1. Contexto.

Hemos decidido hacer la práctica sobre alguna página que proporcionara información sobre actividad sísmica. En un primer momento queríamos hacer web scrapping sobre la página del instituto geográfico nacional de España, dependiente del Ministerio de Transportes, Movilidad y Agenda Urbana.

Al echar un vistazo a esta web vimos que se proporcionaba información sísmica entre los años 1370 y la actualidad, lo que nos pareció un buen banco de datos.

A continuación analizamos el archivo robots.txt, cuyo contenido era:

```
User-agent: Googlebot
Disallow: /?*

User-agent: Baiduspider
Disallow: /?*

User-agent: YandexBot
Disallow: /?*

User-agent: ichiro
Disallow:  /?*

User-agent: sogou spider
Disallow:  /?*

User-agent: Sosospider
Disallow: /?*

User-agent: YoudaoBot
Disallow: /?*

User-agent: YetiBot
Disallow: /?*

User-agent: bingbot
Crawl-delay: 2
Disallow: /?*

User-Agent: Yahoo! Slurp
Crawl-delay: 2
Disallow: /?*

User-agent: rdfbot
Disallow: /?*

User-agent: Seznambot
Request-rate: 1/2s
Disallow: /?*

User-agent: ia_archiver
Disallow:

User-agent: Mediapartners-Google
Disallow:

Exclusión de todos los robots:
User-agent: *
Disallow: /
```

Y al ver que se excluían todos los robots, descartamos el utilizar esta página y seguimos buscando páginas que proporcionaran el tipo de información que queríamos. Podríamos haber intentado utilizar algún método para prevenir el web scraping, pero decidimos que sería mejor buscar otras páginas.

Así, llegamos a la página www.volcanodiscovery.com que tenía información de todo tipo centrada en los terremotos: cuándo habían ocurrido, la magnitud, localización, la fuente de la información, imágenes, comentarios, etc.

Analizando el archivo robots.txt

```
User-agent: *
Disallow: /tmp/
Disallow: /typo3temp/
#Disallow: /uploads/
#Disallow: /fileadmin/

User-agent: *
Crawl-delay: 2

User-agent: Googlebot
Allow: /

User-agent: Twitterbot
Allow: /

User-agent: Slurp
Allow: /

#User-Agent: msnbot
#Disallow: /

#User-Agent: MauiBot
#Disallow: /

#Baiduspider
#User-agent: Baiduspider
#Disallow: /

#Yandex
#User-agent: Yandex
#Allow: /
```

vimos que se restringía el acceso a todos los robots a unos determinados directorios, pero no lo restringía totalmente, así que decidimos realizar la práctica sobre está página.

En el archivo también se puede ver cómo el parámetro craw delay (utilizado por los propietarios de las páginas para determinar el tiempo (en segundos) que el robot tiene que esperar entre dos peticiones sucesivas) estaba en 2 segundos.

2. Título para el dataset.

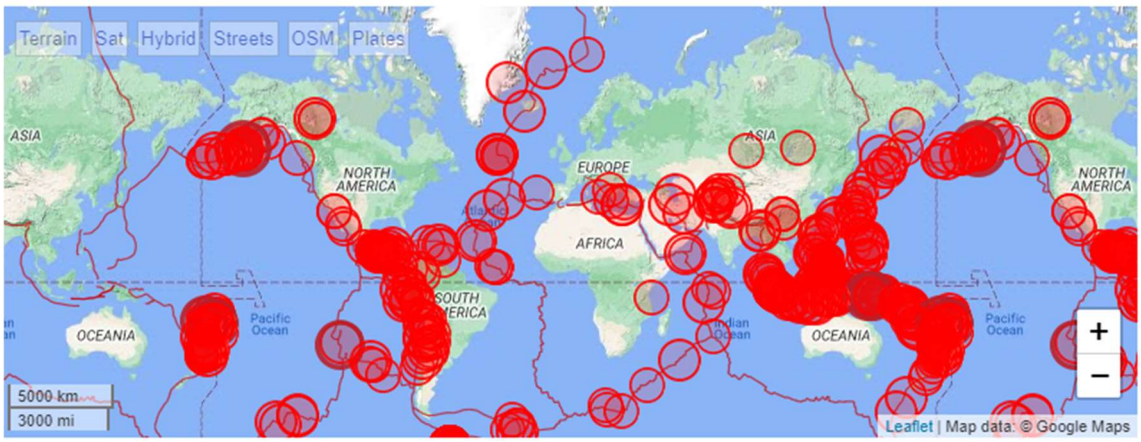
Earth tremors.

