Práctica 1: Web scraping

Autores:

Manuel Ed. Escobar Ubreva

Elena Cantón García

Asignatura:

Tipología y ciclo de vida de los datos.

Contribuciones	Firma
Investigación previa	Manuel / Elena
Elección del tema para web scraping	Manuel / Elena
Redacción de las respuestas	Manuel / Elena
Desarrollo del código	Manuel / Elena
Creación repositorio	Manuel / Elena

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

El contexto de este script de web scraping es el siguiente:

Los integrantes del grupo de trabajo nos planteamos averiguar qué importancia puede tener las previsiones meteorológicas en actividades diversas, como puedan ser el turismo, el ocio y actividades comerciales.

De esta reflexión e interés, nos propusimos realizar un script que nos permitiera obtener las previsiones meteorológicas, y posteriormente, cruzar estos datos con diversos sistemas de información. Esto nos permitiría analizar la correlación entre actividades concretas con las previsiones meteorológicas, a nuestro entender muy interesante.

En este escenario, se implementa un script en Python para extraer de la página: https://www.meteosat.com/tiempo/ las previsiones meteorológicas. El desarrollo se realiza desde la perspectiva que sea escalable. Esto quiere decir que, realizando pequeñas modificaciones, permita incorporar nuevas ciudades a la lista, o ampliar el rango de previsiones por hora extraídas, por tanto, que sea un ejemplo inicial de carácter pedagógico y escalable.

Adicionalmente, en el script, hay una función, ¨get_CitiesAvailable¨, que devuelve un listado completo de las ciudades de España disponibles en la https://www.meteosat.com/tiempo/ , por si se quisiera consultar para añadir otras ciudades a la lista del script principal. La función implementada genera el fichero csv “citiesAvailable.csv”.

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

"Previsiones meteorológicas de las próximas x horas en una lista de ciudades de España.”

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 titulo elegido).

En el dataset “data” se han extraído las previsiones meteorológicas de las próximas 12 horas para 4 ciudades. Se genera un registro por hora en el que se informan los siguientes atributos:

- Ciudad, se trata de un texto con el nombre de la ciudad

- Fecha, se trata de un date con el formado DD/MM/YYYY

- Hora, la hora del formato texto.

- Temperatura, temperatura prevista en grados centígrados.

- Dirección del viento, dirección del viento mediante texto.

- Velocidad del viento, texto con la velocidad del viento en kilómetros por hora

- Precipitaciones, texto con las precipitaciones previstas en litros por metro cuadrado.

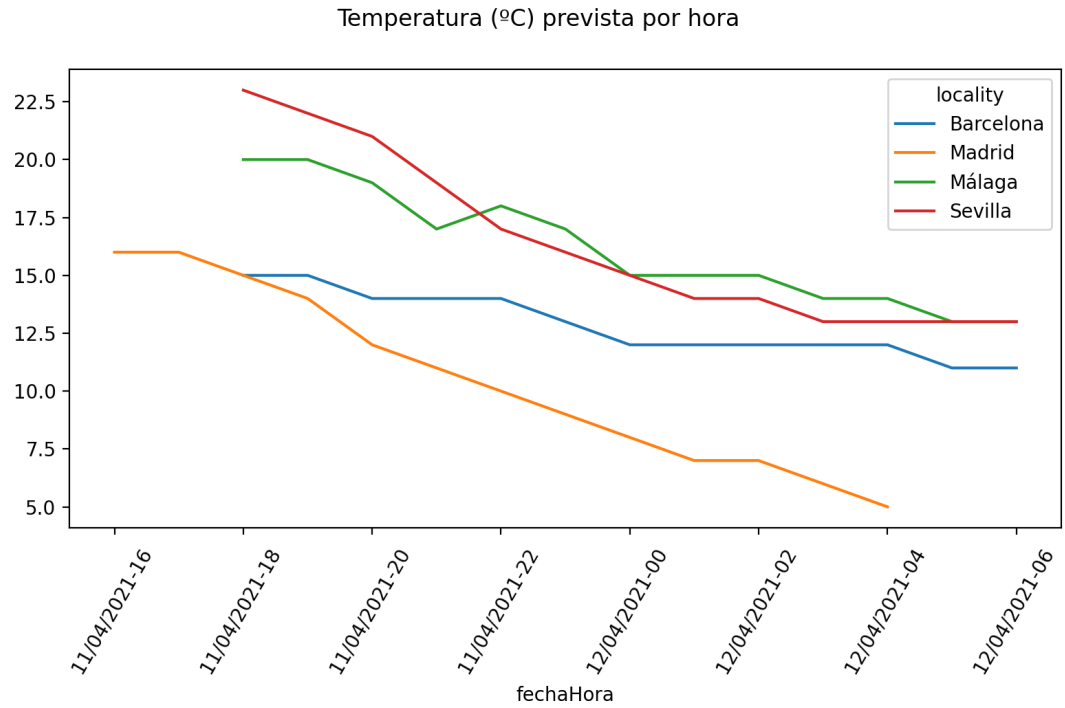
De lo anterior se obtienen tantos registros por ciudades de España que se pase en la lista.

