PRÁCTICA 1

WEB SCRAPING

TIPOLOGÍA Y CICLO DE VIDA DE LOS DATOS

Lucas de Torre Barrio
Abril de 2022

Índice

1. Contexto

La intención de este trabajo es recoger información de las mejores películas de la historia con la intención de, posteriormente, poder observar qué características son habituales entre estas exitosas películas.

Para ello, se hace uso del listado de la página web de Filmaffinity, concretamente, del listado de las 1.000 mejores películas de la historia de las que tienen al menos 1.000 votos (`https://www.filmaffinity.com/es/ranking.php?rn=ranking_fa_movies`). Este listado se elabora con las películas mejor valoradas por los usuarios de la plataforma y, el hecho de necesitar tener al menos 1.000 votos para poder aparecer en el listado, evita que aparezcan películas con una cantidad tan pequeña de votos que no sea significativa.

2. Título

Las 1.000 mejores películas con al menos 1.000 votos de Filmaffinity.

3. Descripción del dataset

El conjunto de datos contiene información de cada una de las películas referida a su proceso de creación, su resultado final y sus valoraciones.

Respecto a la creación aparecen distintas personas o entidades que hayan participado en el desarrollo de la película (como directores o guionistas entre otros).

Respecto a su resultado final aparece su duración.

Respecto a sus valoraciones aparece su valoración en Filmaffinity y el número de votos que ha tenido.

