DATA SCRAPING

Gathering Data from Websites

Presented by **Munyala Eliud.**

You can reach out and see my work below

LinkedIn: https://www.linkedin.com/in/eliud-munyala/
Github: https://github.com/meaLuda
Project Link: https://github.com/meaLuda/DE-WebScrapping-Presentation

Web Scraping - What We'll Cover

1. A brief hands-on introduction into HTML parsing
2. A brief discussion on the ethics of scraping
3. Live scraping session with Scapy

1. A brief hands-on introduction into HTML parsing

HTML

HyperText Markup Language

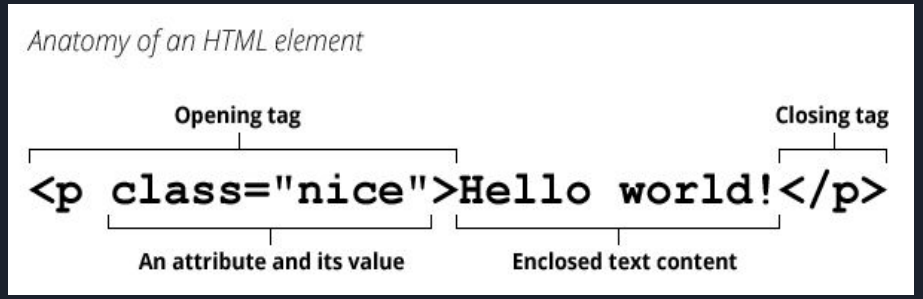
The standard markup language on the Web

As the web evolves so does the proliferation of technical wrappers surrounding the visible content of websites (text and data)

```
<!DOCTYPE html>
<html>
<head>
    <title> My First Page </title>
</head>
<body>
    <p> Welcome to <em> Simplilearn!! </em></p>
    <p>
        This is the <b>HTML</b> tags <u>article.k/u>
        <a href="https://www.simplilearn.com/"> This is the link to
        Simplilearn website</a>


    </p>


</body>
</html>
```



This is not a programming workshop, but...

1. We will discuss Python, Scrapy and Pandas
2. We will not learn Python in the workshop
3. However, some automation tools are used in this workshop
4. Web Scraping is about deconstructing websites. Effective scraping requires learning about technical infrastructure as well as subject content
5. Data/Text Analysis with pandas
6. Data cleaning

Definitions

Scraping

Using tools to gather data you can see on a webpage

A wide range of web scraping techniques and tools exist. These can be as simple as copy/paste and increase in complexity to automation tools, HTML parsing, APIs and programming

But not this kind of scrapping….

Definitions...

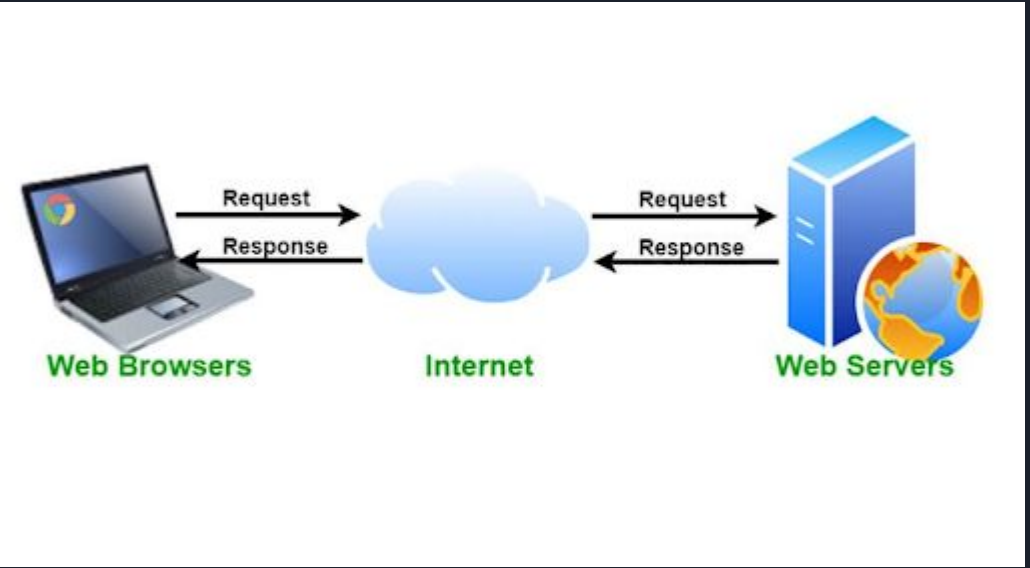
Scraping

HTTP

HyperText Transfer Protocol

Machine interchange information transported over the Internet to enable multimedia data exchange, aka WWW. The protocol defines aspects of authentication, requests, status codes, persistent connections, client/server request/response. etc.

