

# Práctica 1 - Memoria

## Tipología y ciclo de vida de los datos

---

Lucas Gómez Torres

Joan Amengual Mesquida

22 de Noviembre del 2022

---

<b>Contexto</b>	<b>2</b>
<b>Título</b>	<b>2</b>
<b>Descripción del dataset</b>	<b>3</b>
<b>Representación gráfica</b>	<b>4</b>
<b>Contenido</b>	<b>6</b>
<b>Propietario</b>	<b>7</b>
<b>Inspiración</b>	<b>8</b>
<b>Licencia</b>	<b>9</b>
<b>Código del proyecto</b>	<b>10</b>
<b>Dataset</b>	<b>10</b>
<b>Vídeo</b>	<b>11</b>
<b>Complementos realizados</b>	<b>12</b>
Log In y User Agent	12
Uso de Docker	12
<b>Contribuciones de los integrantes</b>	<b>13</b>

## Contexto

**Contexto.** Explicar en qué contexto se ha recolectado la información. Explicar por qué el sitio web elegido proporciona dicha información. Indicar la dirección del sitio web.

Imdb (Internet Movie Database) es la base de datos de referencia del cine y la televisión en Internet, siendo un espacio que permite dejar valoraciones, comentarios y críticas sobre las películas y series que hayan visto y así ayudar al resto de la comunidad cinéfila por ejemplo a descubrir nuevos géneros o a estar al día de las películas mejor valoradas por el resto de usuarios. En esta base de datos, los miembros de la comunidad pueden recibir alertas de seguimiento por email de los títulos que les interesen, las películas y las series están clasificadas por popularidad, número de votos y puntuación (ya sea nuestra y de los demás usuarios), así como otras funcionalidades que puede ofrecer Imdb.

Como el tema elegido para nuestro proyecto era el de extraer información sobre películas y series, sin duda Imdb estaba entre las mejores opciones para recopilar información sobre cine y televisión. En concreto, nuestro dataset recolectado recoge información sobre las películas y programas de televisión más populares extraídos por género. En la web, Imdb ofrece una opción de explorar las películas por género, mostrándonos una gran variedad de películas y series junto con sus características como el título, la duración, la sinopsis y su valoración, entre muchas otras.

La dirección del sitio web de Imdb es: <https://www.imdb.com>.

## Título

**Título.** Definir un título que sea descriptivo para el dataset.

**Características de películas y programas de televisión más populares en la base de datos de Imdb**

## Descripción del dataset

**Descripción del dataset.** Desarrollar una descripción breve del conjunto de datos que se han extraído. Es necesario que esta descripción tenga sentido con el título elegido.

Se ha realizado una extracción de datos a través de técnicas de web scraping en la web Imdb, con las películas y programas de televisión más populares distribuidos por género.

El dataset cuenta con información referente a las películas y programas de televisión más populares según la comunidad cinéfila de Imdb. Esta información se puede utilizar para clasificar estas películas entre las más votadas, las mejores valoradas, las que más actores aparecen, las se pueden enmarcar en más tipos de géneros o incluso saber el género que presenta las películas peor valoradas. Además, el dataset se ha construido con solo las primeras cincuenta películas más populares de cada género ya que la web contiene más de 2 millones de títulos y no nos interesa tener un dataset tan grande para su posterior tratamiento.

Toda la información que se ha recogido se presenta en un fichero CSV para facilitar su posterior limpieza y análisis en la siguiente práctica.

## Representación gráfica

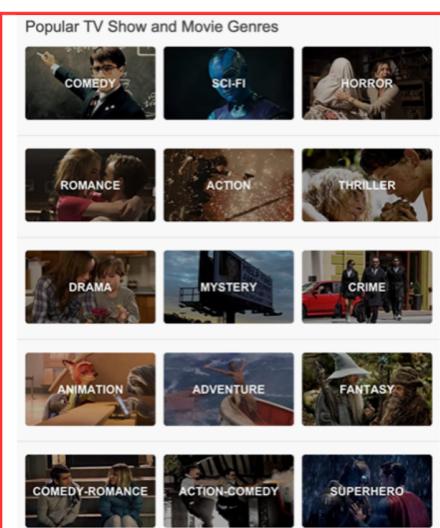
**Representación gráfica.** Dibujar un esquema o diagrama que identifique el dataset visualmente y el proyecto elegido.

La extracción de datos de la Web Imdb (<https://www.imdb.com>) empieza en la página de clasificación en géneros de las películas: <https://www.imdb.com/feature/genre/>.

En dicha página extraemos el nombre y la URL de la página de categorías de las películas para posteriormente acceder a los datos de las películas.

