

Práctica 1

Javier Díaz Domínguez.

Contexto

Para la obtención de la información, se ha optado por buscar un tema de actualidad, y del que se pudieran obtener datos **libres** y de **fácil acceso**. Por ello, la decisión final fue la de optar por explorar los diferentes conjuntos de datos que existían en la web del **Instituto Geográfico Nacional** (www.ign.es), en busca de un conjunto de datos que pudiera estar relacionado de alguna manera con el accidente volcánico que sucede ahora mismo en la isla de Palma. Finalmente se encuentra un apartado denominado **vigilancia volcánica**, donde se indexan diferentes recursos relacionados con este tema. Entre ellos, aparece un apartado de información de los últimos terremotos, donde se ofrece una visualización en formato de tabla con los datos sobre los terremotos ocurridos en los últimos diez días en las Islas Canarias.

Título

El título escogido para el dataset es el de “**Seísmos registrados en los últimos 10 días en las Islas Canarias**”.

Descripción

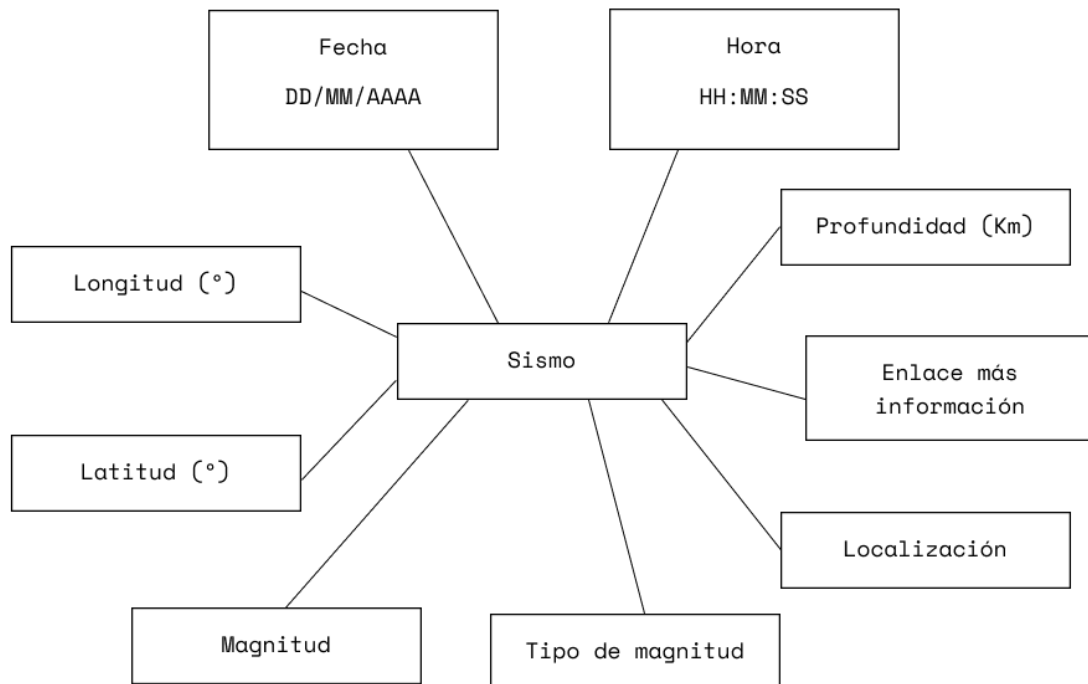
Los datos recogidos consisten en una lista de entradas, donde cada una de ellas representa un terremoto ocurrido en las Islas Canarias durante los últimos 10 días.

Para cada uno de los terremotos, se recogen los siguientes datos:

- Fecha en la que ocurrió el evento
- Hora en la que ocurrió el evento
- Latitud del epicentro del terremoto
- Longitud del epicentro del terremoto

- Profundidad del hipocentro del terremoto
- Magnitud del terremoto
- Localización del terremoto
- Enlace para información detallada

Representación gráfica



Contenido

En primer lugar, procedemos a explicar cada uno de los campos que contiene el dataset.

