Webi Scrapper

Introduction

Webi Scrapping is a technique employed to extract large amounts of data from websites whereby the data is extracted and saved to a local file in your computer or to a database in table (spreadsheet) format. The data on the websites are unstructured. Web scraping helps collect these unstructured data and store it in a structured form. There are different ways to scrape websites such as online Services, APIs or writing your own code. In this article, we'll see how to implement web scraping with python.

Our Project Idea

- Design & code a web scraper in python to traverse the data in our database.
- The database used will be Mongo DB. It is classified as a NoSQL database program, MongoDB uses JSON-like documents.
- The Project will be operated in a cloud environment named Heroku. Heroku is a cloud platform as a service supporting several programming languages.
- The python code will render the data of a particular website in the database in the form of tables.
- The process will be automated which means the code will execute in a cloud environment in an interval of 6 hrs to render the data of websites.
- Data Visualizations Computations will be performed to study the data extracted from the website.

- Once this process is complete the cloud will automatically send a notification on Slack conforming the data extraction is complete.

Tools used in our project

1. Atom.io
2. Heroku Cloud
3. Slack
4. Github

Atom. Io:
Atom is a free and open-source text and source code editor for macOS, Linux, and Microsoft Windows with support for plug-ins written in Node.js, and embedded Git Control, developed by GitHub. Atom is a desktop application built using web technologies. Most of the extending packages have free software licenses and are community-built and maintained.[9] Atom is based on Electron (formerly known as Atom Shell), a framework that enables cross-platform desktop applications using Chromium and Node.js. It is written in CoffeeScript and Less.

Installing Atom.io on PC :

1. Download atom.io on your machine, Download Link
2. install some Important packages for react native development in atom.io
   File > **Settings**.
   This will bring up the Settings View.
   click on **Install** and type the name of the **package** you are planning to **install**.

Package name:

1. Ide Python, Download Link
2. Data Explorer, Download Link

Heroku Cloud:

Heroku is a cloud platform as a service supporting several programming languages. One of the first cloud platforms, Heroku has been in development since June 2007, when it supported only the Ruby programming language, but now supports Java, Node.js, Scala, Clojure, Python, PHP, and Go.

Link: Heroku Cloud

MongoDb:

MongoDB is a cross-platform document-oriented database program. Classified as a NoSQL database program, MongoDB uses JSON-like documents with optional schemas. MongoDB is developed by MongoDB Inc. and licensed under the Server Side Public License.

Link: Mongo Db

Slack:

Slack is a proprietary business communication platform developed by American software company Slack Technologies. Slack offers many IRC-style features, including persistent chat rooms organized by topic, private groups, and direct messaging.

Link: Slack Link

Github:

GitHub, Inc. is an American multinational corporation that provides hosting for software development and version control using Git. It offers the distributed version control and source code management functionality of Git, plus its own features.

Link: Github Link

What we use in python script:

1. Pip: Pip is a standard package-management system used to install and manage software packages written in Python.
   Installation Command: `python get-pip.py`

2. Pymongo : Pymongo is a Python distribution containing tools for working with MongoDB, and is the recommended way to work with MongoDB from Python.

Installation Command: python -m pip install pymongo

1. **Bs4**: It is a Python library for pulling data out of HTML and XML files. It works with your favorite parser to provide idiomatic ways of navigating, searching, and modifying the parse tree. It commonly saves programmers hours or days of work.
   Installation Command: pip install beautifulsoup4

3. **Requests**: Requests is a Python HTTP library, released under the Apache License 2.0. The goal of the project is to make HTTP requests simpler and more human-friendly.
   Installation Command: pip install requests

4. **Pymongo[srv]**: The +**srv** indicates to the client that the hostname that follows corresponds to a DNS **SRV** record. The driver or **mongo** shell will then query the DNS for the record to determine which hosts are running the mongod instances.
   Installation command: pip install pymongo[srv].

5. **Pandas**: Pandas is a high-level data manipulation tool developed by Wes McKinney. It is built on the Numpy package and its key data structure is called the DataFrame. DataFrames allow you to store and manipulate tabular data in rows of observations and columns of variables.
   Installation command: pip install pandas

