Cesar Nicolás Ramos Milián
María Rollán Martínez Herrera

Práctica 1: ¿Cómo podemos capturar los datos de la web?

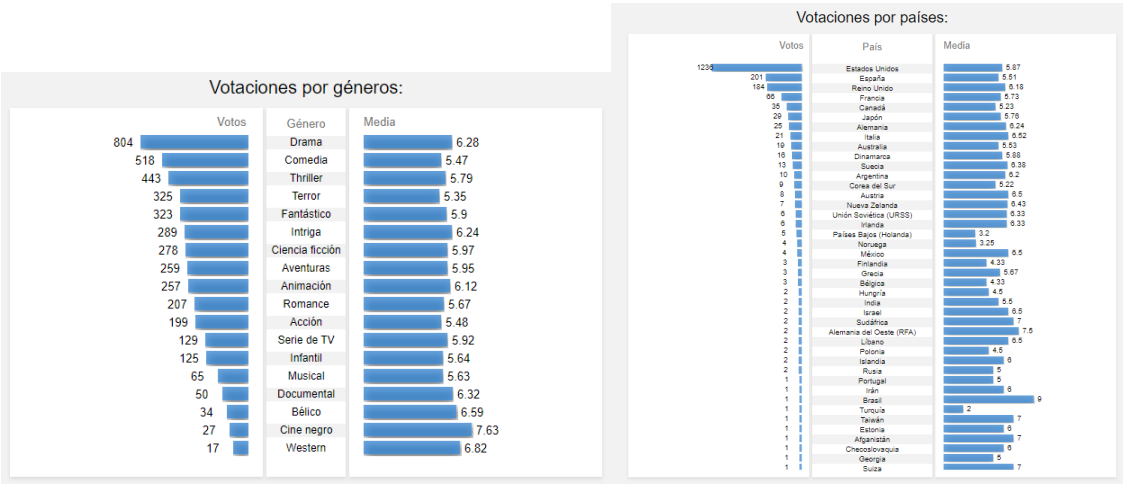
Contexto

Filmaffinity es una web en la cual, entre otras cosas, los usuarios pueden votar las películas que han visto, y dichos votos quedan registrados. La lista de votaciones de cada usuario es pública, y cualquiera puede acceder a ella.

Nuestro objetivo en esta práctica es, aplicando web scraping, un programa mediante el cual un usuario puedadescargar la información de todas las películas que ha votado. Del dataset resultante se podrán obtener diferentes estadísticas, como cuál es el director que mejor valora el usuario, cual es el guionista que mejor valora el usuario. Esto es útil si se quieren buscar nuevas películas basándose en los gustos del propio usuario. Más que un ejercicio de investigación es una herramienta orientada para que cada usuario de la web pueda analizar su propio perfil de "Cinéfilo" y obtener todas las características de su historial de visualizaciones.

Filmaffinity tiene buenas herramientas para el análisis de los metrajes con mejor puntuación a nivel global. También permite a cada usuario ver sus propias estadísticas, sin embargo, estás son limitadas: votaciones por valoración, votaciones por género, votaciones por países y votaciones por año.

Las votaciones por género solo incluyen el género principal, pero no los secundarios, por ejemplo, si una película es Drama, Comedia Dramática, Sátira, solo incluye Drama. Además, hay muchas más estadísticas posibles que se podrían extraer como votaciones por director, votaciones por guionista, votaciones por fotografía. Y diferentes maneras de ordenarlas, en este caso están ordenadas por número de votaciones, pero se podrían ordenar por nota media. También se podrían hacer gráficos tipo boxplot, ya que la nota media no refleja del todo bien como se comportan todas las películas de un género, por ejemplo.

