

Práctica 1: web scraping

<https://github.com/fnunezs/web-scraping>

Contexto

Primera práctica de la asignatura Tipología y Ciclo de Vida de los Datos del Máster de Ciencia de Datos de la Universitat Oberta de Catalunya (UOC), en la que se propone al alumno elaborar un caso práctico orientado a aprender a identificar los datos relevantes para un proyecto analítico y usar las herramientas de extracción de datos.

La actividad que se desarrolla en esta práctica consiste en la creación de un *dataset* mediante *web scraping* que contendrá las fichas técnicas, clasificación y valoración de las películas más populares entre la crítica y usuarios del sitio web de revisión y reseñas cinematográficas **Rotten Tomatoes**.

Se trata de uno de los agregadores de opiniones de películas y programas de televisión más veteranos y populares de Internet, cuyo sistema de puntuación o "*tomatometer*" se basa en las revisiones de cientos de críticos de cine y televisión.

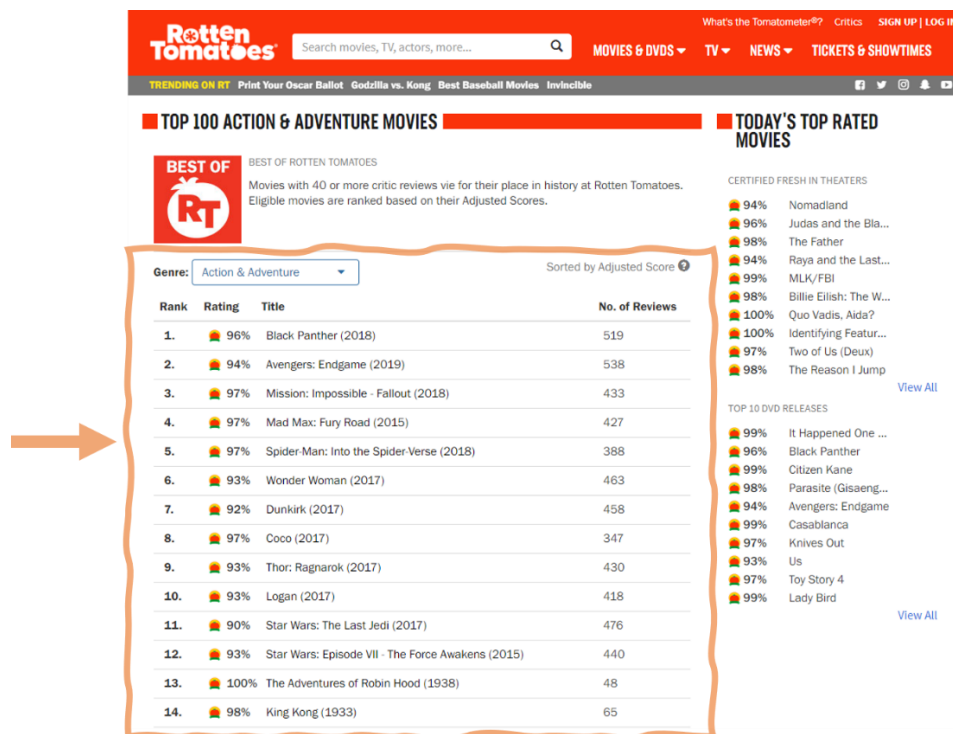
La siguiente imagen muestra la **página** de este sitio que se tomará como punto de partida para extraer los datos, en ella se ha señalado el listado que da acceso a las películas mejor valoradas en diferentes categorías:

The screenshot shows the Rotten Tomatoes homepage. At the top is a red navigation bar with the logo, a search bar, and links for MOVIES & DVDS, TV, NEWS, and TICKETS & SHOWTIMES. Below the navigation bar is a grey bar with trending movies like 'Print Your Oscar Ballot', 'Godzilla vs. Kong', 'Best Baseball Movies', and 'Invincible'. The main content area is divided into several sections:

