# Contexto.

Hemos decidido hacer la práctica sobre alguna página que proporcionar 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 excluía todos los robots, descartamos el utilizar esta página y seguimos buscando páginas que proporcionaran el tipo de información que queríamos.

Así, llegamos a la página [www.volcanodiscovery.com](http://www.volcanodiscovery.com) que tenía información de todo tipo centrada en los terremotos: cuándo habían ocurrido, la magnitud, localización, la fuente de la información, 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.

Yout tuve Sara Ruiz – web scrapping <https://www.youtube.com/watch?v=zoYEKL3-Gq0>

# Título para el dataset.

Catálogo de terremotos

# Descripción del dataset.

# Representación gráfica.

# Contenido

# Agradecimientos

VolcanoDiscovery GmbH

Dr. Tom Pfeiffer

# Inspiración

# 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 animo de lucro, oficiales, gubernamentales….), a la hora de elegir el tipo de licencia, hemos utilizado 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.



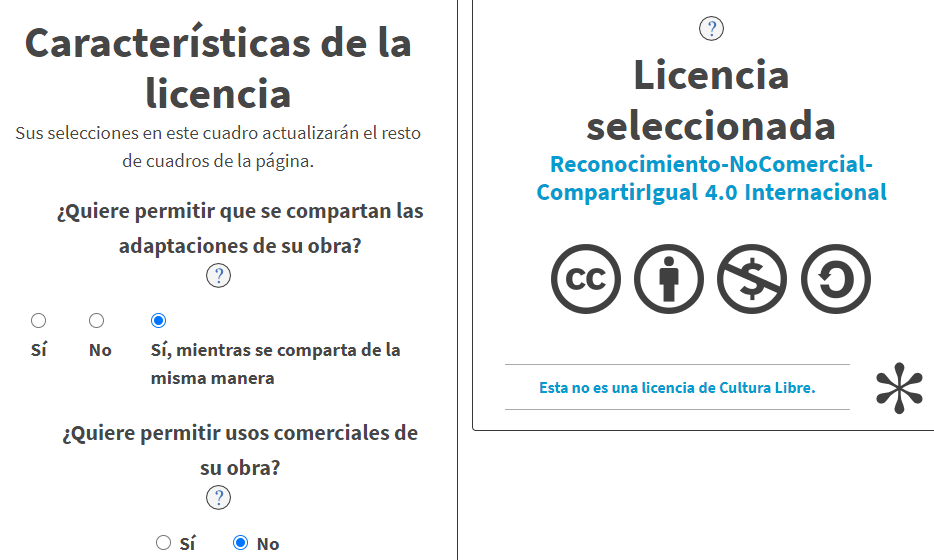
* NC: (no comercial): se prohíbe la utilización de la obra con fines comerciales



* 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/>

# Publicación de Dataset en Zenodo