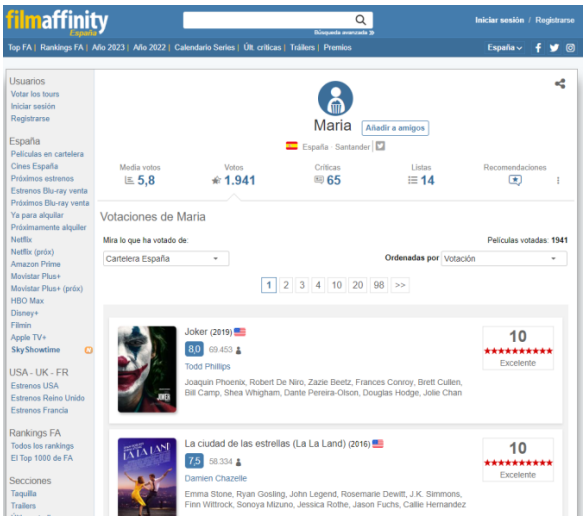


Esta es la razón por la cual creamos este programa que solventa las carencias de esta página web para analizar preferencias por géneros, directores, guionistas...

La página web desde donde se ha extraído la información es filmaffinity. Es un repositorio de películas, series y documentales con más de 800.000 usuarios registrados, 147 millones de votaciones y más de 740.000 críticas escritas. En los perfiles de cada metraje podemos ver datos como el año de estreno, directores o sinopsis entre otros.

La dirección de la página web desde donde extraemos nuestro dataset es https://www.filmaffinity.com, más en concreto desde la url del perfil de usuario. Este perfil de usuario podrá ser modificado antes de ejecutar las distintas celdas de código.

https://www.filmaffinity.com/es/userratings.php?user_id=<**USER_ID**>&p=<**PAGINA**>&orderby=0



Título del dataset

El título del dataset lo hemos llamado "puntuacionesFA.csv". Nuestro análisis o estudio de los datos no va a basarse solo en las puntuaciones, pero debido a que Filmaffinity es una web orientada la votación de los distintos metrajes por parte de los usuarios hemos pensado que sería un nombre adecuado.

Descripción del dataset y contenido.

El datasetextraidoes el del usuario 968973, que corresponde a la cuenta de uno de nosotros dos, pero se podría haber obtenido de cualquier usuario, únicamente modificando el id de usuario que aparece al inicio del código. El dataset obtenido se compone de 14 columnas y 1941 filas. Las filas del dataset recogen las 1941 películas valoradas por el usuario y las columnas los siguientes datos referentes a cada metraje:

1. titulo: Es el título de la película dado para el mercado español.
2. año: El año de estreno de la película.
3. duración: La duración en minutos del film.
4. pais: País donde se ha producido el metraje.
5. directores: Lista de director/es que realizaron la película.
6. guionistas: Lista de guionista/s que guionizaron la película.
7. musicos: Lista de musico/s implicado/s en la banda sonora del metraje.
8. fotografos: Lista de fotógrafo/s que participaron en la película.
9. actores: Lista de intérprete/s que actúan en la película.
10. productores: Lista de compañía/s que financian el metraje.
11. generos: Lista de genero/s entre los que se clasifica el metraje.
12. nota: La nota media de todos los usuarios de Filmaffinity que votaron el film.
13. votos: Número de votos que ha recibido la película por parte de los usuarios.
14. votacion: Votación del usuario del perfil del dataset.

El periodo de tiempo al que pertenecen los datos no es relevante, ya que son datos inmutables aunque no fijos en el número de registros. Este periodo comprenderá el límite temporal desde el día de la primera crítica a un film hasta el día de la última.

Se podría hacer un filtrado por fechas, ya que Filmaffinity registra cuando se ha votado cada película, pero no nos parecía que tuviera sentido, ya que no nos imaginamos en qué contexto el usuario puede estar interesado únicamente en unas fechas concretas en las que haya votado las películas.

Propietario

El propietario del conjunto de datos que trabajamos es la propia plataforma de Filmaffinity, la cual genera y almacena los datos.

