رواد مصر الرقمية

وزارة الاتصـــــالات
وتكنولوجيا المعلومات

Python Project for Data Science

17/04/2024

Web Scraping

Course Overview

You will perform specific data science and data analytics tasks such as extracting data, web scraping, visualizing data and creating a dashboard. This project will showcase your proficiency with Python and using libraries such as Pandas and Beautiful Soup within a Jupyter Notebook. Upon completion you will have an impressive project to add to your job portfolio.

Session Content

- What is Web Scraping?
- How do we do?
- Tools to use
- Ethics for scraping
- Demo

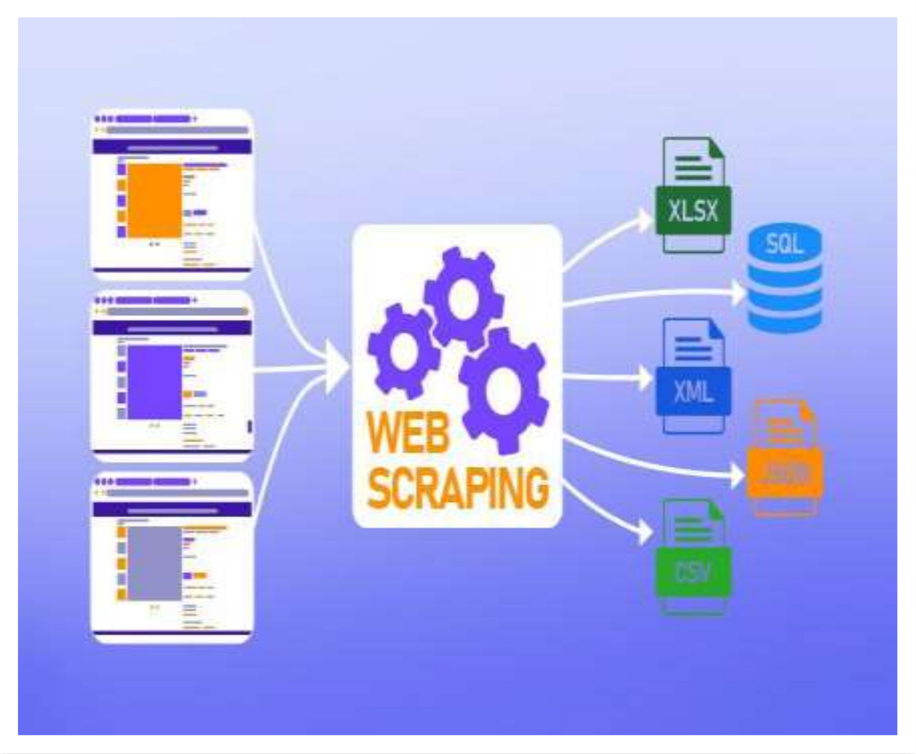
What is Web Scraping?

- **Web scraping:** is technique for gathering data or information on web pages.

- **Web scraping:** is method to extract data from a website that does not have an API, or we want to extract LOT of data which we can not do through an API due to rate limiting.

