TIPOLOGÍA Y CICLO DE VIDA DE LOS DATOS
PRACTICA 1

Componentes del grupo:
- Gabriel Villalba Pintado
- Leire Iturregui Inoriza
- Miguel Rodríguez Olmos

1. Título del dataset: filmaffinity
2. Subtitulo: scraping de la web https://www.filmaffinity.com. Realización de un programa que descargue, organice y almacene datos sobre cada una de las películas almacenadas en la web en forma de tabla .csv.
3. Imagen identificativa:

Imagen corporativa de filmaffinity.com

4. Contexto: Filmaffinity es originalmente una web de recomendación de películas. Posteriormente incorporó a su negocio y estructura el almacenamiento de críticas de los usuarios, de forma similar a Rotten Tomatoes y IMBD. De la página sobre filmaffinity en la Wikipedia encontramos:

   ""FilmAffinity fue creado en Madrid en el año 2002 por el crítico de cine Pablo Kurt Verdú Schumann y el programador Daniel Nicolás.[3] Desde un principio la página constaba de un sistema recomendador de películas llamado "Almas gemelas", el cual mostraba las personas más afines en función de las puntuaciones que se dan a las películas. Tres años después se lanzó la sección de críticas, en donde los usuarios expresaban su opinión sobre una película."

   Filmaffinity recoge una ficha técnica de cada película almacenando varios datos, como título, año, director, etc.. así como un sistema de puntuación de películas basado en la media de las puntuaciones otorgadas por cada usuario (fuente: Wikipedia)

5. El dataset que hemos preparado se basa en la recolección de varios campos pertenecientes a la ficha técnica y la puntuación media de cada película. Notar que aunque la ficha técnica es inmutable, la puntuación es dinámica y puede cambiar a lo

largo del tiempo. De todas formas hemos decidido incorporar este dato también en nuestro dataset. Los campos que se han recogido de cada película son:

- Id
- Título
- Año
- Duración
- País
- Dirección
- Guión
- Música
- Fotografía
- Productora
- Reparto
- Género
- Sinopsis
- Nota
- Votaciones
- Web

El dominio temporal de los datos no es fijo, ya que se van añadiendo continuamente nuevas películas, tanto porque se ruedan más, evidentemente, pero también porque filmaffinity va incorporando en su base de datos películas antiguas que, o bien son reeditadas, o bien no estaban aún indexadas por la web. Podemos decir que su dominio temporal es la historia del cine y la TV, ya que además de películas también se incluyen series, documentales, etc… El proceso de recolección de los datos ha sido por medio de un software que primero explora todas las paginas del índice (https://www.filmaffinity.com/es/allfilms.html) y extrae la id de cada película, almacenarlas en una tabla (movies_id) auxiliar que también es descargada. Posteriormente se descarga cada página correspondiente a los índices almacenados y realiza un scraping para extraer los campos que estamos interesados en conservar. Precisamente por la naturaleza dinámica de la página, que constantemente incorpora nuevas películas, el código que hemos realizado es reutilizable, en el sentido de que si queremos continuar la descarga en un momento posterior, bien por haber sido baneados de ella o bien porque queremos actualizar la tabla descargada, se puede retomar el proceso en el punto que se abandonó en la última ejecución.

La tabla csv que hemos subido a Github corresponde a una version de test reducida, que contiene solo las películas indexadas por las letras X, Y, Z. Adicionalmente se adjunta la tabla auxiliar correspondiente. La ejecucion del codigo pregunta primero al usuario si se quiere descargar el conjunto de test o la web completa. En todos los casos, y tal como se ha comentado, el proceso se puede realizar en un numero arbitrario de etapas.

6. El propietario de los datos es la web filmaffinity.com. Los propietarios del material gráfico son las productoras de cada una de las películas (o series o documentales). No nos hemos basado en análisis anteriores para esta práctica.

7. La motivación de esta elección para realizar la práctica es por un lado técnica, ya que nos ha permitido enfrentarnos a un web scraping de una web completa de grandes dimensiones. Por otro lado, desde un punto de vista estadístico, el dataset que hemos descartado puede permitir analizar la historia del cine como por ejemplo:

- Realizar agregaciones del número de películas por ciertos campos como director, género, etc...
- Estudiar cómo se relacionan las valoraciones de las películas por género, año, número de actores, etc…
- Cualquier tipo de estudios similares que necesiten acceder de forma rápida a una recopilación de las fichas técnicas de toda la historia del cine y la TV.

8. Licencia: El software se encuentra bajo licencia MIT. Esta licencia no tiene restricciones, permite el uso, copia, modificación, integración con otro software, publicación, distribución, sublicenciamiento y uso comercial del código. Estos derechos están sujetos a la condición de que se incluya la nota de copyright y la parte de los derechos en todas las copias o partes sustanciales del Software. Esta condición invalida la licencia en caso de no cumplirse.

   El dataset obtenido mediante el uso del programa se encuentran bajo licencia creative commons by-nc-sa 4.0. Esta licencia permite copiar y redistribuir el material en cualquier medio o formato y adaptarlo o modificarlo bajo las condiciones de: otorgar el crédito apropiado e indicar si se realizaron cambios en los datos, no realizar un uso comercial de ellos y que los proyectos en los que sean utilizados dispongan de la misma licencia. Para más información se puede visitar el sitio web de creative commons:

   https://creativecommons.org/licenses/by-nc-sa/4.0/

9. El código ha sido realizado en Python 3 y es accesible en el enlace de Github:

   https://github.com/gabvilpi/film-scrap

10. El dataset se encuentra en el mismo enlace de Github.