En la siguiente imagen de la izquierda podemos visualizar los géneros de películas y programas de Imdb. Por otro lado, en la imagen de la derecha visualizamos los campos de información que extraemos de cada una de las películas.

**Géneros populares de películas y programas de televisión**



Popular TV Show and Movie Genres

1. COMEDY 2. SCI-FI 3. HORROR

4. ROMANCE 5. ACTION 6. THRILLER

7. DRAMA 8. MYSTERY 9. CRIME

10. ANIMATION 11. ADVENTURE 12. FANTASY

13. COMEDY-ROMANCE 14. ACTION-COMEDY 15. SUPERHERO

**Géneros populares de películas y programas de televisión**

**Top 50 Comedy Movies and TV Shows**

1-50 of 2,009,141 titles. | Next » View Mode: Compact | Detailed

Sort by: Popularity▲ | A-Z | User Rating | Number of Votes | US Box Office | Runtime | Year | Release Date | Date of Your Rating | Your Rating

Rank	Title	Year	Rating	Genre	Plot Summary
1	The White Lotus	(2021–2022)	7.6	Comedy, Drama	Set in a tropical resort, it follows the exploits of various guests and employees over the span of a week.
2	Stranger Things	(2016–2022)	8.5	Science Fiction, Mystery, Thriller, Drama	A science fiction drama television series... (more)
3	BoJack Horseman	(2014–2020)	8.2	Animation, Comedy, Drama, Fantasy	An adult animated series... (more)
4	Breaking Bad	(2008–2013)	8.8	Crime, Drama, Thriller	A crime drama television series... (more)
5	Game of Thrones	(2011–2019)	8.2	Science Fiction, Fantasy, Drama	A fantasy drama television series... (more)
6	House of Cards	(2013–2018)	8.2	Political, Drama, Thriller	A political drama television series... (more)
7	True Blood	(2008–2014)	7.8	Horror, Mystery, Thriller, Drama	A horror drama television series... (more)
8	Mad Men	(2007–2015)	8.2	Drama, Thriller, Mystery	A historical drama television series... (more)
9	Breaking Bad	(2008–2013)	8.8	Crime, Drama, Thriller	A crime drama television series... (more)
10	House of Cards	(2013–2018)	8.2	Political, Drama, Thriller	A political drama television series... (more)

**Extracción del nombre y URL para el acceso a la información de cada género**

Cada película se compone de una diversidad de campos que definen sus características con precisión.

El contenido audiovisual también será tratado en esta práctica y por ello se ha extraído la imagen y el trailer de cada película o programa de televisión. Hay dos posibilidades en cuanto a la extracción de datos audiovisuales:

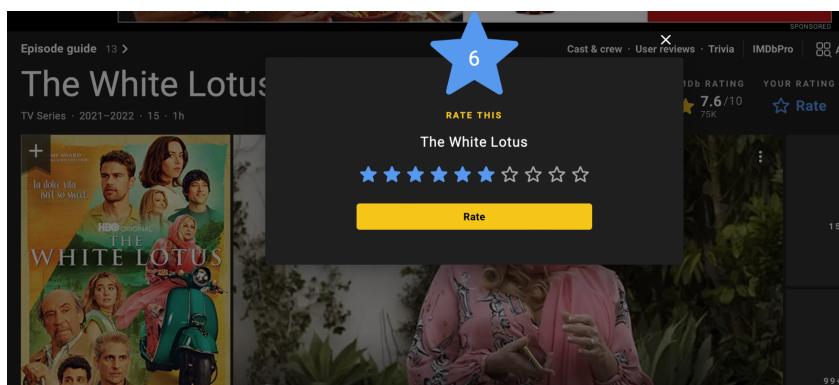
1. Extracción de la imagen en formato *.jpg* y el vídeo del trailer en formato *.mp4*
2. Extracción de la URL del contenido almacenado en el cloud de Amazon Web Services (AWS). De esta forma, se evita tener un exceso de datos en nuestra máquina para almacenar todo el contenido audiovisual.

Para la extracción de estos datos debemos introducirnos en cada una de las películas o programas de televisión para así extraer el contenido.

Véase la siguiente captura de pantalla para clarificar aquellos campos que serán extraídos.



Para finalizar, hemos decidido añadir un tratamiento totalmente dinámico del campo de valoración de películas. Por ello, usando librerías recomendadas en esta práctica, como por ejemplo Selenium, hemos podido realizar votaciones de películas abriendo una sesión del navegador (Firefox) y estableciendo un tratamiento automatizado de valoración de películas colaborando así en el crecimiento de la plataforma y en el número de usuarios que interactúan con esta web.



## Contenido

**Contenido.** Explicar los campos que incluye el dataset y el periodo de tiempo de los datos.

Los campos que conforman el dataset de esta práctica se pueden visualizar en la siguiente tabla con la descripción de cada uno de estos respectivamente.

