PRÁCTICA 1. WEBSCRAPPING

Autores:

César Fernández Domínguez

Isabel Fernández Esparza

César Fernández Domínguez
Isabel Fernández Esparza

Contents

César Fernández Domínguez
Isabel Fernández Esparza

Contexto

Poder predecir el tiempo ha sido siempre una de las necesidades recurrentes del ser humano puesto que la climatología condiciona nuestras actividades y forma de vida. Basándonos en esta premisa hemos pensado en desarrollar un proyecto de recopilación de datos meteorológicos históricos de los últimos 50 años tanto a nivel nacional como a nivel mundial y poder utilizarlos para realizar predicciones meteorológicas a corto y medio plazo. Pensamos que estos datos pueden ser de gran utilidad para poder planificar viajes o cualquier otro tipo de evento al que le influya la climatología, y que por tanto, podrían servir a un cliente potencial como puede ser una agencia de viajes o una empresa organizadora de eventos.

En este caso, la descarga de datos se ha hecho desde páginas que ya los utilizan para hacer su propio negocio. Esto se ha realizado así, simplemente, para probar el uso de técnicas de *WebScraping*. Sin embargo, después de profundizar un poco más en el contexto de estudios climáticos, hemos visto que existen servicios públicos que ofrecen datos de observaciones climáticas de forma pública y gratuita. Como por ejemplo:

1. AEMET – OpenData: http://www.aemet.es/es/portada
2. NOAA- Agencia estadounidense https://www.noaa.gov/
3. WMO: Organización meteorológica mundial: https://public.wmo.int/en

En el caso de la página de la *AEMET-OpenData*, ofrece una interfaz *API-Rest* para acceder a los datos observados por cada una de sus estaciones repartidas por el territorio español. El hecho de que estos datos sean abiertos, hace posible el uso de la información obtenida según lo establecido en la Ley 18/2015 que regula el uso de la información elaborada y custodiada por la *AEMET* por personas físicas o jurídicas con fines comerciales o no comerciales.
Por su parte, en la página de la administración americana del National Oceanic and Atmospheric Administration (*NOAA*) se pueden descargar datos de su base de datos meteorológica. Esta dispone de datos de más de 80000 estaciones meteorológicas repartidas por todo el planeta. En algunos casos, sobre todo en estaciones estadounidenses, estos datos son tomados desde hace más de 200 años. En este caso, el acceso a los datos se realiza a través de un servicio ftp: http://www1.ncdc.noaa.gov/pub/

El análisis y procesamiento de estos datos climáticos históricos puede traducirse en un proyecto en el que se genere una herramienta que permita una correcta toma de decisiones en cualquier tema relacionado con el clima e incluso con el cambio climático.

Son muchos los datos que vamos a recopilar en este proyecto: datos de temperaturas máximas y mínimas, nivel de precipitaciones, velocidad del viento, calidad del aire,…de cada una de las estaciones en cada una de las provincias españolas y en las diferentes estaciones distribuidas por el mundo. Con un proyecto de *Big Data* podremos optimizar la gestión y el análisis de estos datos para poder extraer las conclusiones necesarias.

Pensamos que son muchos los clientes potenciales que pueden beneficiarse de este estudio:

1. Agencias de viajes o empresas encargadas de la organización de eventos: Podrán servirse de los datos analizados para poder predecir con mayor exactitud qué lugares

son más recomendables en cada fecha para viajes turísticos o para organización de eventos.

2. Organismos nacionales o internacionales que pueda utilizar los resultados de estos análisis para obtener información del cambio climático y poder llevar a cabo políticas o medidas para paliar el impacto. Se pueden realizar estudios para poder prevenir desastres naturales atendiendo a los datos históricos de los que disponemos.

En nuestro caso, hemos realizado las técnicas de rastreo desde dos páginas web diferentes, que ya utilizan estos datos para uso comercial. Estas páginas son:

1. Datos nacionales: https://datosclima.es/
2. Datos internacionales: https://www.tutiempo.net/clima

Definir un título para el Dataset

Hemos elegido el título: **Histórico de datos climáticos para operadores turísticos**

Descripción del Dataset

Como hemos indicado anteriormente, el objeto de esta práctica es recolectar datos meteorológicos, recogidos en distintas estaciones meteorológicas, tanto de España como del resto de países del mundo. Inicialmente pensábamos obtener ambas informaciones de la misma página: https://datosclima.es/ pero la información mundial gratuita disponible estaba limitada a pocas fechas y pensamos que para la idea que tenemos de un proyecto de Big Data es importante contar con una gran cantidad de información de datos históricos. Por eso decidimos recuperar los datos mundiales de la página: https://www.tutiempo.net/clima

Esto hace que la información recuperada a nivel nacional no sea exacta a la recuperada a nivel internacional. Tenemos unos datos comunes para ambos casos para cada día:

- Temperatura máxima: Valor numérico medido en grados centígrados
- Temperatura mínima: valor numérico medido en grados centígrados
- Temperatura media: valor numérico medido en grados centígrados
- Velocidad media del viento: valor numérico medido en km/h
- Velocidad máxima y mínima del viento
- Nivel de precipitaciones: valor numérico medido en mm.

