Práctica 1: Web Scraping

Objetivo

El objetivo de este documento es describir la práctica 1 “Web Scraping” de la asignatura *Tipología y ciclo de vida de los datos*, asignatura del Máster en Ciencia de Datos de la Universitat Oberta de Catalunya.

En la práctica se han utilizado técnicas de web scraping y consultas a API públicas mediante el lenguaje de programación Python para extraer datos de registros diarios de diferentes web meteorológicas como Ogimet (https://www.ogimet.com/) y el API de provisión de datos AEMET OpenData (https://opendata.aemet.es/) y generar un dataset con las climatologías diarias de todas las estaciones de España.

Miembros del equipo

La práctica ha sido realizada por los estudiantes del Máster en Ciencia de Datos **Francisco Jesús Cárdenas Ruiz** y **Jesús Manuel Montero Garrido.**

Contexto

El objetivo de la ***observación meteorológica*** consiste en determinar o estimar de manera manual o mediante sensores el valor de diferentes parámetros físicos que permiten conocer el estado de la atmósfera y preparar análisis, predicciones y avisos meteorológicos, así como realizar la vigilancia del clima. Las observaciones meteorológicas desempeñan un papel fundamental en muchas aplicaciones sobre inundaciones, sequías, medio ambiente y recursos hídricos. Aunque las observaciones de la lluvia son las más utilizadas y difundidas, otros parámetros de interés son la temperatura del aire, la humedad y la velocidad del viento.

La competencia para realizar estas observaciones meteorológicas está asignada a los diferentes servicios meteorológicos nacionales y regionales.

En el caso de España uno de los organismos que realiza observación meteorológica es la ***Agencia Estatal de Meteorología*** (http:/www.aemet.es) que despliega, mantiene y opera diferentes tipos de redes de observación, que permiten medir las variables meteorológicas in situ, con instrumentos convencionales, o a distancia, mediante técnicas de teledetección. Estas redes de observación están constituidas por ***estaciones meteorológicas***. Según wikipedia, una estación climatológica es un área o zona determinada que ha sido destinada a la obtención, medición y procesamiento de los datos de los distintos fenómenos meteorológicos que se producen en la atmósfera. Existen diferentes tipos de estaciones meteorológicas y operadas por personal totalmente (***manuales***), automatizadas totalmente (***automáticas***) y con operación mixta (***semiautomáticas***).

Figura 1: Estación meteorológica semiautomática de AEMET consistente en un jardín meteorológico en el que se despliegan diferentes sensores para la medición de múltiples parámetros meteorológicos (Fuente: AEMET)

Asimismo, las estaciones meteorológicas pertenecen a diferentes redes de observación con funcionalidades y requerimientos muy diferentes y están identificadas unívocamente por un ***indicativo*** de la estación. En el caso de AEMET y para atender a los requerimientos de observación y predicción meteorológica a escala nacional, así como los compromisos internacionales, se dispone de una red de estaciones meteorológicas de alta densidad espacial en las que se llevan a cabo programas de medida. Esta red se denomina **red sinóptica de referencia (RS).** Las estaciones de de la RS están identificadas por un ***indicativo sinóptico*****.** Estos indicativos sinópticos son gestionados y asignados por OMM a nivel global. Además de la RS se dispone también de una red de estaciones meteorológicas de alta densidad espacial y con un nivel de resolución superior en las que se llevan a cabo programas de medida más completos que los realizados en la RS. Esta red se denomina **red climatológica (RC).**

Las estaciones de la RS están identificadas por un ***indicativo climatológico*****.** Las estaciones que no tienen indicativo sinóptico pero sí indicativo climatológico son las que constituyen la ***red mesoscalar.***

La observación meteorológica ha sido materia de colaboración internacional desde el siglo XIX. Los diferentes servicios meteorológicos nacionales son coordinados por la Organización Meteorológica Mundial, en adelante OMM (https://public.wmo.int/es), que opera el Sistema Global de Observación (GOS, por sus siglas en inglés). Los diferentes servicios meteorológicos nacionales contribuyen al GTS mediante la operación de las diferentes redes basadas en tierra, la realización de programas de medida normalizados y la difusión y el intercambio estandarizados de sus datos.

