Juan Pablo Botero Suaza
Juan Carlos Ghiringhelli
41024416Y

Tipología de vida y ciclo de los datos

Practica 1 – Web scraping

Contenido

El objetivo de esta actividad será la creación de un dataset a partir de los datos contenidos en una web. Para su realización, se deben cumplir los siguientes puntos:

1. Contexto. Explicar en qué contexto se ha recolectado la información. Explique por qué el sitio web elegido proporciona dicha información.

La calidad del aire hoy día se ha convertido en un factor crítico para la salud pública de todos los ciudadanos, es por ello que los sistemas de medición de la calidad del aire son una estrategia fundamental para la toma de decisiones en el ámbito gubernamental. Para desarrollar este concepto, emplearemos como fuente de datos la información del sitio web https://www.eltiempo.es/calidad-aire, el cual reporta información sobre el índice de calidad del aire para varias regiones de España.

"El índice de **calidad del aire** (ICA) es un indicador genérico de la calidad del aire y sus efectos sobre la salud en un lugar determinado. Indica el **nivel de contaminación** existente en un lugar, sus potenciales efectos para la salud y las recomendaciones que se deben seguir para protegerla".

Para calcular el índice de calidad del aire, el Eltiempo.es aplica la fórmula propuesta por la EPA (United States Environmental Protection Agency), con la que se calcula el índice de cada contaminante, el cual depende de su concentración.

Los contaminantes son los siguientes: Partículas (PM), Monóxido de carbono (CO), Ozono (O3), Dióxido de nitrógeno (NO2), Dióxido de azufre (SO2).

Imagen tomada de https://en.wikipedia.org/wiki/Air_quality_index#Computing_the_AQI

Computing the AQI [edit]

The air quality index is a piecewise linear function of the pollutant concentration. At the boundary between AQI categories, there is a discontinuous jump of one AQI unit. To convert from concentration to AQI this equation is used:[36]

$$I = \frac{I_{high} - I_{low}}{C_{high} - C_{low}}(C - C_{low}) + I_{low}$$

where:

I = the (Air Quality) index,
C = the pollutant concentration,
C_{low} = the concentration breakpoint that is ≤ C,
C_{high} = the concentration breakpoint that is ≥ C,
I_{low} = the index breakpoint corresponding to C_{low},
I_{high} = the index breakpoint corresponding to C_{high}.

EPA's table of breakpoints is:

O_3 (ppb)	O_3 (ppb)	$PM_{2.5}$ (µg/m^3)	PM_{10} (µg/m^3)	CO (ppm)	SO_2 (ppb)	NO_2 (ppb)	AQI	AQI
C_{low} - C_{high} (avg)	C_{low} - C_{high} (avg)	C_{low}- C_{high} (avg)	C_{low} - C_{high} (avg)	C_{low} - C_{high} (avg)	C_{low} - C_{high} (avg)	C_{low} - C_{high} (avg)	I_{low} - I_{high}	Category
0-54 (8-hr)	-	0.0-12.0 (24-hr)	0-54 (24-hr)	0.0-4.4 (8-hr)	0-35 (1-hr)	0-53 (1-hr)	0-50	Good
55-70 (8-hr)	-	12.1-35.4 (24-hr)	55-154 (24-hr)	4.5-9.4 (8-hr)	36-75 (1-hr)	54-100 (1-hr)	51-100	Moderate
71-85 (8-hr)	125-164 (1-hr)	35.5-55.4 (24-hr)	155-254 (24-hr)	9.5-12.4 (8-hr)	76-185 (1-hr)	101-360 (1-hr)	101-150	Unhealthy for Sensitive Groups
86-105 (8-hr)	165-204 (1-hr)	55.5-150.4 (24-hr)	255-354 (24-hr)	12.5-15.4 (8-hr)	186-304 (1-hr)	361-649 (1-hr)	151-200	Unhealthy
106-200 (8-hr)	205-404 (1-hr)	150.5-250.4 (24-hr)	355-424 (24-hr)	15.5-30.4 (8-hr)	305-604 (24-hr)	650-1249 (1-hr)	201-300	Very Unhealthy
-	405-504 (1-hr)	250.5-350.4 (24-hr)	425-504 (24-hr)	30.5-40.4 (8-hr)	605-804 (24-hr)	1250-1649 (1-hr)	301-400	Hazardous
-	505-604 (1-hr)	350.5-500.4 (24-hr)	505-604 (24-hr)	40.5-50.4 (8-hr)	805-1004 (24-hr)	1650-2049 (1-hr)	401-500	

Ilustración 1 - Rangos de contaminación

De los índices calculados para cada contaminante se selecciona el que presenta un valor más desfavorable. Ese valor es el que se considera como índice de calidad del aire para ese momento.