En este caso el período de tiempo no está definido para el conjunto de datos obtenido, ya que la extracción de las películas más destacadas no está influido por un rango de tiempo.

Campo del dataset	Descripción
<b>Movie Title</b>	Título de la película/programa de televisión.
<b>Movie Duration</b>	Duración de la película/programa de televisión (en minutos).
<b>Movie Genres</b>	Conjunto de géneros donde se incluye la película/programa de televisión.
<b>Movie Rating</b>	Valoración de la película/programa de televisión determinada por los usuarios (de 1 a 10).
<b>Our Rating</b>	Valoración de la película/programa de televisión realizada en base a nuestro criterio (de 1 a 10).
<b>Movie Description</b>	Descripción de la película/programa de televisión.
<b>Movie Stars</b>	Estrellas de cine que aparecen en la película/programa de televisión.
<b>Movie Votes</b>	Cantidad de votaciones realizadas a la película/programa de televisión.
<b>Movie Image</b>	URL de la imagen de la película/programa de televisión.
<b>Movie Video</b>	URL del vídeo del trailer de la película/programa de televisión.

Para la extracción de los campos anteriores se han utilizado las siguientes librerías:

- **Requests**: se usó para realizar peticiones http y poder posteriormente parsear código HTML.
- **BeautifulSoup**: librería para parsear el código y poder obtener diferentes datos de las películas.
- **Urllib.parse**: se utilizó para manejar fácilmente URLs y gestionar URLs relativas y absolutas, en este caso, las URLs del contenido audiovisual.
- **Selenium**: facilitó la creación de acciones automatizadas de una manera sencilla. Nosotros lo usamos para automatizar el inicio de sesión y la valoración personalizada de algunas películas.

## Propietario

**Propietario.** Presentar al propietario del conjunto de datos. Es necesario incluir citas de análisis anteriores o, en caso de no haberlas, justificar esta búsqueda con análisis similares. Justificar qué pasos se han seguido para actuar de acuerdo a los principios éticos y legales en el contexto del proyecto.

El propietario del conjunto de datos es Amazon.com, Inc: una corporación estadounidense de comercio electrónico y servicios de computación en la nube. Los derechos de propiedad intelectual de las críticas corresponden a los medios de comunicación de los que han sido extraídos. Algunas características de las películas como las carátulas, fotografías, tráilers y banda sonora original (BSO) pertenecen a las productoras.

De todos ellos, se hace un buen uso con fines exclusivamente académicos. Para actuar de acuerdo a los principios éticos y legales del proyecto, se miró el archivo robots.txt donde se indican las páginas a las que según el propietario del sitio, en este caso Amazon, expresa que no está permitido realizar web scraping. En la mayoría de estas páginas no se puede realizar web scraping, sin embargo, para reducir las posibilidades de ser bloqueados y evitar problemas legales en el futuro, una de las cosas que se hizo fue modificar la cabecera del User-Agent. Además, nosotros en este proyecto no incumplimos los términos y las condiciones de la web ya que se hace un uso legítimo y razonable para fines puramente académicos e informativos, y al crear una cuenta en Imdb con que la que poder iniciar sesión, se ha tenido que aceptar el acuerdo de los términos y de las condiciones de la web.

Algunos análisis similares son los siguientes:

- <https://github.com/ShehzadaAlam/IMDb-Web-Scraping-and-Data-Analysis>
- <https://www.freecodecamp.org/news/web-scraping-sci-fi-movies-from-imdb-with-python/>
- <https://binalkagathara.medium.com/web-scraping-with-python-imdb-rating-f4bb946662cd>

Los análisis anteriores muestran cómo se hace la extracción de algunas características de las películas y de los programas de televisión de Imdb. Los 2 primeros son los más completos, donde a parte de la extracción de datos realizan un análisis exhaustivo con el estudio de la correlación entre varias variables. Además, en el primero se aplican técnicas de análisis de sentimientos basados en las reseñas de los usuarios para saber si son positivos o negativos. Sin embargo, en el tercero no explican resultados, sólo muestran el código de cómo hacer la extracción de datos mediante web scraping.

## Inspiración

**Inspiración.** Explicar por qué es interesante este conjunto de datos y qué preguntas se pretenden responder. Es necesario comparar con los análisis anteriores presentados en el apartado 6.

Este dataset puede resultar de interés para todo aquel aficionado al mundo del cine y de la televisión o para aquellos que quieren empezar a descubrirlo ahora. En él, se almacenan las películas más populares extraídas por género, resultando ser un contenido de alta calidad para poder consultarse a la hora de buscar buenas películas por diferentes ítems como valoraciones, votaciones, actores, géneros o duración. Gracias a este conjunto de datos con sus diferentes atributos, se pueden sacar estadísticas interesantes con las que poder buscar películas que se adapten a los gustos de cada uno de nosotros.

