Web Scraping Datos Coronavirus

Contexto

En estos días en los que nos vemos agitados por una pandemia, como no pasaba desde hace 100 años, tenemos suerte de que la sociedad actual esté hiper informatizada. Por esto es posible encontrar páginas web donde se recopilan los datos, en este caso sobre la pandemia de la Covid-19. Esta gran disponibilidad de datos nos ha hecho visitar diferentes páginas como la del instituto Carlos III o la del ministerio de salud de la República Checa, donde tenemos la posibilidad de directamente descargarnos los datos.

Pero como el ánimo de esta actividad es recuperar nosotros mismo desde la web sin usar métodos de acceso estandarizados o descargas de datasets.

La página web escogida ha sido https://www.worldometers.info/coronavirus/ y debo indicar que no posee de archivo robots.txt.

Esta página es interesante debido a las diferentes formas de visualización de los datos que emplean y por la constante actualización que provee.

Para hacer la tarea de extracción un poco más compleja se han realizado 3 tareas de extracción, dos a tablas que se van actualizando periódicamente y la otra a las gráficas lineales que muestran el número de casos que han salido a la luz cada día desde el 22 de enero

Títulos de datasets

Data_countries: tabla que recoge los datos de los países

Data_continents: tabla que recoge los datos de los continentes

Data_series_cases_countries: tabla que recoge los datos de casos/día de los países

Descripción del dataset

Data_countries: En este dataset tenemos información sobre 210 países. La información proporcionada es Número total de casos, Nuevos casos del día, Muertes totales, Nuevas muertes del día, Recuperaciones totales, Casos activos, Pacientes críticos, Ratio de casos por millón de población, Ratio de muertes por millón de población, Tests totales y Tests por millón de población.

Data_continents: Presenta la misma información que el caso anterior, pero para los 5 continentes y América partida en dos. En este caso, por ahora no están disponibles los datos relativos al millón de habitantes ni al número de tests

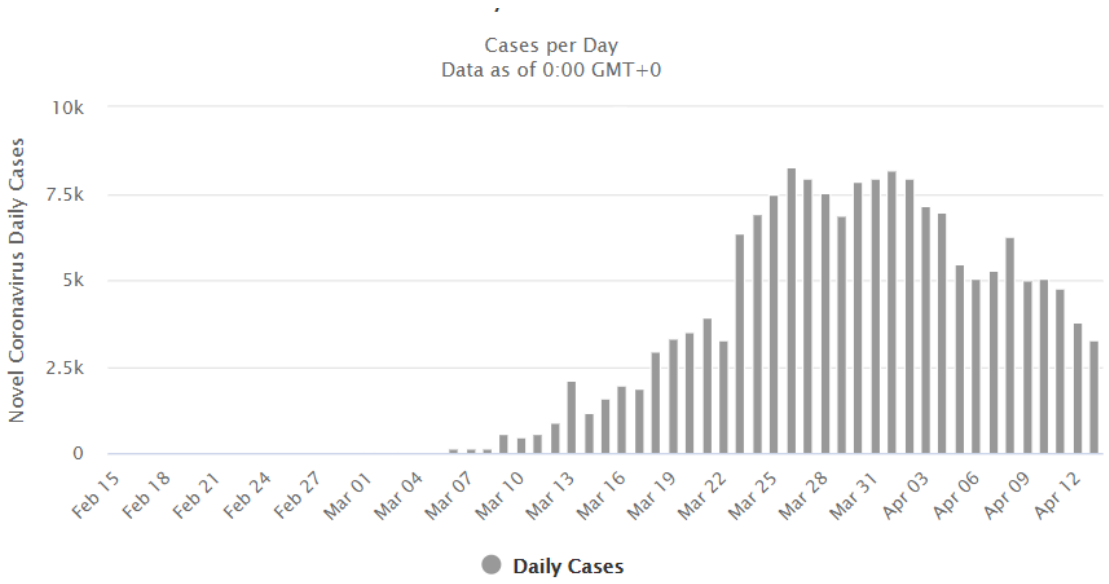
Data_series_cases_countries: Compuesto por los datos referentes a nuevos casos para los 210 países disponibles y almacenados atendiendo a la distribución temporal

Javier Pérez Córdova 2020-04-14

Representación gráfica

Country, Other	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
Europe	923,280	+15,894	81,787	+1,798	240,179	601,314	29,269				
Spain	172,541	+2,442	18,056	+300	67,504	86,981	7,371	3,690	386	600,000	12,833
Italy	159,516		20,465		35,435	103,616	3,260	2,638	338	1,046,910	17,315
France	136,779		14,967		27,718	94,094	6,821	2,095	229	333,807	5,114
Germany	130,434	+362	3,220	+26	68,200	59,014	4,288	1,557	38	1,317,887	15,730

1. Representación gráfica de los valores existentes en tablas



2. Representación gráfica de los datos temporales relacionados con el número de casos

Indicar que todos estos dats han sido almacenados en tablas en formato csv

Contenido datasets

Para los casos de las tablas:

- **Country-other**: el nombre de país o continente del que corresponden los datos.
- **TotalCases**: número total de casos acumulados desde el primer contagio registrado hasta el 13 de abril de 2020
- **NewCases**: nuevos casos que van apareciendo en el día
- **TotalDeaths**: número total de muertes acumulados desde el primer contagio registrado hasta el 13 de abril de 2020
- **NewDeaths**: nuevas muertes que se producen a lo largo del día
- **TotalRecovered**: número total de pacientes recuperados acumulados desde el primer contagio registrado hasta el 13 de abril de 2020
- **ActiveCases**: número de casos activos a fecha actual
- **Serious-Critical**: número de pacientes en condiciones graves a fecha actual
- **Tot Cases/1M pop**: Extrapolación del número de casos totales al millón de habitantes
- **Deaths/1M pop**: Extrapolación del número de muertes totales al millón de habitantes
- **Total Tests**: Número total de tests realizados a la población del país a fecha actual

- **Tests/1M pop**: Extrapolación del número de tests totales al millón de habitantes

Para el caso en el que se extrae de la gráfica:

- En un eje tenemos los días desde el 22 de enero hasta la actualidad
- En el otro eje tenemos los 210 países
- Entre medias tenemos el número de casos que se han producido cada día

Agradecimientos

Todos los datos han sido obtenidos de la página web https://www.worldometers.info/coronavirus/ que recoge la información de las diferentes agencias gubernamentales de los países presentados, así como información anónima y de redes sociales

Inspiración

Lo interesante de estos datasets es tener la información necesaria correspondiente a las principales métricas tomadas durante esta pandemia, aunque se podría haber incluido alguna más como la tasa de hospitalización o la tendencia de muertes por día, y habrían sido buenos ejemplos para extraer esos datos.

Con los datos disponibles, y si en un futuro se extraen más métricas podríamos aplicar modelos y representaciones para conocer o predecir como se desarrollará esta pandemia en base a la información actual

Licencia

Esto lo hago porque el dataset es muy sencillo, y como he dicho en otros apartados sería interesante enriquecerlo en un futuro

Código

Todo el código empleado para esta práctica está almacenado en este GitHub (https://github.com/Javipercor/WebScrapping-Python)

Hay un archivo principal (main) y un archivo auxiliar donde se almacenan las diferentes calases y funciones.

Los datasets generados se almacenan en la carpeta datasets del mismo GitHub y en el enlace de Zenodo que se muestra a continuación

Publicación de los datasets

Enlace a Zenodo: https://zenodo.org/record/3751756#.XpXldsgzZEY