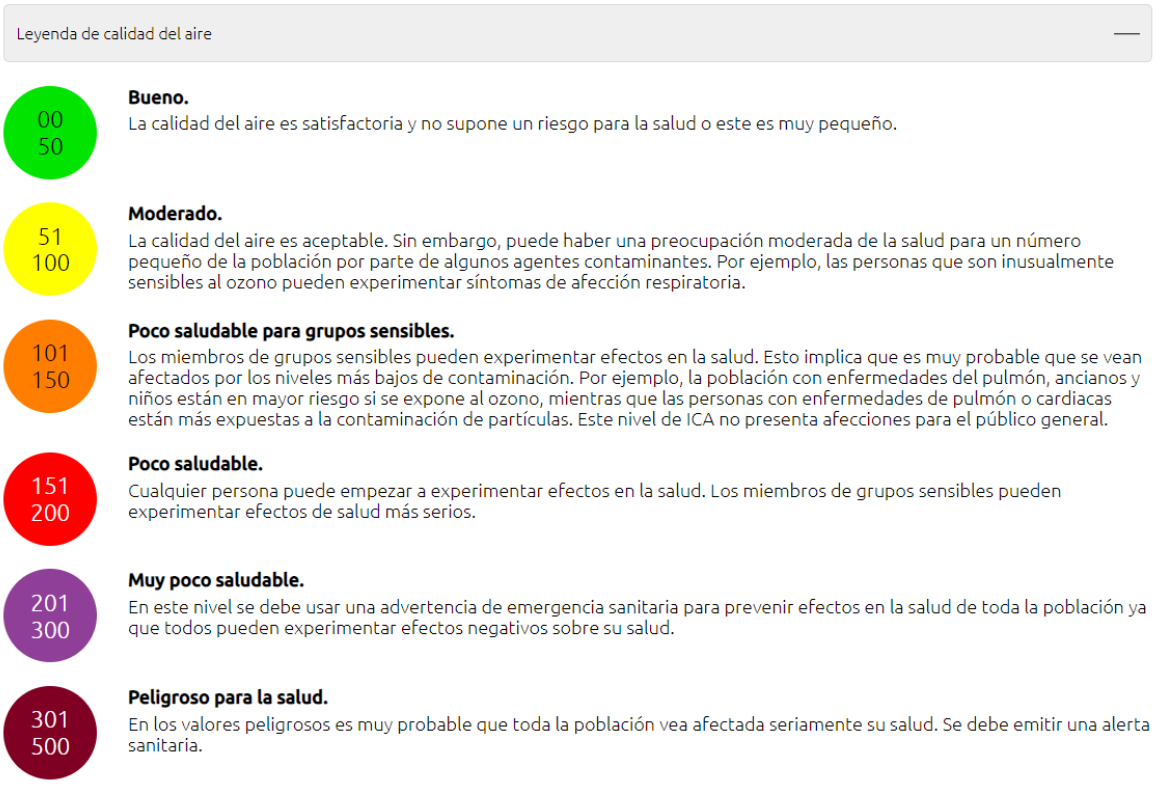

Ilustración 2 - Escalas de gravedad

2. Definir un título para el dataset. Elegir un título que sea descriptivo.

Dataset: Índice calidad del aire (**ICA**) en España por regiones.

3. Descripción del dataset. Desarrollar una descripción breve del conjunto de datos que se ha extraído (es necesario que esta descripción tenga sentido con el título elegido).

El conjunto de datos recolectado agrupa información de la concentración de contaminantes promedio (ug/m3) en el aire y el valor final ICA obtenido en cada estación para diferentes regiones de España, publicados en intervalos de tiempo variable.

Los tipos de datos de asociados a las cantidades de concentración son de naturaleza numérica, sin embargo, en el proceso de recolección, incluye las unidades (ug/m3), en un proceso de limpieza de datos posterior habría que tener presente esta condición.

El valor del ICA es un tipo de dato numérico que obedece al contaminante con mayores niveles de concentración dentro de la muestra recolectada.

Los valores del nombre de la región, el nombre de la estación de medición y la hora son tipos de datos texto como parte del proceso de recolección, en un fase posterior de limpieza, se podrían convertir las variables de acuerdo a su naturaleza como categóricas y continuas en el caso de la hora de publicación, un *timestamp* por ejemplo.

