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 (IGN), 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, con lo que procedimos a analizar el archivo robots.txt de ese sitio:

```
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, y aunque podríamos haber intentado utilizar algún método para prevenir el web scraping, descartamos el utilizar esta página (principalmente cuestiones éticas) y seguimos buscando otras páginas que proporcionaran el tipo de información que queríamos.

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.

Analizamos 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 esta 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.

Hemos considerado que sería una muy buena fuente de información (en cuanto a calidad y fiabilidad) porque que recoge los datos, para un mismo seísmo, de varios organismos.

Así, por ejemplo, para España las fuentes de información son el EMSC (European Mediterranean Seismological Centre) y el IGN que es el organismo que hay detrás de la página que queríamos utilizar en primer lugar para realizar la práctica. O, por ejemplo, para el caso de Ecuador las fuentes son el IGEPN (Instituto Geofísico de la Escuela Politécnica Nacional) y también el EMSC.

2. Título para el dataset.

Earthquakes.

3. Descripción del dataset.

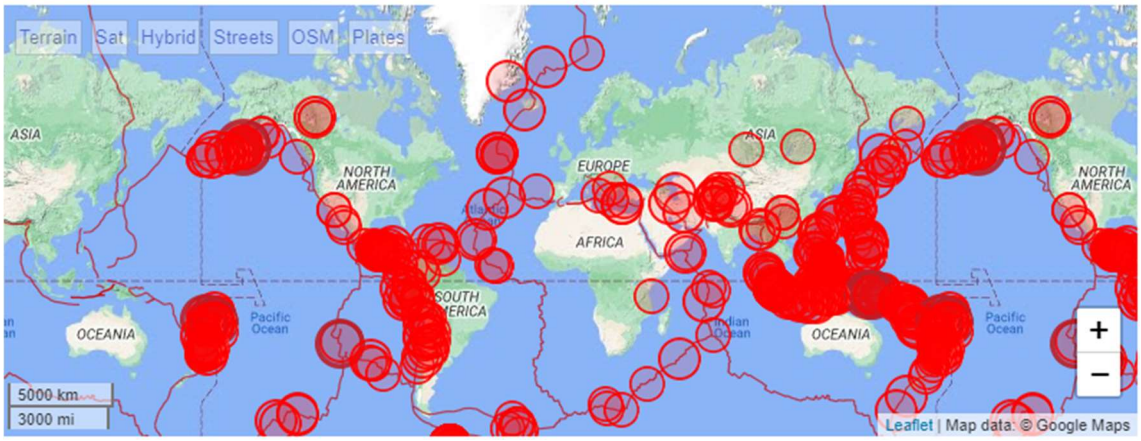
Con el data set que hemos obtenidos, disponemos de datos asociados a todos los sismos que se han producido a nivel mundial durante un período de tiempo concreto. En el caso de la práctica

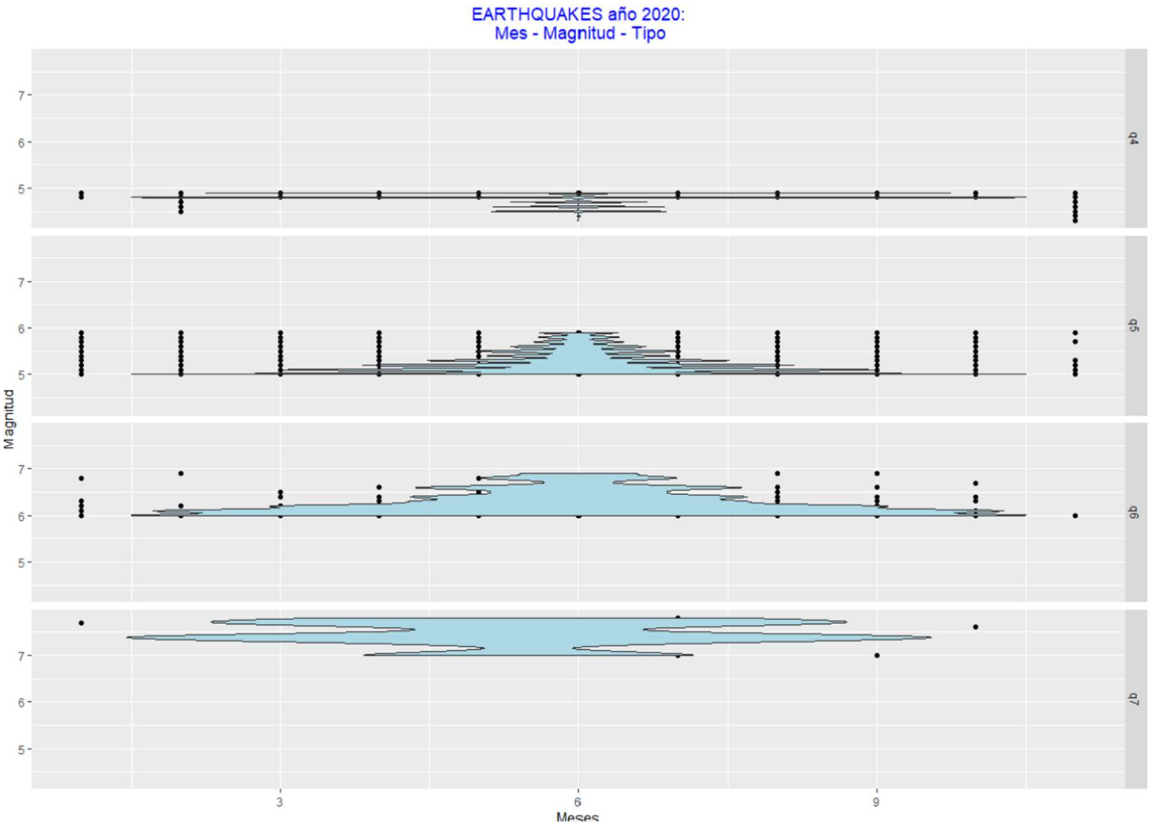
hemos obtenido todos los que se han producido desde enero del 2020 hasta la fecha (inicio de noviembre).

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.

Hemos decidido que esta imagen, conjunto de dos imágenes diferentes es la que podría representar nuestro data set: una imagen visual del tema que hemos trabajado y una imagen analítica de los datos obtenidos.





