Nitty-gritty of Data and Exploratory Analysis with Python 3

Dev Skill
Class 4 – 5 - 6

Jayanta Sarker Shuva

Web Scrapping:

Web scraping is nothing but the process of collecting data from the web. It involves automating the process of fetching data from the web. In order to fetch the web data, all we need is the URL or the web address that we want to scrape from. The fetched data will be found in an unstructured form. In order to make use of the data or collect useful insights from it, we transform it into a structured form. Once converted into a structured form, we need to store the data for further processing. The whole process is called web scraping.

There are mainly 3 types of web scrapping –

1. Scrape data from a web site with an API, which return data in JSON Format.
2. Scrape data from a Raw HTML tags.
3. Scrape data from a Raw XAML tags.

Main Python3 Libraries for Data Scrapping:

1. Urllib (use to open a URL in python)
2. JSON (use to process JSON data)
3. Beautifulsoup or BS4 (use to interpret the HTM into text)
4. Selenium (use to make a browser interpreter for any programming language)

Urllib:

Urllib makes it easy to interact with web services. It is used to construct a URL, send a GET request to a server, and then parse the response.

URL stands for Uniform Resource Locator. Let's breakdown an example –

http://example.com/path/to/page?name=ferret#Nothong

1. **Protocol**: *Http* or *HTTPS*.
2. **Host Name**: *example.com*.
3. **Port**: Some time you will see a colon followed by number. For HTTP port is 80 and for HTTPS port is 443.
4. **Path**: *path/to/page*.
5. **Query String**: the part after "?" mark. *name=ferret*.
6. **Fragment**: Used to jump into a section of a web page. *#Nothong*.

Now let's break down Urllib Module. Urllib has mainly 4 modules or functions.

1. **Request**: Open a URL.
2. **Response**: this module works internally works with Request module. You will not work with it directly.
3. **Error**: Contains different types of errors made by Request module.
4. **Parse**: this module have different functions to break up the URL.

JSON:

JSON stands for Javascript Object Notation. It is a lightweight data format. JSON data packets are small and quickly parsed by browsers.

There are 4 main modules which is mainly used.

1. **json.load** : Load JSON data from a file.
2. **json.loads** : Load JSON data from string.
3. **json.dump** : Write JSON object into a file .
4. **json.dumps** : Ouptputs JSON object as a string.

Scrape data from a web site with an API, which return data in JSON Format:

Let's scrap today's maximum and minimum forecasted temperature of Tangail from –

https://developer.accuweather.com/

Scrape data from a Raw HTML tags:

Let's scrap different types of product information from –

https://www.newegg.com/