Access a server on port 80; the declarative Document Type Definition (HTML, XML, JSON, etc.)



2. A brief discussion on the ethics of scraping

The concept of web scraping is surrounded by inherent paradoxes around legal and ethical aspects.

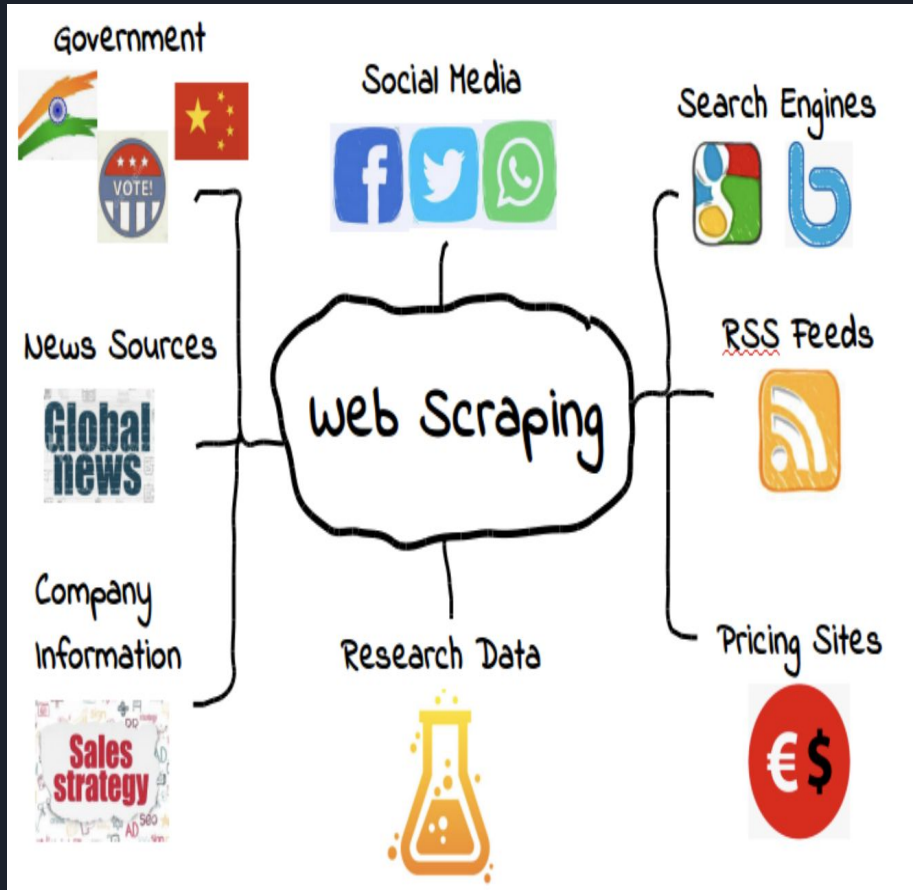
On one end, openness and sharing is supposed to help the public at large. On the other hand, website data could be viewed as proprietary, private assets needing protection by their owners.

There are also questions on who owns the data presented in the website and is it truly the website owner.

The use of web scraped data poses ethical questions about research ethics including consent, privacy, anonymity, trust, and transparency. Individual's confidential and company sensitive information could be exposed.

The methods used for storing harvested data may not be in compliance. API's like Dice, LinkedIn, Facebook, Twitter & Craigslist could be bypassed to retrieve more than approved volumes of information and use it with malicious intent.

src



Web Scraping project Scenario

HOUSE HUNTING

You have now reached an aged that you want to settle down and create a family.

Having done a few good life choices and investment you have some money in the bank that will enable you to buy a good house to raise your kids in.

Having this in mind can you find a way to get information about houses in your area in terms of location, house description, number of rooms and price to help you make an informed decision before going > kwa ground.

https://www.buyrentkenya.com/houses-for-sale

4. Live scraping session with Scapy