



Universitat  
Oberta  
de Catalunya

MASTER UNIVERSITARIO DE CIENCIA DE DATOS

UNIVERSITAT OBERTA DE CATALUNYA

Tipología y ciclo de vida de los datos

Práctica (I)

*Autor: Miguel Pérez Rego*

A Coruña, Marzo 2020

# Índice

1. Contexto	2
2. Definición de un título para el dataset	2
3. Descripción del dataset	2
4. Representacion gráfica	2
5. Contenido	3
6. Agradecimientos	4
7. Inspiración	4
8. Licencia	5
9. Código	6
10.Firma integrantes	7

## 1. Contexto

Para la realización de esta práctica, se ha buscado una web que actualice sus datos constantemente, para que se pueda ir generando y almacenando información nueva continuamente, creándose un *dataset* cada cierto periodo de tiempo, con el objetivo de tener un histórico de la información almacenada.

Para ello, navegando por la red, se ha encontrado una web que muestra en una de sus páginas la información de todos aquellos terremotos con una magnitud mayor a 4.0 en la escala sismológica de Richter en los últimos treinta días. La URL dónde se encuentran los datos es la siguiente: <http://ds.iris.edu/seismon/eventlist/index.phtml>.

Por lo tanto, programando una ejecución del *script* que obtiene la información cada treinta días, se conseguirá almacenar en diferentes datasets todo el histórico de terremotos.

## 2. Definición de un título para el dataset

El título del dataset es: “Histórico de terremotos en los últimos treinta días”.

## 3. Descripción del dataset

El dataset generado contiene la información de los terremotos ocurridos a lo largo del mundo en los últimos treinta días y que superan una magnitud por encima de 4.0 en la escala sismológica de Richter.

## 4. Representacion gráfica

En este apartado, se presenta una imagen capturada de la página web comentada en el primer apartado y que identifica el dataset visualmente.

Missing quake?  
Wrong magnitude?

## Latest Earthquakes Worldwide

690 earthquakes of magnitude > 4.0, for uniform distribution

**TIP** To sort by multiple columns hold shift key and click on second and even third column header.

Seleccionar idioma ▼

DATE and TIME (UTC)	LAT	LON	MAG	DEPTH km	LOCATION (Shows interactive map)	IRIS ID (Other info)
21-MAR-2020 15:40:55	-6.55	104.34	4.4	71	<a href="#">SUNDA STRAIT, INDONESIA</a>	<a href="#">11200613</a>
21-MAR-2020 12:33:11	9.13	126.83	4.7	46	<a href="#">MINDANAO, PHILIPPINES</a>	<a href="#">11200566</a>
21-MAR-2020 11:40:28	11.12	138.76	5.0	49	<a href="#">W. CAROLINE ISLANDS, MICRONESIA</a>	<a href="#">11200556</a>
21-MAR-2020 09:34:28	-29.17	-69.04	4.1	16	<a href="#">CHILE-ARGENTINA BORDER REGION</a>	<a href="#">11200529</a>
21-MAR-2020 06:42:34	-12.34	44.90	5.0	10	<a href="#">NORTHWEST OF MADAGASCAR</a>	<a href="#">11200492</a>
21-MAR-2020 03:52:46	1.67	127.24	4.5	116	<a href="#">HALMAHERA, INDONESIA</a>	<a href="#">11200453</a>
21-MAR-2020 03:37:55	41.66	141.99	4.3	80	<a href="#">HOKKAIDO, JAPAN REGION</a>	<a href="#">11200458</a>
21-MAR-2020 03:12:13	39.33	20.49	4.3	10	<a href="#">GREECE-ALBANIA BORDER REGION</a>	<a href="#">11200413</a>
21-MAR-2020 02:23:08	37.08	71.45	4.3	111	<a href="#">AFGHANISTAN-TAJIKISTAN BORD REG.</a>	<a href="#">11200380</a>
21-MAR-2020 01:33:35	39.11	-119.74	4.5	8	<a href="#">NEVADA</a>	<a href="#">11200350</a>
21-MAR-2020 00:49:51	39.37	20.63	5.7	10	<a href="#">GREECE-ALBANIA BORDER REGION</a>	<a href="#">11200342</a>
21-MAR-2020 00:22:57	22.78	123.99	4.6	10	<a href="#">SOUTHEAST OF TAIWAN</a>	<a href="#">11200349</a>
21-MAR-2020 00:03:51	-6.45	131.06	4.3	60	<a href="#">TANIMBAR ISLANDS REG., INDONESIA</a>	<a href="#">11200333</a>
20-MAR-2020 21:38:30	39.25	20.46	4.6	10	<a href="#">GREECE-ALBANIA BORDER REGION</a>	<a href="#">11200201</a>
20-MAR-2020 18:39:36	-4.29	151.95	4.5	11	<a href="#">NEW BRITAIN REGION, P.N.G.</a>	<a href="#">11200181</a>
20-MAR-2020 16:09:27	-23.63	-179.78	4.6	519	<a href="#">SOUTH OF FIJI ISLANDS</a>	<a href="#">11200052</a>
20-MAR-2020 13:44:21	-8.21	150.40	4.7	10	<a href="#">EASTERN NEW GUINEA REG., P.N.G.</a>	<a href="#">11200483</a>
20-MAR-2020 09:54:18	-48.46	31.35	5.2	10	<a href="#">SOUTH OF AFRICA</a>	<a href="#">11199936</a>
20-MAR-2020 09:03:33	16.82	-95.23	4.8	19	<a href="#">OAXACA, MEXICO</a>	<a href="#">11199930</a>
20-MAR-2020 07:53:26	56.48	-148.62	5.1	10	<a href="#">GULF OF ALASKA</a>	<a href="#">11199911</a>

## 5. Contenido

En este apartado se va a explicar los campos que incluye el dataset, el periodo de tiempo de los datos y cómo se han recogido.