5. Contenido

Campos que incluyen el data set:

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

Tipo: identifica el tipo del sismo en función de la magnitud. Leer más sobre esta clasificación en el apartado de magnitud. Este dato es el que nos ha servido para realizar la iteración con el web scraping.

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: la definición de magnitud es: *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" hasta la "Magnitud épica" . Concretamente la clasificación existente es:

2.0-3.0 Micro Magnitud – No son perceptibles.
3.0-3.9 Menor Magnitud – Perceptibles con poco movimiento y sin daño.
4.0-4.9 Ligera Magnitud – Perceptibles con movimiento de objetos y rara vez produce daño.
5.0-5.9 Moderada (o Mediana) Magnitud – Puede causar daños mayores en construcciones débiles o mal construidas.
6.0-6.9 Fuerte Magnitud – Pueden ser destructivos.
7.0-7.9 Mayor Magnitud – Pueden ser destructivos en zonas extensas.
8.0-9.9 Gran Magnitud – Catastróficos, provocando destrucción total en zonas cercanas al epicentro.
10 o + Magnitud Épica – Jamás registrado, puede generar una extinción local.

Y esta clasificación da lugar a las diferentes "q's" en las que (para nuestro conjunto de datos del año 2020 solamente se han recogido terremotos de magnitud ligera, moderada, fuerte y mayor).

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

Ubicación: 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 se produce el terremoto

Período de tiempo

Se han recogido los datos correspondientes a todo el año 2020 (lógicamente, hasta el mes de noviembre). El código del programa está preparado para poder volver a hacer web scrapping y recoger el año entero, simplemente ejecutándolo.

Con una sencilla variación de este, podría aumentarse la funcionalidad para poder recoger información de un período de tiempo más amplio e incluso que en la ejecución, se solicitaran las fechas de inicio y de fin.