Además, con respecto a los datos nacionales tenemos la siguiente información:

- Dirección del viento
- Horas de presión máxima y mínima
- Horas de temperatura máxima y mínima
- Horas de racha máxima y mínima
- Horas de sol
- Presión máxima
- Presión mínima

Para el caso de los datos internacionales, tenemos los siguientes:

- Presión atmosférica a nivel del mar
- Humedad relativa media
- Visibilidad media
- Indicador de si hubo lluvia, nieve, niebla o tormenta (valor lógico, indica solo si la hubo o no en el día señalado)

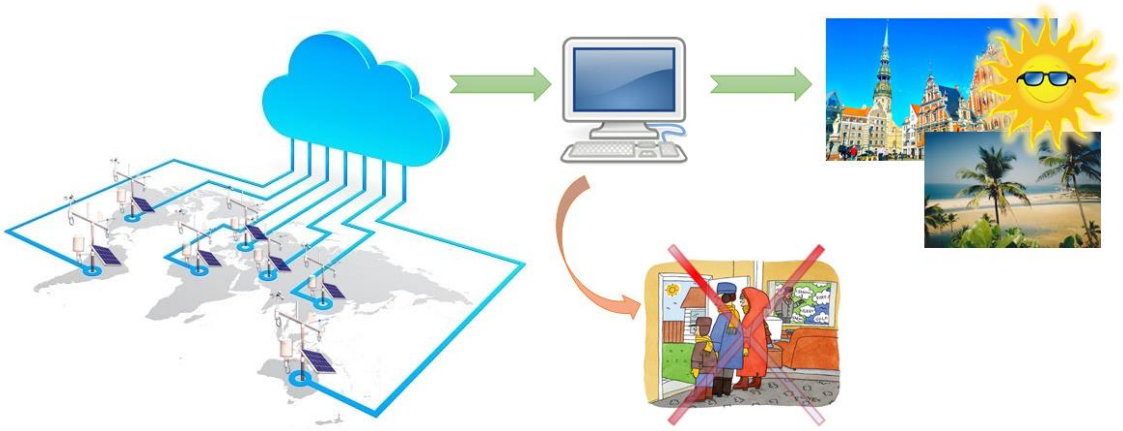
A la hora de llevar a cabo nuestro proyecto tendremos que hacer una limpieza de nuestros datos para analizar si aquellos que no son comunes pueden ser relevantes o no para nuestros intereses. En un primer lugar se puede pensar que es mejor almacenar todos los datos disponibles aunque puedan ocupar más espacio puesto que nos pueden servir para realizar mejores predicciones y para hacer análisis desde distintos puntos de vista.

Como resultado de la extracción tendremos tres ficheros csv, uno con los datos climáticos nacionales, otro con los datos de las estaciones nacionales y otro con los internacionales. Los datos nacionales estarán agrupados por provincias y en cada una de ellas incluiremos todos los datos disponibles. En el caso del csv con los datos mundiales los agruparemos por continentes, países y dentro de cada país por estaciones. Estamos duplicando la información de las estaciones por cada día de cada mes pero hemos pensado hacerlo así para tener siempre la referencia de cuál es la estación de la que estamos analizando los datos. Dado que para cada estación se dispone de datos de una antigüedad diferente tendremos que realizar un proceso de limpieza posterior para eliminar aquellas filas para las que no existen registros y tendremos también que analizar aquellos casos en los que algún valor de todos los recuperados esté vacío ver cuál es la mejor estrategia a seguir. La idea es que el conjunto de datos que manejemos contenga, al menos, un valor en cualquiera de las dimensiones que estamos tratando.

Una vez limpiados los datos nacionales e internacionales tendremos que realizar un formateo de los mismos para agrupar aquellos que sean comunes y poder tener una visión global de las condiciones climatológicas por día, mes y año entre los diferentes países o provincias.