Los campos que contiene el dataset son los siguientes:

- **DATE and TIME (UTC)**: La fecha y la hora en la que se ha producido el terremoto.
- **LAT**: Coordenada geográfica correspondiente a la latitud terrestre donde se ha producido el terremoto.
- **LONG**: Coordenada geográfica correspondiente a la longitud terrestre donde se ha producido el terremoto.
- **MAG**: La magnitud en la escala sismológica de Richter del terremoto.
- **DEPTH km**: La profundidad en kilómetros donde se ha producido el origen del terremoto.
- **IRIS ID**: Identificador que permite tener registrado como un elemento único al terremoto.

Por otro lado, el periodo de tiempo, es una extracción completa de la información actual que hay en la web, es decir, de todos los terremotos que se han producido en los últimos treinta días.

Finalmente, los datos cargados son extraídos de la página oficial de IRIS. Realizando una labor de investigación se ha podido observar que muestran en una tabla en su web el listado de los terremotos que se han producido en los últimos treinta días. De este origen se extraerá la información que será almacenada en el dataset de la práctica.

## 6. Agradecimientos

En este apartado, se realiza una presentación del propietario del conjunto de datos del que se va a extraer la información.

El propietario de los datos es IRIS (*Incorporated Research Institutions for Seismology*). Se trata de es un consorcio de investigación universitaria dedicado a explorar el interior de la Tierra a través de la recopilación y distribución de datos sismográficos con sede en Washington DC.

Sus funciones principales son proporcionar la gestión y acceso a datos observados y derivados para la comunidad global de ciencias de la tierra, incluyendo datos del movimiento del suelo, atmosféricos, infrasónicos, hidrológicos e hidroacústicos.

Es importante tener en cuenta forma parte de re3data.org (Registry of Research Data Repositories) y que consiste en una herramienta de Open Science que ofrece a los investigadores, organizaciones de financiación, bibliotecas y editores una visión general de los repositorios internacionales existentes para datos de investigación.

## 7. Inspiración

En este apartado se va a explicar por qué es interesante este conjunto de datos y qué preguntas se pretenden responder al analizar el dataset generado.

Este conjunto de datos es interesante porque refleja en tiempo real todos aquellos terre-

motos que pueden suponer algún peligro o catástrofe a nivel mundial. Como únicamente registra los datos de los últimos treinta días, es interesante ir almacenando el histórico de estos registros dado que la web no los proporciona, con el objetivo de poder llevar a cabo un análisis de mayor temporalidad que la ofrecida.

Con la información de este dataset se puede realizar múltiples estudios acerca de los terremotos existentes a nivel mundial. Diversos estudios pueden ser los siguientes:

- Si almacenas la información a lo largo del tiempo, permite conocer en que momento del año son más frecuentes los terremotos.
- En que latitudes y longitudes se localizan un mayor número de terremotos.
- En qué localización se sitúan los terremotos, cuáles son sus profundidades y magnitudes máximas y medias.

Este tipo de estudios, puede ayudar a la prevención y a la preparación por parte de los gobiernos de medidas que eviten los mayores daños posibles en sus comunidades.

## 8. Licencia

En este apartado, se realiza una selección de la licencia para el dataset y el motivo que ha llevado a su elección. Para su elección se ha estudiado las licencias que se muestran en la página web: <https://creativecommons.org/licenses/>.

La licencia que más se adapta en el caso expuesto es la *Released Under CC BY-SA 4.0 License*. Los motivos son los siguientes:

- Se reconoce adecuadamente la autoría, proporcionar un enlace a la licencia e indicar si se han realizado cambios. En el ejemplo expuesto los autores es *Incorporated Research Institutions for Seismology* y se reconoce que no se ha hecho ningún cambio en los datos, dado que trata de extracción pura de lo que se muestra en su web.
- Se permite la comercialización de los datos. Dado que cualquier empresa podría hacer uso de la extracción generada para almacenar la información en cualquier de sus proyectos.
- Dado que si se remezcla, transforma o crea a partir del material, se difunde sus contribuciones bajo la misma licencia que el original.

## 9. Código

```
import requests
import csv
from bs4 import BeautifulSoup as bs
import re

url = requests.get("http://ds.iris.edu/seismon/eventlist/index.
    phtml")
soup = bs(url.content, 'html.parser')

filename = "Terremotos.csv"
csv_writer = csv.writer(open(filename, 'w'))

for tr in soup.find_all("tr"):
    data = []
    for th in tr.find_all("th"):
        data.append(re.sub(r'(\s+|\n)', ' ', th.text.strip()))

    if data:
        print("Inserting headers : {}".format(', '.join(data)))
        csv_writer.writerow(data)
        continue

    for td in tr.find_all("td"):
        data.append(td.text.strip())

    if data:
        print("Inserting data: {}".format(', '.join(data)))
        csv_writer.writerow(data)
```

## 10. Firma integrantes

En este apartado se recoge la firma de los integrantes que han hecho cada uno de las tareas de la práctica.

Contribuciones	Firma
Investigación previa	miguelpre
Redacción de las respuestas	miguelpre
Desarrollo código	miguelpre