Hay que destacar el hecho de que, para 10 meses del año, ya ha sido posible recoger casi 119.000 registros.

Cómo

Al realizar el web scrapping lo que hemos hecho ha sido: para cada mes del año 2020, buscar la página en la que aparecían los terremotos producidos durante ese mes. Una vez tenemos la URL, hemos iterado por cada uno de los tipos (clases) en los que el propietario de la página web ha clasificado los terremotos (lo que hemos llamado tipo y que se identifica por q4, q5, q6 y q7). Para cada uno de estos tipos, extraemos cada uno de los terremotos y trabajamos con el string obtenido para quedarnos con los datos que van a formar parte del data set.

Una vez tenemos todos los datos de un terremoto concreto, pasamos al siguiente terremoto del tipo correspondiente; cuando ya tenemos hemos finalizado la obtención de los datos de un tipo de terremoto, pasamos al siguiente tipo de terremoto. Y cuando se finaliza la obtención de los datos de todos los tipos de terremoto para un mes, pasamos al mes siguiente.

Hay que decir que lo indicado en el párrafo anterior se explica con más detalle en los comentarios que hemos añadido en el código del programa.

6. Agradecimientos

Al Dr. Tom Pfeiffer, el vulcanólogo mánager de la empresa que hay detrás de la página de Volcano Discovery: 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.

Y a continuación, hacemos referencia a las páginas que hemos ido visitando para obtener información más detallada sobre los terremotos y que nos ha permitido entender mejor nuestro data set:

- https://www.biobiochile.cl/noticias/2011/03/29/como-se-clasifican-los-sismos-segun-su-magnitud-e-intensidad.shtml
- http://www.ssn.unam.mx/jsp/reportesEspeciales/Magnitud-de-un-sismo.pdf
- http://www.proteccioncivil.es/riesgos/terremotos/faq
- https://www.lavanguardia.com/vida/20201102/49168109714/las-sacudidas-de-torrevieja-nos-recuerdan-que-vivimos-en-zona-de-terremotos.html
- https://www.upo.es/biblioteca/servicios/pubdig/propiedadintelectual/tutoriales/derechos_autor/htm_12.htm
- https://creativecommons.org/

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.

Como dicen los expertos, si estamos en zona de terremotos, es normal que se produzcan pequeñas sacudidas, lo que ocurre es que no las notamos, pero cuando se produce algún terremoto cercano a un núcleo de población lo sentimos y nos asustamos, aunque *deberíamos*

entender que los movimientos sísmicos no son nada negativo, simplemente un recuerdo de que estamos en un territorio donde ocurren terremotos (palabras de José Delgado, responsable de la Red sísmica de la UA)

Teniendo en cuenta que un terremoto o sismo es una sacudida o movimiento brusco (o no tan brusco, ya hemos hablado de ello en la magnitud) del terreno causado por disturbios volcánicos ¿deberíamos preocuparnos si la actividad sísmica sufre importantes variaciones con respecto a momentos anteriores en el tiempo? ¿qué conclusiones podemos obtener si analizamos con técnicas de minería de datos toda la información que tenemos ahora disponible? La continuidad de los terremotos para una misma ubicación geográfica, ¿podría ayudar a prevenir posibles desastres? ¿o establecer unas medidas o pautas para las edificaciones y, en general, cualquier infraestructura de las zonas de terremotos y así estar más preparados?

Teniendo en cuenta que para que se produzca un terremoto "importante" es necesario liberar mucha energía y que las pequeñas sacudidas son liberaciones de esa energía poco a poco ¿la ausencia de sismos (datos) podría ponernos en sobre aviso de que está por venir una sacudida importante?

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 data set 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).

9. Código

Se adjunta el archivo PRA1.jnpy y el .py para su ejecución.

10. Publicación de Dataset en Zenodo

El DOI registrado es: 10.5281/zenodo.4252300

Enlace: https://zenodo.org/deposit/4252300

11. Tabla de contribuciones al trabajo:

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