- TOP MOVIES**: A section with a sub-header 'All lists are sorted by Adjusted Score'. It contains two lists:
 - BEST MOVIES OF 2020**: A list of 10 movies with their adjusted scores (e.g., One Night in Miami at 98%, Portrait of a Lady on Fire at 98%, Ma Rainey's Black Bottom at 98%, Minari at 98%, Soul at 95%, 1917 at 89%, The Invisible Man at 91%, Never Rarely Sometimes Always at 99%, Sound of Metal at 97%, and Hamilton at 98%).
 - BEST MOVIES OF 2019**: A list of 10 movies with their adjusted scores (e.g., Parasite (Gisaengchung) at 98%, Avengers: Endgame at 94%, Knives Out at 97%, Us at 93%, Toy Story 4 at 97%, The Irishman at 95%, Little Women at 95%, Booksmart at 96%, The Farewell at 97%, and A Beautiful Day in the Neighborhood at 95%).
- TOP 100 MOVIES BY GENRE**: A central list of 20 movie genres, each with a link to a 'Top 100' list (e.g., Top 100 Action & Adventure Movies, Top 100 Animation Movies, Top 100 Art House & International Movies, Top 100 Classics Movies, Top 100 Comedy Movies, Top 100 Documentary Movies, Top 100 Drama Movies, Top 100 Horror Movies, Top 100 Kids & Family Movies, Top 100 Musical & Performing Arts Movies, Top 100 Mystery & Suspense Movies, Top 100 Romance Movies, Top 100 Science Fiction & Fantasy Movies, Top 100 Special Interest Movies, Top 100 Sports & Fitness Movies, Top 100 Television Movies, and Top 100 Western Movies). This section is highlighted with an orange border.
- BEST MOVIES OF ALL TIME**: A list of 20 movies ranked by their Tomatometer Score (e.g., It Happened One Night (1934), Black Panther (2018), Citizen Kane (1941), Parasite (Gisaengchung) (2019), Avengers: Endgame (2019), Casablanca (1942), Knives Out (2019), Us (2019), Toy Story 4 (2019), Lady Bird (2017), Mission: Impossible - Fallout (2018), BlackKkKlansman (2018), The Wizard of Oz (1939), The Irishman (2019), Get Out (2017), The Godfather (1972), Mad Max: Fury Road (2015), Spider-Man: Into the Spider-Verse (2018), Moonlight (2016), and All About Eve (1950)).
- MOVIE AWARD WINNERS**: A section at the bottom right.

An orange arrow points from the 'BEST MOVIES OF 2019' list to the 'TOP 100 MOVIES BY GENRE' list, indicating the source of the data for the scraping exercise.

Los enlaces del listado anterior dan a su vez acceso a los denominados "Top 100" de la categoría elegida. En la siguiente imagen se puede ver, por ejemplo, el de las películas clasificadas como "Action and Adventure":



TOP 100 ACTION & ADVENTURE MOVIES

BEST OF ROTTEN TOMATOES
Movies with 40 or more critic reviews vie for their place in history at Rotten Tomatoes. Eligible movies are ranked based on their Adjusted Scores.

Genre: **Action & Adventure** Sorted by Adjusted Score

Rank	Rating	Title	No. of Reviews
1.	96%	Black Panther (2018)	519
2.	94%	Avengers: Endgame (2019)	538
3.	97%	Mission: Impossible - Fallout (2018)	433
4.	97%	Mad Max: Fury Road (2015)	427
5.	97%	Spider-Man: Into the Spider-Verse (2018)	388
6.	93%	Wonder Woman (2017)	463
7.	92%	Dunkirk (2017)	458
8.	97%	Coco (2017)	347
9.	93%	Thor: Ragnarok (2017)	430
10.	93%	Logan (2017)	418
11.	90%	Star Wars: The Last Jedi (2017)	476
12.	93%	Star Wars: Episode VII - The Force Awakens (2015)	440
13.	100%	The Adventures of Robin Hood (1938)	48
14.	98%	King Kong (1933)	65

TODAY'S TOP RATED MOVIES

CERTIFIED FRESH IN THEATERS

- 94% Nomadland
- 96% Judas and the Bla...
- 98% The Father
- 94% Raya and the Last...
- 99% MLK/FBI
- 98% Billie Eilish: The W...
- 100% Quo Vadis, Aida?
- 100% Identifying Featur...
- 97% Two of Us (Deux)
- 98% The Reason I Jump

