DATA ANALYSIS AND VISUALIZATION

INSTRUCTOR: UMME AMMARAH

WEB SCRAPPING

INTRODUCTION

WHAT IS WEB SCRAPING

- The process of gathering information from the Internet, usually refer to a process that involves automation.

USES

- Financial Data Analysis
- Marketing and Sales
- Academic
- Journalism
- Real Estate
- Machine Learning
- Brand Monitoring and Competition Analysis
- Social Media Analysis

WEB CRAWLING VS WEB SCRAPING

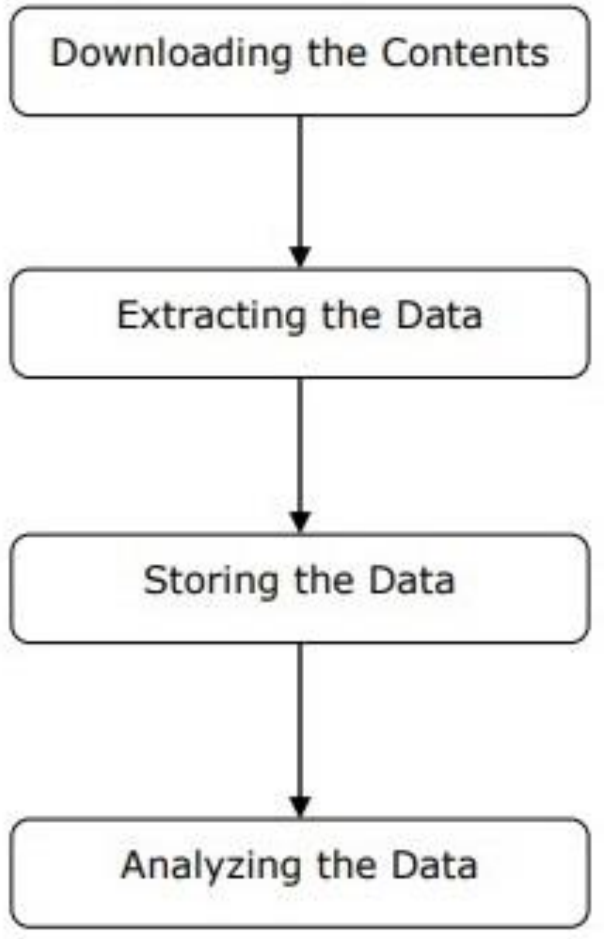
- Used to index the information on the page using bots called crawlers.
- Downloading and storing the contents of a large number of websites.
- Mostly done on large scale.
- Yields generic information.
- Google, Bing, Yahoo. **Googlebot** is an example of a web crawler.

- Automated way of extracting the information using bots called scrapers
- Extracting individual data elements from the website by using a site-specific structure.
- Can be implemented at any scale.
- Yields specific information.
- Can be used for data analysis

COMPONENT OF WEB SCRAPER

- Web Crawler Module
- Extractor (Parser)
- Data Transformation and Cleaning Module
- Storage Module

WORKING OF A WEB SCRAPER



Downloading the Contents

Extracting the Data

Storing the Data

Analyzing the Data

CHALLENGES OF WEB SCRAPING

- Variety
- Durability

ALTERNATE TO WEB SCRAPING

Application Programming Interfaces (APIs)

Some websites allows to access their data in a predefined manner. With APIs, parsing HTML can be avoided. Instead, the data can be accessed directly using formats like JSON and XML.

LET’S GET STARTED…

INSTALLATIONS

- Install python

OR

- Any IDE (PyCharm, Spyder)

OR

- Anaconda

INSTALLING PYTHON LIBRARIES

- Requests (*pip install requests*)
- BeautifulSoup (*pip install beautifulsoup4*)

Run these command on anaconda prompt or cmd to install both libraries.

STEPS OF WEB SCRAPING

- Inspect your data source
    - Explore website
    - Decrypt url
    - Use developers tool
- Scrap content
    - Request library
    - Resquest-html
    - Selenium
- Parse content
    - beautifulSoup

PRACTICE SITES

HTTPS://WEBSCRAPER.IO/TEST-SITES