4. Representación gráfica. Presentar una imagen o esquema que identifique el dataset visualmente.

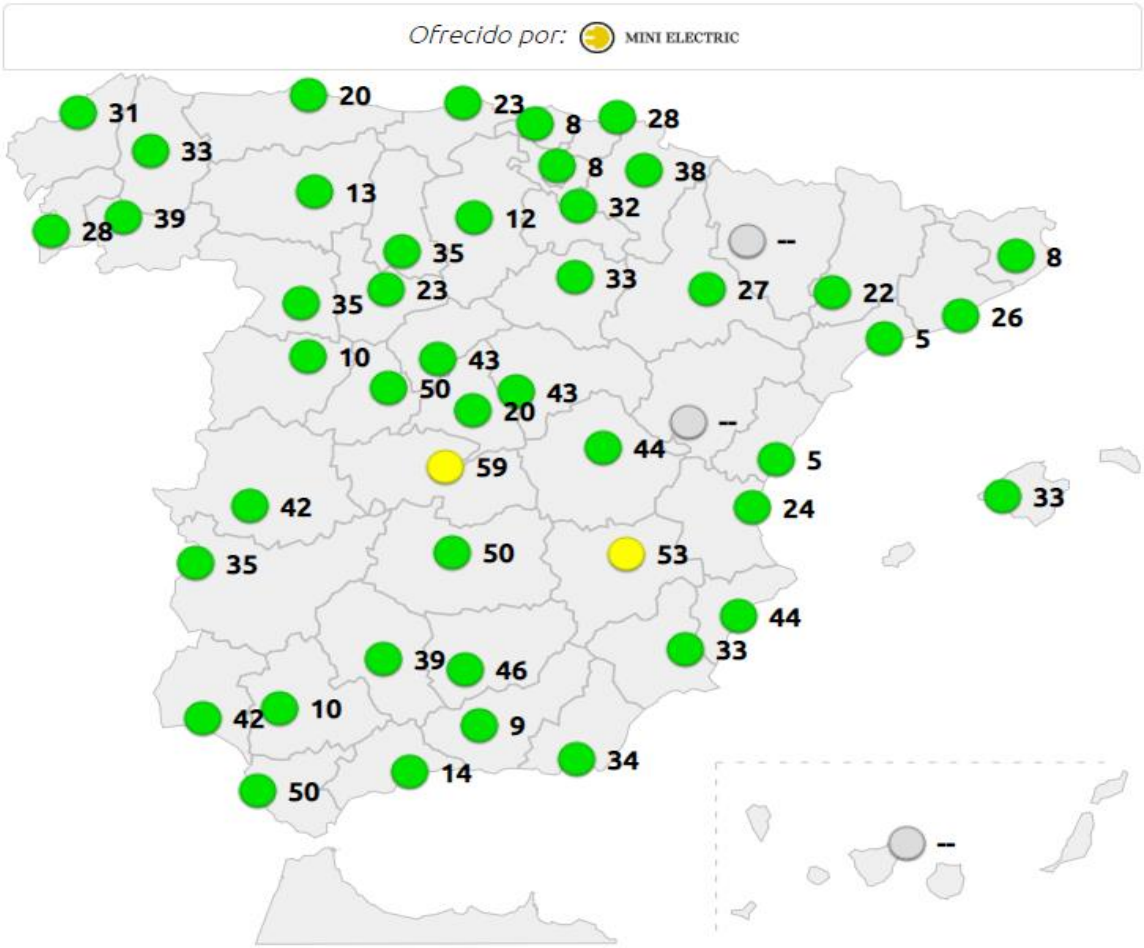

Ilustración 3 - Valor general de ICA por estación en mapa

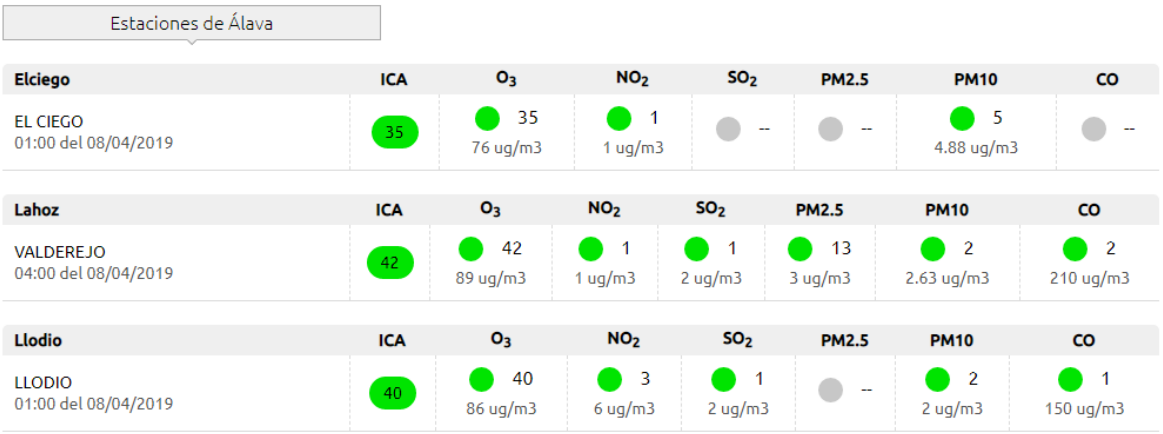

Ilustración 4 - Detalle de valores por región y estación

5. Contenido. Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y cómo se ha recogido.

El dataset definido contiene las cantidades de concentración en el aire para cada uno de los contaminantes descritos en el apartado 1), el valor del ICA para la estación, el nombre de la región donde se mide el ICA, el nombre de la estación en cada región; encargada de la toma de muestras para cada contaminante y la hora de publicación de los datos en el sitio web:

Ejemplo:

Fecha publicación: 14:00 del 03/04/2019 Región: ALBACETE Estación: ALBACETE ICA Estación: 53 O3: 91 ug/m3	NO2: 13 ug/m3 SO2: 2 ug/m3 PM2.5: 13.13 ug/m3 PM10: 25.25 ug/m3 CO: 630 ug/m3

El crawler empieza obteniendo la página principal, donde encuentra los vínculos a la página de cada región. Posteriormente se obtiene cada página, y en esta, se lee cada table de cada estación. Estas tablas se cargan en objetos que representan las tablas y las filas en el scraper. Finalmente se pasan estos objetos a la clase encargada de persistir, que guarda los datos anexándolos a un archivo CSV, previo codificado en UTF-8 para conservar los caracteres especiales del castellano.

6. Agradecimientos. Presentar al propietario del conjunto de datos. Es necesario incluir citas de investigación o análisis anteriores (si los hay).

Agradecemos a Pelmorex Weather Networks, que publica el conjunto de datos. El análisis del archivo robots.txt nos dice que no hay restricciones en la captura automática de datos. El uso del

comando *whois* de Python solo devuelve datos vacíos, por lo que no sabemos a ciencia cierta quién es el maestro de la página.

Análisis anteriores:

SIATA: Sistema de Alerta Temprana de Medellín y el Valle de Aburrá (Colombia) https://siata.gov.co/sitio_web/index.php/calidad_aire

AirNow: Environmental Protection Agency US: https://airnow.gov/

7. Inspiración. Explique por qué es interesante este conjunto de datos y qué preguntas se pretenden responder.