Campos

- **Fecha (String):** la fecha en la que ocurre el seísmo. El conjunto de datos recoge los datos sobre los seísmos ocurridos en los últimos 10 días, por lo que los valores de este campo estarán dentro de el rango de valores correspondiente.
- **Hora (String):** la hora en la que ocurre el seísmo. La hora corresponde con la hora oficial en Canarias, UTC + 2 horas en verano y UTC + 1 en invierno.
- **Latitud y longitud (Double):** dupla de medidas (separadas en dos campos), expresadas en grados que permiten localizar geográficamente el epicentro del seísmo.
- **Profundidad (Int):** la distancia vertical entre el epicentro y el hipocentro del terremoto.
- **Magnitud (Double):** la medida que permite conocer la intensidad de un seísmo. Su unidad viene determinado por **el siguiente campo**.
- **Tipo de Magnitud (String):** este campo determina el tipo de magnitud con el que se mide el seísmo. Para la región de Canarias, existen diversos tipos de magnitud. El más común en este dataset es mbLg. ([Descripción del tipo de magnitud](#)).
- **Localización (String):** el lugar donde se registra el seísmo.

Proceso de recolección de datos

Para la recolección de los datos, se crea un script de Python que realiza las siguientes tareas:

1. Petición HTTP a la dirección de la página suministrada, para obtener el HTML sin procesar.
2. Procesa el HTML mediante el uso de BeautifulSoup.

3. Obtiene todas las instancias de la etiqueta 'tr' (table row), es decir, todas las filas de la tabla.
4. Para cada una de las filas, se extrae la información valiosa para elaborar el dataset, la que se corresponde con cada uno de los campos mencionados anteriormente.
5. Por último, se procede con la elaboración del documento CSV, manteniendo el formato de separación por comas (en este caso puntos y comas).

Agradecimientos

El propietario del conjunto de datos extraído es el Instituto Geográfico Nacional. Para actuar de acuerdo a los principios éticos y legales en el contexto del proyecto, se han emprendido dos acciones principales. La primera ha sido la **visualización del fichero robots.txt** (<https://www.ign.com/robots.txt>). Del análisis de este fichero se concluye que la localización del directorio que se explota en el proyecto no se encuentra especificada en la lista de directorios restringidos por este fichero. Además, el bot empleado para realizar el Scrapping ('Python-urllib/3.9') no se encuentran entre los 'User agents' con acceso restringido.

La segunda acción fue la lectura del fichero que contiene la información sobre la licencia de los servicios de datos producidos por el IGN (https://www.ign.com/resources/licencia/Condiciones_licenciaUso_IGN.pdf), para conocer bajo qué licencia se publicará el dataset.

Inspiración

Para entender la decisión de realizar una recopilación de datos sobre los últimos terremotos en las islas canarias, es necesario conocer el contexto volcanológico actual de las Islas Canarias, y es que como probablemente sea conocido por todos, el volcán de Cumbre Vieja ha acabado con la normalidad de miles de ciudadanos en el este de la Isla. Antes y durante todo el proceso eruptivo, la monitorización de los diferentes indicadores y la identificación de los patrones permitieron la prevención de la erupción y la evacuación de los pueblos que fueron afectados por las coladas. En especial, los seísmos registrados los días antes de la erupción fueron indicadores determinantes para que las autoridades tomaran medidas para prepararse para una probable erupción.

Es por todas estas razones que se decide elaborar un dataset tomando información sobre los últimos seísmos registrados en la isla, que puede ser útil de cara a la **prevención y el estudio** tanto de la actual erupción como de posibles erupciones futuras.

Licencia

La licencia bajo la que se publicará este conjunto de datos, como se indica en el documento suministrado por el IGN, será **CC BY 4.0 License**.

Código

El código desarrollado para completar la especificación de la práctica se puede encontrar en el siguiente repositorio: <https://github.com/javidiazdom/py-scrapper>.

Dataset

El dataset está publicado en el siguiente enlace. <https://doi.org/10.5281/zenodo.5652158>

Contribuciones

Contribuciones	Firma
Investigación previa	Javier Díaz Domínguez
Redacción de las respuestas	Javier Díaz Domínguez
Desarrollo del código	Javier Díaz Domínguez