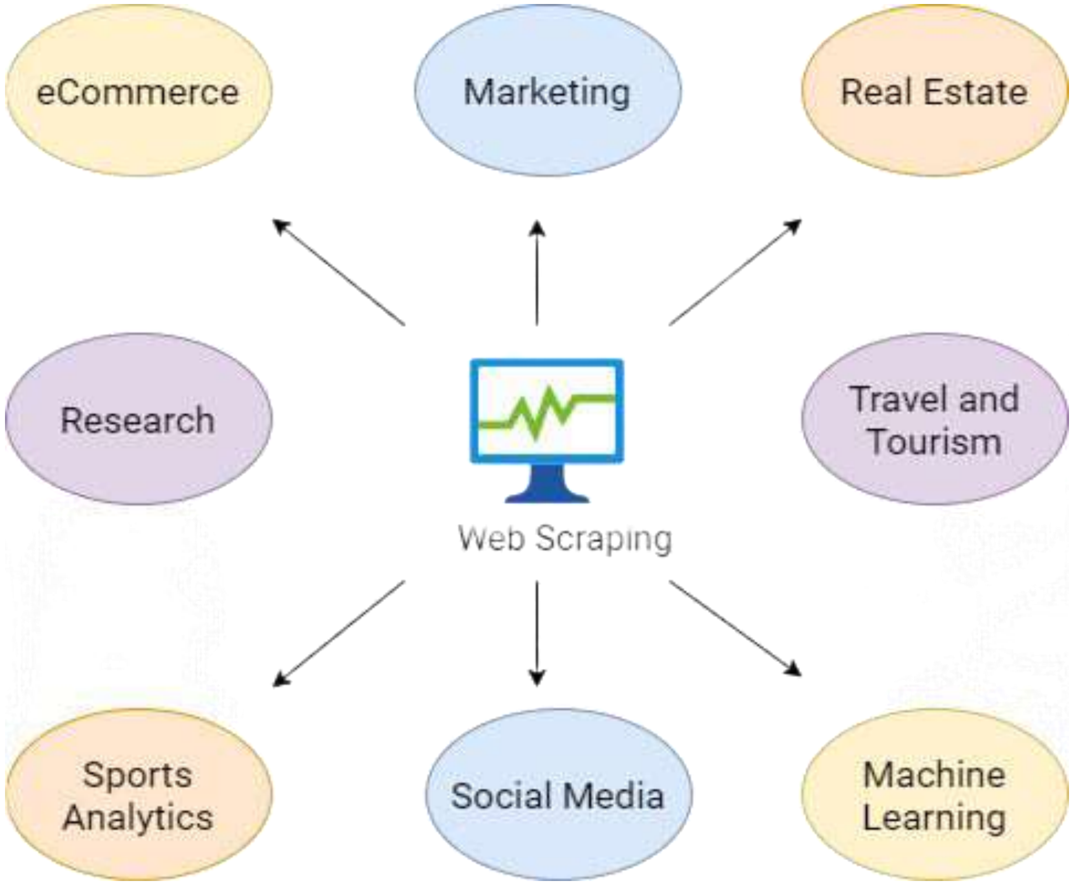
- Through web scraping we can extract any data which we can see while browsing the web.

- You could revisit your favorite website every time it updates for new information, Or you could write a web scraper to have it



Web Scraping in Real Life

- Extract products information
- Extract job posting and internships
- Extract offers and discount from deal of the day website
- Extract date to make search engine
- Gathering weather data
- Etc.

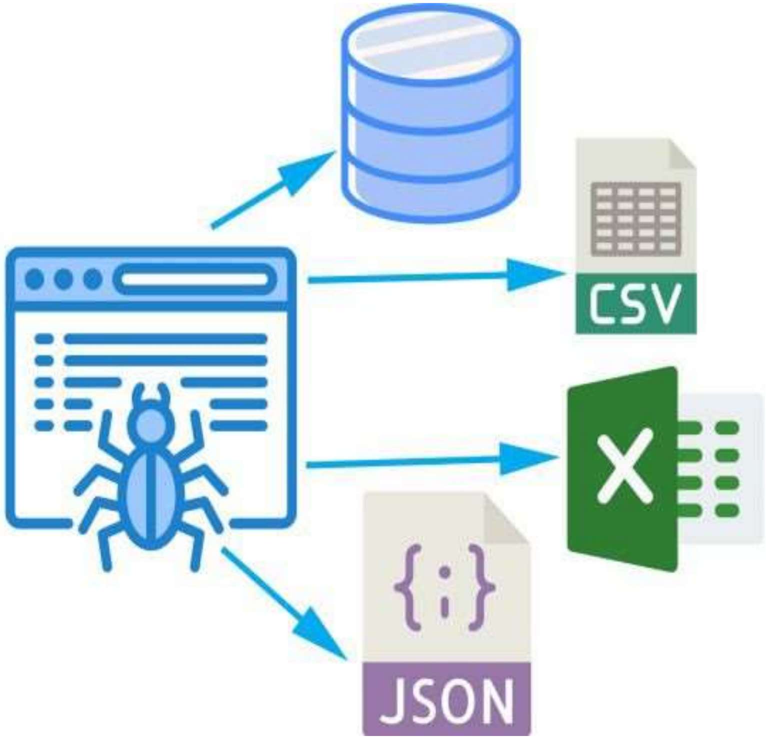


Advanced Web Scraping Vs. API

- Web scraping is not rate limited
- Anonymously access the website and gather data
- Some website don’t have API
- Some data is not accessible through an API

Workflow

- Web scraping follows this workflow:
- Get the website – using HTTP library
- Parse the html document – using any parsing library
- Store the results – either a db , csv, txt file etc.



Libraries

- BeautifulSoup (bs4)
- Lxml
- Selenium
- Re
- scrapy

Scrapy
Selenium
Se
BeautifulSoup
CSV
Web Scraping

Is Web Scraping Legal?

In short, the action of web scraping is not illegal. However, some rules need to be followed. Web scraping is illegal when non-publicly available data is extracted.

Demo 1
Beautiful Soup

Demo 2
Selenium

Questions & Answers

Thank you!