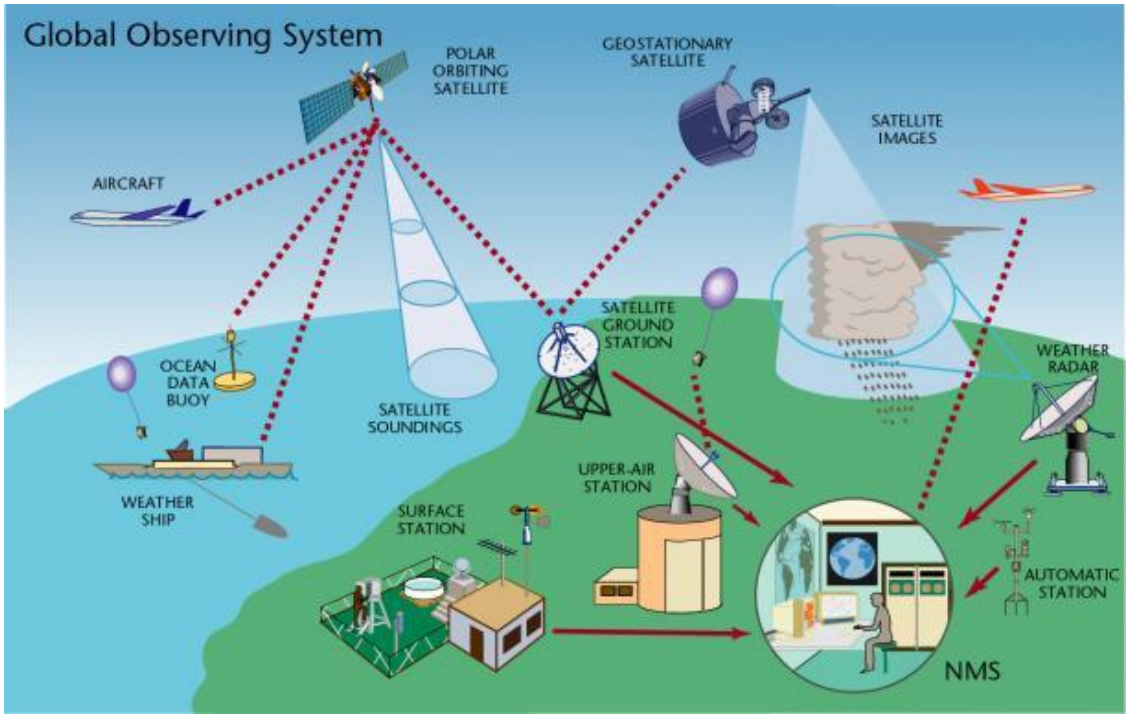


Figura 2: Esquema de las diferentes redes de observación que constituyen el Sistema Global de Observación de OMM, GOS por sus siglas en inglés (Fuente: Organización Meteorológica Mundial)

Los datos recopilados por los servicios meteorológicos a través de sus estaciones meteorológicas con indicativo sinóptico son difundidos internacionalmente por GOS para que estén disponibles para todos los servicios meteorológicos del mundo. Los datos recopilados por las estaciones meteorológicas y que no tienen asignado indicativo sinóptico no son difundidas internacionalmente y se difunden a nivel nacional por diferentes canales implementados por el servicio meteorológico.

A su vez, diferentes aficionados recolectan esta información y la publican en portales sectoriales. Ejemplo de uno de estos portales es OGIMET, que tal y como se indica en su página web, utiliza datos disponibles en la red de forma pública, fundamentalmente de la National Oceanic and Atmospheric Administration del Gobierno de Estados Unidos, (NOAA, https://www.noaa.gov/). El objetivo de este sitio es proporcionar a los usuarios información meteorológica actualizada procedente del GOS.

Por todo lo anterior, los datos de las estaciones meteorológicas climatológicas de AEMET que tienen indicativo sinóptico se encuentran disponibles en OGIMET. Por otra parte, un subconjunto de las estaciones meteorológicas sin indicativo sinóptico de AEMET están disponibles en el API AEMET OpenData. Este sistema es un API REST (Application Programming Interface. REpresentational State Transfer) a través del cual se pueden descargar gratuitamente los datos explicitados en el Anexo II de la resolución de 30 de diciembre de 2015 de AEMET.

