TIPOLOGÍA Y CICLO DEL DATO.

Web Scraping

MEMORIA PRÁCTICA 1

Autor: Juan Ramón Tonda Barberá

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

   La crisis provocada por el coronavirus y la consecuencia del confinamiento de la población desde el día 14 de marzo de 2020 ha dado lugar a una reducción drástica del tráfico de vehículos a motor en las ciudades. Analizar el impacto de la contaminación atmosférica, especialmente la relativa al dióxido de nitrógeno que aparece en la combustión de los vehículos a motor, es interesante ver como evoluciona este contaminante durante el periodo de alarma.

   En este caso, el ayuntamiento de Valencia es la institución que publica estos datos, extraídos a su vez de los diferentes sensores que tiene repartidos por toda la ciudad y publicados en su página web. El análisis se realizará sobre una única estación, la relativa a la entrada a la ciudad por la llamada "Pista de Silla", una de las principales arterias de acceso a la ciudad.

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

   Contaminación por Dióxido de Nitrógeno en la ciudad de Valencia para el periodo 2019 – 2020 (febrero)

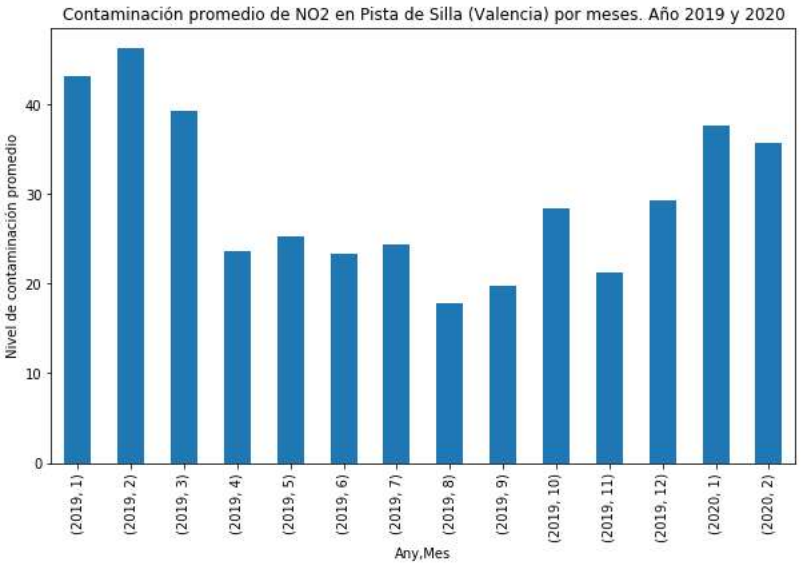
3. Descripción del dataset. Desarrollar una descripción breve del conjunto de datos que se ha extraído.

   El Dataset contiene toda la información suministrada por la estación que está ubicada en la salida de Valencia, en la denominada "Pista de Silla" para cada uno de los días del año 2019 y para cada uno de los días del año 2020 para los meses de enero y febrero, por lo tanto, tenemos un campo fecha con el formato dd/mm/yyyy. A continuación, se relaciona una serie de contaminantes atmosféricos, en total once atributos que recogen datos sobre el Benceno, el Tolueno, el Ozono, el NOx (óxidos de nitrógeno en general), el NO2(dióxido de nitrógeno), partículas varias, Xileo, dióxido de azufre o monóxido de carbono. También destaca el set de datos con un dato sobre los decibelios diarios.

   Entiendo que el conjunto de datos debe ser un resumen de actividad del día, ya que los sensores deben tener una periodicidad de recogida de datos diezminutal o similar.

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

   Este sencillo gráfico de barras recoge los principales atributos incluidos en el test: valor de contaminación del NO2, fecha (en este caso el mes) y el lugar en el que se han recogido los datos.



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

El conjunto de datos analizados se centra en el contaminante relativo al dióxido de nitrógeno por ser uno de los principales contaminantes emitidos por los vehículos de motor y que en el caso de Valencia representa el 90% de dichas emisiones en la ciudad. Dicho conjunto de datos está formado por los principales campos:

- Fecha.- El recorrido temporal del conjunto de datos corresponde a todo el año 2019 y a los dos primeros meses del año 2020 en escala diaria. He esperado hasta última hora para ver si el ayuntamiento subía el pdf del mes de marzo para analizar el impacto que realmente ha tenido el confinamiento durante el estado de alarma en la contaminación atmosférica pero desgraciadamente no ha sido posible.
- Niveles de Dióxido de nitrógeno (N02) en ug/m3.
- Any.- Se extrae el año a partir del campo fecha
- Mes- Se extrae el mes a partir del campo fecha.
- Sem.- Se extrae el número de semana del año a partir del campo fecha.