El conjunto de datos propuesto pretende dar a conocer la influencia que tienen algunos de los contaminantes más conocidos sobre el aire que respiramos, por eso la importancia de identificar cuáles son los efectos que puede traer consigo este tipo de fenómenos sobre la salud humana, con el fin de poder actuar con oportunidad y tomar medidas que disminuyan el riesgo al que nos exponemos cuando los índices de contaminación en el aire son bastante elevados.

La actividad diaria de las ciudades genera una gran cantidad de sustancias que modifican la composición natural del aire que respiramos tanto en el exterior como en interiores. La quema de combustibles fósiles para el transporte y la generación de energía, tanto a nivel industrial como doméstico, produce miles de toneladas de contaminantes que diariamente se quedan en la atmósfera. Los vehículos son la principal fuente de emisión de contaminantes del aire, le siguen la industria, los hogares y las emisiones de fuentes naturales.

El deterioro de la calidad del aire por la presencia de sustancias contaminantes tiene un efecto negativo en la salud humana y del medio ambiente. Diversos estudios realizados en España en los últimos 3 años han demostrado que existe una relación entre el incremento en la concentración de los contaminantes del aire y el aumento de enfermedades respiratorias y cardiovasculares como el asma, bronquitis o diversas cardiopatías, por mencionar algunas. Los contaminantes como es el caso las partículas suspendidas están asociados además con una mayor cantidad de reincidencia a las salas de urgencia y con casos de mortandad.

https://www.miteco.gob.es/es/calidad-y-evaluacion-ambiental/temas/atmosfera-y-calidad-del-aire/informeevaluacioncalidadaireespana2017_tcm30-481655.pdf

https://spip.ecologistasenaccion.org/IMG/pdf/informe-calidad-aire-2016.pdf

https://www.miteco.gob.es/es/calidad-y-evaluacion-ambiental/temas/atmosfera-y-calidad-del-aire/libroaire2015_tcm30-187887.pdf

https://www.miteco.gob.es/es/calidad-y-evaluacion-ambiental/temas/atmosfera-y-calidad-del-aire/calidad-del-aire/documentacion-oficial/Analisis-CA.aspx

Una manera de proteger la salud de la población es a través del monitoreo y la difusión continuos del estado de la calidad del aire. En la ciudad de Medellín (Colombia), el Sistema de Alerta Temprana de Medellín y el Valle de Aburrá, es el responsable de la medición permanente de los principales contaminantes del aire.

