DS&AI 2023
DATA SCIENCE AND ARTIFICIAL INTELLIGENCE CONFERENCE 2023
1ST - 3RD FEBRUARY 2023

Effective Web Scraping for Data Science

Victor Ashioya

Sponsored by Artificial Intelligence www.elsevier.com/locate/artint JOURNAL
Google TensorFlow

Background

- Web scraping, also known as web data extraction or web harvesting, is the process of automatically extracting large amounts of data from websites. This data can then be used for a variety of purposes, such as creating a database, analyzing trends, or even using it for machine learning.
- Web scraping is commonly performed using programming languages such as Python or Java, and there are many libraries and frameworks available to make the process easier, such as Scrapy, Beautiful Soup, and Selenium.
- Web scraping is a powerful tool, but it is important to be aware of the legal and ethical implications of using it. Many websites have terms of service that prohibit the use of web scraping, and it is important to respect these terms. Overall, web scraping can be a valuable tool for extracting data from the web, but it should be used responsibly and with respect for the rights of website owners.

DS&AI 2023

Problem

Analysis of African Energy Resources

Study / Project Objectives

- To map Africa's energy resources against population to aid in planning energy consumption

Background Literature

- "Web Scraping: The Art of Collecting Data from the Web" by Ryan Mitchell
- "Web Scraping: A Practical Guide" by David C. Vergnaud
- "Web scraping and crawling are perfectly legal – as long as you don't break the terms of service" by Mike James, a legal expert explains the legal aspects of web scraping and the differences between scraping and crawling
- General Data Protection Regulation (GDPR) and the Computer Fraud and Abuse Act (CFAA) in the United States.

Methodology

- Identify the target website and the specific information to be extracted. This includes determining the URLs of the pages to be scraped and identifying the HTML elements that contain the desired information.
- Inspect the website's structure and HTML code to understand how the data is organized and how to locate the specific elements that contain the desired information.
- Choose a programming language and scraping library or framework. In Python libraries such as Beautiful Soup, Scrapy, and Selenium are commonly used for web scraping.
- Write the scraping script, using the chosen library or framework to navigate the website and extract the desired information.
- Test the script on a sample to ensure it’s working fine.

Results / Outputs

- A CSV file that contains the following outputs;
  country name
  energy source
  population

Discussion / Implications

The goal of this project was to extract data from the CIA World Factbook on a variety of topics for African nations. The data extracted included information like climate, terrain, natural resources, population, GDP, labor force etc

In terms of the data obtained, we found that the GDP per capita was positively correlated with the Human Development Index (HDI) in most of the countries, this may indicate that the GDP is a good indicator of a country's standard of living.

Limitations of the project include the fact that the data is only as current as the date of last update on the CIA World Factbook, and also some of the data was not available for certain countries. In future work, it would be valuable to extract data on a regular basis to track changes over time and also to extract data from other sources to cross-reference.

Overall, web scraping is very useful when extracting large amounts of data

Conclusions

In conclusion, this web scraping project successfully extracted a large amount of data from the CIA World Factbook on a variety of topics for every country in the world. The data extracted provided valuable insights into the economic and social development of countries around the world. However, it was noted that some of the data was incomplete or inaccurate, highlighting the need for additional validation and quality checking of the data. Despite these limitations, the project has demonstrated the potential of web scraping as a tool for gathering and analyzing large amounts of data.

Future Work / Directions

Focus on regular data extraction and cross-referencing data from other sources to improve the accuracy and completeness of the data.

Project

https://github.com/ashioyajotham/Webscraping/tree/main/CIA%20Factbook

THANK YOU!