¿CÚAL PONEMOS HOY?

Las 1000 mejores películas de la historia del cine, según *Filmaffinity*

Diana Celine Pérez Rojas, Ariadna Romero Montero
Tipología y ciclo de vida de los datos
Universitat Oberta de Catalunya
Abril 2023

Índice

1. Contexto

1.1. ¿Qué es filmaffinity?

film affinity se define como "la principal web prescriptora de cine en Internet con una completa base de datos" [1]. Es una página de votación y recomendación de películas y series la cual se basa en en la afinidad entre los usuarios tal como su nombre indica. Permite encontrar a personas más afines de los usuarios y, así, poder generar recomendaciones.

Fue fundada en 2002 por los españoles Pablo Kurt y Daniel Nicolás, con el objetivo de ser una web en referencia en español para los "cinéfilos" [2]. Dos décadas más tarde, los recién nombrados siguen siendo los socios y propietarios únicos. Sin embargo, la proyección ha aumentado. Aunque Filmaffinity tuvo sus inicios en España, en el año 2016 iniciaron una expansión en varios países de Latinoamérica con versiones personalizadas para cada región [3]. En el mundo anglosajón, *Metacritic* se había lanzado el año anterior, *Rotten Tomatoes* en 1998 y *IMDB* en 1990. En solo dos años de la creación de la web, la edición estadounidense de PC Magazine la incluyó entre las 100 mejores webs por descubrir del mundo [4].

Filmaffinity es una red social de cine en la que los usuarios pueden escribir sus críticas de películas y ayudar a la comunidad. En la ficha de cada película/serie que se muestra, se pueden ver fragmentos de reseñas de críticos profesionales, así como la nota media de los usuarios, lo que hace de Filmaffinity un lugar donde se puede combinar la crítica amateur y profesional, y compartir conocimientos sobre cine [4]. En cuanto a los filtros que deben pasar las críticas, la plataforma cuenta con procesos de validación para evitar textos sin sentido, así como un sistema para evitar "los fusiladores", personas que votan negativamente en todas las críticas. Además, han implementado nuevos servicios entre los que se incluye una sección en la que se pueden visualizar las últimas cincuenta críticas de usuarios registrados, ya sea de películas recientes o antiguas y de diferentes géneros y nacionalidades. Por la banda de red social, también puede mostrar las críticas de los usuarios conectados con el usuario en la red. De esta manera se consigue también un ambiente de confianza el cual, si una persona sabe que tiene los mismos gustos que otra, podrá ser capaz de ver sus opiniones y recomendaciones siempre que lo necesite.

El éxito de *filmaffinity* se debe principalmente a su sistema de recomendación personalizado y la combinación efectiva de información y opiniones. A la vez, la web es muy valiosa para la publicidad de estrenos de cine debido a su perfil de usuario, que coincide con los consumidores de cine. Otro de los factores clave del éxito del sitio web es que los cinéfilos apasionados a menudo se aficionan al sistema de votación de la plataforma pues, al acumular votaciones, pueden superar a sus contactos en cantidad de votos y competir con otros conocidos cinéfilos agregados a la red.

En comparación con otras webs de cine de referencia como es el caso de IMDb, Filmaffinity se encuentra en una posición más alta en el ranking mundial de sitios web visitados, seguramente debido a la mayor interacción entre usuarios y medios de recomendación como veníamos comentando. Además, ofrece la posibilidad de llevar un registro personalizado de todo lo que se ha visto, lo que se quiere ver en el futuro, y la opinión de amigos sobre lo

que deberíamos ver. Esto incluye una base de datos personal de todas las películas vistas y la posibilidad de comparar el número de votaciones con amigos y otros usuarios de la plataforma [5].

El hecho de que Filmaffinity no pertenezca a ningún medio o grupo de comunicación y, además, no sea una página de e-commerce, refuerza su imagen de credibilidad. En 2022, en el vigésimo aniversario de su lanzamiento, contaba con un millón de usuarios registrados y cerca de 800 000 críticas [6].

1.2. Datos de películas, en español

Uno de los puntos fuertes de *filmaffinity* es la elaboración de listas por parte de los usuarios.

En el contexto de este trabajo, se pretende capturar información sobre las "Mejores películas de la historia del cine, que en esta página son las 1000 películas mejor valoradas en Filmaffinity con más de 1000 votos.

En la ficha de cada título aparece información como el año de publicación, la duración, el país de producción, género, entre otra información.