Representación gráfica

Hemos pensado en el siguiente diagrama para representar de forma gráfica lo que buscamos con este trabajo. A partir de los datos climáticos recogidos para cada una de las estaciones meteorológicas repartidas por el mundo y accesibles públicamente a través de páginas web se construye un *dataset* que facilite el análisis de cuáles pueden ser las mejores fechas para planificar actividades turísticas en determinados lugares.



Contenido

Para la recolección de datos, desde las páginas web mencionadas anteriormente, se ha realizado un programa en *Python* utilizando el *Framework* de rastreo web de código abierto *Scrapy*. Este *framework* permite realizar rápidamente programas bastante complejos de rastreo de páginas web utilizando *Python* analizando cada página web y navegando a través de su estructura html para extraer los datos que necesitamos. Funciona tanto en plataformas *Linux*, *Windows* o *MacOS*.
Hemos usado scrapy 1.6 (https://scrapy.org/download/) para realizar el rastreo. Además, este framework tiene también una consola interactiva en ipython que permite ir haciendo pruebas con la extracción de los datos. Nos hemos servido de la documentación https://docs.scrapy.org/en/latest/ para aprender a manejarnos con la herramienta.

Para esta práctica se han programado los "spiders" necesarios para acceder a las dos páginas web mencionadas arriba. Para una de ellas se accede a las sucesivas páginas de datos a través de llamadas a un formulario *PHP*. En la otra página, se va accediendo a distintas páginas que finalmente nos llevan a los datos finales.

Una vez encontrada la inspiración y el objetivo de nuestra práctica y una vez elegidas las fuentes de datos, lo primero que hemos tenido que hacer ha sido profundizar en cada una de las páginas web que hemos rastreado para poder entender cuál es la forma en la que se puede acceder a los datos que cada una de ellas nos ofrecían. Para ello, nos hemos servido de las funcionalidades de inspección de los navegadores *Chrome* y *Firefox*. Estas herramientas nos permiten explorar el formato y el contenido de las páginas web visualizadas y poder hacer el seguimiento de las comunicaciones entre cada uno de los enlaces que nos van proporcionando los datos necesarios.

Una vez entendida la estructura de las páginas web analizadas, nos hemos servido de la documentación https://docs.scrapy.org/en/latest/ para aprender a manejarnos con la

herramienta. Tal y como se explica en esta documentación hemos generados dos spiders para la extracción de los datos mundiales como nacionales y hemos definido en el fichero ítems.py aquella información que queremos extraer. Además, dado que, sobre todo en el caso de la información mundial los datos estaban bastante desperdigados por la web, nos hemos servido del pipeline para procesar esos datos y extraerlos en el formato que necesitamos en nuestro csv.

A continuación pasamos a describir el contenido de cada uno de los conjuntos de datos extraídos de las páginas web rastreadas.

Conjunto de datos de datosclima.es

A partir de los datos recopilados de la página de "datosclima.es" se generan dos conjuntos de datos diferentes.

- ✓ Un primer conjunto de datos que contiene información de localización y disponibilidad de muestras para cada una de las estaciones meteorológicas del estado español.
- ✓ El segundo conjunto de datos contendrá las medidas de temperatura mínima y máxima, velocidad del viento, presión atmosférica, horas de sol y precipitación observadas en cada una de las estaciones para cada una de las fechas en las que se dispone de datos.

Las siguientes tablas muestran los atributos de cada uno de estos conjuntos de datos:

Campo	Descripción	Tipo de dato	Unidades
Provincia	Provincia donde se ubica la estación meteorológica	Texto	
estación	Nombre o ubicación de la estación meteorológica en la que se han recogido los datos	Texto	
Día	Día al cual se refieren los siguientes datos	dd/mm/aaaa	Día, mes, año
tempMax	Temperatura máxima registrada en el día	Doble precisión	ºC
horaTempMax	Hora en la cual se registra la temperatura máxima en la	hh:mm	Hora y minuto
tempMin	Temperatura mínima registrada en el día	Doble precisión	ºC
horaTempMin	Hora en la cual se registra la temperatura mínima en la estación meteorológica	hh:mm	Hora y minuto
presMax	Presión atmosférica máxima registrada durante el día	Doble precisión	mb estación meteoroló gica
horaPresMax	Hora en la cual se registra la presión atmosférica máxima	hh	hora
presMin	Presión atmosférica mínima registrada durante el día	Doble precisión	Mb

horaPresMin	Hora en la cual se registra la presión atmosférica mínima	hh	Hora
rachaMax	Racha máxima de viento registrada en el día	Doble precisión	m/s
horaRachaMax	Hora a la cual se registra la máxima racha de viento	Hh:mm	Hora y minuto
dirViento	Dirección del viento	entero	
velMedia	Velocidad media del viento durante todo el día	Doble precisión	m/s
horasSol	Horas de sol registradas durante todo el día	Doble precisión	Horas
precipitacion	Precipitación registrada durante el día	Doble precisión	l/m2