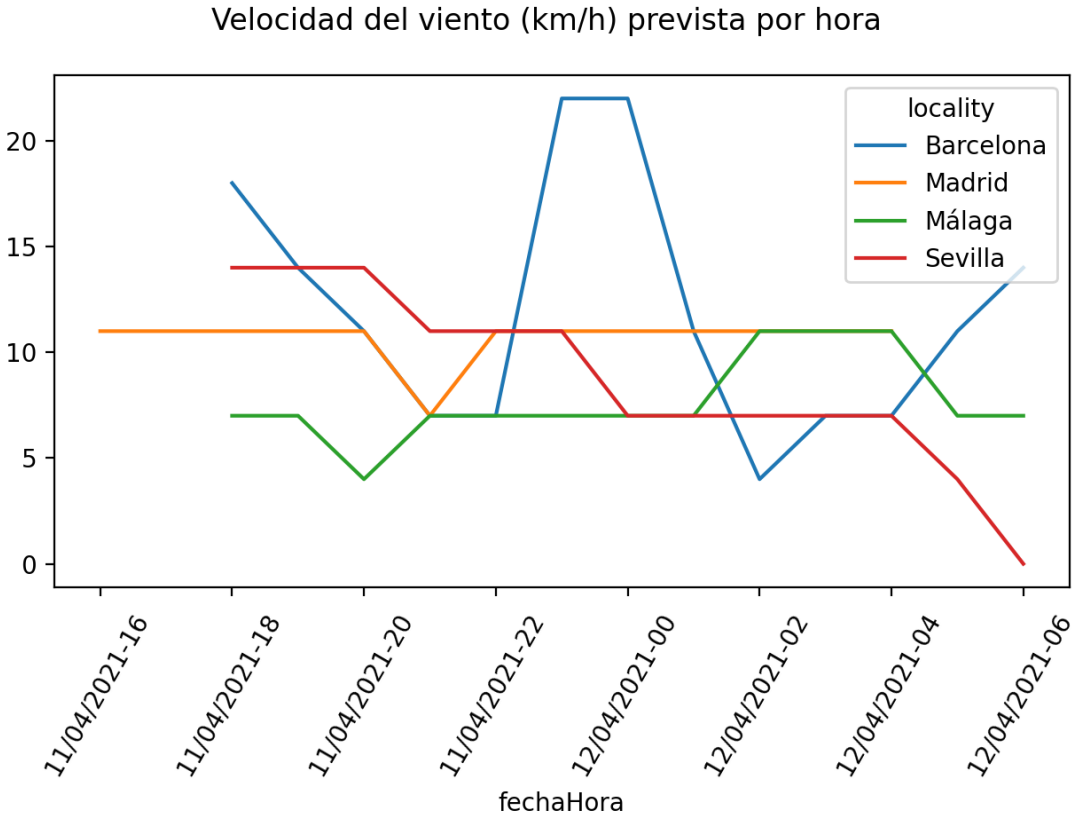
En el caso del script subido en github se obtienen 12 registros y un total de 4 ciudades, esto es, se obtienen 48 registros.totales.