Hemos escogido esta web de *Filmaffinity*, porque representa un sitio agregador de películas de referencia en español, tanto para España como para América Latina que, además, nos permite recopilar mucha información cinematográfica sobre los diferentes títulos.

Si bien es cierto que Filmaffinity no era nuestra idea principal en la cual recopilar información, nos hemos encontrado con otros sitios web que expresamente indican que no desean que se utilicen robots o rastreadores (Crawlers), como es el caso de IMDb.

2. Descripción del dataset

El dataset contiene 1000 filas, en adición a la columna de encabezado, y 10 columnas, que representan las siguientes variables:

- Título
- Título original
- Año de lanzamiento
- País(es) de producción
- Director(es/as)
- Compañías de producción
- Género(s)
- Reparto
- Sinopsis

- Calificación/Rating

Esta información proviene de la ficha de cada película:

Figura 1: Ficha de la película “El padrino (1972)”

Puesto que se recoge esta información para cada película, el conjunto de datos tendrá un total de 10 000 “casillas”.

3. Contenido

Por la parte del contenido, nuestro dataset recoge las 1000 películas mejor valoradas en Filmaffinity con más de 1000 votos. Las películas comprenden toda la historia del cine hasta día de hoy, 25 de abril de 2023. No obstante, al tratarse de un ranking que se actúa en función de las votaciones de los usuarios, el dataset está sometido a variaciones según cuándo damos la orden del scrapy crawl.

El conjunto de datos que hemos realizado tiene 10 columnas, como se ha mencionado previamente:

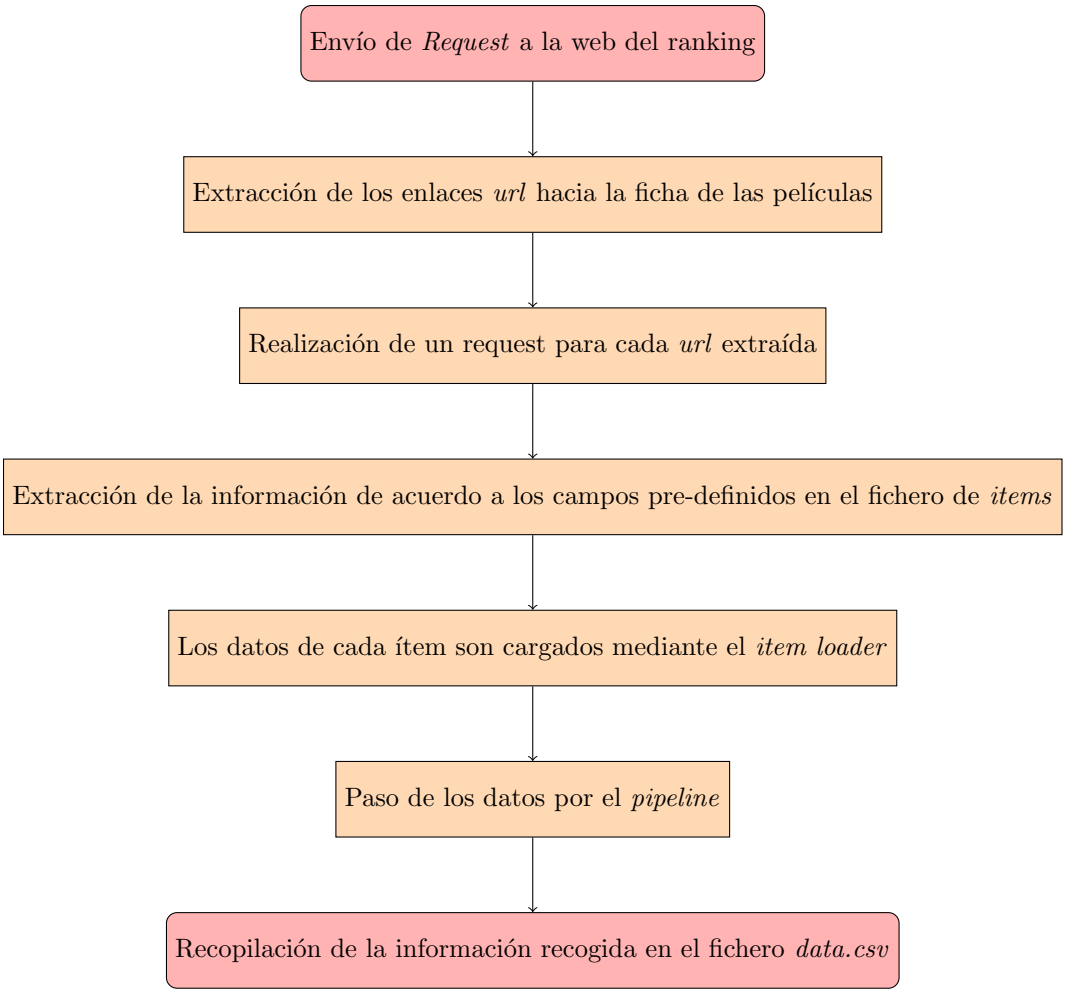
- Cast – “character”: Conjunto de actores que conforman el reparto de la película.

