Daniel Bagan Martínez
Rafael García Henríquez

Índice

Listado de películas de Rotten Tomatoes

Contexto

Debido a la situación actual de pandemia mundial, los servicios de streaming han tenido un auge excepcional. Las personas, al tener que cumplir con confinamientos y por ende pasar mucho más tiempo en sus hogares, están consumiendo con mayor velocidad series, películas, documentales, etc. Esta velocidad de consumo puede llevar a una búsqueda cada vez más exhaustiva (y quizás agotadora después de horas realizando "zapping" en aplicaciones de streaming) de material filmográfico que se ajuste mucho más a nuestros gustos y preferencias.

En este contexto, contamos con páginas web como **Rotten Tomatoes** [5], con más de 20 años prestando un servicio de reviews y puntuación de películas (tanto profesional como de los usuarios), con un catálogo de más de 10 mil películas. Aunque cuenta con un sistema de filtrado de películas, este carece de la posibilidad de realizar búsquedas más concretas como:

- Filtrar por fecha de estreno (se pueden ordenar las películas por fecha, pero no se puede seleccionar, por ejemplo, películas de un año en específico).

- Filtrar por idioma de origen.

- Realizar filtros más avanzados como: tiempo de duración de la película, idioma de origen y puntuación de la audiencia.

Por último, es importante destacar la significativa robustez que poseen los datos concernientes a las puntuaciones de cada película en la página; todo esto debido al tiempo y reconocimiento que posee su sistema de evaluación el cual se divide en:

- **Tomatometer:** Basado en la opinión de críticos de cine profesionales. Este comienza a calcularse cuando la película posee por lo menos 5 críticas profesionales y se subdivide a su vez en “fresh status” (si más del 60% de la crítica es positiva) o “rotten status” (si más del 60% de la crítica es negativa).

- **Audience Score:** Es la puntuación de la audiencia que, según su página web, “se pueda verificar que han adquirido el ticket”.

Descripción del dataset

Rotten Tomatoes es una web de reseñas para cine y televisión. El conjunto de datos generado en esta práctica contiene el listado e información de todas las películas disponibles en la web Rotten Tomatoes .

Representación gráfica

	Title	Tomatometer	Audience score	Rating	Genre
0	On The Rocks	86%	52%	R(SomeLanguage|SexualReferences)	comedy
1	The Sounding	63%	NotYetAvailable		mysteryandthriller,drama
2	Wolfman's Got Nards	100%	86%		horror,documentary
3	The Witches	52%	39%	PG(Language|ThematicElements|ScaryImages/Moments)	comedy,adventure,fantasy
4	J.R. 'Bob' Dobbs & The Church Of The SubGenius	100%	NotYetAvailable		comedy,documentary
5	The Place Of No Words	69%	NotYetAvailable		drama
6	Nomad: In The Footsteps Of Bruce Chatwin	91%	64%		documentary
7	Terra Willy: Unexplored Planet (Astro Kid)	100%	83%	PG(SomeMildPeril)	animation,adventure,kidsandfamily,s
8	The True Adventures Of Wolfboy	75%	NotYetAvailable	PG-13(SomeStrongLanguage|Drinking|SexualReferences|MatureThematicContent|Violence)	drama

Imagen 1: campos del dataset Title, Tomatometer, Audience Score, Rating y Genre.

Director	Producer	Writer
SofiaCoppola	SofiaCoppola,YoureeHenley	SofiaCoppola
EricStyles		CharlesSavage
StephenSchible		
AndreGower	NicholasCaprio,SarbaDas,RitaDoumar,MatthewCharlesDucey,AndreGower,CarlenaGower,AaronKunkel	AndreGower,HenryDarrowMcComas
AndreaDiStefano	WayneMarcGodfrey,JamesHarris,BasilIwanyk,RobertJones,MarkLane,EricaLee	MatthewCook,RowanJoffe,AndreaDiStefano
MarkWebber	DustinHughes,KaiLillie,TeresaPalmer,MarkWebber	MarkWebber
WernerHerzog	SteveO'Hagan,LuckiStipetic	WernerHerzog
EricTosti	Jean-FrançoisTosti	DavidAlaux,EricTosti,Jean-FrançoisTosti
MartinKrejcí	DeclanBaldwin,LaurenBeck,BenjaminBlake,JoshGodfrey,KimberlySteward	OliviaDufault
GavinMichaelBooth	GavinMichaelBooth,DavedWilkins	DavedWilkins,GavinMichaelBooth

Imagen 2: campos del dataset Director, Producer y Writer.

