Web Scrapping

Web Scrapping also called as web data mining or web harvesting is the technique of constructing an agent which can extract, parse, download and organize useful information from the web automatically.

In [ ]:
```
1
```

Web Crawling vs Web Scrapping

```
**web crawling**                                    **Web Scrapping**


1) Refers to downloading and storing                1) Refers to extracting individual data elements
from the
   the contents of a large number of                   website by using a specific site structure.
   website.

2) It's mostly done on a large scale                2) It can be implemented at any scale


3) It gives generalized information                 3) It gives specific information.


4) It is typically used by search-engines           4) It is typically used by any size companies and
mostly
   like Google,Yahoo,Microsoft                         used in the process of data analytics(data
acquiring).
```

In [ ]:
```
1
```

In [ ]:
```
1
```

working of web scraper

Web scraper can be defined as a software(program) script used to download the contents of multiple pages and extracting data from it.

In [ ]:
```
1
```

1) Visting the Website : The scrapper will visit the website

2) Downloading the contents : A web scrapper will download the requsted content from multiple pages

3) Extracting the Data : The data on the website is typically in HTML and mostly not in a structured format. Hence in this step web scrapper will parse and extract structured data from the downloaded content.

4) Storing the Data : Here a web scrapper will store and save the extracted data in any format (CSV, JSON, DB,PDF,HTML)

5) Analyzing the Data: After all the steps are successfully done, the scrapper will analyze the data

In [ ]:
```
1
```

In [ ]:
```
1
```

requests

pip install request

In [1]:
```
1 import requests
```

In [10]:
```
1 data = requests.get('https://www.amazon.in/')
```

In [11]:
```
1 data
```

Out[11]:
```
<Response [200]>
```

In [12]:
```
1 data.text
```

```
11rRjDLdAVL.js,51H19hJRYrL.js,11kWu3cNjYL.js,11g2BPXNtrL.js,11URENu1epL.js,11wcWdhrhDL.js,21r53SJg7LL.js,
0190vxtlzcL.js,511VNbag2QL.js,31NShmNbJyL.js,01Gf12ogmOL.js,01ezj5Rkz1L.js,11+RxVdhNcL.js,31o2NGTXThL.js,
01rpauTep4L.js,01rvjRKHRNL.js_.js?AUIClients/AmazonUI&gOUTNJqP#page_type-Gateway.372963-T1.354901-T1.3514
11-T1\');\n  (window.AmazonUIPageJS ? AmazonUIPageJS : P).load.js(\'https://images-eu.ssl-images-amazon.c
om/images/I/51xaFbd-18L.js?AUIClients/CardJsRuntimeBuzzCopyBuild\');\n});\n</script>\n<!-- sp:end-featur
e:aui-assets -->\n<!-- sp:feature:nav-inline-css -->\n<!-- NAVYAAN CSS -->\n\n<style type="text/css">\n.n
av-sprite-v1 .nav-sprite, .nav-sprite-v1 .nav-icon {\n  background-image: url(https://images-eu.ssl-image
s-amazon.com/images/G/31/gno/sprites/nav-sprite-global-1x-hm-dsk-reorg._CB405936311_.png);\n  background-
position: 0 1000px;\n  background-repeat: repeat-x;\n}\n.nav-spinner {\n  background-image: url(https://i
mages-eu.ssl-images-amazon.com/images/G/31/javascripts/lib/popover/images/snake._CB485935600_.gif);\n  ba
ckground-position: center center;\n  background-repeat: no-repeat;\n}\n.nav-timeline-icon, .nav-access-im
age, .nav-timeline-prime-icon {\n  background-image: url(https://images-eu.ssl-images-amazon.com/images/
G/31/gno/sprites/timeline_sprite_1x._CB439943932_.png);\n  background-repeat: no-repeat;\n}\n</style>\n<l
ink rel="stylesheet" href="https://images-eu.ssl-images-amazon.com/images/I/41KBYOkTjIL._RC|71NRoBcc4sL.c
ss,31g1ibO5ySL.css,31CdpXAsWCL.css,313Ydl5aIRL.css,21MKjoYL8wL.css,41OiMQkB+EL.css,01yCq3WXEcL.css,11kO7y
AgiQL.css,31OvHRW+XiL.css,01XHMOHpK1L.css,01ucgi+I44L.css,31jJwAF+yLL.css_.css?AUIClients/NavDesktopUberA
sset&eRSuagD1#desktop.in.310484-T1" />\n<!-- sp:end-feature:nav-inline-css -->\n<!-- sp:feature:host-asse
ts -->\n<style>\n#gw-desktop-herotator,#gw-desktop-herotator .a-carousel-viewport{height:300px}#gw-deskto
p-herotator.tall{z-index:0}#gw-desktop-herotator.tall,#gw-desktop-herotator.tall .a-carousel-controls{max
-height:230px}#gw-desktop-herotator.tall .a-carousel-viewport{height:auto!important}#gw-desktop-herotato
```

In [ ]:
```
1
```