4. Representación gráfica

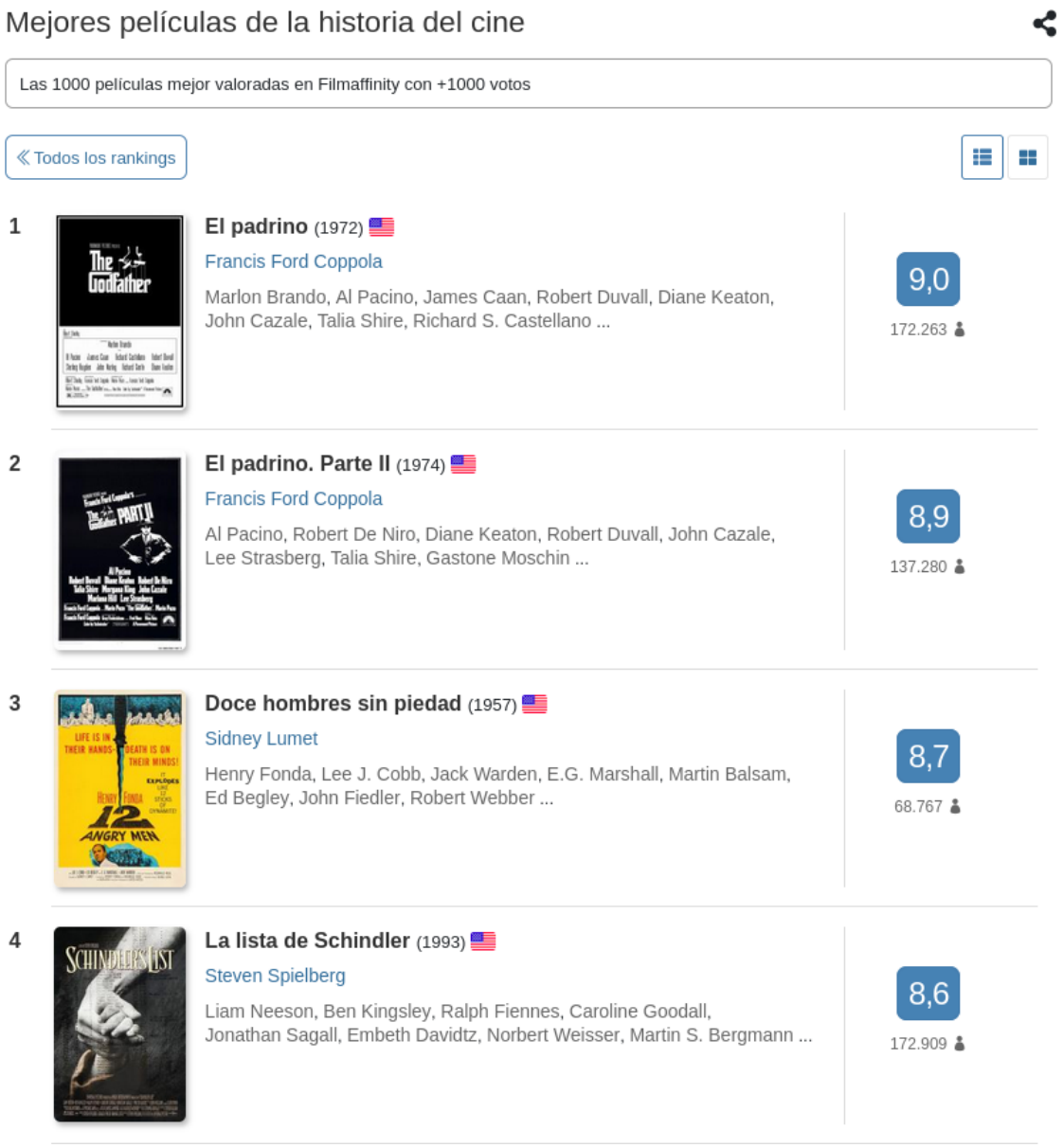

Figura 1: Ranking de Filmaffinity

5. Contenido

Cada fila del dataset se corresponde con una película. Los campos recogidos han sido:

- **titulo**: Contiene el título original de la película.
- **año**: Contiene el año de publicación de la película.
- **pais**: Contiene el país de producción de la película.

- **valoración**: Contiene la valoración media de la película en Filmaffinity.
- **votos**: Contiene el número de votos de la película en Filmaffinity.
- **direccion1**: Contiene el nombre de la persona encargada de dirigir la película.
- **direccion2**: Contiene el nombre de la segunda persona encargada de dirigir la película en caso de haber más de una.
- **direccion3**: Contiene el nombre de la tercera persona encargada de dirigir la película en caso de haber más de dos.
- **guion1**: Contiene el nombre de la persona encargada de escribir el guion de la película.
- **guion2**: Contiene el nombre de la segunda persona encargada de escribir el guion de la película en caso de haber más de una.
- **guion3**: Contiene el nombre de la tercera persona encargada de escribir el guion de la película en caso de haber más de dos.
- **produccion1**: Contiene el nombre de la persona o compañía encargada de producir de la película.
- **produccion2**: Contiene el nombre de la segunda persona o compañía encargada de producir la película en caso de haber más de una.
- **produccion3**: Contiene el nombre de la tercera persona o compañía encargada de producir la película en caso de haber más de dos.
- **genero1**: Contiene el género la película.
- **genero2**: Contiene el segundo género en caso de haber más de uno.
- **genero3**: Contiene tercer género en caso de haber más de dos.

Los datos se corresponden con todas las películas existentes en la historia y han sido extraídos mediante la página web de Filmaffinity (`https://www.filmaffinity.com/es/main.html`).

6. Agradecimientos

Agradecemos a Filmaffnity poner a disposición os datos de todas las películas y a todos sus usuarios el haber participado en las votaciones para poder elegir estas 1000 mejores películas.

Para verificar el cumplimiento legal, hemos consultado el archivo robots.txt de la página (`https://www.filmaffinity.com/robots.txt`) y hemos comprobado que los links consultados mediante herramientas de web scraping no están incluidos entre los que consideran como no permitidos.

7. Inspiración

Este conjunto de datos puede tener distintas utilidades. Se puede utilizar para intentar predecir qué películas pueden ser un éxito según la crítica (por ejemplo si resulta que se encuentra que hay películas dirigidas por D

y guionizadas por G que están en esta lista, una película que vuelva a estar dirigida or D y guionizada por G es probable que pueda ser un éxito de crítica).

También se puede utilizar para buscar directores, guionistas o géneros en los que las películas están muy bien valorados. Se podría encontrar (no es el caso) que no hay películas de Drama entre las 1000 mejores películas, o de que un director ha dirigido 20 de las películas de la lista.

8. Licencia

La licencia escogida para la publicación de este conjunto de datos ha sido CC BY-SA 4.0 License. Los motivos que han llevado a la elección de esta licencia tienen que ver con la idoneidad de las cláusulas que esta presenta en relación con el trabajo realizado:

- Se debe proveer el nombre del creador del conjunto de datos generado, indicando los cambios que se han realizado. De esta manera, se reconoce el trabajo ajeno y en qué medida se han realizado aportaciones en relación con el trabajo original.
- Se permite un uso comercial. Esto haría que incrementen las probabilidades de que una empresa utilice los datos generados y realicen trabajos de calidad que reporten cierto reconocimiento al autor original.
- Las contribuciones realizadas a posteriori sobre el trabajo publicado bajo esta licencia deberán distribuirse bajo la misma. Esto hace que el trabajo del autor original continúe distribuyéndose bajo los términos que él mismo planteó.

9. Código

El código en python utilizado para la extracción del dataset está disponible en `https://github.com/ldetorreUOC/Practica1-Web-scraping/blob/main/web_scraping.py`.

10. Dataset

El csv con los datos extraídos se encuentra en `https://github.com/ldetorreUOC/Practica1-Web-scraping/blob/main/web_scraping.py`. Los campos están separados por ‘,‘.

El **DOI** correspondiente es: **10.5281/zenodo.6429855**