Web Scraping and Social Media Scraping

Students

Muhammad Usman (437790)

Project Description

Topic: CoronaVirus

Link: https://www.worldometers.info/coronavirus/

This report contains an explanation regarding the working mechanics of Web scraping methods; Beautiful soup, Scrapy and Selenium. It explains briefly the steps involved in extracting desirable information form web sources. As the world is going through one of the worst pandemics it has ever experienced this website contains worldwide information related to corona virus cases. It has been updated on daily basis. It contains really valuable information e.g. total active cases, total daily deaths and total tests etc. One could use this valuable information for numerous purposes such as for comparative analysis just to see which countries or regions have been the most affected ones in terms of the total active cases or total deaths etc.

Description of Scraper Mechanics

For this project to get the desired output I used three scrapers; Beautiful soup, scrapy and selenium. The goal was to generate the same output using these three methods. Web scrapping allows us to automatically extract information from the internet. All three methods have been evaluated in terms of their functionality, complexity and easiness in writing and implementing. There exist few differences among these scrapers. Such as scrapy and selenium are more API friendly in comparison to Beautiful soup. All three features have some commonalities as well. All three are flexible in terms of HTML parsing and programing. In addition, description about the functionality of codes has been thoroughly explained in all three scrapers.

Performance Comparison

I evaluated the performance of all three scrappers by calculating the total time taken by each scrapper to generate the same output. It turns out that scrapy is the most efficient in terms of generating the output. It took just three seconds to generate the desired output. Beautiful Soup on the other hand took only 4.3 seconds. And least efficient in terms of generating same output is selenium. Selenium took around 118 seconds to generate the same output. However this is

just the one of the comparisons. In terms of writing with ease and implementation I found scrapy the most feasible one. However this could vary from person to person.

Description of the output

My output contains information related to thirteen (13) features that have been used by various stakeholders to spread information about the spread and impact of corona virus on various countries. The features are Country, total cases, new cases, total deaths, new deaths, total recovered, active cases, serious critical, tot cases/1M pop, deaths/1M pop, total tests, tests/1M pop and population. The information or the output has been arranged in an ascending order. Therefore one could easily see which country has the highest number of active cases and total deaths etc. And which region of the world has the lowest number of reported cases and deaths. Below is the version of the output.

#	Country,Other	TotalCases	NewCases	TotalDeaths	NewDeaths	TotalRecovered	NewRecovered	ActiveCases	Serious,Critical	...	TotalTests	Tests/\n1M pop\n	Population
1	USA	33456913	2332	595619	31	26407793	1922	6453501	9002	...	457389227	1374971.0	332653797
2	India	22585749	289838	245057	2659	18591134	279636	3749558	8944	...	302275471	217225.0	1391531080
3	Brazil	15150628	0	421484	0	13677668	0	1051476	8318	...	46834128	219010.0	213844242
4	France	5767959	0	106277	0	4832981	0	828701	5005	...	78568784	1201422.0	65396516
5	Turkey	5016141	0	42746	0	4691224	0	282171	3175	...	49192673	577990.0	85109837
6	Russia	4880262	8419	113326	334	4496132	7517	270804	2300	...	131400000	900075.0	145987781
7	UK	4434860	1770	127605	2	4248211	2467	59044	163	...	163171777	2392919.0	68189424
8	Italy	4111210	8292	122833	139	3604523	14416	383854	2192	...	61097024	1011774.0	60386030
9	Spain	3567408	0	78792	0	3248010	0	240606	2183	...	47213067	1009469.0	46770209