Los datos son resúmenes de actividad procedentes de la Red Automática de Control de la Contaminación Atmosférica del ayuntamiento de Valencia a partir de estaciones sensorizadas repartidas por toda la ciudad de Valencia y que recogen las señales de contaminación en tiempo real.

El Ayuntamiento de Valencia en su página web publica un documento pdf de manera mensual en el que se recogen los datos de los contaminantes con un espacio temporal diario, sin embargo, la publicación del mes de marzo de 2020, a 13 de abril de 2020 todavía no se ha realizado.

Dicho documento pdf es extraído de su página web y almacenado en el equipo donde se va a realizar el análisis. A continuación, se utiliza una API para realizar la conversión desde pdf a csv. Por último, dicho fichero csv se carga en el "Query Editor" de Excel para realizar una breve transformación con el fin de eliminar sobre todo todas las filas que no se correspondan con datos, sin embargo, la conversión del campo fecha a un formato de tipo fecha y que no dé problemas en el posterior análisis se realizará con librerías phython.

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

Todos los conjuntos de datos que ofrece el Ayuntamiento de Valencia, si no se indica lo contrario, se publican bajo los términos de la licencia Creative Commons – Reconocimiento (CC-By 4.0) que permite lo siguiente:

- Copia, distribuir y divulgar públicamente
- Modificar, transformar o adaptar
- Desarrollar obras derivadas
- Utilizar con fines comerciales o no comerciales

También esta licencia obliga a mencionar la autoría del Ayuntamiento de Valencia.

El fichero robotx.txt de la página del ayuntamiento deshabilita el acceso a algunos directorios y excluye el acceso a cualquier robot excepto sistrix.

```
User-agent: *
Disallow: /ayuntamiento2/
Disallow: /valencia/
User-agent: sistrix
Disallow: /
```

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

El análisis de la contaminación atmosférica en estos momentos es de capital interés por cuanto la sociedad está muy sensibilizada respecto de estos temas, así que los análisis sobre los contaminantes atmosféricos, y además desde una perspectiva temporal, pueden resultar muy útiles para realizar un seguimiento sobre las políticas medioambientales realizadas. Por lo tanto, la pregunta que pretende responder el presente conjunto de datos es la de la evolución de la contaminación de dióxido de nitrógeno en un punto concreto de la ciudad de Valencia y por supuesto, ver su evolución. Es cierto que el objetivo de este análisis se queda un poco corto, sería necesaria una perspectiva temporal más larga, y sobre todo, falta el mes de marzo con el fin de visualizar el impacto del estado de alarma en la contaminación de la ciudad, por no decir que también faltaría parte del mes de abril para poder analizar, ya no por meses ni por semanas, sino por días, la evolución de la contaminación. Por lo tanto, otro de los problemas a los que nos tendremos que enfrentar para el análisis de los datos es la oportunidad en la entrega de datos por parte de las fuentes autorizadas.

8. Licencia.

La licencia es un instrumento legal mediante el cual el titular de los derechos permite a terceras personas realizar determinados usos de los datos sin infringir dichos derechos.

Para el dataset del presente estudio utilizaría la Creative Commons Zero (CC0 1.0) . Implica una liberación de los derechos de propiedad intelectual al ofrecer los datos en dominio público y se permite su utilización sin ningún tipo de restricción. Los datos se pueden copia, modificar, distribuir y hacer públicos incluso para fines comerciales, sin solicitar autorización. En este caso no vería necesario que se reconociera la autoría del dataset, pero si así fuere, se debería utilizar la licencia Creative Commons CC BY-SA 4.0

9. Código. Adjuntar el código con el que se ha generado el dataset.

En el repositorio del Github.

10. Dataset. Publicación del dataset en formato CSV en Zenodo con una pequeña descripción.

ZENODO: https://zenodo.org/deposit/3748912

11. Entrega. Presentar el trabajo con el DOI del dataset en Github.

DOI: 10.5281/zenodo.3748912