Release Date (Theaters)	Release Date (Streaming)	Runtime	Production Co
Oct2,2020 limited	Oct23,2020	1h36m	A24,AmericanZoetrope
	Nov3,2020	1h32m	GSPStudios
Jul6,2018 limited	Apr24,2019	1h41m	AVROTROS,NHK
	Oct27,2020	1h31m	
	Nov6,2020	1h53m	MaddemFilms,thefyzz,ImaginationParkEntertainment,ThunderRoadPictures
Oct23,2020 limited	Oct23,2020	1h35m	MythicalFilms,WideAwakeCinema
Apr8,2020 limited	Oct25,2020	1h29m	
Oct23,2020 limited	Oct27,2020	1h29m	FranceTélévisions,Procirep,France3Cinéma,OCS,CentreNationalduCinémaetdeL'imageAnimée,AgenceNationaledeGestiondes?uvresAudiovisuelles,BacFil
	Oct30,2020	1h28m	BigIndiePictures,KPeriodMedia
Sep25,2020 limited	Oct27,2020	1h17m	MimeticEntertainment
Oct30,2020 limited	Nov6,2020	1h21m	NationalFilmBoardofCanada,DocumentaryChannel,JohnWalkerProductions
	Oct23,2020	1h42m	SightUnseenPictures

Imagen 3: campos del dataset Release Date (Theaters y Streaming), Runtime y Producer Co.

Contenido

El dataset es un fichero CSV con información sobre todas las películas disponibles en la página web Rotten Tomatoes el día **9 de noviembre de 2020 a las 00:30**. Cada película contiene la siguiente información:

- **Title**: título de la película.
- **Tomatometer**: puntuación (sobre 100%) que otorga la propia página a la película, basada en la opinión de cientos de críticos.
- **Audience score**: porcentaje de usuarios de la web que han valorado la película positivamente.
- **Rating**: clasificación por edades de la película [6] y motivo de la clasificación. Por ejemplo, *R(SexualContent|SomeDrugMaterial)* indicaría una clasificación de "Restringido" (los menores de 17 años acompañados de un adulto) por contenido sexual y drogas.
- **Genre**: género o géneros de la película.
- **Original Language**: lenguaje original de la película.
- **Director**: director de la película.
- **Producer**: productor de la película.
- **Writer**: escritor del guión de la película.
- **Release Date (Theaters)**: fecha de lanzamiento en cines.
- **Release Date (Streaming)**: fecha de lanzamiento en streaming.
- **Runtime**: duración de la película.
- **Production Co**: compañía de producción.

Funcionamiento

El fichero *prueba_tomates.py* contiene el código encargado de realizar el Web Scrapping.

- Creación de un *WebDriver* con Selenium.
- Modificación del User Agent.
- Ingreso con el driver de Firefox en la página web https://www.rottentomatoes.com/browse/dvd-streaming-all/, donde se muestran 32 películas.
- Obtención del número de películas totales disponibles (un error en la página inutiliza este paso).

    - Existe un error al mostrar el total de películas disponibles en la página. Inicialmente aparece el mensaje "Showing 32 of 22505" (aproximadamente). Al hacer clic en *Show more*, el número total de películas baja a 17953, al siguiente clic a 17167, y así sucesivamente, hasta llegar al número real de, aproximadamente, 10000. Este error en la página impide que se pueda usar el número de películas mostrado para calcular cuántas veces se pulsa el botón *Show more*. Por lo que se ha tenido que usar otro método.

- Localización del botón *Show more*.
- Mientras el botón exista, se hace *scroll* hasta el botón y se clica en él hasta que aparezcan en la página todas las películas disponibles.
    - Cada vez que hay un error en la búsqueda o uso del botón, adición de un *time.sleep* exponencial entre solicitudes.
- Una vez se muestran todas las películas en la página, obtención de la lista de todas las etiquetas *h3* de clase "movieTitle", que contienen el nombre de la película, y *div* de clase "movie_info", que contienen su url.
- Para cada película:
    - Creación de un diccionario para almacenar su información.
    - Descarga de su correspondiente página con BeautifulSoup.
    - Obtención de su título.
    - Obtención de la puntuación de la propia página y de sus usuarios.
    - Obtención de la información del apartado de la página "movie info": género, director, fecha de estreno, duración, etc.
    - Almacenamiento del diccionario de la película en una lista.
- Creación de un DataFrame a partir de la lista de diccionarios de películas y creación del CSV a partir de este.