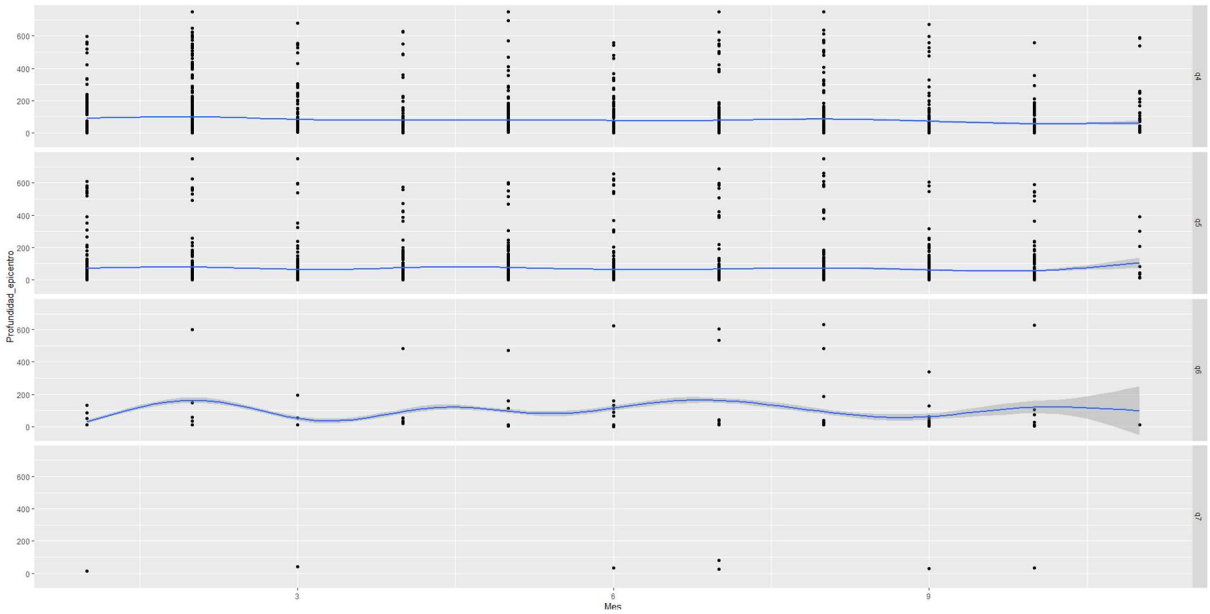
3. Descripción del dataset.

Con este dataset podemos tener una relación de información asociada a todos los seismos que se han producido a nivel mundial durante un período de tiempo concreto. En el caso de la práctica hemos obtenido todos los que se han producido desde enero del 2020 hasta la fecha (inicio de noviembre)

En una primera aproximación hemos trabajado sobre la extracción de datos en una única fecha y una vez viendo que nos funcionaba bien el web scrapping hemos decido hacerlo un poco más complejo y realizar iteraciones sobre varias fechas.

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





5. Contenido

Los datos que se incluyen en el datasatet son:

Id sismo: es un número que lo identifica de manera unívoca.

Tipo: identifica el tipo de la zona en la que se produce el sismo. En este caso los posibles valores de tipo son:

- q4
- q5
- q6
- q7

Coordenadas: localización exacta del epicentro del seísmo

Fecha: fecha en la que se produce el terremoto (formato aaa-mm-dd)

Hora: hora en la que se produce el terremoto (formato 24 horas)

Magnitud: es un número que permite establecer las características del tamaño y la energía sísmica liberada en el terremoto. Existen diversos tipos de magnitud que pueden ser utilizadas para expresar esta información y en este caso, la magnitud es la que se conoce como Magnitud Richter. Se trata, por tanto, de la medida de la energía liberada por el terremoto y esta clasificación va desde la "Micro magnitud" en que los terremotos no son perceptibles hasta la "Magnitud épica" en el que se puede generar una extinción local y que por ahora no ha sido registrado.

Fuente:

- http://www.ssn.unam.mx/jsp/reportesEspeciales/Magnitud-de-un-sismo.pdf
- http://www.proteccioncivil.es/riesgos/terremotos/faq

Distancia: indica el punto de la Tierra desde donde se libera la energía del terremoto: cuanto más superficial, más destructivo.

