TIPOLOGÍA Y CICLO DE VIDA DE DATOS

-

Practica 1

Por Guillermo Orts y Nicola Bafundi

Inspiración

Actualmente, el mundo está viviendo una situación excepcional debido a la irrupción de un nuevo virus que está poniendo a prueba de la capacidad de actuación y supervivencia de muchos países. Al ser una situación nueva al nivel global, no se dispone de experiencia previa ni de un guión a seguir para reducir el impacto social y económico del virus lo más bajo posible. Por eso, creemos necesario recolectar y relacionar datos que permitan ver el impacto de la pandemia para facilitar posteriores estudios para entender mejor e identificar cuáles fueron las consecuencias de esta pandemia global.

En este sentido, hemos relacionado la evolución del avance del COVID-19 en diversos países (medida en función de los nuevos casos de infectados, decesos y recuperados) con el nivel de calidad del aire de las ciudades más importantes.

La calidad del aire que se respira es un factor de riesgo muy importante en la salud de las personas, por lo que medidas actuales que se están aplicando como el aislamiento podrían volver en un futuro para reducir la cantidad de gases contaminantes en episodios de alta contaminación atmosférica.

Contexto

Los datos se han recogido en pleno desarrollo de la pandemia del COVID-19 cuando la mayoría de países ya han aplicado medidas de contención, entre los que destaca el aislamiento en el hogar, para frenar el avance del virus y evitar el colapso sanitario. Debido a esto, se han reducido los traslados y la actividad industrial por lo que se espera una reducción del nivel de polución en las ciudades más importantes de cada país. Con ello, se espera ver cuál es la relación entre el número de infectados en un país con la evolución de la cantidad de gases contaminantes presentes en la ciudad.

Para la elaboración del dataset, se ha recogido información de diferentes páginas web. En primer lugar, se ha accedido a la página web https://www.worldometers.info/ , cuya función es recoger y mantener actualizadas diversas estadísticas globales. Para la pandemia del COVID-19, dispone de una sección especializada en la cual recoge los datos de infectados, decesos y recuperaciones de cada país, así como la variación de cada una de las estadísticas respecto al día anterior.

La información de los datos de contaminación se ha extraído a partir del portal https://aqicn.org/, destinado a mostar un informe de los datos de contaminación de ese día, así como una previsión de los datos de contaminación en el aire para los próximos tres días, pudiendo visitar los datos de las principales ciudades del mundo. Este portal, además incluye una sección para ver la evolución histórica de los valores de contaminación durante los ultimos 6 años.

Descripción y contenido del Dataset

El título del dataset es **COVID19_Pollution_Dataset**.

El conjunto de datos contiene la información de la evolución de los casos detectados de COVID-19 por día para 8 países diferentes repartidos por todo el mundo. En concreto, se dispone de los registros de casos activos totales acumulados, de casos nuevos detectados, de decesos y de recuperaciones de COVID-19. Además, se recogen también el nivel de diversas partículas contaminantes en el aire para una de las ciudades más pobladas de los países seleccionados.
Los 8 países con su ciudad correspondiente son los siguientes:

- España (Madrid)
- Argentina (Buenos Aires)
- Alemania (Berlín)
- Inglaterra (Londres)
- Italia (Roma)
- China (Pekín)
- Francia (París)
- Estados Unidos (Nueva York)

Las partículas seleccionadas para determinar el nivel de calidad del aire son las siguientes:

- PM2,5: partículas de 2,5 µm de diámetro o menor que pueden incluir sustancias químicas orgánicas, polvo, hollín y metales[1]
- PM10: partículas sólidas o líquidas de polvo, cenizas, hollín, partículas metálicas, cemento o polen, dispersas en la atmósfera, y cuyo diámetro varía entre 2,5 y 10 µm [2]
- O3 (Ozono): gas tóxico que a concentraciones elevadas puede tener efectos en la salud humana, afectando principalmente al aparato respiratorio e irritando las mucosas, pudiendo llegar a producir afecciones pulmonares.[3]
- NO2 (dióxido de nitrógeno): compuesto químico gaseoso de color marrón amarillento formado por la combinación de un átomo de nitrógeno y dos de oxígeno. Es un gas tóxico e irritante.[4]
- SO2 (dióxido de azufre): es un gas que se origina sobre todo durante la combustión de carburantes fósiles que contienen azufre (petróleo, combustibles sólidos). Tiene efectos adversos sobre la salud.[5]

