Práctica1: Web scraping

Datos meteorológicos de capitales de las provincias españolas

Unai Mateos Corral, Mikel Laburu Haro

9 de noviembre de 2018

Descripción

Este conjunto de datos dispone de los datos meteorológicos de las capitales de todas las provincias españolas. Se ofrecen distintos datos meteorológicos, como la previsión de la temperatura, la velocidad del viento, la humedad, etc. de cada hora a 72 horas vista, desde la hora en la que se ejecuta el *script* para cargar los datos.

Imagen

En la Figura 1, se puede observar la ilustración seleccionada para este *dataset*.

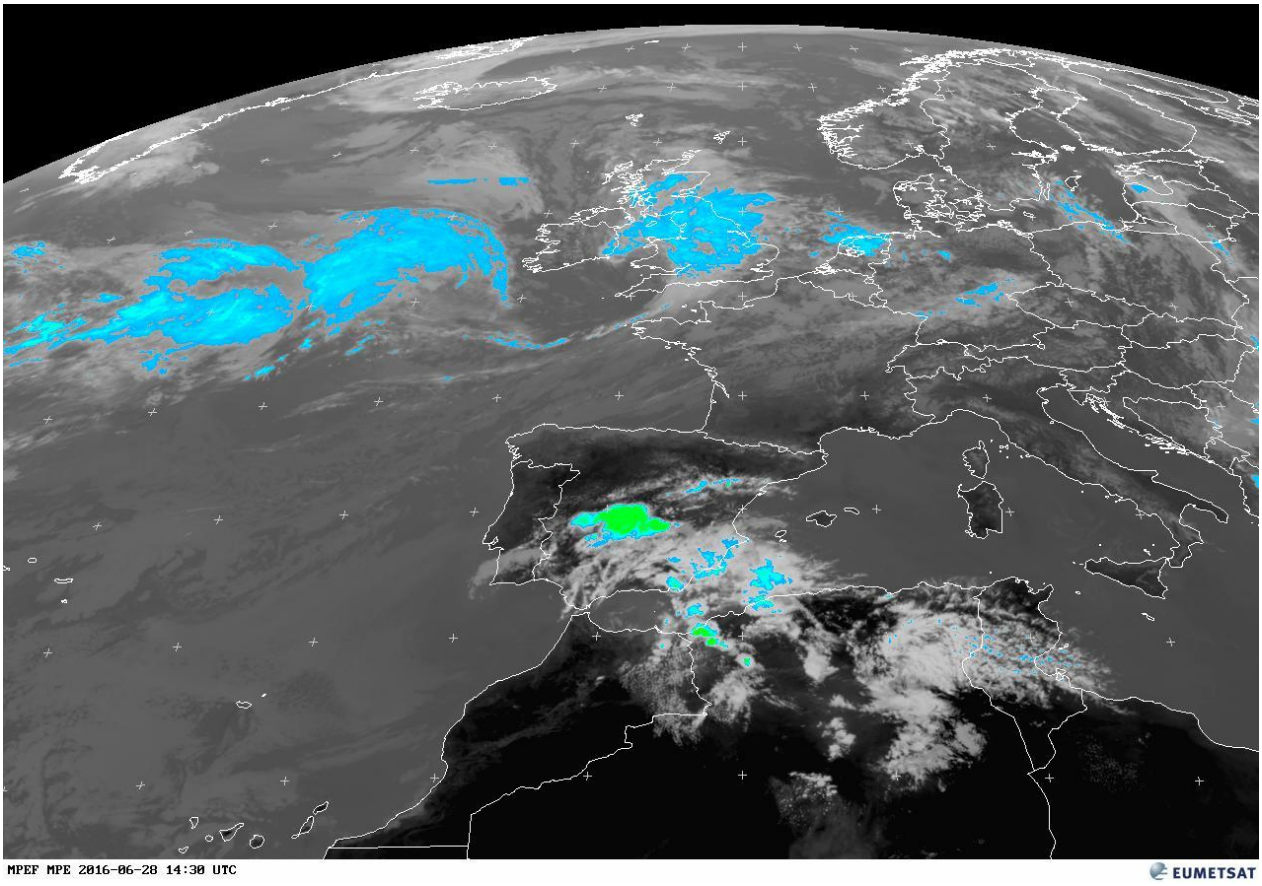

Figura 1: Imagen representativa del *dataset* (ref. meteosat.es)

Contexto

Este conjunto de datos ha sido tomado de la página eltiempo.es, concretamente de la sección que ofrece el tiempo por días y horas de una ciudad determinada. En este caso en lugar de seleccionar una única ciudad, se aplican técnicas de *web scraping* en las páginas de todas las ciudades que son capitales de provincias españolas, tal y como se ha comentado anteriormente, de esta forma se logra un conjunto de datos mucho más completo de lo que sería uno que simplemente dispusiese de una sola ciudad.

