12-July-2019

PYTHON PROJECT REPORT

Web Scrapping and Web Development - Indeed

Submitted By:

Abhigna DC (Masters in Data Science and Analytics)

Krishna Teja KANCHERLA (Masters in Data Science and Analytics)

Lucy (Masters in Software Engineering)

ABSTRACT

Data plays an important role in the decision making of a business or any sectors. It is very essential in today's competitive world to realize the necessities of your customers and their preferences. It is a tedious task to read and analyse the content from different websites. Web scraping is a one-step solution where the user can collect and organize data from various websites. With the collected information, one may be able to predict the way market trends would work and change as per the analysis.

In this project, we have taken the indeed website which has a massive dataset. Indeed, is an American worldwide employment-related search engine for job listings launched in November 2004. It is a subsidiary of Japan's Recruit Co. Ltd. and is co-headquartered in Austin, Texas and Stamford, Connecticut with additional offices around the world. As a single-topic search engine, it is also an example of vertical search. Indeed, is currently available in over 60 countries and 28 languages. In October 2010, Indeed.com passed Monster.com to become the highest-traffic job website in the United States.

The site aggregates job listings from thousands of websites, including job boards, staffing firms, associations, and company career pages. In 2011, Indeed began allowing job seekers to apply directly to jobs on Indeed's site and offering resume posting and storage. Using web scraping, data such as Job title, Job Seekers, Roles, Contract Types and others are collected. The scraped data is then exported to csv and json format for better readability and understanding.

A simple web application is created using Flask and SQLAlchemy. The scraped data is used to build the SQLite database. A front-end website is designed to display the book information, search a book based on the title, author, to add, edit or delete a book detail.

Web Scraping

Web scraping, web harvesting, or web data extraction is data scraping used for extracting data from websites. Web scraping software may access the World Wide Web directly using the Hypertext Transfer Protocol, or through a web browser. While web scraping can be done manually by a software user, the term typically refers to automated processes implemented using a bot or web crawler. It is a form of copying, in which specific data is gathered and copied from the web, typically into a central local database or spreadsheet, for later retrieval or analysis.

Web scraping a web page involves fetching it and extracting from it. Fetching is the downloading of a page (which a browser does when you view the page). Therefore, web crawling is a main component of web scraping, to fetch pages for later processing. Once fetched, then extraction can take place. The content of a page may be parsed, searched, reformatted, its data copied into a spreadsheet, and so on. Web scrapers typically take something out of a page, to make use of it for another purpose somewhere else. An example would be to find and copy names and phone numbers, or companies and their URLs, to a list (contact scraping).

Scraper may be defined as a software or script used to download the Dataset.

We are analysing a web scraper in simple steps as shown in the diagram given above.

Step 1: Downloading Contents from Web Pages

In this step, a web scraper will download the requested as shown

Below, Dataset Scraped from different URLs

https://www.indeed.fr/Paris-(75)-Emplois-job-etudiant

https://resumes.indeed.com/search?q=java&l=&searchFields=jt

Since the script sends a greater number of requests simultaneously to the indeed server, it is less likely to get our IP address blocked. We also avoid disrupting the activity of the website we scrape by allowing the server to respond to other users' requests too. To control the loop rate, we have used sleep() function from time module. It will pause the execution of the loop for a specified amount of time. To mimic human behaviour, we'll vary the amount of waiting time between requests by using the randint() function from the Python's random module. randint() randomly generates integers within a specified interval.

Using argparse library, genre input is collected from the user by which the script will scrape only the book details of the opted genres.

Step 2: Extracting Data using beautiful soup library and using panda's library converting it to JSON and CSV file.

The data on Indeed website is HTML and mostly unstructured. Hence, in this step,

- web scraper will parse and extract structured data from the downloaded contents.
- The Good Reads websites thousands of books. Our project extracts the data and save that into CSV file and JSON and then build a database to read JSON and displays them in the console

We are done with front end we have created add, edit, update and search buttons to search for particular Jobs .

In the front-end user can search for Job roles and view details about the Jobs.

Step 3: Analysing the Data

After all these steps are successfully done, the web scraper will analyse the data thus obtained shown below in detail.