Resolución de obstáculos

En el código utilizado para realizar el *web scrapping* y generar el fichero CSV se han usado los siguiente métodos para evitar el colapso del servidor o el baneo de la web:

- Inspección de robots.txt.
- Modificación del User Agent en la cabecera de las peticiones para evitar baneo.
- Adición de un tiempo de espera exponencial cada vez que se produce un error.

Además, aunque no ha sido necesario su uso en la generación del dataset final, se han probado con éxito los siguientes métodos en listas pequeñas para eludir un posible baneo [4]:

- Generación de User Agent aleatorio.
- Uso de direcciones proxy [2].

Agradecimientos

Los datos y puntuaciones de las películas se han extraído de la web Rotten Tomatoes [3].

Inspiración

La web Rotten Tomatoes ofrece un listado de todas las películas disponibles. Estas películas se pueden filtrar por *tomatometer*, género y plataforma proveedora y ordenar por *tomatometer* o fecha de lanzamiento.

La creación de este dataset permite seleccionar otros criterios para filtrar, como por duración de la película y productor, o ordenar por otros criterios, como la puntuación de la audiencia. Por ejemplo, si quiero ver una película de terror o suspense que esté disponible en *Netflix* o *Amazon* y dure menos de dos horas, a partir de este dataset se me podría proporcionar un listado de las películas con estas características ordenado por la puntuación de la audiencia.

Además, los datos disponibles sobre cada película pueden utilizarse en un proyecto de minería de datos; por ejemplo, se podría crear un modelo de predicción que estime la aceptación por la crítica que va a tener una película según sus características, o se pueden establecer clasificaciones o agrupamientos.

Licencia

La selección de licencia se realizó siguiendo respondiendo las preguntas de la página web License Selector [3], aconsejada por la red de bibliotecas CSIC [1], sobre el dataset generado. Para ello se consideró lo siguiente:

- El dataset se encuentra dentro del alcance del Copyrights y derechos relacionados.

- No se poseen derechos de autor ni derechos similares sobre el conjunto de datos ni sus partes constitutivas.

- Todos los elementos del conjunto de datos están al alcance del dominio público.

- Se permite a terceros realizar trabajos derivados.

- Se le requerirá a terceros que compartan sus trabajos derivados de este dataset bajo una licencia compatible.

- No se permitirá realizar un uso comercial de la data.

Con todas estas observaciones, la licencia seleccionada será la Creative Commons Attribution-NonCommercial-Share-Alike (CC-BY-NC-SA 4.0).

Este dataset ha sido publicado en Zenodo con el DOI: **10.5281/zenodo.4265051** y puede ser visualizado en el siguiente link: https://zenodo.org/record/4265051#.X6mORGgReUk

Recursos

[1] Bibliotecas CSIC [en línea][Fecha de consulta: noviembre de 2020]. Disponible en: http://bibliotecas.csic.es/es/node/300

[2] Free Proxy List [en línea] [Fecha de consulta: noviembre de 2020]. Disponible en: https://free-proxy-list.net/

[3] License Selector [en línea][Fecha de consulta: noviembre de 2020]. Disponible en: https://ufal.github.io/public-license-selector/

[4] Medium. The Art Of Not Being Blocked: How I Used Selenium And Python to Scrape Facebook, And Tiktok [en línea] [Fecha de consulta: noviembre de 2020]. Disponible en: https://medium.com/analytics-vidhya/the-art-of-not-getting-blocked-how-i-used-selenium-python-to-scrape-facebook-and-tiktok-fd6b31dbe85f

[5] Rotten Tomatoes [en línea] [Fecha de consulta: octubre de 2020]. Disponible en: https://www.rottentomatoes.com

[6] Wikipedia. Sistema de calificaciones de la Asociación Cinematográfica de Estados Unidos [en línea] [Fecha de consulta: noviembre de 2020]. Disponible en: . https://es.wikipedia.org/wiki/Sistema_de_calificaciones_de_la_Asociaci%C3%B3n_Cinematogr%C3%A1fica_de_Estados_Unidos#Adici%C3%B3n_de_la_clasificaci%C3%B3n_PG-13

Contribuciones	Firma
Investigación previa	D.B.M, R.G.H
Redacción de respuestas	D.B.M, R.G.H
Desarrollo código	D.B.M, R.G.H