Relacionamos algunos estudios previos que han utilizado la analítica de datos para estudiar este tipo de fenómenos y como los modelos estadísticos propuestos puede apoyar la toma de decisiones desde los ámbitos gubernamentales, industrial y sector salud:

https://ourworldindata.org/air-pollution

https://www.epa.gov/outdoor-air-quality-data

https://www.researchgate.net/publication/320669285_Statistical_analysis_of_air_pollution_with_specific_regard_to_factor_analysis_in_the_Ciuc_basin_Romania

https://doee.dc.gov/sites/default/files/dc/sites/ddoe/service_content/attachments/AQ%20TRENDS%20Report%20for%20DDOEwebsite_finalDraft_2014Oct29.pdf

Aspectos interesantes que pueden analizarse a partir del set de datos pueden ser los siguientes:

- ¿Cuáles son los contaminantes que predominan y en qué periodos de tiempo? ¿Como pueden identificarse las fuentes que los generan?

- Modelo predictivo de regresión/clasificación sobre el índice de calidad del aire esperado, simulando condiciones atmosféricas con contaminantes para periodos futuros, con el fin de plantear mecanismo de alerta temprana que minimicen los riegos asociados a la salud de las comunidades. Esto permite tomar medidas preventivas como impedir circulación de coches e informar a la población para que ciudadanos con alto factor de riesgo sean prevenidos.

- Identificar en que regiones debe existir mayor foco de atención por parte de autoridades del sector salud por sus altos índices de contaminantes en el aire. Esto permite identificar altas fuentes de contaminación, como industrias, y tomar medidas correctivas.

- Cuáles son las principales enfermedades que se derivan a partir de la contaminación del aire para cada región de España, y si de ellas se derivan eventos de mortalidad, como pueden ser tratadas estas causas con programas de atención y prevención de la salud.

8. Licencia. Seleccione una de estas licencias para su dataset y explique el motivo de su selección:

- Released Under CC0: Public Domain License
- Released Under CC BY-NC-SA 4.0 License
- Released Under CC BY-SA 4.0 License
- Database released under Open Database License, individual contents under Database Contents License
- Other (specified above)
- Unknown License

Al tratarse de un estudio netamente académico con fines no lucrativos, cuyo objetivo es la búsqueda del bienestar para las comunidades en general, a través de la identificación y análisis de fenómenos que pueden comprometer la salud de las personas, optamos por elegir la licencia pública sin restricciones *Released Under CC0: Public Domain License*, con el fin de que pueda estudiarse desde cualquier lugar, institución académica, empresa centros de ciencias, por analistas y

profesionales de los datos que deseen compartir ideas y mejorar el dataset con información que se considere relevante.

9. Código. Adjuntar el código con el que se ha generado el dataset, preferiblemente en Python o, alternativamente, en R.

https://github.com/jghiringhelli/scrapping-pec1-tcvd/, directorio “src”.
.

10. Dataset. Presentar el dataset en formato CSV.

https://github.com/jghiringhelli/scrapping-pec1-tcvd/ , directorio “data”, archivo: “calidad_aire.csv”.

Contribuciones	Firma
Investigación Previa	JPBS, JCG
Redacción de las respuestas	JPBS, JCG
Desarrollo de código	JPBS, JCG

Bibliografía

- Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 2. Scraping the Data.
- Masip, D. (2010). El lenguaje Python. Editorial UOC.
- Subirats, L., Calvo, M. (2019). Web Scraping. Editorial UOC.
- Air quality index. Wikipedia. [Consulta: 2 de abril de 2019]. https://en.wikipedia.org/wiki/Air_quality_index.
- AirNow, Publicaciones. Airnow. [Consulta: 2 de abril de 2019]. https://airnow.gov/index.cfm?action=pubs_spanish.index.
- Creative Commons license. Wikipedia. [Consulta: 6 de abril de 2019] https://en.wikipedia.org/wiki/Creative_Commons_license.
- Índice de calidad del aire en España. Wikipedia. [Consulta: 30 de marzo de 2019]. https://www.eltiempo.es/calidad-aire.