- Country – “character”: país de realización del film.
- Directors – “character”: se corresponde con la persona o las personas encargadas de la dirección de la película.
- Genre – “character”: se corresponde al género o géneros en el que se clasifica la película. P.e. “drama”, “comedia”, “terror”, etc.
- Original_title – “character” se corresponde con el título original de la película. Es decir, el título sin haber sido sometido a traducciones. P. e. el título original de “El Padrino” es “The Godfather”
- Production – “character”: en este apartado se referencias las productoras con las que se ha llevado a cabo el filme. P.e. “Paramount Pictures”.
- Rating – “numeric”: valoración/puntuación en una escala del 1 al 10 obtenida de la media de votaciones en Filmaffinity.
- Sinopsis – “character”: Pequeño resumen de la película con el que se presenta la trama.
- Title – “character”: se corresponde con el título de la película en su versión en español. P.e. “El Padrino”

4. Representación gráfica del proyecto

Una copia de los datos que se recopilarán estarán disponibles en el archivo “data.csv”, dentro del repositorio.

El ciclo que siguen los datos desde su captura a su procesamiento es el siguiente:



5. Código

El código se basa fundamentalmente en el paquete *scrapy*, un entorno de *crawling* y *scraping* [7] [8]. También hacemos uso de los paquetes *scrapy_selenium*, *selenium*, y *random* y textittime en algunos casos concretos.

Un proyecto estándar de *Scrapy* contiene los siguientes archivos/directorios:

- Fichero de configuración del despliegue: *scrapy.cfg*
- Módulo Python del proyecto: *spider/*
  - Fichero de definición de items del proyecto: *items.py*
  - Fichero de carga de items del proyecto: *itemsloader.py*
  - Archivos middleware del proyecto: *middlewares.py*
  - Archivos pipeline del proyecto: *pipelines.py*
  - Archivo de configuración del proyecto: *settings.py*
  - Directorio donde se encuentra la "araña": *spiders/*

- Archivo de la araña: *filmspider.py*

Puesto que la página del ranking requiere que se presione un botón para cargar las películas, se hace uso de un *webdriver*, del paquete Selenium [9]. Para poder usar esta funcionalidad, es necesario tener instalado en el ordenador de ejecución la última versión disponible del navegador Chrome y tener dentro del directorio de archivos el archivo del Chrome Web Driver

A continuación se comentarán puntos claves sobre algunos de los archivos dentro del proyecto:

5.1. items.py

Se opta por estructurar la información extraída de las páginas web en un ítem, en este caso, como diccionarios, donde las claves son los nombres de las variables/columnas, y los valores son los que corresponda a cada película.

5.2. itemsloader.py

Como se indica en la documentación, si bien los elementos se pueden rellenar directamente, los cargadores de elementos o *items loaders* permiten mucha más flexibilidad: se pueden procesar algunos valores de los elementos, como por ejemplo, en el caso de los títulos, podemos indicar que se deben eliminar los caracteres de espacio, o quitar los signos de puntuación en la categoría de producción, entre otros.

5.3. settings.py

Este archivo permite personalizar el comportamiento de todos los componentes de Scrapy, incluidos el núcleo, las extensiones, las canalizaciones y las propias arañas.

En el contexto de la araña que se utiliza en este proyecto, destacan las siguientes "configuraciones":

- FEEDS: Indicamos que el *output* será en formato *csv* y que, en el caso de que la araña se ejecute, sobreescribirá el fichero.
- ROBOTSTX_OBEY: De esta forma indicamos a la araña que debe obedecer las indicaciones que recoje este fichero para la url en cuestión.
- DOWNLOADER_MIDDLEWARES y ITEM_PIPELINES: para configurar el uso de los *middlewares* y *item_pipelines*

5.4. pipelines.py

Los *pipelines* - o tuberías, en castellano - son clases de Python que implementan un único método. Se usan para limpiar los datos, que es la tarea para la que se emplean en esta araña.