Título

La práctica consiste en la construcción de un dataset que se denomina *Valores Climatológicos Diarios de España*, VCDE.

Descripción del Dataset VCDE

Los registros del VCDE están constituidos por los siguientes campos:

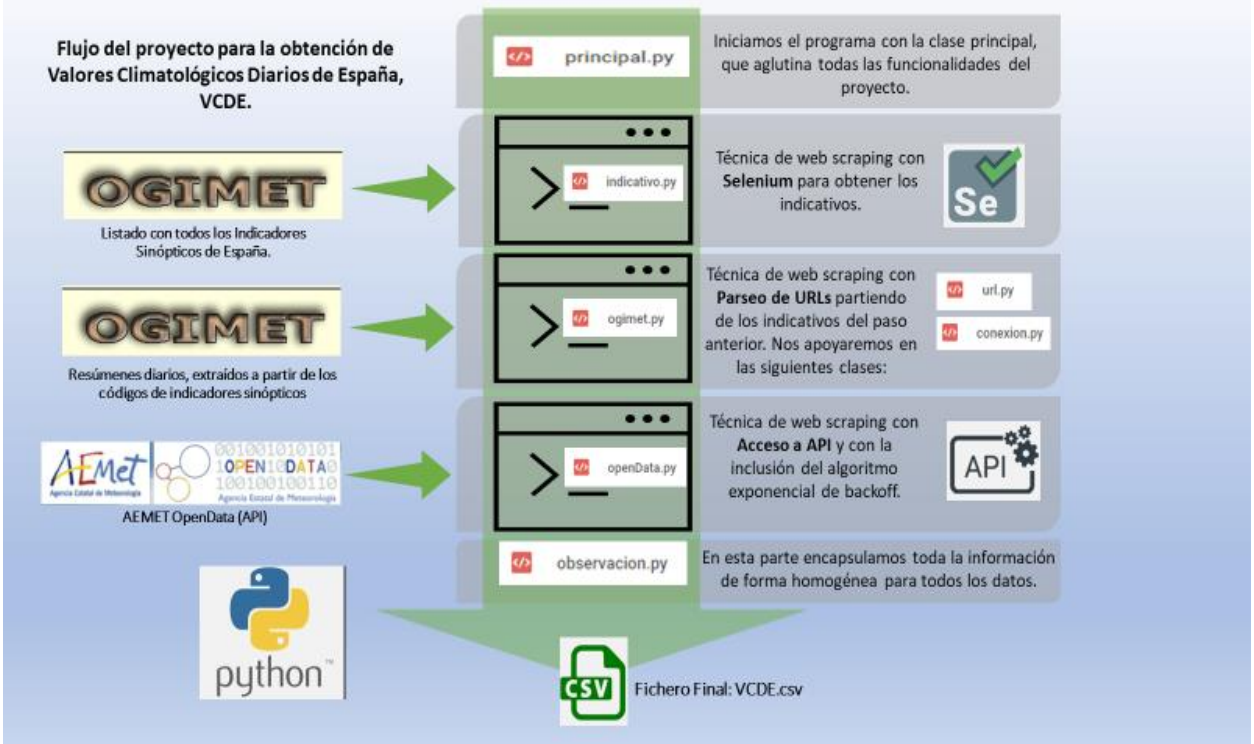
ID de campo	**Descripción**	**Tipo de dato**	**Formato**	**Unidad de medida**
Indicativo	Indicativo sinóptico o climatológico de la estación meteorológica	String		NA
Fecha	Día al que se refiere el valor climatológico	Date	AAAA-MM-DD	NA
Temperatura(C)_Max	Temperatura máxima	Float	Valor decimal con dos dígitos enteros y dos dígitos decimales	Grado centígrado