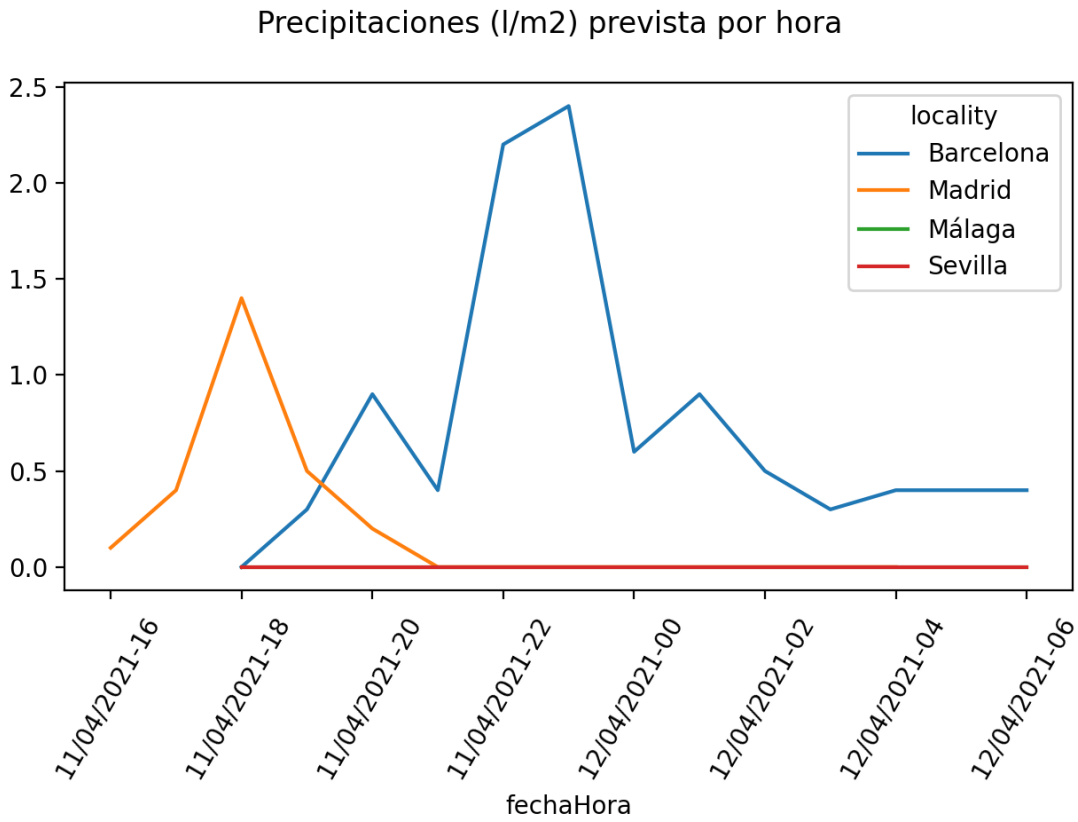
4. Representación gráfica. Presentar esquema o diagrama que identifique el dataset visualmente y el proyecto elegido.

En el script se generan unos gráficos con la librería matplotlib sobre:

- Temperaturas por hora. Una línea para cada ciudad de la lista.
- Velocidad del viento por hora. Una línea para cada ciudad de la lista.
- Precipitaciones por hora. Una línea para cada ciudad de la lista.


Temperatura (ºC) prevista por hora
locality
Barcelona
Madrid
Málaga
Sevilla
22.5
20.0
17.5
15.0
12.5
10.0
7.5
5.0
11/04/2021-16
11/04/2021-18
11/04/2021-20
11/04/2021-22
12/04/2021-00
12/04/2021-02
12/04/2021-04
12/04/2021-06
fechaHora


Velocidad del viento (km/h) prevista por hora
locality
Barcelona
Madrid
Málaga
Sevilla
20
15
10
5
0
11/04/2021-16
11/04/2021-18
11/04/2021-20
11/04/2021-22
12/04/2021-00
12/04/2021-02
12/04/2021-04
12/04/2021-06
fechaHora



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

Actualmente extrae los siguientes atributos:

- Ciudad
- Fecha
- Hora
- Temperatura
- Dirección del viento
- Velocidad del viento
- Precipitaciones

Cada vez que se ejecuta, el script toma los atributos anteriores para las ciudades pasadas en la lista, que en el caso de la práctica son: Barcelona, Madrid, Sevilla y Málaga. Tomando la previsión meteorológica prevista por horas de las próximas 12 horas desde el momento de ejecución del script.

