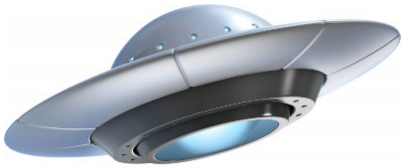
Tipología y ciclo de vida de los datos
Práctica 1

Alumno: José Carlos García Pérez

1. Título del dataset
MUFON (Mutual UFO Network) UFO Reports

2. Subtítulo del dataset
Avistamientos OVNI reportados por testigos de todo el mundo a una de las organizaciones más reconocidas globalmente, dedicada a la investigación de este tipo de fenómenos.

3. Imagen



UFO data from Mutal UFO Network

4. Contexto
El conjunto de datos está compuesto por información anónima reportada por testigos de avistamientos de objetos voladores no identificados o de encuentros con entes extraterrestres por todo el mundo. MUFON cuenta con un sistema de reporte que permite a los testigos de estos fenómenos informar de los mismos, compartiendo la información con el público general a través de su web (previa anonimización de los datos).

5. Contenido
El programa realizado para la recopilación de los datos permite la selección de informes por diferentes criterios.

Por un lado, cada informe dispone de un identificador autoincremental, de forma que es posible descargar los informes usando un rango de valores, es decir, especificando un valor mínimo y uno máximo.

Por otro, es posible seleccionar informes en base a un término, de forma que sólo son seleccionados los informes que incluyen ese término. Este caso es particularmente útil si estamos interesados sólo en un tipo de informes, por ejemplo, por país o por forma del objeto.

MUFON recoge los datos vía telefónica, email o directamente rellenando un formulario a través de su web, pasando posteriormente a estar disponibles para el público general, una vez anonimizados. Para este trabajo hemos analizado la web **ufostalker.com**, y aunque en un principio se optó por hacer *web scraping*, finalmente se optó por llamar directamente a dos servicios que permiten acceder a la información de interés, cuyos *endpoints* se indican a continuación. Esta solución parece más robusta, ya que se ha dado el caso de que se ha cambiado el *front-end* de la aplicación, lo que seguramente habría obligado a cambiar el proceso de recorte web.

En este caso se han generado dos datasets de prueba, abarcando reportes desde el 10 de octubre de 1890 hasta el 15 de abril de 2018.

Endpoint	Descripción	Interfaz
http://www.ufostalker.com:8080/event	Permite seleccionar informes por id	XML
http://ufostalker.com:8080/search	Permite seleccionar informes por término	JSON

Cada informe recogido consta de los siguientes atributos:

Atributo	Descripción
id	Identificador del informe
sighted_at	Fecha del fenómeno. Es la fecha real en la que ocurrió el evento
reported_at	Fecha del informe. Es la fecha en la que se comunicó el evento
location	Ubicación del evento. Tiene la forma “ciudad (país)”
shape	Forma del objeto avistado
duration	Duración del evento
description	Descripción del evento
latitude	Latitud
longitude	Longitud
case_number	Referencia del número de caso en MUFON

6. Agradecimientos

Especial agradecimiento a la agencia **MUFON (Mutual UFO Network)**, quien desde el año 1961 se encarga de recopilar e investigar casos de avistamientos OVNI, haciendo que la información esté disponible para el público general a través de sus webs.

Igualmente, especial agradecimiento al autor de la web **ufostalker.com**, que permite acceder a la información de MUFON a través de la misma.

Sin ninguna de estas dos fuentes no podríamos haber realizado este trabajo.

En la propia web de la organización hay disponibles trabajos de investigación realizados con estos datos:

http://www.mufon.com/research.html

Igualmente, en **ufo-hunters.com** existen mapas de calor y gráficas con la evolución del número de avistamientos por año generados a partir de estos datos y otros de organismos similares.

http://www.ufo-hunters.com/stats

7. Inspiración

Cada vez más avistamientos y encuentros con entes extraterrestres son reportados por todo el mundo, hecho que llama mucho la atención. Este conjunto de datos podría ser interesante para realizar *topic modeling*, de forma que podamos clasificar los reportes por áreas temáticas relacionadas.

Además sería bastante interesante descubrir si hay determinados tipos de avistamientos o encuentros que son más frecuentes en ciertas áreas, o zonas determinadas donde se den más casos de abducciones.

8. Licencia

La licencia seleccionada es **CC BY-NC-SA**, ya que permite la copia, distribución y el uso del dataset siempre y cuando se reconozca y cite al autor, con fines no comerciales. Asimismo, permite que otros puedan contribuir creando trabajos derivados siempre y cuando lo hagan bajo una licencia idéntica.

9. Código

El código está disponible en:

https://github.com/jcarlosgarcia/mufon-crawler

10. Dataset

Se han generado dos datasets, uno por id y otro por término (en este caso, informes que incluyen el término “triangular”). Están disponibles en:

https://github.com/jcarlosgarcia/mufon-crawler