Los datos de la web no son datos de carácter personal, por lo que no incumple la legalidad. En la misma política de privacidad de la página indican que no proveen datos personales de los usuarios y su base de datos está legalizada en la Agencia Española de Protección de Datos bajo el código 2021140167.

Al ser una herramienta de uso personal solo está orientada al uso del dataset del usuario final. En el caso de que tome los datos de otro usuario no hay información que comprometa la privacidad del mismo.

Ejemplos de análisis similares de uso público:

https://github.com/gestur1976/filmaffinity-movie-info-scrapper

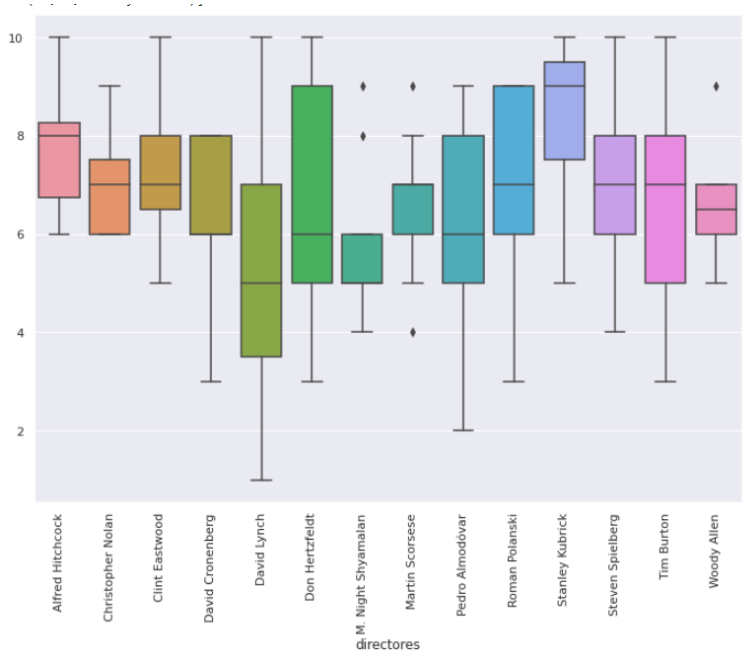
https://github.com/rubzip/FilmAffinity-s-lists-platforms

Inspiración

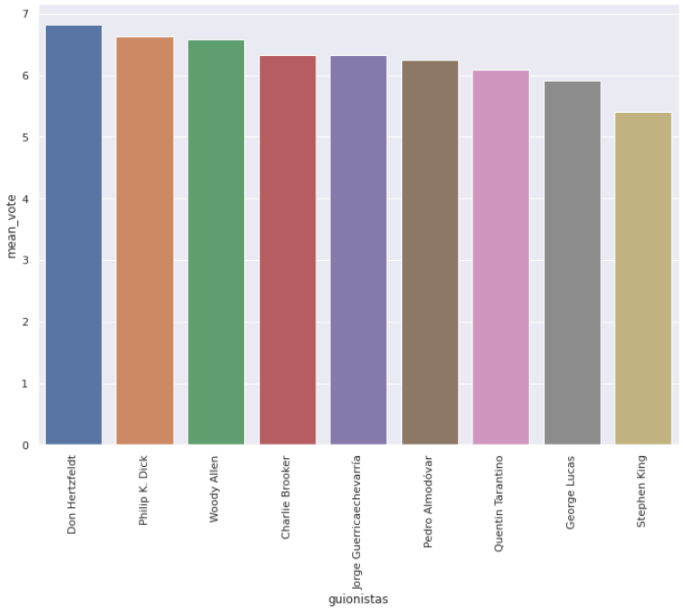
La idea de este método surgió porque uno de nosotros siempre ha querido tener un registro exhaustivo de las películas que había visto, y poder analizar los diferentes aspectos de las mismas, para así conocer mejor cuales son sus gustos y poder buscar futuros visionados que le vayan a gustar.

