**WEB SCRAPING**

## **WORKSHEET – 1**

## **Solutions**

1. B) Web scraping
2. C) Requests
3. A) Browser based applications
4. B) Crawler C) Boat D) Spider
5. B) tag.name
6. C) lxml
7. D) to download any content from a webpage
8. C) the list of all web elements associated with the ‘given xpath’
9. D) **‘a’** number of pixels vertically
10. A) <a> B) <b>

[**Web Scraping**](https://www.geeksforgeeks.org/introduction-to-web-scraping/) :

1) It is used for **downloading** information .

2) It **need not visit all the pages** of website for infomation.

3) A Web Scraper **doesn’t obey robots.txt** in most of the cases.

4) Data de-duplication is not necessarily a part of Web Scraping.

**Web Crawling** :

1. It is used for **indexing** of Web pages .
2. It **visits each and every page**, until the last line for information.
3. It always **obeys robots.txt**.
4. Data de-duplication is and integral part of Web Scraping.
5. A **robots.txt** file tells search engine crawlers which pages or files the crawler can or can’t request from your site. A search engine bot (like Googlebot) will read the robots.txt file prior to crawling your site to learn what pages it should deal with.

The use of **‘robots.txt’** file is to block bots, spiders, and other crawlers from accessing any site’s pages. When bots, spiders and other crawlers hit a site, they can utilize extensive resource (memory and CPU) usage and can also lead to high load on the server and slow down a site. Robots.txt files are used to limit server load from bots, spiders, and other crawlers, and they also tell search engines what content on your site they should and should not index.

1. **Static webpage** - Contents don't change unless manually edited by someone.

(uses file extensions .html, .htm etc. for its pages)

**Dynamic webpage**- Contents update and change on their own without human intervention.

(uses file extensions .php, .jsp .asp etc.)

**14)**



**15)**

