Web Scraping

using Python

What is Webscraping ?

- Web scraping is a automated way of extracting information from websites. This technique mostly focuses on the transformation of unstructured/semi-structured data (HTML format) on the web into structured data (database or spreadsheet).

Websites

Webscrapper

XML

TXT

Why we use Web Scraping ?

- Lead generation

- Monitoring prices on e-commerce sites

- Retrieving social profiles

- Scraping blogs & forums to keep up with trends and news

- Maintaining up-to-date real estate database

- Aggregating favourites from various media services (Delicio.us, Xmarks, Chrome bookmarks, Youtube, Spotify, 9gag, etc.)

- Scraping deep-web (Tor)

How can we use Web Scraping ?

- We can perform web scraping in various ways.

- We would resort to Python because of its ease and rich eocsystem.

- It has a library known as 'BeautifulSoup' which assists this task

- Some dynamic websites will need different other packages (e.g: Selnium)

- Some websites block automated scraping through several ways
  - Captcha
  - Blocking IPs which hit the website frequently