Temperatura(C)_Min	Temperatura mínima	Float	Valor decimal con dos dígitos enteros y dos dígitos decimales	Grado centígrado
Temperatura(C)_Med	Temperatura media	Float	Valor decimal con dos dígitos enteros y dos dígitos decimales	Grado centígrado
TdMed(C)	Temperatura de rocío media	Float	Valor decimal con dos dígitos enteros y dos dígitos decimales	Grado centígrado
Hr.Med(%)	Humedad relativa del aire	Float	Valor decimal con dos dígitos enteros y dos dígitos decimales	Grado centígrado
Vel._Dir.	Dirección del viento medio	String		Octante de dirección
Vel._Viento (km/h)	Velocidad media del viento	Float	Valor decimal con dos dígitos enteros y dos dígitos decimales	Km/h
Vel._Rch.	Racha máxima del viento	Float	Valor decimal con dos dígitos enteros y dos dígitos decimales	Km/h
Prec.(mm)	Precipitación	Float	Valor decimal con dos dígitos enteros y dos dígitos decimales	mm o l/m2

Pres.n. mar(Hp)	Presión reducida al nivel del mar	Float	Valor decimal con hasta cuatro dígitos enteros y dos dígitos decimales	HectoPascales
Pres.n. est(Hp)	Presión al nivel de la estación meteorológica	Float	Valor decimal con hasta cuatro dígitos enteros y dos dígitos decimales	HectoPascales
NubTotOct	Cantidad total de nubes	Int	Número entero que representa las octas de cielo cubiero por nubes de todo tipo	Octantes de cielo
NubbajOct	Cantidad total de nubes bajas	Int	Número entero que representa las octas de cielo cubiero por nubes bajas	Octantes de cielo
VisKm	Visibilidad atmosférica desde la estación	Float		Kilómetros
SolD-1(h)	Insolación Diaria. Número de horas de sol	Float		Horas
Esp.Nie.(cm)	Espesor de nieve acumulado en la estación	Float		Centímetros

Tabla 1: Campos del dataset VCDE

Representación gráfica



Contenido

Los campos del dataset VCDE están descritos en la Tabla 1. Los registros del VCDE están formados por los valores climatológicos diarios registrados por las estaciones de las redes sinóptica y climatológica (estaciones con indicativo sinóptico) y de las red mesoscalar (estaciones sin indicativo sinóptico). El período de tiempo es el comprendido entre el 1 de enero de 2021 hasta el 31 de diciembre de 2021.

La metodología de obtención de los datos es la indicada en el apartado *representación gráfica.*

Requerimientos software

Driver chrome (explicar configuración)

- Versión al menos de Google Chrome 100

Python:

- Intérprete python (probado con 3.10)

Librerías python:

- Selenium (4.1.3 o superior)
- Requests (2.27.1 o superior)
- Beautifulsoup4 (4.10.0 o superior)
- Backoff (1.11.1 o superior)

Descripción de los objetos creados:

src/principal.py: punto de entrada al programa. Inicia los diferentes procesos de scraping.

src/indicativo.py: contiene clase indicativo que encapsula la funcionalidad de conexión a través de Selenium a la siguiente dirección https://ogimet.com/indicativos.phtml. Automatiza la entrada de datos a un formulario para la obtención de los indicativos sinópticos de las estaciones meteorológicas de Agencia Estatal de Meteorología (AEMET). El resultado es un listado con los indicativos sinópticos de las estacions meteorológicas.

src/url.py: contiene clase url que encapsula la funcionalidad de la técnica de webscraping denominada variación de parámetros en una URL. En particular, esta clase encapsula la construcción y variación de parámetros de esta URL:

https://ogimet.com/cgi-bin/gsynres?ind=<indicativo>&ndays=<num_dias>&ano=<anio>&mes=<mes>&day=<dia>&hora=<hora>&ord=REV&enviar=Ver

Donde:

Indicativo es el indicativo sinóptico.

