ESTUDIANTES: PAULA LEÓN GIL-GILBERNAU (PLG) de aula 1
SILVINA GUIJARRO DOMINGO (SGD) de aula 2

Práctica 1: WEBSCRAPING NIVELES POLENES

El objetivo de esta actividad es la creación de un dataset a partir de los datos contenidos en una web. Para su realización, se ha realizado el trabajo de forma conjunta entre Paula León Gil-Gibernau (PLG) del aula 1 y Silvina Guijarro Domingo (SGD) del aula 2 y hemos seguido los siguientes puntos:

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

Este conjunto de datos proporciona información sobre los niveles actuales y una predicción de los niveles ambientales de pólenes y esporas alergénicos , recogidos en las diferentes estaciones de agrobiología repartidas por Cataluña. Estos datos están recogidos semanalmente en la web del Punto de Información Aerobiológica de Cataluña

Esta web ha sido creada por la Red Aerobiológica de Cataluña, cuyo principal objetivo es publicar y divulgar el estado actual y la predicción de futuro de los niveles de pólenes alergógenos en diferentes puntos de Cataluña.

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

Niveles de pólenes ambientales en Cataluña

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 título elegido).

En este dataset se muestra los resultados de los niveles actuales para los principales pólenes y esporas ambientales en la semana en curso y una previsión de cómo se modificará la polinización para cada planta en las próximas semanas; en diferentes puntos de Cataluña.

Para ello se han usado los datos extraídos de forma específica para cada polen y espora en las diferentes estaciones de la red de aerobiologia de Cataluña,

localizadas en Barcelona, Bellaterra, Girona, Lleida, Manresa, Roquetes, Tarragona, Vielha y Planes de Son.

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

Imagen 1: Desprendimiento y liberación ambiental de polén. Imagen extraída del proyecto ATMOSENV-Aerobiologia

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

Esta base de datos ha sido realizada gracias a la extracción de datos proporcionados en la web: https://lap.uab.cat/aerobiologia/es/forecast/catalunya. Esta web proporciona datos actualizados semanalmente sobre los niveles de los principales pólenes y esporas alergénicos contenidos en el aire ambiental para diferentes puntos de Cataluña. Proporcionándonos un registro actualizado de los niveles ambientales para cada alérgeno y una predicción de como será en los próximos días la polinización ambiental para cada espora y polen, en las diferentes estaciones de Barcelona, Bellaterra, Girona, Lleida, Manresa, Roquetes, Tarragona, Vielha y Planes de Son.

ESTUDIANTES: PAULA LEÓN GIL-GILBERNAU (PLG) de aula 1
SILVINA GUIJARRO DOMINGO (SGD) de aula 2

Los datos extraídos en este dataset muestran los resultados para la semana en curso, en nuestro caso la semana del 15 al 22 de Abril del 2019, para cada estación y están distribuidos en los atributos:

nombre de taxón: se especifica nombre del polen o espora (con la planta de origen entre paréntesis), codificada como una variable categórica nominal.

niveles ambientales para cada taxón en cada estación: codificados como una variable ordinal numérica, recogidos en un rango de 0 (nivel ambiental nulo) al 4 (nivel ambientales máximo),

predicción para cada estación: codificados como una variable ordinal alfanumérica, definidos como aumenta(A), estable (=), descenso(D) o situación excepcional(¡).

y estación de aerobiologia: recoge el nombre de localización de cada estación y está codificado como una variable nominal en formato cadena.

Los datos han sido extraídos de la página web: https://lap.uab.cat/aerobiologia/es/, usando técnicas de Web Scraping mediante el lenguaje de programación Python.

Para ello primero se ha evaluado el fichero robots.txt mediante librería robotparser, y se ha consultado el permiso de acceso (aunque no hay robots.txt), posteriormente se ha analizado el mapa de la web de estudio, se ha evaluado el propietario de la web mediante librería whois y para finalizar se ha creado un crawler que extrae los datos contenidos en la tabla de la web y que los importa a un dataset en formato csv. Para realizar dicho crawler se han usado las librerías Pandas, BeautifulSoup y Urllib.

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

El propietario de los datos presentados pertenecen al Punto de Información Aerobilógica (PIA) de la Red Aerobiólogica de Cataluña. Todos los contenidos

de dicha web son del PIA y permite el uso a terceros en base a la licencia de Creative Commons Reconocimiento-NoComercial-CompartirIgual 4.0 Internacional (CC BY-NC-SA 4.0). Así mismo, la imagen usada para la representación de este trabajo pertenece al grupo de investigación ATMOSENV-Aerobiología

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

En los últimos años se ha producido un aumento constante en la prevalencia de enfermedades alérgicas a nivel mundial, de forma que entre un 30-40% de la población mundial está afectada por una o más enfermedades alérgicas. Entre los diferentes tipos de alergias, sabemos que las alergias ambientales es una de los tipos más frecuentes y es causa de diferentes problemas de salud como angioedema, reacciones cutáneas, rinitis, conjuntivitis y enfermedades respiratorias como el asma y descompensaciones de enfermedad pulmonar obstructiva crónica.

Diferentes estudios, postulan que este incremento de reacciones alérgicas ambientales es debido al incremento de contaminación y a los cambios climatológicos que produce la contaminación y el cambio climático. Así los datos proporcionados en esta base de datos, combinados con otros datos, puede ayudar a evaluar como los cambios en las condiciones climatológicas y los cambios de polución ambiental afectaran a la polinización, al recuentos de polen ambiental, la presencia de insectos y la presencia de hongos asociados a las enfermedades alérgicas ambientales.

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

La licencia seleccionada es la CC BY-NC-SA 4.0 License, puesto que permite copiar, distribuir, exhibir y representar la obra y hacer obras derivadas si se cumple:

- se cita al autor de la fuente original,

- se usa sin fines comerciales, ya que nuestra página está destinada a la provisión de información y a la investigación y por tanto no tiene un fin comercial sino meramente informativo y de investigación
- y se distribuye la obra derivada bajo las mismas licencias que la fuente original.

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

Se adjunta el código en Python

10. Dataset. Presentar el dataset en formato CSV

Se adjuntan los en formato CSV. Los datos se separan ‘,’

Contribución a la práctica

CONTRIBUCIONES	FIRMAS
Investigación previa	PLG aula 1, SGD aula 2*
Redacción de las respuestas	PLG aula 1, SGD aula 2*
Desarrollo del codigo	PLG aula 1, SGD aula 2*

*cada integrante ha firmado con sus iniciales

Recursos

Para la realización de esta PEC se ha usado la siguiente bibliografía:

- información sobre alergias:

Libro Blanco sobre Alergia de la WAO.https://www.worldallergy.org/UserFiles/file/WWBOA_Executive-Summary_Spanish.pdf

- información sobre licencias corporativas:

https://es.wikipedia.org/wiki/Licencias_Creative_Commons#Licencias
https://creativecommons.org/licenses/?lang=es_ES
https://opendatacommons.org/licenses/odbl/1-0/index.html
https://en.wikipedia.org/wiki/Open_Database_License

- Información sobre websacraping

Subirats, L., Calvo, M. (2018). Web Scraping. Editorial UOC.

Masip, D. El lenguaje Python. Editorial UOC. Tipología y ciclo de vida de los datos Práctica 1

Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 2.Scraping the Data.

Simon Munzert, Christian Rubba, Peter Meißner, Dominic Nyhuis. (2015). Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining. John Wiley & Sons.