Tomando la información de la página:

https://www.meteosat.com/tiempo/

.

6. Agradecimientos. Presentar al propietario del conjunto de datos. Es necesario incluir citas de analisis anteriores o, en caso de no haberlas, justificar esta búsqueda con análisis similares.

Algunas referencias usadas:

https://www.aprendemachinelearning.com/ejemplo-web-scraping-python-ibex35-bolsa-valores/

https://www.youtube.com/watch?v=bBbiLsnU24M

https://jarroba.com/scraping-python-beautifulsoup-ejemplos/

https://guides.github.com/activities/citable-code/

Los datos han sido recolectados a partir de la url de Meteosat.

Hemos usado código Python para la realización de web scraping.

7. Inspiración. Explique por qué es interesante este conjunto de datos y que preguntas se pretenden responder. Es necesario comparar con los análisis anteriores presentados en el apartado 6.

Como se ha comentado en apartados anteriores, las previsiones meteorológicas pueden estar relacionadas con multitud de otros datos, con los datos obtenidos podemos enriquecer la información de muchos sistemas de información disponiendo de las previsiones meteorológicas.

En este caso específico, se ha planteado para intentar prever datos sobre el turismo en ciudades españolas. Asimismo, también es interesante conocer las previsiones del tiempo para decidir a dónde viajar. Es por ello que consideramos que el factor del tiempo que hará, es clave a la hora de decidir un destino.

Indicar adicionalmente que en la página de la que se toman las previsiones meteorológicas existe una amplia diversidad de datos meteorológicos, para este ejercicio nos hemos basado principalmente en tres inputs del dataset: temperaturas, viento y precipitaciones.

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

La licencia escogida para esta publicación ha sido ¨Released Under CC0: Public Domain License¨.

Los motivos de la elección, principalmente, han sido:

- Engloba el patrimonio intelectual que está libre de toda exclusividad en su acceso y utilización.
- No se requiere ningún permiso o licencia para un trabajo en el dominio público.
- Las obras que están en el dominio público pueden ser utilizadas, adaptadas, traducidas o modificadas por distintos autores para crear nuevas obras sin pedir permiso de ningún tipo a nadie. A partir de ese momento, a la nueva obra se la denomina obra derivada.

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

Para acceder al código y/o al conjunto de datos que el código genera, ir a este enlace:

Script python: https://github.com/manu2eu/DS_M2851_practica1/blob/main/web-scraping.py

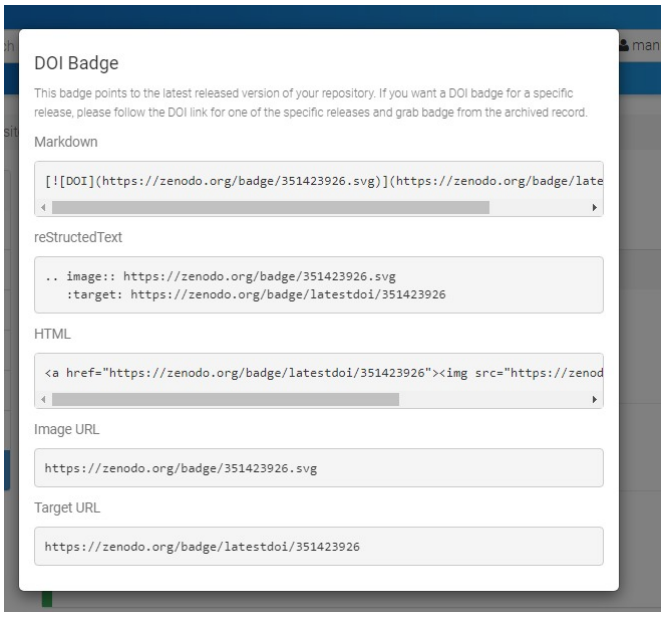
Dataset: https://github.com/manu2eu/DS_M2851_practica1/tree/main/csv (data.csv)