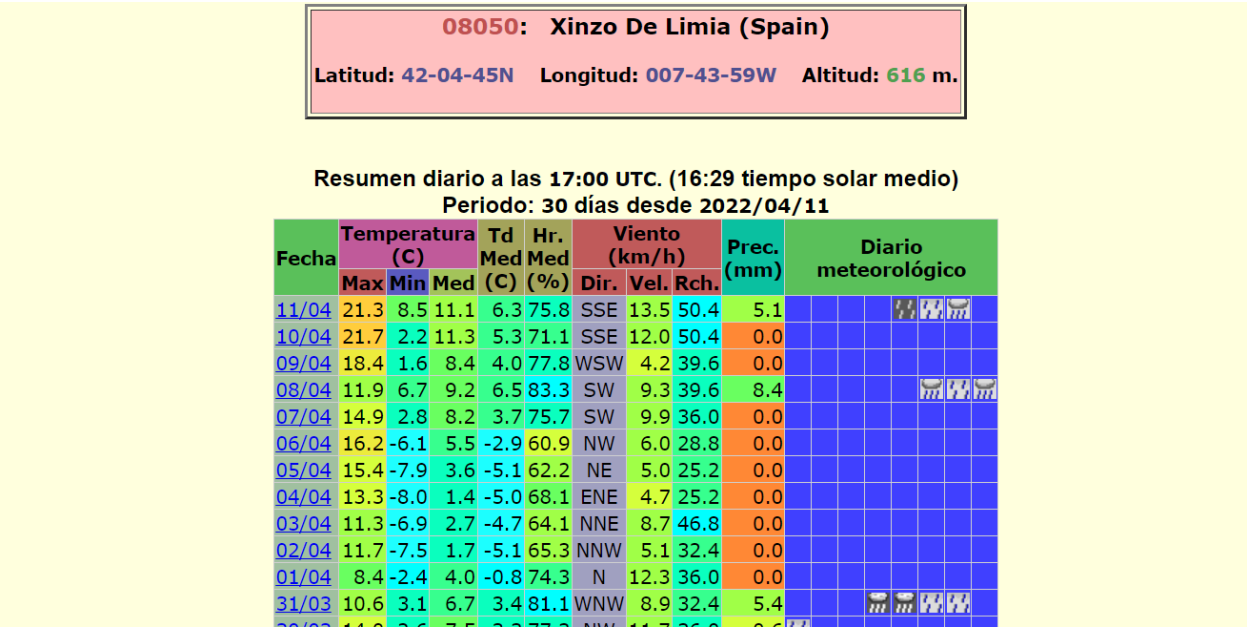
num_dias: número de días para los que se piden datos.

anio: año del que se piden datos.

mes: mes del que se piden datos.

hora: hora desde la que se piden datos.

El resultado de una petición a través de este formulario es la siguiente página web que es parseada para obtener los datos de los valores climatológicos.



08050: Xinzo De Limia (Spain)

Latitud: 42-04-45N Longitud: 007-43-59W Altitud: 616 m.

Resumen diario a las 17:00 UTC. (16:29 tiempo solar medio)
Periodo: 30 días desde 2022/04/11

Fecha	Temperatura (C)			Td Med (C)	Hr. Med (%)	Viento (km/h)			Prec. (mm)	Diario meteorológico
	Max	Min	Med			Dir.	Vel.	Rch.		
11/04	21.3	8.5	11.1	6.3	75.8	SSE	13.5	50.4	5.1	
10/04	21.7	2.2	11.3	5.3	71.1	SSE	12.0	50.4	0.0	
09/04	18.4	1.6	8.4	4.0	77.8	WSW	4.2	39.6	0.0	
08/04	11.9	6.7	9.2	6.5	83.3	SW	9.3	39.6	8.4	
07/04	14.9	2.8	8.2	3.7	75.7	SW	9.9	36.0	0.0	
06/04	16.2	-6.1	5.5	-2.9	60.9	NW	6.0	28.8	0.0	
05/04	15.4	-7.9	3.6	-5.1	62.2	NE	5.0	25.2	0.0	
04/04	13.3	-8.0	1.4	-5.0	68.1	ENE	4.7	25.2	0.0	
03/04	11.3	-6.9	2.7	-4.7	64.1	NNE	8.7	46.8	0.0	
02/04	11.7	-7.5	1.7	-5.1	65.3	NNW	5.1	32.4	0.0	
01/04	8.4	-2.4	4.0	-0.8	74.3	N	12.3	36.0	0.0	
31/03	10.6	3.1	6.7	3.4	81.1	WNW	8.9	32.4	5.4	

src/conexion.py: contiene la clase conexion que encapsula la funcionalidad para acceder a las diferentes URL. Tiene implementado el algoritmo Exponential Backoff (https://en.wikipedia.org/wiki/Exponential_backoff) con el objeto de separar los sucesivos intentos de conexión fallidos y evitar una posible denegación de servicio (DoS) de los recursos de OGIMET.