Running the Webapp

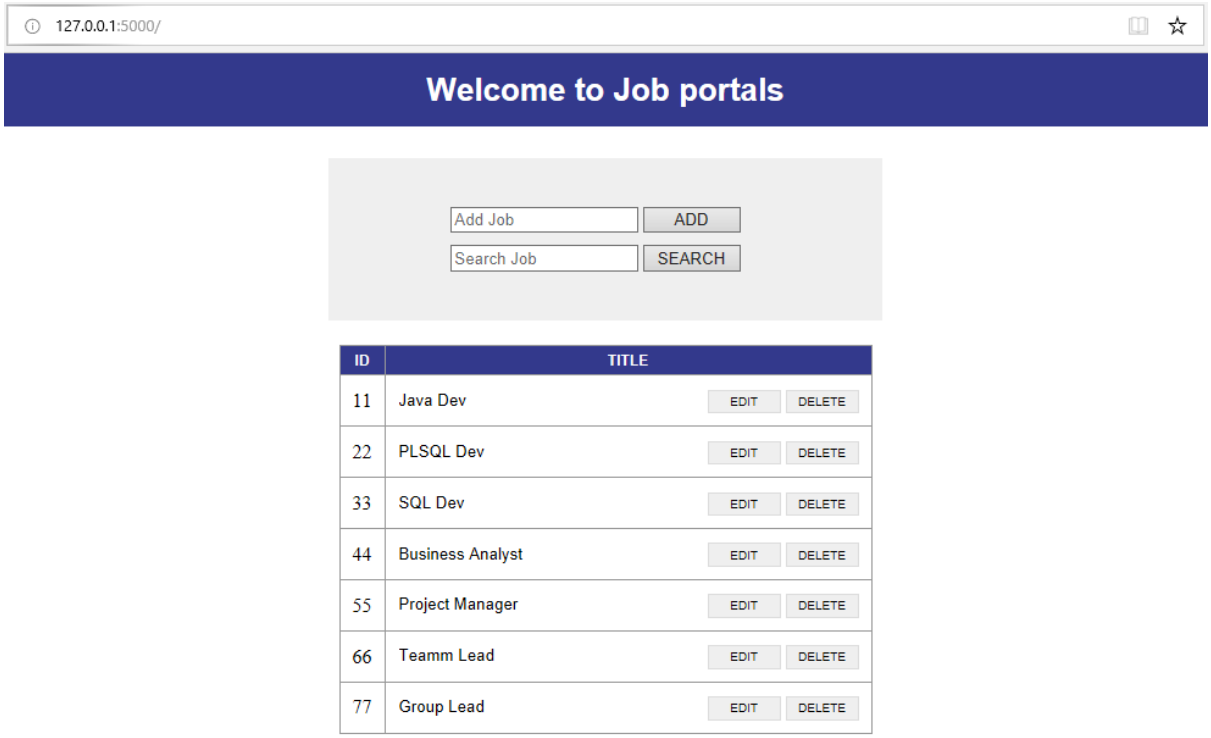
Note: The ***home.html*** file is saved in templates folder

Run the following command in the terminal

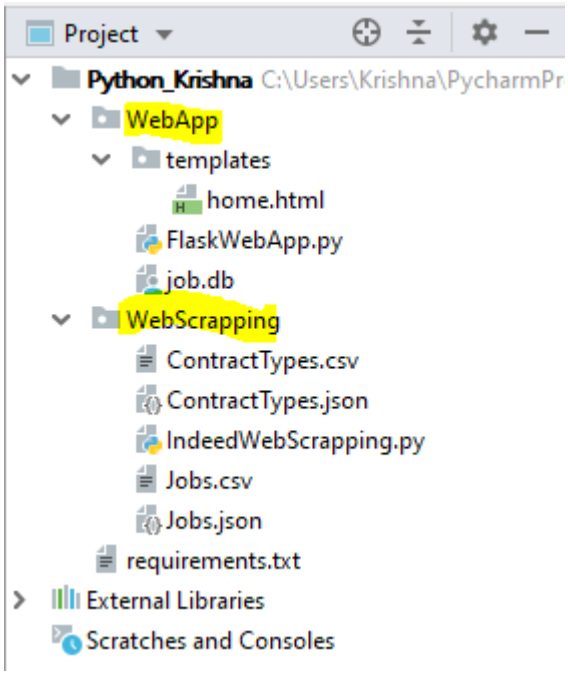
```
$ python FlaskWebApp.py
```

Copy the url from the command http://127.0.0.1:5000 and paste in the browser.

You can view the html file.



Project Structure:



CSV File

Python has a vast library of modules that are included with its distribution. The csv module gives the Python programmer the ability to parse CSV (Comma Separated Values) files. A CSV file is a human readable text file where each line has several fields, separated by commas or some other delimiter. You can think of each line as a row and each field as a column. The CSV format has no standard, but they are similar enough that the csv module will be able to read most CSV files. You can also write CSV files using the csv module. The data read from Csv by using CSV module's read function and the data can written into CSV file using CSV module's write function.

JSON

The process of encoding JSON is usually called **serialization**. This term refers to the transformation of data into a *series of bytes* (hence *serial*) to be stored or transmitted across a network. You may also hear the term **marshalling**, Naturally, **deserialization** is the reciprocal process of decoding data that has been stored or delivered in the JSON standard.

Python Supports JSON Natively!

Python comes with a built-in package called json for encoding and decoding JSON data.