[1] Fuente: https://oehha.ca.gov/calenviroscreen/indicator/pm25
[2] Fuente: http://www.prtr-es.es/Particulas-PM10,15673,11,2007.html
[3] Fuente: http://www.aragonaire.es/ozone.php?n_action=health
[4] Fuente: https://www.saludgeoambiental.org/dioxido-nitrogeno-no2
[5] Fuente: https://www.miteco.gob.es/es/calidad-y-evaluacion-ambiental/temas/atmosfera-y-calidad-del-aire/calidad-del-aire/salud/dioxido-azufre.aspx

- CO (monóxido de carbono): es un gas tóxico, inodoro, incoloro e insípido, parcialmente soluble en agua, alcohol y benceno, resultado de la oxidación incompleta del carbono durante el proceso de combustión.[6]

A continuación, se muestra una captura de como quedaría el dataset recogido:

Country	City	Date	Active Cases	Daily New Cases	Daily New Deaths	Newly Recovered	PM2.5	PM10	O3	NO2	SO2	CO
Spain	Madrid	29/02/2020	56	25	0	0	88	47	26	18	2	0
Spain	Madrid	01/03/2020	82	26	0	0	64	28	27	10	2	0
Spain	Madrid	02/03/2020	118	36	0	0	29	26	30	13	2	0
Spain	Madrid	03/03/2020	162	45	1	0	25	17	21	21	1	0
Spain	Madrid	04/03/2020	224	63	1	0	47	14	24	17	2	0
Spain	Madrid	05/03/2020	276	54	1	1	38	15	33	17	2	0
Spain	Madrid	06/03/2020	387	119	5	3	38	9	32	17	2	0
Spain	Madrid	07/03/2020	485	124	2	24	23	13	31	21	2	0
Spain	Madrid	08/03/2020	625	149	7	2	32	16	25	22	3	0
Spain	Madrid	09/03/2020	1169	557	13	0	48	14	30	24	3	0
Spain	Madrid	10/03/2020	1524	464	6	103	37	13	26	17	3	0
Spain	Madrid	11/03/2020	2039	582	19	48	36	32	20	28	3	0

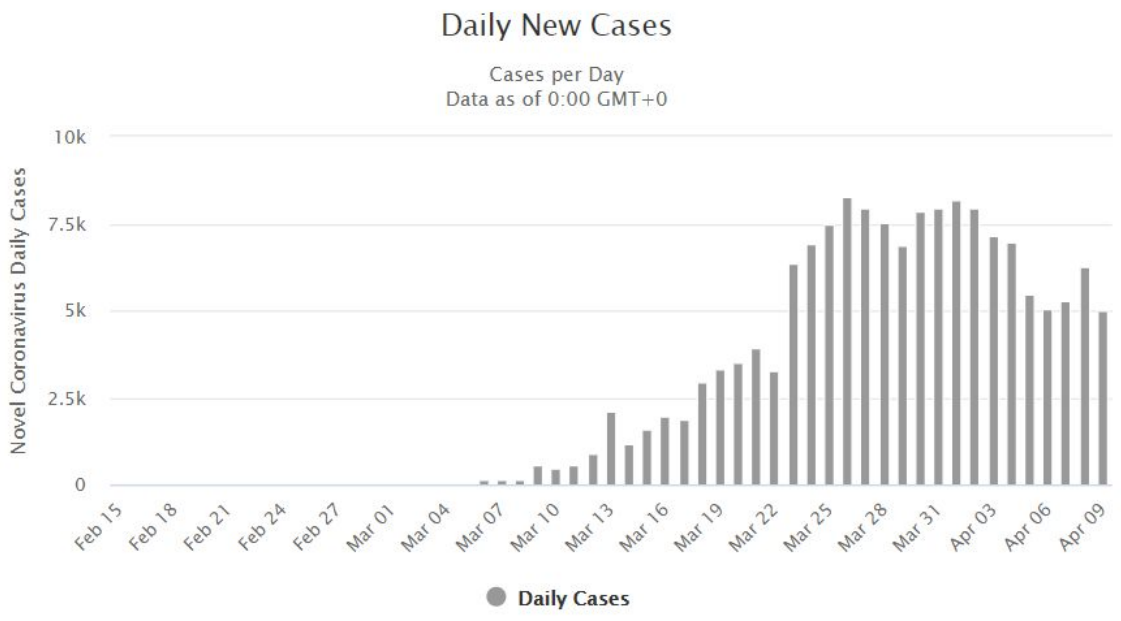
Como se puede ver en la captura anterior, el dataset estará formado por los siguiente campos:

- Country: Nombre del pais, en inglés, empezando en mayúscula.
- City: Nombre de la ciudad, en inglés, empezando en mayúscula.
- Date: fecha en formato DD/MM/YYYY.
- Active Cases: número entero que indica los casos activos totales registrados de COVID-19 en el día en concreto.
- Daily New cases; número entero que indica el incremento de casos positivos en el día en concreto.
- Daily New Deaths: número entero que indica la cantidad de muertes por COVID-19 registradas el día en concreto.
- Newly recovered: número entero que indica la cantidad de pacientes recuperados de COVID-19 ese día.
- PM2.5: medición en ug/m3 de partículas de 2,5 um de diámetro o menor.
- PM10: medición en ug/m3 de partículas de 10 um de diámetro o menor.
- O3: medición en ug/m3 de moléculas de ozono.
- NO2: medición en ug/m3 de moléculas de dióxido de nitrógeno.
- SO2: medición en ug/m3 de moléculas de dióxido de azufre.
- CO: medición en ug/m3 de moléculas de monóxido de carbono.

Los datos relacionados con la evolución del COVID-19 en cada país, se han obtenido mediante web-scraping de la página web https://www.worldometers.info/ . En concreto, se ha accedido a cada URL de la página que contiene los datos de COVID-19 de los países indicados: https://www.worldometers.info/coronavirus/country/x/ donde "x" es el país seleccionado.
En cada URL, los datos se muestran en forma de gráfico por el navegador:

[6] Fuente: http://www.crana.org/es/contaminacion/mas-informacion_3/monaxido-carbono

Daily New Cases in Spain

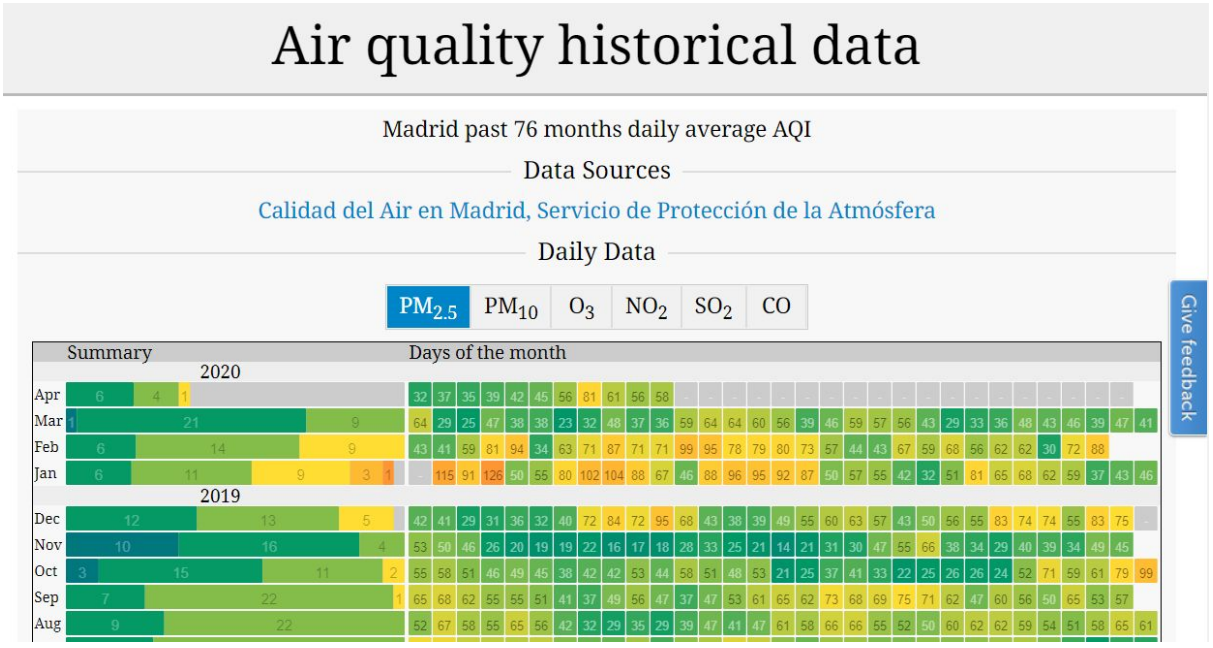


En el fichero html, los datos a extraer se encuentran en una etiqueta <*script*> cuyo atributo <*type*> es '*text/javascript*'.
Se ha utilizado la librería *BeautifulSoup* de python para obtener el segmento de código que contiene los datos. Para obtener los días y la cantidad de casos en un script concreto se ha utilizado la librería *re* de python para identificar el patrón que contiene los datos que queremos extraer. A continuación, se muestra una captura del fragmento de código HTML de donde se obtienen los datos numéricos de casos activos:

```
series: [{
    name: 'Currently Infected',
    color: '#00DDDD',
    lineWidth: 5,
    data: [0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,2,8,8,11,16,18,20,30,32,43,51,60,74,92,122,152,151,235,
    246,329,442,505,504,654,709,569,787,853,973,1045]        }
],
```

