PES UNIVERSITY ONLINE

STATISTICS FOR DATA SCIENCE

Web Scraping

D. Uma
Computer Science and Engineering
umaprabha@pes.edu

STATISTICS FOR DATA SCIENCE

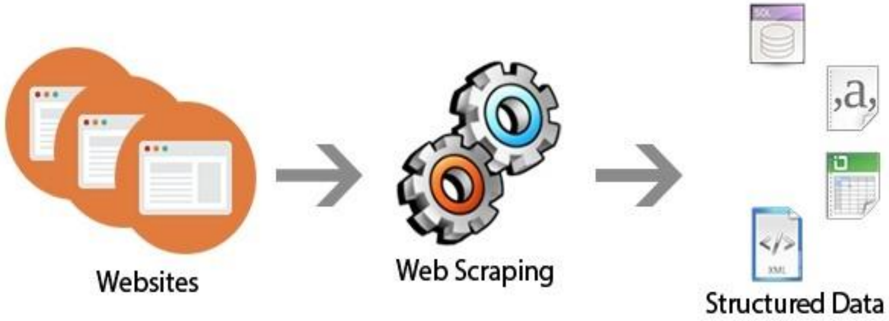
Web scraping

D. Uma

Department of Computer Science and Engineering

What is Web Scraping?

- **Web scraping** is the **process** of **gathering information** from the **Internet**.



Source: realpython
Image Source: Quora.com

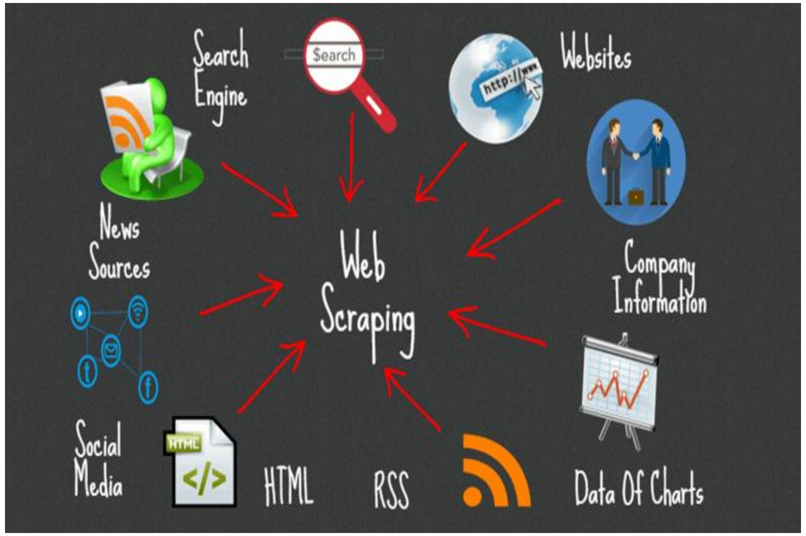
Extraction of Data from Websites

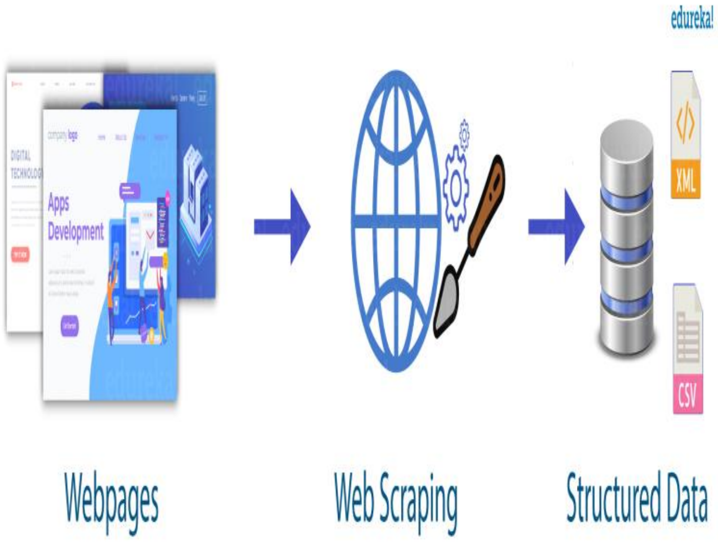
- **Use the API** of the website (if it exists).
- **Access the HTML of the webpage** and extract useful information/data from it.

Source:DataCamp

Why Web Scraping?

Web Scraping is the technique of automating this process.





Source: https://www.webharvy.com/,
Image: octoparse.com,Edureka.com

What is done with Web Scraping!!!!!

- Web scraping is typically used for change detection, market research, data monitoring, and in some cases, theft.
- **Testing** for **broken links** and **images** within each page.
- For **unlawful purposes**, such as copying a website and republishing it under a different name.
- Web scraping is considered **malicious** when data is extracted without the permission of website owners.

Source:scrapingexpert.com

Steps involved in Scraping

The web scraping process follows the below 3 steps.

1. Request-Response
2. Parse and Extract
3. Transform the data

Third-party python library called **Beautiful Soup** is used for pulling data out of HTML and XML files.

Source:RealPython,Image: Digitalocean.com

Is it Legal to Scrape?

- **Not all web scraping** acts are considered as **legal**.
- Python Web scraping services that extract **publicly available data are legal.**
- **Check robots.txt** – displays the pages that can be scraped.

Source: intellipad.com, Image: promptcloud.com

Applications

- Search Engines
- Price Monitoring
- Sales and Marketing
- Content Aggregators
- Sales intelligence
- Training datasets for Machine Learning
- Data for Research

How is Web Scraping Done??

- Web Scraping can be done efficiently using **Python** because it is **flexible** and **powerful.**

- Includes **Python libraries** like **request** and **BeautifulSoup4** which helps us to **fetch URL** and pull out information from web pages.

- In addition, **re**, **numpy** and **pandas** could help us **clean** and **process** the data.

Source: https://www.octoparse.com/

Web Scraping Demonstration

Let’s do it practically.

WEB SCRAPING DEMONSTRATION : AMAZON PRODUCT REVIEWS.docx

url: https://www.amazon.in/Apple-iPhone-XR-64GB-Black/product-reviews/B07JWV47JW

Source: https://www.octoparse.com/

Problem Statement

Do It Yourself!!!!

- With same url: https://www.amazon.in/Apple-iPhone-XR-64GB-Black/product-reviews/B07JWV47JW

- **Scrape the Rating, Review Content and Display the result.**

THANK YOU

D. Uma

Department of Computer Science and Engineering

umaprabha@pes.edu

+91 99 7251 5335