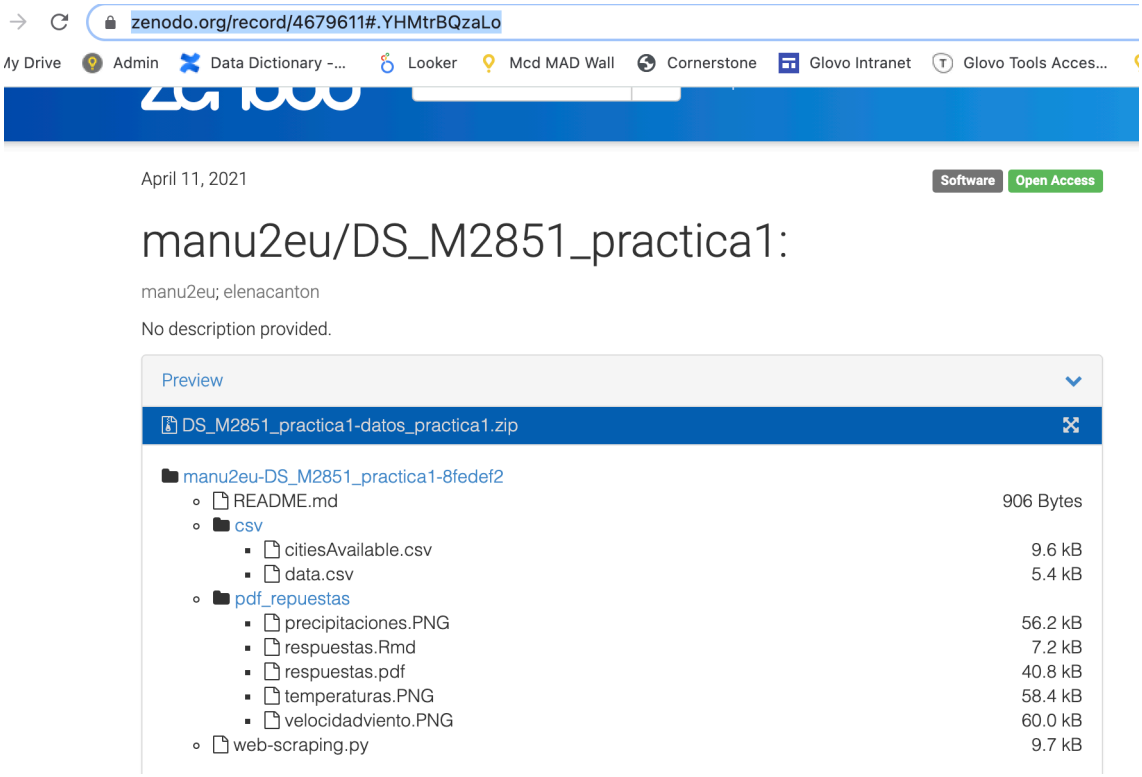
10. Dataset. Publicación del dataset en formato CSV en Zenodo (obtención del DOI) con una breve descripción.

DOI: 10.5281/zenodo.4678420

DOI: identificador de objeto digital. Este identificador está vinculado al repositorio de Github sobre el que hemos trabajador.

Target URL: https://zenodo.org/record/4679611#.YHMtrBQzaLo




zenodo.org/record/4679611#.YHMtrBQzaLo
Admin
Data Dictionary -...
Looker
Mcd MAD Wall
Cornerstone
Glovo Intranet
Glovo Tools Acces...
April 11, 2021
Software
Open Access
manu2eu/DS_M2851_practica1:
manu2eu; elenacanton
No description provided.
Preview
DS_M2851_practica1-datos_practica1.zip
manu2eu-DS_M2851_practica1-8fedef2
README.md 906 Bytes
csv
citiesAvailable.csv 9.6 kB
data.csv 5.4 kB
pdf_repuestas
precipitaciones.PNG 56.2 kB
respuestas.Rmd 7.2 kB
respuestas.pdf 40.8 kB
temperaturas.PNG 58.4 kB
velocidadviento.PNG 60.0 kB
web-scraping.py 9.7 kB