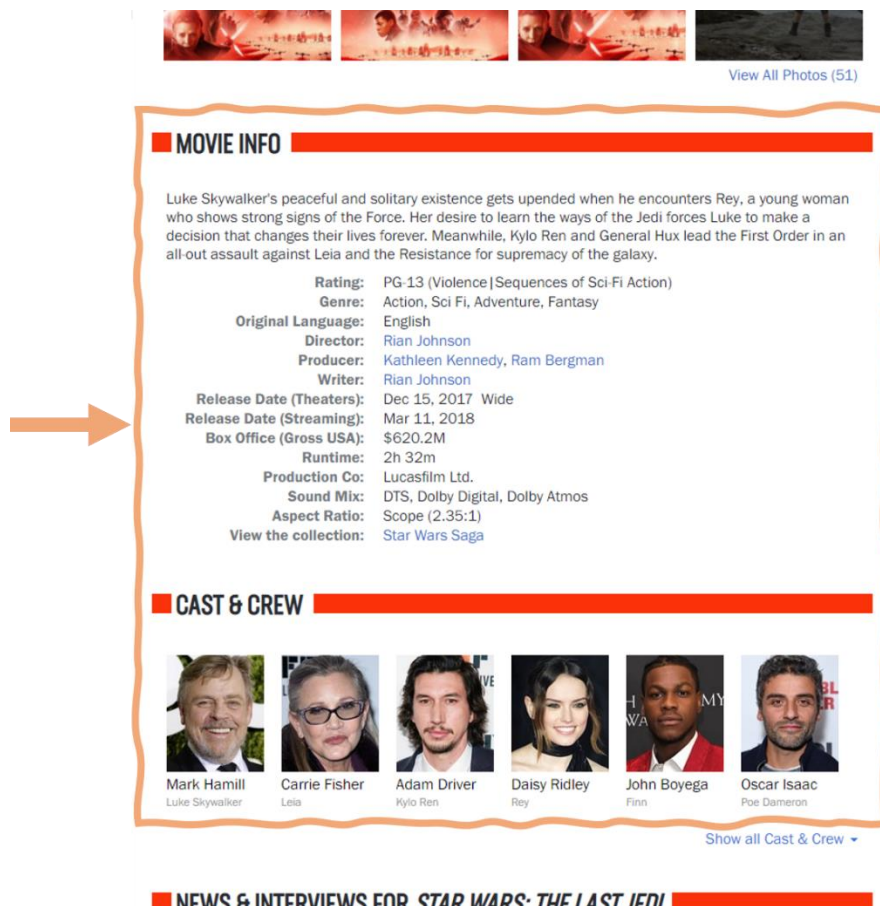
[View All](#)

TOP 10 DVD RELEASES

- 99% It Happened One ...
- 96% Black Panther
- 99% Citizen Kane
- 98% Parasite (Gisaeng...
- 94% Avengers: Endgame
- 99% Casablanca
- 97% Knives Out
- 93% Us
- 97% Toy Story 4
- 99% Lady Bird

[View All](#)

Seleccionando una de las entradas del listado anterior se accede a la ficha de la película escogida, en la que se muestran, entre otros datos, su valoración y datos técnicos, tal y como se puede ver en esta imagen:



MOVIE INFO

Luke Skywalker's peaceful and solitary existence gets upended when he encounters Rey, a young woman who shows strong signs of the Force. Her desire to learn the ways of the Jedi forces Luke to make a decision that changes their lives forever. Meanwhile, Kylo Ren and General Hux lead the First Order in an all-out assault against Leia and the Resistance for supremacy of the galaxy.

Rating: PG-13 (Violence) Sequences of Sci-Fi Action
Genre: Action, Sci-Fi, Adventure, Fantasy
Original Language: English
Director: Rian Johnson
Producer: Kathleen Kennedy, Ram Bergman
Writer: Rian Johnson
Release Date (Theaters): Dec 15, 2017 Wide
Release Date (Streaming): Mar 11, 2018
Box Office (Gross USA): \$620.2M
Runtime: 2h 32m
Production Co: Lucasfilm Ltd.
Sound Mix: DTS, Dolby Digital, Dolby Atmos
Aspect Ratio: Scope (2.35:1)
View the collection: [Star Wars Saga](#)

CAST & CREW

Mark Hamill Luke Skywalker
Carrie Fisher Leia
Adam Driver Kylo Ren
Daisy Ridley Rey
John Boyega Finn
Oscar Isaac Poe Dameron

[Show all Cast & Crew](#)

NEWS & INTERVIEWS FOR STAR WARS: THE LAST JEDI

Título del dataset

El título elegido es “*Rotten Tomatoes Top Movies*”.

