**Web Scraping**

# What is Web Scraping

Web Scraping is an automatic method to obtain large amounts of data from websites.

Web Scraping extracts underlying HTML code and, with it, data stored in a database. The scraper can then replicate entire website content elsewhere.

Structured Data Unstructured Data

Excel Spreadsheet Database video , voice image PDF

Most of this data is unstructured data format which is video, voice and image or structured data in a spreadsheet or a database.

Also used for converting unstructured data to structured data.

# Different Types of Web Scrapers

* Self-built web scrapers
* Pre-built web scrapers
* Browser extensions web scrapers
* Software web scraper
* Cloud web scraper

# Python Language for Web Scraping

Scrapy BeautifulSoup Selenium

# Components of Web Scraping

# Legality of Web Scraping

Web Scraping is technically not any kind of illegal process:

* You can avoid any trouble if you consider below points:
* Consider **robots.txt** file before scraping any website.
* Give priority to API if provided.
* Scraping the website with slow rate of requests to fetch the data.
* Follow the rules and standards of the website.

# Get Requests Web Page using Python

import requests

web = requests.get("https://www.icc-cricket.com/")

print(web) output = response 200 read page 🡪 good , 300 redirect like move another page 🡪 average , 400 not read like working on this site 🡪 bad.

web.content

web.url

web.status\_code

# Parsing the Html in web page using python

from bs4 import BeautifulSoup

soup = BeautifulSoup(web.content, "html.parser") # "lxml", "lxml-xml", "html5lib"

print(soup.prettify())

soup.title

soup.title.name

soup.p, soup.a, soup.h2

# Kinds of objects BeautifulSoup in Web Scraping

## Tag

tag = soup.html

type(tag)

tag = soup.p

tag

tag = soup.a

tag

tag = soup.h1

tag

## NavigableString

tag = soup.p.string

print(tag)

tag = soup.a.string

print(tag)

tag = soup.h1.string

## BeautifulSoup

soup.name

soup.title

soup.head

print(soup.head.prettify())

soup.find('h1')

soup.find\_all('h1')

## Comments

com = soup.p.string

com

# Finding Elements in Web Page using Python

## Finding Elements by Class

class\_data = soup.find("div", class\_ = "container flex items-center h-full")

print(class\_data)

class\_data.find\_all('p')

## Finding Elements by id

id\_data = soup.find('nav', id = "nav")

print(id\_data)

id\_data.find\_all('p')

# Extracting Text from the Tags in Web Page using Python

web1 = requests.get("https://www.tutorialsfreak.com/")

soup1 = BeautifulSoup(web1.content, "html.parser")

print(soup1.prettify())

lines = soup1.find\_all("p")

for l in lines:

print(l.text)

s = soup1.find('div', class\_='d-md-flex align-items-md-center why-choose-left-section')

lines\_1 = s.find\_all('p')

for l1 in lines\_1:

print(l1.text)

# Extracting Links in Web Page using Python

soup1.find\_all('a')

for a in soup1.find\_all("a"):

print(a.get("href"))

# Extracting Image Information in Web Page using Python

soup1.find\_all("img")

for i in soup1.find\_all("img"):

print(i.get("src"))

for i in soup1.find\_all("img"):

print(i.get("alt"))