src/ogimet.py: contiene la case ogimet que encapsula los diferentes métodos que posibilitan el parseado del HTML de la página donde están los valores climatológicos diarios. El resultado de este análisis es una lista de valores climatológicos diarios a partir de los indicativos sinópticos obtenidos por la clase indicativo.

src/openData.py: contiene la clase openData que encapsula las peticiones al API AEMET OpenData (https://opendata.aemet.es). Tiene implementado el algoritmo Exponential Backoff (https://en.wikipedia.org/wiki/Exponential_backoff) con el objeto de separar los sucesivos intentos de conexión fallidos y evitar una posible denegación de servicio (DoS) de los recursos de AEMET OpenData.

src/observacion.py: contiene la clase *observación* que se trata de un objeto plano que contiene todas las propiedades que una observación diaria puede tener. Es utilizada para encapsular los valores climatológicos diarios obtenidos por la clase ogimet y por la clase opendata. Posibilita la homogeneización de todas los valores procedentes de todas las estaciones.

src/utilidades.py: contiene funciones auxiliares utilizadas por el resto de clases. Se trata de una especie de cajón desastre. En particular y entre otras funcionalidades, contiene una rutina para escribir las instancias de la clase observación al fichero de salida.

src/vcde.py: clase que contiene un método denominado obtención_registros_diarios del que se obtienen todos los valores climatológicos diarios de España procedentes de OGIMET y AEMET OpenData.

Agradecimientos

Los datos del dataset VCDE han sido obtenidos de los siguientes orígenes:

- OGIMET
- Agencia Estatal de Meteorología de España

Inspiración

La web y el API de OpenData permiten obtener datos meteorológicos libres en virtual de la resolución 40 de la Organización Meteorológica Mundial. Pueden ser reutilizados para cualquier tipo de análisis o exploración.

El conjunto de datos es interesante por las siguientes razones:

- Permite realizar estudios climatológicos a partir de datos obtenidos de las estaciones.
- Evolución del cambio climático.
- La fiabilidad del dato está respaldada por entes públicos y mundiales como los diferentes servicios climatológicos nacionales.

Licencia

Se ha optado por “Creative Commons Zero v1.0 Universal” por ser una de las licencias más utilizadas que permiten compartir, adaptar y modificar a partir del material publicado, siempre y cuando se dé crédito al creador original y no se apliquen restricciones adicionales ya sean tecnológicas o legales a los posteriores proyectos derivados del original.

Código

El código, dataset y documentación se encuentran en GitHub en el siguiente enlace:

https://github.com/fjcardenasuoc/Valores_Climatologicos_Diarios

https://github.com/jmmonterog/valores_climatologicos_diarios

Dataset

El dataset ha sido publicado en Zenodo en el siguiente enlace:

https://zenodo.org/record/6450171#.YlSLG8hByUk

Vídeo

El vídeo de la práctica está disponible en el siguiente enlace de Google Drive de la UOC:

https://drive.google.com/file/d/1R6ccEuP9qNtRx5AHImEbr7cpaBTIVQpU/view?usp=sharing

Contribuciones	Firma
Investigación previa	FJCR, JMMG
Redacción de las respuestas	FJCR, JMMG
Desarrollo del código	FJCR, JMMG

Bibliografía

- Subirats, L., Calvo, M. (2018). Web Scraping. Editorial UOC.
- Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 2 Scraping the Data.
- Sene, K. (2016). Meteorological Observations. In: Hydrometeorology. Springer, Cham. https://doi.org/10.1007/978-3-319-23546-2_2
- Redes de observación de superficie y en altura: http://www.aemet.es/es/idi/observacion/observacion_convencional
- OGIMET: https://ogimet.com/
- AEMET OpenData: https://opendata.aemet.es
- Resolución de 30 de diciembre de 2015, de la Agencia Estatal de Meteorología, por la que se establecen los precios públicos que han de regir la prestación de servicios meteorológicos y climatológicos.
- Wikipedia
- WMO Global Observing System: https://public.wmo.int/en/programmes/global-observing-system
- Resolución 40 de la OMM: https://community.wmo.int/resolution-40
- Nota Legal de AEMET: http://www.aemet.es/es/nota_legal
- Intervalos de grados sexagesimales correspondientes a rumbos. https://www.todoababor.es/historia/maniobras-de-un-buque-de-vela-conceptos-basicos/