Python project

January 9, 2019

Concept used

- Data structures: list and dict

- List and dict comprehension

- functions and Class (You can make one class or more by module)

- Manage every cases and handle errors (use exceptions).

Web Scraping: static or dynamic

Modules used: Urllib, Requests, BeautifulSoup, Scrapy, Selelium.

Data sample

You can choose any massive dataset. For each step explain your strategy (which tags you scrap and why ?)

NBA

- Scrape name of all basketball player since 1950

- Get players name and detail link of each.

- Get detail information for one player.

- Get detail information for all players.

- Get image for one player.

- Get image for all players.

IMDB

- Scrape list of best films ever as text.

- Scrape list of best film ever with link to detail.

- scrape poster get link contain full size poster

- scrape poster image for one film

- scrape poster image for each film in list film.

- scrape all poster (exercise).

- scrape list of current popular film and all poster.

Amazon

- Scrape python books list on one page search.

- scrape detail information for one book.

- scrape detail information for a book list

- scrape most value comment.

- scrape recently review for specific product

- Get list name of 100 best seller book

Indeed Job

- get list of people with specific skill

- get list of people with specific skill, login script

- get list of people with specific skill, search script

- get list of people with specific skill, scrape link

- expand project 1 in to 10 page search

- get name and skill on one publish profile

Save data to json and csv files

Create csv folder and json folder.

- Create functions to save output into csv files. Put these files into csv folder.

- Create functions to save output into json files. Put these files into json folder.

Web Development: Flask and sqlAlchemy

Modules used: Flask, SqlAlchemy, Jinja2

Front-end

- Create forms to search each main information (player name, movie title...etc).

- Display information in a html table (¡table¿). Limit to 50 first items.

- Add button to add each information details to database.

- Add button to delete each information details to database.

- Add button to modify any field of information.

Back-end

Create the SqlAlchemy Database models

- Create tables and columns for your topic of data scraping.

- Create at least 5 columns of numerical and string data.

Create Database operations

- Create each function to manage creating, reading, updating and deleting information in database.

Parser for command line

Module used: argparse. See documentation.

- Create every command for scraping and print on the terminal

- Create command to run the server with front-end and backend.

Results

- Create an zip archive of your projects.

- Include the requirements.txt which contains a simple list of all the packages in the current environment, and their respective versions: The command is pip freeze > requirements.txt (See documentation of virtualenv).