Aclarar que, en la imagen del dataset, para clarificar, esta información está identificado como Profundidad del epicentro.

Ubicación: lugar en el que se produce el seismo. región a la que pertenece el lugar en el que se produce el terremoto. Permitirá hacer agregaciones de datos.

País: indican el país en el que que se produce el terremoto

6. Agradecimientos

Al Dr. Tom Pfeiffer, el vulcanólogo manager de la empresa que hay detrás de la página de VolcanoDiscovery: le escribimos un correo indicándole que íbamos a realizar web scrapping sobre su página web. Aparte de un correo automático no hemos recibido por su parte ninguna contestación con lo que hemos supuesto que no ha tenido ningún inconveniente en que hiciéramos el trabajo.

7. Inspiración

Parece que solamente existen los terremotos que salen por la televisión o, en general, por los medios de comunicación, pero no, esto no es así: realmente hay mucha más actividad sísmica de la que no somos consciente porque no son terremotos con el impacto suficiente para que puedan tener un hueco en los informativos, pero existir existen. Y son sacudidas

Teniendo en cuenta que un terremoto o seísmo es una sacudida o movimiento brusco del terreno causado por disturbios volcánicos ¿deberíamos preocuparnos si la actividad sísmica sufre importantes variaciones con respecto a momentos anteriores en el tiempo? Y aunque parece ser que no existe ningún método que sea capaz de predecir el tiempo, lugar y magnitud de un terremoto ¿sería posible llegar a predecir los terremotos a través del análisis de datos?

8. Licencia

El propietario del sitio sobre el que hemos hecho web scraping indica que existen textos e imágenes, principalmente fotos, que están protegidas por copyright y que existen limitaciones

en cuanto a la utilización de (determinadas) fotos sin previo permiso. No hemos encontrado ninguna información adicional sobre tipo de licenciamiento.

Teniendo en cuenta esta información y que el dataset que vamos a publicar no contiene ninguna foto, ni tampoco textos, sino que se tratan de datos que han sido proporcionados por otros entidades (de diferentes tipos: sin ánimo de lucro, oficiales, gubernamentales….), a la hora de elegir el tipo de licencia, hemos optado por una licencia tipo Creative Commons puesto que son las que permiten derechos de autor (se da permiso para usarla con una condición: citar la autoría de la obra) aunque con algunas condiciones (lo que se conoce como "algunos derechos reservados"). Centrándonos en este tipo de licencias y teniendo en cuenta que:

- BY: (reconocimiento) hace falta siempre (para cualquier uso o explotación) reconocer el uso de la autoría de la obra. Esta condición se exige en todas las licencias CC y no puede ser excluida.

- NC: (no comercial): se prohíbe la utilización de la obra con fines comerciales

- CC BY-NC-SA 4.0: no se permite un uso comercial de la obra original ni de las posibles obras derivadas y la distribución se debe hacer con una licencia igual a la que regula la obra original

- CC BY-SA: este tipo de licencia permite el uso comercial de la obra y de las posibles obras derivadas que deberán ser explotadas bajo la misma licencia

Hemos decidido utilizar la licencia CC BY-NC- SA 4.0, teniendo en cuenta cómo esta página se nutre de información.

Hemos utilizado de la página Creative commons el seleccionador de licencias para ver si íbamos a licenciar de la manera que queríamos:

Aunque no es una licencia de cultura libre (al no permitir el uso comercial de la obra), hemos preferido poner este tipo de licenciamiento puesto que, como decimos al principio de este apartado, en la página web no se concreta mucho sobre la licencia y aunque preguntamos por email a los propietarios de la página sobre este asunto y la posibilidad de realizar web scrapping (el día 22 de octubre), no hemos recibido respuesta (salvo una respuesta automática).

Información de licenciamiento recogida principalmente de:
https://www.upo.es/biblioteca/servicios/pubdig/propiedadintelectual/tutoriales/derechos_autor/htm_12.htm

https://creativecommons.org/

9. Código

Se adjunta el archivo PRA1.py comentado.

10. Publicación de Dataset en Zenodo

El DOI registrado es: 10.5281/zenodo.4252300

11. Tabla de contribuciones al trabajo:

Contibuciones	Firma
Investigación previa	BCO y PBE
Redacción de las respuestas	BCO y PBE
Desarrollo código	BCO y PBE