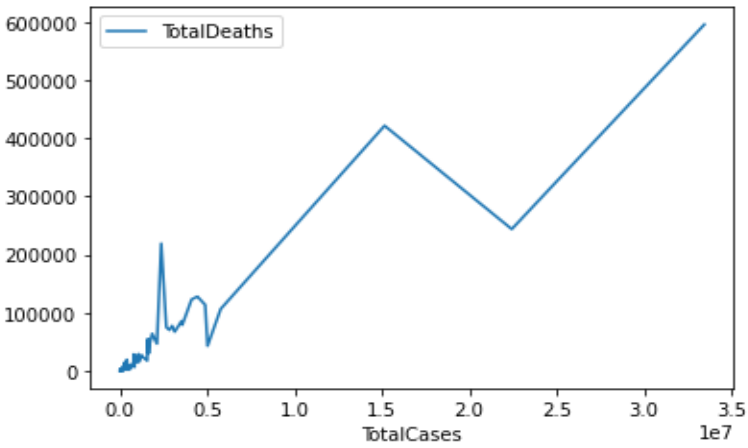
Elementary Data Analysis

One of the simplest analyses is the descriptive analysis. They provide important information ranging from mean, standard deviation, minimum and maximum value etc. Below figure shows the descriptive statistics. This information is quite useful in analysing the distribution patterns of certain features.

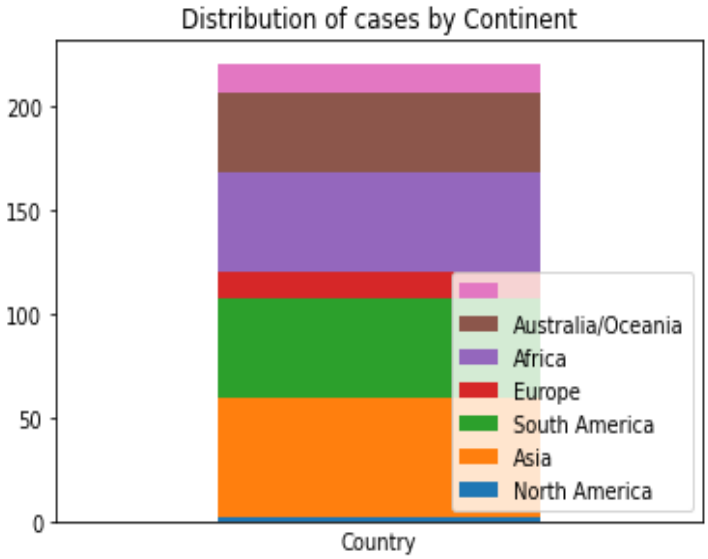
]:

	TotalCases	NewCases	TotalDeaths	NewDeaths	TotalRecovered	NewRecovered	ActiveCases	Serious,Critical	Tot Cases/1M pop	Deaths/1M pop
count	2.210000e+02	221.000000	221.000000	221.000000	2.210000e+02	221.000000	2.210000e+02	221.000000	221.000000	221.000000
mean	7.178632e+05	1936.212670	14921.040724	26.438914	6.199604e+05	2095.389140	8.298173e+04	485.647059	31709.678733	525.089593
std	2.989378e+06	19535.032045	56553.451596	184.446228	2.452616e+06	18899.451583	5.097497e+05	1405.613303	36294.022549	695.481164
min	1.000000e+00	0.000000	0.000000	0.000000	1.000000e+00	0.000000	0.000000e+00	0.000000	0.000000	0.000000
25%	4.654000e+03	0.000000	83.000000	0.000000	3.682000e+03	0.000000	1.450000e+02	0.000000	1748.000000	25.000000
50%	4.069200e+04	0.000000	683.000000	0.000000	3.602200e+04	0.000000	3.263000e+03	14.000000	15992.000000	170.000000
75%	3.175470e+05	178.000000	5093.000000	3.000000	2.761980e+05	222.000000	2.597900e+04	205.000000	59382.000000	839.000000
max	3.345691e+07	289838.000000	595619.000000	2659.000000	2.640779e+07	279636.000000	6.453501e+06	9002.000000	173491.000000	2967.000000

Using this data a lot of analysis could be performed. I have selected and outlined few options that are important to visualize and explain. One of the most important comparisons is to see Total Cases against Total Deaths. We can see in the line graph that both features are strongly related with each other and both show a strong upward trend.



One could even visualize proportion of cases by continent. Following figure on the right hand side shows a continent wise distribution of cases. One could see that Asia accounts for majority of the cases in the world as we have seen a recent dramatic surge in Covid cases in India. Furthermore, various Comparison (Correlation) and Forecasting Analysis (Predictive analysis) could be performed depending upon the interest of the researcher.



Participant

Muhammad Usman (*Beautiful Soup, Scrapy and Selenium*)