**WEB SCRAPING**

## **WORKSHEET – 1**

**In Q1 to Q9, only one option is correct, Choose the correct option.**

1. B) Web scraping
2. C) Requests
3. A) Browser based applications
4. D) Spider
5. B) tag.name
6. A) HTML. Parser
7. C) execute tests on HTML Unit browser
8. C) the list of all web elements associated with the ‘given xpath’
9. D) **‘a’** number of pixels vertically

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

10. A) <a>

B) <b>

**Q11 to Q13 are subjective answer type questions, Answer them briefly.**

**11.** What is the main difference between a web scraper and a web crawler?

Ans.

**1. Webscraping :**  
Web Scraping is a technique used to extract a large amount of data from websites and then saving it to the local machine in the form of Xml excel or SQL. The tools used for web scraping are known as **web scrapers**. On the basis of the requirements given, they can extract the data from any website in a fraction of time. This automation of tasks is very helpful for developing data for machine learning and other purpose.

**2. Web Crawling :**

Web Crawling is analogous to a spider crawling but the place of crawling here is the web It basically visits a website and read web pages for the purpose of building entries for search engine index. The tools that are used for web crawling are known as **web crawlers** or **spiders**. A series of web pages are analyzed and links to the pages on them are then followed for even more links thus it does a deep search for extracting of information. Famous search engines such as Google, Yahoo and Bing do web crawling and use this information for indexing web pages. Examples are Scrapy and Apache nut.

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

Ans :- A **robots.txt** file tells crawlers which pages or files the crawler can or can't request from your site. This is used mainly to avoid overloading your site with requests. Robots.txt is not a mechanism of keeping your weppage out of the reach of crawlers or spiders. Robots.txt is mainly used for traffic management of a webpage.

13. What are static and dynamic web pages?

Ans:- **Static Web pages:** Static Web pages are very simple. They are 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. eg: Blogs, Documentation sites.

**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. eg: stock prices, weather information, etc.