Descripción del dataset

El *dataset* que se generará en esta práctica consiste en una colección de 1.610 registros con 24 atributos que representan a cada una de las entradas de los listados “*Top 100*” de películas mejor valoradas del sitio web *Rotten Tomatoes*.

Representación gráfica

Consultar el fichero `rottentomatoes.ipynb` directamente en el [repositorio github](#) o mediante el siguiente enlace:

<https://nbviewer.jupyter.org/github/fnunezs/web-scraping/blob/main/rottentomatoes.ipynb>

Contenido

Los atributos registrados para las películas del dataset son los siguientes:

- Título
- Año
- Sinopsis
- Categoría
- Valoración de la crítica profesional ("*tomatometer*")
- Opinión consensuada de la crítica profesional
- Total de críticas profesionales recibidas
- Valoración de los usuarios
- Total de valoraciones de usuarios recibidas
- Ficha técnica: clasificación por edades, género, lenguaje, director, productor, guionista, fecha de estreno en cines, fecha de estreno en plataformas de *streaming*, ingresos generados, duración, sonido, relación de aspecto, colección de películas asociada y los nombres de los actores, director, productores y guionistas.
- Enlace a la página de *Rotten Tomatoes* de la que se ha extraído la información.

Consultar el fichero `rottentomatoes.ipynb` directamente en el [repositorio github](#) o mediante el siguiente enlace:

<https://nbviewer.jupyter.org/github/fnunezs/web-scraping/blob/main/rottentomatoes.ipynb>

Agradecimientos

Rotten Tomatoes es probablemente el agregador de reseñas de películas y programas de televisión más popular de Internet, cuyo sistema de puntuación o "*tomatometer*" está respaldado por las opiniones de cientos de críticos de cine y televisión. Además, cuenta con una extensa comunidad de usuarios que realizan sus propias reseñas y valoraciones lo que convierte en más valiosa si cabe la información de este sitio web.

Los *datasets* generados a partir de las reseñas de *Rotten Tomatoes* son muy populares entre la comunidad de usuarios de [Kaggle](#), una plataforma que permite encontrar y publicar conjuntos de datos, explorar y construir modelos o participar en competiciones y desafíos en un entorno de ciencia de datos basado en la web. Uno de los *dataset* más completos de esta plataforma sobre críticas cinematográficas es *Rotten Tomatoes movies and critic reviews dataset* creado por [Stefano Leone](#) a partir de las reseñas de más de 17.000 películas.

Inspiración

Tanto el proceso de generación como el propio *dataset* resultado de esta práctica tienen como objetivo facilitar la comprensión de las técnicas de *web scraping* y la familiarización con el uso de herramientas como *Beautiful Soup* o *Selenium WebDriver*. Los pasos seguidos están planteados de manera similar a la descrita por autores como [Isabella Benabaye](#) en su artículo "[Step-by-step Scraping Epsiode IMDb Ratings](#)".

Entre los usos prácticos que se le pueden dar al *dataset* generado en esta práctica está el servir como fuente de datos para analizar distintos algoritmos de clasificación, e incluso para desarrollar un sistema de recomendación de películas sencillo, gracias a la información que contiene sobre las películas mejor valoradas por la crítica y usuarios en diferentes categorías.

Dadas sus limitaciones de tamaño y atributos recopilados, no es adecuado, sin embargo, para algoritmos de inteligencia artificial más complejos, como los de análisis de sentimientos a partir de las críticas de películas. Para estas y otras actividades más sofisticadas es recomendable un *dataset* más amplio y completo como el de Stefano Leone mencionado anteriormente.

Licencia

El *dataset* generado en esta práctica se distribuirá mediante licencia *CC0: Public Domain License*, dado que los datos extraídos del sitio web están disponibles para el público en general sin restricciones conocidas.

Código

El código Python que genera el *dataset* está disponible en el fichero **rottentomatoes.ipynb** en el [repositorio github](#) y se puede visualizar directamente mediante el siguiente enlace:

<https://nbviewer.jupyter.org/github/fnunezs/web-scraping/blob/main/rottentomatoes.ipynb>

Dataset

El *dataset* generado ha sido publicado en formato CSV en Zenodo con los siguientes datos:

Target URL: <https://doi.org/10.5281/zenodo.4682107>

DOI: 10.5281/zenodo.4682107