Con el estudio de este dataset se pueden obtener datos muy interesantes para el cinéfilo. Por ejemplo, una estadística interesante podría ser cuales son los directores más vistos y dentro de estos cuales son los mejor valorados.

Por ejemplo, a continuación, tendríamos los directores con más de 10 películas vistas y con sus valoraciones representadas mediante un box-plot. Podemos ver, por ejemplo, que David Lynch tiene unas valoraciones muy irregulares por parte del usuario, mientras que por ejemplo Hitchcock tiene todas las valoraciones entre el 6 y el 10. Así el usuario podría saber que si busca nuevas películas de Hitchcock es muy probable que estas le gusten.



También se podría hacer una gráfica más sencilla como las que presenta Filmaffinity de serie, pero con los guionistas.



Aquí tendríamos los guionistas con más de 10 votos, y la nota media de todas las películas votadas por el usuario a cada uno. Aquí el usuario puede saber que si busca películas guionizadas por Don Hertzfeldt es probable que le gusten.

Licencia

La licencia a escoger para el juego de datos sería la ReleasedUnder CC BY-SA 4.0 License. Las principales razones son:

-Los datos de Filmaffinity son de carácter público al consultar los usuarios.

-Queremos que se puedan hacer análisis de datos complementarios según las necesidades del usuario al usar la herramienta. Hemos creado varios gráficos que según nuestro criterio parecen los más interesantes, pero quizás otras personas tengan interés en analizar otros campos.

-Es una licencia que se atribuye al autor original.

Esta elección también justifica que incluyamos algunas columnas que no vamos a utilizar pero pueden ser útiles a posteriori.

Cesar Nicolás Ramos Milián
María Rollán Martínez Herrera

Diagrama de flujo del proyecto de data scraping

filmaffinity

Se plantean el proyecto a realizar a partir del web scraping de la web filmaffinity.

Selección entre los diferentes repositorios de la web el url para extraer los datos necesarios para el proyecto.

Creación de las funciones necesarias para el web scraping

Funciones:

-web

- number_of_pages

- extract_movie_info

- extract_df_per_page

Aplicación de las funciones con un bucle y creación del dataframe

titulo	año	duracion	pais	directores	guionistas	musicos	fotografos	actores	productores	generos	nota	votos	votacion
Joker	2019	121 min.	Estados Unidos	[Todd Phillips]	[Todd Phillips, Scott Silver]	[Hildur Guðnadóttir]	[Lawrence Sher]	[Joaquin Phoenix, Robert De Niro, Zazie Beetz,...	[DC Comics, DC Entertainment, Warner Bros., Vi...	[Thriller, Drama, Crimen, DC Comics, Cómic, Pa...	8.0	69408.0	10
La ciudad de las estrellas (La La Land)	2016	127 min.	Estados Unidos	[Damien Chazelle]	[Damien Chazelle]	[Justin Hurwitz]	[Linus Sandgren]	[Emma Stone, Ryan Gosling, John Legend, Rosema...	[Summit Entertainment, Gilbert Films, Impostor...	[Musical, Romance, Comedia, Drama, Drama román...	7.5	58319.0	10
Halt and Catch Fire (Serie de TV)	2014	60 min.	Estados Unidos	[Christopher Cantwell, Christopher C. Rogers, ...	[Christopher Cantwell, Christopher C. Rogers, ...	[Paul Haslinger]	[Nelson Cragg, Evans Brown, Jeff Jur]	[Lee Pace, Scoot McNairy, Mackenzie Davis, Ker...	[AMC Studios]	[Serie de TV, Drama, Internet / Informática, T...	7.6	4517.0	10

Exportar a formato csv:

- puntuacionesFA.csv

Contribuciones

Contribuciones	Firma
Investigación previa	MR,CR
Redacción de las respuestas	MR,CR
Desarrollo del código	MR,CR
Participación en el vídeo	MR,CR