Datos climáticos nacionales

Campo	Descripción	Tipo de Dato	Unidades
Provincia	Provincia donde se ubica la estación meteorológica	texto	
Estación	Nombre o ubicación de la estación meteorológica en la que se han recogido los datos	texto	
Indicativo	Identificador de la estación. Útil para relacionar este conjunto de datos con los anteriores	texto	
Latitud	Latitud a la que se encuentra la estación meteorológica	Doble precisión	grados
longitud	Longitud a la que se encuentra la estación meteorológica	Doble precisión	Grados
Altitud	Altitud a la que se encuentra la estación meteorológica	Doble precisión	Metros
inicioDatos	Fecha de inicio de disponibilidad de datos meteorológicos para la estación	dd/mm/aaaa	Día, mes y año
finDatos	Fecha de fin de disponibilidad de datos meteorológicos para la estación	dd/mm/aaaa	Día, mes y año

Datos estaciones nacionales

Conjunto de datos de tutiempo.net

Para el caso de los datos climáticos mundiales, recopilados usando *scrapy* desde la página web ("tutiempo.net") generaremos un único conjunto de datos en el que incluiremos tanto los datos de la estación meteorológica como los datos climáticos recogidos. Estos datos serán:

Campo	Descripción	Tipo de dato	Unidades
FG	Indica si hubo niebla	Boolean	True or false
H	Humedad relativa media	Numérico	porcetaje
PP	Precipitación total de lluvia y/o nieve derretida	Doble precisión	mm
RA	Indica si hubo lluvia o llovizna	Boolean	True or false
SLP	Presión atmosférica a nivel del mar		
SN	Indicador de nieve		
T	Temperatura media	Doble precisión	ºC
TM	Temperatura máxima	Doble precisión	ºC
TS			
Tm	Tempertura mínima	Doble precisión	ºC
V	Velocidad media del viento	Doble precisión	Km/h
VG	Velocidad de ráfagas máximas de viento	Doble precisión	Km/h
VM	Velocidad máxima sostenida del viento	Doble precisión	Km/h
VV	Visibilidad media	Doble precisión	Km
Altitud	Altitud a la que se encuentra la estación meteorológica	Doble precisión	metros
Continente	Continente de la estación	Texto	
Día	Día de la medición	Numérico	
Estación	Estación meteorológica	Texto	
latitud	Latitud a la que se encuentra la estación meteorológica	Doble precisión	grados
longitud	Longitud a la que se encuentra la estación meteorológica	Doble precisión	grados
mes	Mes de la medición	Texto	

Datos estaciones/clima mundiales

Agradecimientos

Damos las gracias al gestor de la página "datosclima.es" por haber ido recopilado diariamente datos meteorológicos ofrecidos por la Agencia Española de Meteorología (AEMET) para más de 800 estaciones meteorológicas repartidas por todo el territorio español.
Esta información es ofrecida por el AEMET en abierto en la página web: (http://www.aemet.es/es/eltiempo/observacion/ultimosdatos). En esta página solamente es posible consultar los datos recogidos en cada estación en los últimos días, no siendo posible acceder a un histórico de los mismos. Por supuesto, también debemos de dar las gracias al AEMET por ofrecer estos datos en abierto a través de su página web.
En el caso del sitio web la página web www.tutiempo.net vemos que este ofrece un servicio similar al planteado para esta práctica: https://www.tutiempo.net/viajar/. Este servicio es ofrecido para los principales destinos turísticos mundiales. Este ejemplo muestra claramente cuál es el potencial de la propuesta que se ha pretendido llevar a cabo con esta práctica.
Sin embargo, como ya se ha mencionado anteriormente, este proyecto debería partir de las fuentes originales de los datos, para así disponer de datos completos y libres de restricciones.