Los datos relacionados con la contaminación en las ciudades de cada país, se han obtenido mediante web-scraping de la página web https://aqicn.org . En concreto, se ha accedido a cada URL de la página que contiene los datos de contaminacion de las ciudades indicados: https://aqicn.org/city/x/ donde “x” es la ciudad seleccionado.

Esta página web contiene una tabla en la sección *Air quality historical data* donde se muestra por cada tipo de magnitud, una tabla con la información día a día por mes y año.



El problema es que esta página genera el contenido *HTML* de forma dinámica, de tal modo que la sección *Air quality historical data* solo se genera cuando se hace scroll hasta el tag *<h1></h1>*. Por otro lado, para que el contenido de la tabla se genere, se tiene que hacer click sobre el botón de cada magnitud. Para poder navegar la página web y simular las interacciones del usuario se ha utilizado la librería Selenium de Python, de este modo se extrae el código *HTML* de cada una de las tablas.

Posteriormente se analiza el código *HTML* que se ha extraído de la página con la biblioteca BeautifulSoup de Python, para indexar la información en un diccionario.

Agradecimientos

Como se ha comentado en los puntos anteriores, los datos se han obtenido mediante dos dominios de página web diferentes. Los datos relacionados con la evolución de los infectados por COVID-19 en los países seleccionados se han obtenido del dominio worldometer.info.
El dominio worldometer.info pertenece a un equipo internacional de desarrolladores, investigadores y voluntarios con el objetivo de hacer que las estadísticas mundiales estén disponibles en un formato de reflexión y tiempo relevante para una amplia audiencia de todo el mundo. Forma parte de una pequeña compañía independiente de medios digitales con sede en los Estados Unidos. Asimismo, indican que no tienen afiliación política, gubernamental o corporativa.
Para los datos de COVID-19, recopilan los datos de informes oficiales, directamente de los canales de comunicación del Gobierno o indirectamente, a través de fuentes de medios locales cuando se consideran confiables. Las actualizaciones diarias son realizadas con la

participación de usuarios de todo el mundo y a la dedicación de un equipo de analistas e investigadores que validan los datos de una lista de más de 5,000 fuentes.[7]

Por otro lado, el dominio https://aqicn.org/ pertenece a la World Air Quality Index project, un proyecto sin ánimo de lucro iniciado en 2007. Su misión es promover la concienciación de la contaminación en la ciudadanía proveyendo una información mundial sobre la calidad del aire.
El proyecto tiene su sede en china y es independiente de cualquier gobierno. Cuenta con un equipo compuesto científicos medioambientales, ingenieros de sistemas, data science y diseñadores. Las actualizaciones diarias son realizadas por más de 14481 colaboradores en 132 países.

Licencia

El dataset será lanzado con la licencia "Released Under CC BY-NC-SA 4.0 License" y por lo tanto con los siguientes derechos de divulgación:

- Compartir: copiar y redistribuir el material en cualquier medio o formato.
- Adaptar: transformar y construir sobre el material.

Bajo los siguientes términos:

- Atribución: se debe otorgar el crédito apropiado, proporcionar un enlace a la licencia e indicar si se realizaron cambios. Se puede hacer de manera razonable, pero de ninguna manera que sugiera que el licenciante lo respalda.
- No comercial: no se puede utilizar el material con fines comerciales.
- Compartir igual: si se transforma o desarrolla el material, se debe distribuir sus contribuciones bajo la misma licencia que el original.

Al no contar con el permiso expreso de las página web de las que se extrae la información, se decide aplicar una licencia de libre uso no comercial al dataset generado.

Repositorio de Git

https://github.com/guillermorts/COVID19_Dataset.git

Tabla de contribuciones:

Contribuciones	Firma
Investigación previa	G.O.T. y N.B.A.
Redacción de respuesta	G.O.T. y N.B.A.
Desarrollo código	G.O.T. y N.B.A.

[7] Fuente: https://www.worldometers.info/about/