En el fichero aparece un único *pipeline*, que usamos para eliminar los espacios en blanco que se pueden crear en algunos campos derivado de las funciones usadas por los *item loaders*.

5.5. filmspider.py

Este código de Python utiliza, como se ha comentado previamente, el marco de rastreo web Scrapy, además del paquete scrapy_selenium para recuperar el contenido dinámico de la página, y el *web driver* para poder interactuar con la página web del ranking.

La araña empieza definiendo una lista de agentes de usuario, que se utilizarán para realizar solicitudes al sitio web, y una función get_random_agent(), que devuelve un agente de usuario aleatorio de la lista.

A continuación se define la clase FilmspiderSpider, que hereda de la clase scrapy.Spider. Tiene el atributo name establecido en "filmspider" y el atributo allow_domains establecido en "filmaffinity.com/es". Se establece la página url que contiene el ranking en *start_url*.

Luego, la araña define el método *start_requests*, que es desde donde se hará la petición a la url de interés. Con el método GET , el *driver* que ha sido iniciado, navegará a la página del ranking y a continuación realizará una comprobación simple: si en la página se encuentra el botón para mostrar más películas, lo que indica que la totalidad de la lista no se ha mostrado, se hará click en ese botón. Esta acción se realizará sucesivamente hasta que este botón no se encuentre más.

Cuando esto suceda, se extrae el contenido de la página web. Mediante el *xpath*, se extraen todos los links que llevan a las fichas de cada película y, posteriormente, se hará un SeleniumRequest a cada una de estas URLS, con el *User-Agent* siendo escogido de manera aleatoria entre los que se han incluido en la lista. El uso de un agente de usuario aleatorio ayuda a evitar la detección y evita que la araña sea bloqueada. Simultáneamente se hace una llamada a la función *parse item*.

SpiderItemLoader se utiliza para extraer los datos de la página, incluidos el título de la película, el director, el año, la clasificación y el número de votos. Los datos extraídos luego se cargan en un SpiderItem.

5.6. Dificultades

Una de las principales dificultades que se encontraron en este trabajo fue el uso del *webdriver* para poder obtener la totalidad de las películas incluidas en el ranking, ya que inicialmente el *output* obtenido reflejaba solo las primera 30 películas.

En adición, debido a la gran cantidad de peticiones que se hizo, la araña no pudo finalmente recoger la totalidad de la lista: en el dataset aparecen menos de 800 películas, aunque esto representa una cantidad considerable de datos. Esto se debió a que el servidor captó la utilización del *scraper* y ya no recibía respuesta en sus requests:

Lo cual sugiere una mejor utilización de las configuraciones (dentro de *settings.py*) para poder limitar el número de peticiones concurrentes.

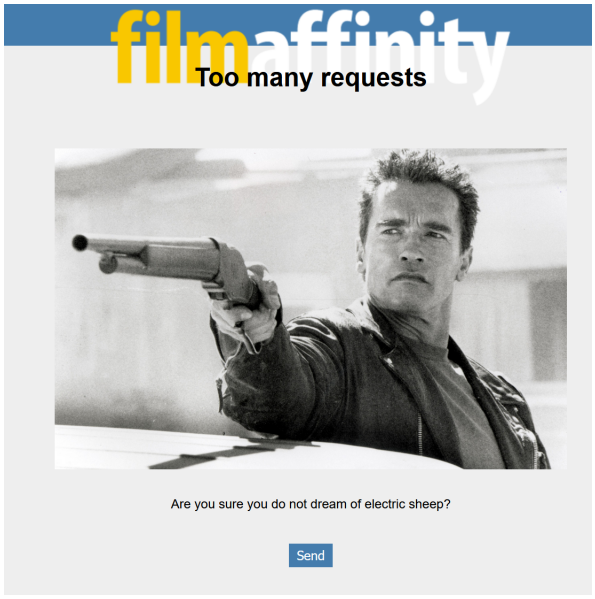


Figura 2: Mensaje de filmaffinity

6. Propietario

Filmaffinity es un portal independiente propiedad de Filmaffinity S.L., CIF B84485325, con domicilio social en Madrid, España. Dicha empresa sigue teniendo como accionistas únicos a los creadores de Filmaffinity, Pablo Kurt Verdú Schumann y Daniel Nicolás Aldea.