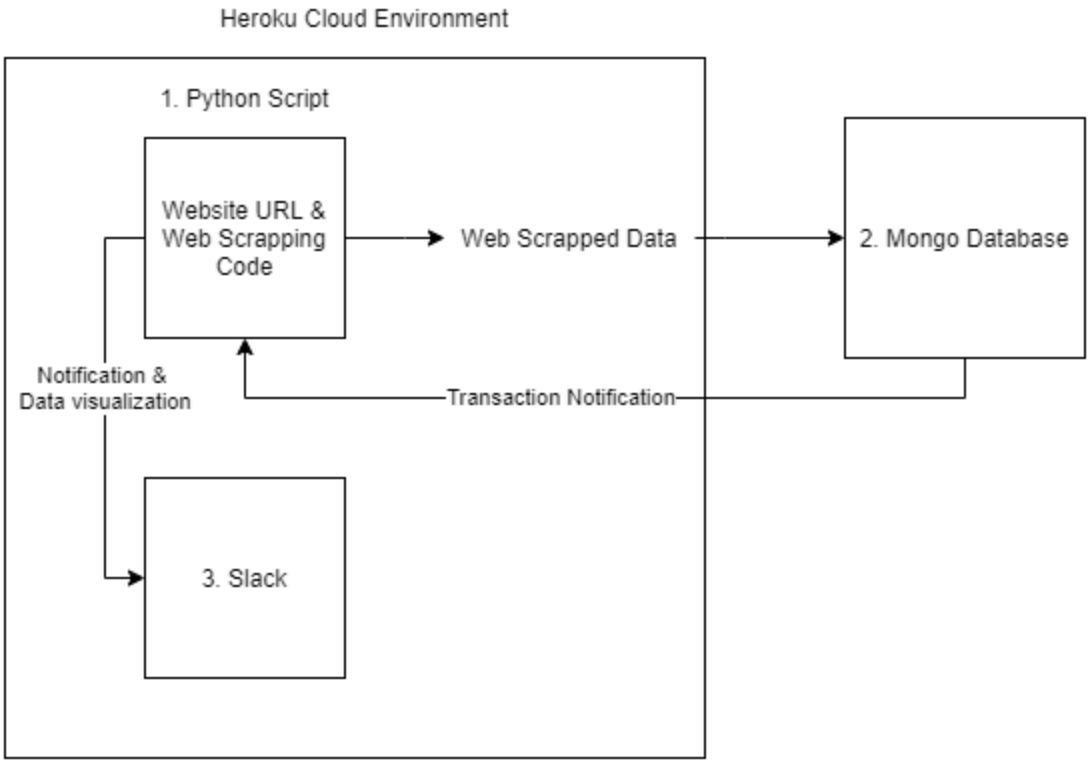
6. **Openpixel**: Openpyxl is a **Python** library to read/write Excel 2010 xlsx/xlsm/xltx/xltm files. It was born from a lack of an existing library to read/write natively from **Python** the Office Open XML format. All kudos to the PHPExcel team as **openpyxl** was initially based on PHPExcel.
   Installation command: pip install pandas
7. **Slack client**: Slack Developer Kit for Python allows you to leverage the flexibility of Python to get your project up and running as quickly as possible.
   Installation command: pip3 install slackclient

8. **Matplotlib**: Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy.
   Installation command: pip install -U matplotlib

How to start the project

1.local host: 1. Go to the project folder
2. Start Cmd
3. Type: Python without_scheduler.py

Data Flow Diagram



Benefits of Web Scraping:

1. It simplifies the process of extracting data.
2. It speeds it up by automating it and creates easy access to the scrapped data by providing it in a CSV format &JSON format.
3. Search engine bots crawling a site, analyzing its content and then ranking it.
4. Market research companies using scrapers to pull data from forums and social media.
5. Extract data from APIs.

Harmful Effects of Web Scraping:

1. Revenue loss: When your competitive advantage is impacted by third-party scrapers and competitor bots, it's quite likely that your customer base will shrink over time. Further, if you monetize your website with advertising, the drop in traffic will significantly impact ad revenue. Eventually, advertisers that partner with you may lower their bids or consider other publishers to place ads with.

2. Poor user experience: Bot traffic performing content and price scraping heavily loads your server infrastructure and slows down page loads and user access to APIs that carry out inventory availability checks, user authentication, location mapping, shopping carts and payment processing. Moreover, scraper bots fill shopping carts and abandon them, rendering products unavailable to genuine users.

3. Form spam and fake leads: Bots are capable of filling web forms with fake data, and it can be difficult to differentiate between actual leads and spam leads.

4. Drop in SEO rankings: Your content is your company's intellectual property of your business, and when it is scraped or misused, it harms your SEO efforts and search engine visibility. Because Google prioritizes original content, scraped content downgrades your search engine rankings, and the scraper using your content can often end up ranking higher than your business in search results.

Data used in our project:

1. NoSql format as input i.e. Mongo Db
2. CSV, EXCEL and JSON format as output.

Data storage concern in our project:

1. As the project code is executed after every 6 hours so the stored data space will increase from time to time.
2. The bigger the data the bigger storage we need.
3. As the database is NoSql and on cloud so it's paid.

Data privacy concern in our project:

1. The project is open source so the data extracted can be accessed by anyone over the internet.
2. Some potential information can be leaked.

Data security concern in our project:

1. Since it’s an open source so security concern can take place anytime.
2. It may attract bad networks.
3. User’s personal information can be accessed and breached.

Online References

https://api.mongodb.com/python/current/installation.html

https://stackoverflow.com/questions/52930341/pymongo-mongodbsrv-dnspython-must-be-installed-error

https://pypi.org/project/slackclient/

https://pypi.org/project/bs4/

https://pypi.org/project/pip/

https://towardsdatascience.com/a-quick-introduction-to-the-pandas-python-library-f1b678f34673

https://www.npmjs.com/package/openpixel

https://docs.anaconda.com/anaconda/navigator/tutorials/pandas/

https://realpython.com/python-requests/

https://matplotlib.org/tutorials/introductory/pyplot.html

https://meet.google.com/linkredirect?authuser=0&dest=https%3A%2F%2Fdevcenter.heroku.com%2Fcategories%2Freference

https://devcenter.heroku.com/