Inspiración

A la hora de recolectar este conjunto de datos meteorológicos se ha pensado, inicialmente, en su utilización en el ámbito de los operadores turísticos, concretamente en una agencia de turismo. Sin embargo, este conjunto de datos podría ser de gran utilidad para otras muchas actividades, como, por ejemplo, la organización de todo tipo de eventos (bodas, conciertos, congresos, etc…) que puedan verse condicionados por factores climatológicos.
La idea es que, mediante la aplicación de técnicas de minería de datos sobre los datos recolectados, la agencia de viajes debería ser capaz de:

- aconsejar a sus clientes sobre cuáles serían las mejores épocas del año para viajar a un determinado país o, dentro del ámbito de España, a una determinada provincia.
- planificar mejor su oferta de destinos turísticos, encontrando la mejor temporada para organizar sus viajes o excursiones, sin que se vean afectados por factores climáticos adversos como: lluvia, viento o temperaturas extremas. Esto también incluye la búsqueda de aquellas épocas del año que, aun no siendo las habituales para el turismo en una determinada zona, podrían ofrecer unas buenas condiciones climatológicas que permitan ofrecer paquetes turísticos a un mejor precio.

Este proyecto podría ampliarse para obtener más información recopilando datos de: establecimientos hoteleros, restaurantes, ferias y fiestas populares en los destinos, oferta turística de los lugares a visitar (museos, puntos de interés turísticos, parques, etc…), así como, información de conflictos abiertos en cada país, rutas aéreas, etc…

El análisis climático ha sido siempre una inquietud entre la población por lo que son bastantes los estudios que se han realizado en este aspecto. Además del estudio que se hace en la propia página tutiempo.net y que ya se ha comentado anteriormente, hemos encontrado algunos artículos en los que se analizan datos climáticos un una región concreta como por ejemplo:

- Análisis estadístico de datos meteorológicos mensuales y diarios para la determinación de variabilidad climática y cambio climático en el distrito metropolitano de Quito (Ecuador)
- Análisis climático de temperaturas de "tu" zona: una página de interés climático
- Tratamiento y estudio de series de temperatura para su aplicación en salud pública. El caso de Castilla-La Mancha
- Análisis meteorológico en Yopal (Colombia)

Pensamos que la idea de nuestro proyecto (extrayendo los datos como ya se ha comentado de las fuentes oficiales) puede ser de interés dado que analiza los datos desde un punto de vista global, no únicamente centrado en una región concreta y pueden establecerse conclusiones más relevantes. Se puede establecer una relación de forma que la climatología de regiones o países pueda repercutir también a regiones con las que hace frontera.

Licencia

- *Released Under CC0*: Public Domain License: Información liberada a nivel mundial. Se puede copiar, modificar, distribuir la obra y hacer comunicación pública incluso para fines comerciales incluso sin pedir permiso.
- *Released Under CC BY-NC-SA 4.0 License*: La información se puede compartir en cualquier medio o formato sin fines comerciales y se puede adaptar reconociendo siempre la autoría y proporcionando enlace a la licencia si se producen cambios.
- *Released Under CC BY-SA 4.0 License*: La información se puede copiar, redistribuir en cualquier medio o formato incluso con fines comerciales. Sería suficiente con mencionar la autoría de la información.
- *Database released under Open Database License*, individual contents under Database Contents License: Permite que los usuarios compartan, modifiquen y usen libremente las bases de datos.
- *Other (specified above)*
- *Unknown License*

Como hemos ido comentando a lo largo de esta memoria, para poder utilizar los datos con fines comerciales deberíamos haberlos obtenido de las fuentes abiertas tales como *AEMET*, *NOAA*, la información que se ofrece en estas páginas es pública y podríamos utilizarla con una licencia de tipo *Attributo Share Alike*, es decir una licencia: **Released Under CC BY-SA 4.0 License** que permite copiar y redistribuir el material en cualquier medio o formato y además posibilita la mezcla, transformación y creación de información a partir de la recopilada para poder utilizarse incluso con fines comerciales. Simplemente tendríamos que mencionar el origen de los datos.

Para nuestro caso, la información recopilada a través de páginas como tutiempo.net está restringida y no puede copiarse, reutilizarse, explotarse o reproducirse para propósitos públicos o comerciales, luego podríamos licenciarla bajo la licencia **Released Under CC BY-NC-SA 4.0 License**

Código

El código fuente de la práctica desarrollada para poder recopilar los datos de información meteorológica, se pueden encontrar en este enlace de GitHub.

Dataset

Los distintos conjuntos de datos en formato csv se encuentran en el siguiente enlace de GitHub

Contribución por participantes

CONTRIBUCIONES	FIRMA
Investigación Previa	César Fernández Domínguez Isabel Fernández Esparza
Redacción de las respuestas	César Fernández Domínguez Isabel Fernández Esparza
Desarrollo del código	César Fernández Domínguez Isabel Fernández Esparza