Su licencia de contenido es "Todos los derechos reservados", por el cual "cualquier uso que afecte a los derechos de propiedad intelectual requiere de la autorización de sus titulares", de acuerdo al Centro Español de Derechos Reprográficos (CEDRO) [10].

7. Inspiración

Con la llegada de las plataformas de *streaming*, el desarrollo de contenido audiovisual ha incrementado considerablemente en los últimos años, por lo puede resultar difícil escoger qué serie o película ver entre tanta oferta. La existencia de páginas como filmaffinity pueden ayudar a facilitar esta "tarea".

Hacer una base de datos con información sobre el *top* 1000 de películas mejor valoradas permite, además de su propósito básico como "recomendación", ver si alguna de las películas lanzadas actualmente han podido "colarse" entre las mejores valoradas, o ver si hay algún género, director o tema predominante en las películas que integran esta lista.

8. Licencia

La licencia que aplicará para este repositorio es MIT License, una licencia simple y permisiva, que es de las más populares en GitHub.

Por una parte, permite la modificación, el uso privado, el uso comercial y su distribución. Por otra parte, una copia de la licencia y de la notificación de copyright debe acompañar al material - en el directorio del repositorio se encuentra una copia de la licencia -.

9. Contribuciones

Contribuciones	**Firma**
Investigación Previa	Diana Celine Pérez Rojas, Ariadna Romero Montero
Redacción de las respuestas	Diana Celine Pérez Rojas, Ariadna Romero Montero
Desarrollo del código	Diana Celine Pérez Rojas, Ariadna Romero Montero
Participación en el vídeo	Diana Celine Pérez Rojas, Ariadna Romero Montero

10. Dataset (Zenodo)

El enlace de Zenodo que permite acceder al dataset es el siguiente: `https://zenodo.org/record/7860540#.ZEf0Fc7P07F`

11. Vídeo (Google Drive)

El enlace para acceder al vídeo es el siguiente: `https://drive.google.com/file/d/165rhk6UH4k8Cvk4LtxUaya53JwsQaacW/view?usp=sharing`

Referencias

[1] film affinity, "¿qué es filmaffinity y cómo funciona?" n.f., acceso el 20-04-2023. [Online]. Available: https://www.filmaffinity.com/es/site-guide.php

[2] J. M. Blanco, "El español que viajó a canadá para montar filmaffinity, el imdb patrio," 2016, acceso el 20-04-2023. [Online]. Available: https://www.eldiario.es/hojaderouter/internet/filmaffinity-pablo-kutz-estrenos-cine-cartelera-recomendaciones-criticas_1_3851937.html

[3] E. de FA, "Sobre filmaffinity," n.f., acceso el 20-04-2023. [Online]. Available: https://www.filmaffinity.com/es/meetus.php

[4] M.-N. G. . F.-L. S. Gavilán, D., "Influencia social en las comunidades de cine: filmaffinity como caso de estudios," *Estudios sobre el mensaje periodístico*, vol. 24, no. 1, pp. 551–565, 2018.

[5] M. Iris, "Análisis crítico sobre la crítica audiovisual: de bazin a filmaffinity," *Universidad Politecnica de Valencia*, 2012.

[6] R. Augusto, "Filmaffinity: dos décadas recomendando cine," 2022, acceso el 20-04-2023. [Online]. Available: https://www.rtve.es/noticias/20220411/filmaffinity-dos-decadas-recomendando-cine/2330186.shtml

[7] S. developers, "Scrapy 2.8 documentation," 2023, acceso el 20-04-2023. [Online]. Available: https://docs.scrapy.org/en/latest/

[8] ScrapeOps, "Scrapy beginners series part 1: How to build your first production scraper," n.f., acceso el 20-04-2023. [Online]. Available: https://scrapeops.io/python-scrapy-playbook/scrapy-beginners-guide/

[9] B. Muthukadan, "Selenium with python," n.d., acceso el 24-04-2023. [Online]. Available: https://selenium-python.readthedocs.io/index.html

[10] CEDRO, "¿qué significa «todos los derechos reservados»?" 2019, acceso el 20-04-2023. [Online]. Available: https://www.cedro.org/blog/articulo/blog.cedro.org/2019/03/19/que-significa-todos-los-derechos-reservados