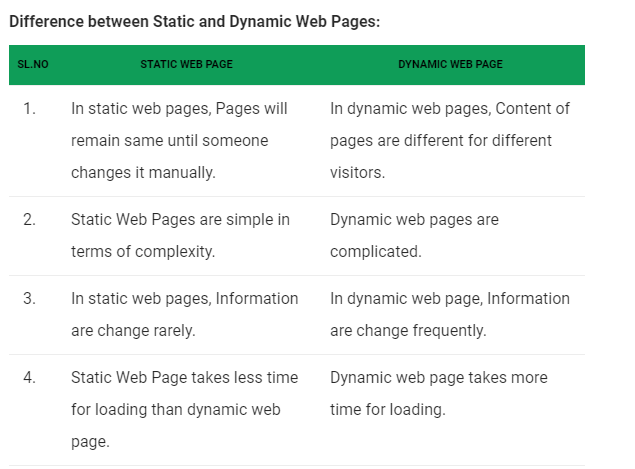
There is a fair bit of overlap between the web crawlers and web scrapers. Web crawlers work by browsing to a series of webpages and analyzing their contents for links to other webpages. The links to the other webpages are then followed and searched for more links. The process of following and recording 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.

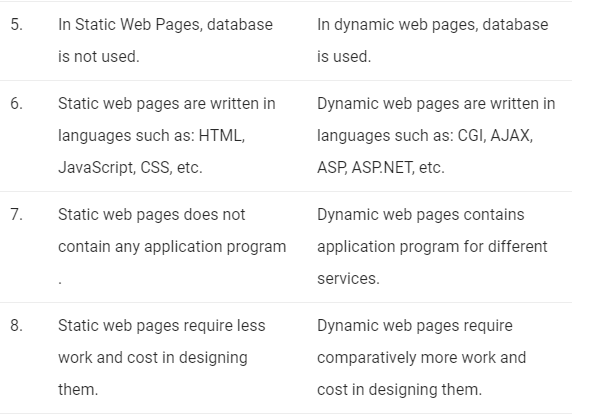
A web scraper's main purpose is to extract data from webpages. Web scrapers often have the ability to browse to different pages and follow links. Though web scrapers can crawl to different pages their primary purpose is scraping the data on those pages, not indexing the web.

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

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.**

1. What are static and dynamic web pages?





**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.



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

