1.B) Web scraping

2.C)Scrapy

3.A) Browser based applications

4.D) to download any content from a webpage

5.B)tag.name

6.C)lxml

7.D) download any content from a webpage

8. C) the list of all webelements associated with the ‘given xpath’

9. D) **‘a’** number of pixels vertically

10.A)<a> and B)<b>

**Ans.11**. 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**.

**So we can conclude that web crawling is basically an internet bot that systematically browses (read crawls) the World Wide Web, usually for the purpose of web indexing,whereas Web scraping is basically extracting data from websites in an automated manner**.

**Ans.12. A robots.txt file tells search engine 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 files control crawler access to certain areas of your site. While this can be very dangerous if you accidentally disallow Googlebot from crawling your entire site (!!), there are some situations in which a robots.txt file can be very handy.

**Some common use cases include**:

1.Preventing duplicate content from appearing in SERPs (note that meta robots is often a better choice for this)

2.Keeping entire sections of a website private (for instance, your engineering team’s staging site)

3.Keeping internal search results pages from showing up on a public SERP

4.Specifying the location of sitemap(s)

5.Preventing search engines from indexing certain files on your website (images, PDFs, etc.)

6.Specifying a crawl delay in order to prevent your servers from being overloaded when crawlers load multiple pieces of content at once

**Ans 13**.1. **Static web pages are generally simple HTML written pages** which serve as response from browser to server in which all the information and data is static in nature and it does not get changed until someone changed it manually. **On other hand Dynamic webpages are the pages written in some more complex language such as ASP.NET** in which data is rendered after some interpretation and capacity to produce distinctive content for different calls

2. As mentioned in above point as **data in static web pages is static and do not require any interpretation before rendering so static web pages are simple in complexity**. **Dynamic web pages on other hand does the interpretation process which make data dynamic in nature** and due to which **dynamic web pages become complex in complexity** as compare to static web pages

3**. Static web pages are generally written in simpler languages such as HTML, JavaScript, CSS**, etc. **On other Dynamic web pages are written** **in** more complex languages such as **CGI, AJAX, ASP, ASP.NET,** etc.

4. **For static web pages data do not changes until someone changes it manually and hence data is static in nature**. On other hand for **Dynamic web page** data is first interoperate at server side and due to which **it does not remain same on every call and this makes data dynamic in nature**

5. **Static web pages due to static data take less time to get load.** **While Dynamic web pages due to dynamic data take comparatively more time** as compare to static web pages

6. **In Static web pages generally no involvement of database for data redecoration**. On other hand in case of **Dynamic web page database is used for data redecoration**.

**Ans14**. **INPUT**

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(url)

if title == None:

return "Title could not be found"

else:

return title

print(getTitle("https://www.samsung.com/in/"))

print(getTitle("https://www.photoshop.com/en"))

**OUTPUT**

None

<h1 class="adobe-hero\_\_title">Photoshop.</h1>

For Samsung page it says none,and for adobe page it shows the title.

**Ans15**. **INPUT**

try:

from googlesearch import search

except ImportError:

print("No module named 'google' found")

# to search

query = "images.google.com"

for j in search(query, tld="co.in", num=10, stop=10, pause=2):

print(j)

**OUTPUT**

<https://www.needpix.com/photo/85661/google-images-image-search-seo-search-engine-media-search-logo-google-inc-google>

<https://en.wikipedia.org/wiki/Google_Images>

<https://www.pcmag.com/how-to/how-to-do-a-reverse-image-search-from-your-phone>

<https://www.popularmechanics.com/technology/a25226475/how-to-do-a-reverse-image-search/>

<https://www.republicworld.com/technology-news/apps/how-to-search-by-image-on-google.html>

<https://www.androidauthority.com/how-to-google-reverse-image-search-800391/>

<https://thenextweb.com/basics/2020/02/17/use-google-image-search-on-iphone/>

<https://www.businessinsider.com/how-to-search-an-image-on-google-phone-computer-reverse>

<https://www.youtube.com/watch?v=opc7aMJszng>

<https://tineye.com/>