Contenido

El periodo de tiempo de los datos que se encuentran en el *dataset* es de 72 horas desde el momento en el que se ejecuta el *script* encargado de realizar el *web scraping*, esto puede abarcar un espacio de tres o cuatro días distintos, en función de la hora en la que se haya probado el *script*. Esto es así puesto que la página web objetivo solamente ofrece la previsión del tiempo en este margen de tiempo.

La página eltiempo.es, da la posibilidad de mostrar el tiempo por horas de cada ciudad, con lo que para poder recoger los registros meteorológicos de todas las capitales españolas, ha sido necesaria la creación de una lista, la cual contiene los nombres de todas estas ciudades, y así hacer la petición a la URL que ofrece esta información variando el campo de la ciudad por cada uno de los que hay en la lista mencionada. De esta forma la forma en la que se ha hecho el *web scraping* es común para las páginas de todas las localizaciones.

Por cada registro del conjunto de datos se pueden identificar los siguientes campos:

- **Ciudad**: Se corresponde a la ciudad a la que pertenece el registro.
- **Día**: Indica la fecha de la predicción que sigue el formato: AAAA/MM/DD.
- **Hora**: Indica la hora del día al que corresponde la predicción. En caso de que sea la hora actual, se indicará con el valor “Ahora”.
- **Previsión**: Describe la temperatura en grados centígrados (°C) de la hora a la que pertenece el registro.
- **Velocidad**: Muestra en kilómetros por hora (km/h), la velocidad del viento.
- **Rachas**: Describe en kilómetros por hora (km/h) la mayor velocidad racha de vientos que pueden alcanzarse.
- **Lluvias**: Indica la cantidad de agua por metro cuadrado que se espera.
- **Nieve**: Representa la cantidad de nieve que puede caer, medida en centímetros.
- **Nubes**: Muestra el porcentaje de nubes, es decir, según este porcentaje se puede detectar si estará, totalmente nublado, parcialmente nublado, etc.
- **Tormenta**: Representa la probabilidad de que suceda una tormenta.
- **Humedad**: Indica el nivel de humedad en el ambiente.
- **Presión**: Índice que representa la fuerza que ejerce el aire que forma la atmósfera sobre la superficie terrestre.

- **Sensación Térmica**: Indica la sensación térmica que hay en el exterior. Temperatura que puede variar en función de distintos factores como la dirección del viento, la velocidad del viento o la humedad.
- **Prob. Precipitación**: Muestra las posibilidades de lluvia. Este campo solo estrá presente en los registros que tengan alguna posibilidad de lluvia, es decir, en caso de que fuese 0 %, el valor que almacenará será *NULL*.

Estos dos atributos son propios del registro que se corresponde con la hora en la que se haya ejecutado el *script*, o lo que es lo mismo, con el primer registro de cada ciudad.

- **Hora observación**: Hora a la que se ha comprobado el pronostico.
- **Visibilidad**: Nivel de visibilidad que hay en el momento en el que se toman los datos.

Agradecimientos

La propiedad de los datos obtenidos corresponde a los propietarios de la página web a la que se le ha aplicado *web scraping*, esto es, eltiempo.es, Pelmorex Weather Networks 2018.

Inspiración

El conjunto de datos, contiene información sobre las predicciones meteorológicas de las diferentes capitales españolas, por tanto, estos datos pueden aportar información útil a la hora de realizar estudios descriptivos sobre temas meteorológicos, como por ejemplo: en qué capital llueve más, cuál es la más soleada, la más fría, la más caliente, etc. En definitiva, es capaz de responder a un amplio rango de preguntas relacionadas con temas meteorológicos.

Por otro lado, se podría enfocar como trabajo futuro, la posibilidad de almacenar los datos con un rango de tiempo mayor (automatizando la ejecución del *script*), por ejemplo, de periodos anuales, lo que facilitaría poder realizar estudios mas completos a lo largo de los años, y observar la variación meteorológica de las ciudades deseadas.

Licencia

La licencia que se ha seleccionado para este conjunto de datos es CC BY-NC-SA 4.0, por los siguientes motivos:

- **Compartir:** Se autoriza la redistribución y la copia del conjunto de datos.
- **Adaptar:** Se permite remezclar, transformar y crear a partir de este Dataset.
- **Reconocimiento:** Se debe reconocer adecuadamente la autoría, proporcionando un enlace a la licencia e indicando si se han realizado cambios.
- **NoComercial:** No se permite la comercialización de este conjunto de datos.
- **CompartiIgual:** Si se remezcla, transforma o crea a partir de este conjunto de datos, esto deberá distribuirse con esta misma licencia.

Código

El código desarrollado se puede obtener del siguiente repositorio, Practica-1 Web Scraping. El archivo correspondiente al código Python es : *scraping.py*

Dataset

El dataset que contiene los datos obtenidos en formato *.csv* es el denominado *SpanishCapitals-Forecast.csv* que se puede descargar del repositorio: Practica-1 Web Scraping.