SUMMARY OF UDEMY COURSE

LEARN WEB SCRAPING WITH PYTHON FROM SCRATCH (3 SECTIONS)

RATINGS: 4/5

SECTION 1- WEB SCRAPING WITH PYTHON BEAUTIFULSOUP AND REQUESTS (8 UNITS)

PART1: Web Scraping Course Overview

- Automated web scraping help get data from web more efficient and effective.
- This course will teach python web scraping libraries and how to install them.
- How to extract URL, text data pieces from one webpage.
- How to crawl multiple webpages and extract data from each of them.
- How to handle navigation links and move to next pages
- How to save your scrapped data into a CSV file
- Quick overview of other popular web scraping frameworks.

PART2: Installing BeautifulSoup and Requests

- Windows: i) Open Command Prompt (Anaconda)
  ii) Type 'conda install beautifulsoup4'
  iii) Type 'conda install requests'
- Confirmation: run 'from bs4 import BeautifulSoup
  Import request'
- Attachment: craiglist-01.py

PART3: URL Extraction

i. Specify the link of the webpage you want to scrape
ii. Get the webpage, creating a response object
iii. Extract source code of webpage: pass it through beautifulsoup ('html parser')
iv. Extract info from the webpage (can use 'for' loop)

- Attachment: craiglist-02.py

PART4: Web Scraping Craig list – Titles

- Right Click on link on webpage – Click on 'inspect' – It shows the class name
- Attachment: craiglist-03.py

PART5: Job Details Craig list – Job Details Wrapper

- Best to extract all the details of the job at once
- 'find' and not 'find_all' if you need only the first of occurrence
- Use IF statement incase tag/text is missing
- Slicing to remove bracket and extra spaces

- Attachment: craiglist-04.py

PART6: Job Scraping Craig list - Job Description Page

- How to open the page each job and extract the full description
- Repeat what we did to parse the main page(URL)
- Attachment: craiglist-05.py

PART7: Crawling and Scraping Next Page

- Attachment: craiglist-06.py

PART8: Saving Output to CSV file

- Using pandas to write into CSV files
- Attachment: craiglist-07.py

SECTION2- FINAL ASSIGNMENT

- Attachment: SCRAPY1.ipynb and ASSIGNMENT.ipynb

SECTION3- OTHER WEB SCRAPING PYTHON FRAMEWORKS OVERVIEW (3 UNITS)

PART1: Scrapy VS Other Python Web Scraping Libraries

- Scrapy is a web crawling framework
- Scrapy supports Python3
- Scrapy is a synchronous framework
- Scrapy is similar to Django
- Python based Scraping Tools:
  i) Urllib2 – included in python standard libraries
  ii) Requests – Beginner friendly
  iii) Beautifulsoup – Beginner friendly
  iv) LXML
  v) Selenium – First of all a tool for writing automated tests for web applications. Beginner friendly. Really extensive.
- Scrapy is not beginner friendly, learning curve is a little steeper than some other tools.

PART2: Scrapy Framework Tutorial: Scraping Craig list

- https://python.gotrained.com/scrapy-tutorial-web-scraping-craigslist/ (further reading)

PART3: Advanced Web Scraping Courses (about $18)