Las preguntas que se pretenden responder son: ¿qué películas están entre las más votadas ?, ¿y las mejores valoradas?, ¿y las que más actores aparecen?, ¿cuáles se pueden englobar en más tipos de géneros?, ¿qué género presenta las películas peor valoradas por los usuarios?, ¿influye el año en que se estrenó en la valoración de la película?, ¿existe una relación entre el género de la película y el número de votos que obtiene?.

Las preguntas anteriores son ejemplos que este dataset podría responder y que a cualquier persona cinéfila le puede parecer interesante.

Si comparamos la información que se ha obtenido en esta práctica formando nuestro conjunto de datos, con los análisis del ejercicio 6, son todos muy parecidos porque entre todos ellos se extraen la mayoría de los atributos de las películas que tenemos en el dataset. Sin embargo, en cuanto al análisis, el que más se parece es el segundo, ya que responde a algunas preguntas como qué año es el que mejor valoración media tiene, cuántas películas de cada calificación hay o calcular el número de películas estrenadas por año.

## Licencia

**Licencia.** Seleccionar una licencia adecuada para el dataset resultante y justificar el motivo de su elección. Ejemplos de licencias que pueden considerarse:

- Released Under CC0: Public Domain License.
- Released Under CC BY-NC-SA 4.0 License.
- Released Under CC BY-SA 4.0 License.
- Database released under Open Database License, individual contents under Database Contents License.
- Otra (especificar cuál).

El dataset resultante se ofrece bajo licencia *Released Under CC BY-NC-SA 4.0 License*. Se ha elegido ya que:

- Permite copiar y redistribuir el material en cualquier medio o formato. De esta forma ayuda a distribuir la información a otros sectores que les interese la temática y así dar a conocer y aportar reconocimiento al creador del dataset.
- Permite que la obra continúe bajo los mismos términos, ya que cualquier modificación o trabajo parecido basado en dicho material que se haga debe ser ofrecido bajo los términos de la misma licencia, indicando el nombre del creador original y las modificaciones realizadas sobre la obra original que permitan diferenciar las distintas contribuciones.

El enlace de la licencia de Creative Commons es:

<https://creativecommons.org/licenses/by-nc-sa/4.0/>

## Código del proyecto

**Código.** Código con el que se ha obtenido el dataset, preferiblemente en Python o, alternativamente, en R.

- El código deberá ubicarse en la carpeta **/source** del repositorio.
- Se deben indicar las librerías y versiones utilizadas. P. ej., en Python pueden obtenerse mediante el comando: pip3 freeze > requirements.txt
- En el documento PDF se deben comentar los aspectos más relevantes sobre cómo el código realiza el proceso de recolección de datos, qué dificultades presenta el sitio web elegido, y cómo las habéis resuelto.

El código de la práctica en el lenguaje Python se encuentra en el repositorio de GitHub: [https://github.com/LucasGomezTorres/Web\\_Scraping](https://github.com/LucasGomezTorres/Web_Scraping).

Las librerías y sus versiones utilizadas se encuentran en el archivo requirements.txt del repositorio: [https://github.com/LucasGomezTorres/Web\\_Scraping/blob/main/requirements.txt](https://github.com/LucasGomezTorres/Web_Scraping/blob/main/requirements.txt).

Una de las dificultades que se ha tenido del sitio web elegido al realizar la extracción de datos ha sido la gestión del contenido audiovisual, al extraer la URL del tráiler de cada película. El problema estaba en que en el contenido del parseo del código html no aparecía la URL del vídeo, solo aparecía accesible después de reproducirlo y esto no era eficiente ya que no nos interesaba reproducir los tráilers de todas las películas, sino sólo guardar la URL de descarga.

Para solucionarlo, se vio que en el html de las páginas webs de los tráilers de cada película se encontraba un JSON que contiene el enlace de la descarga del tráiler, por lo que se optó por obtener desde ahí la URL de cada tráiler disponible.

## Dataset

**Dataset.** Publicar el dataset obtenido en formato CSV en Zenodo, incluyendo una breve descripción. Obtener y adjuntar el enlace del DOI del dataset (<https://doi.org/...>). El dataset también deberá incluirse en la carpeta **/dataset** del repositorio.

Si existe alguna circunstancia que impida publicar abiertamente el dataset real en Zenodo, se deberá: (1) comentar esta circunstancia y justificar el motivo en este apartado; (2) generar un dataset simulado y publicarlo en Zenodo, obteniendo el enlace del DOI; y (3) comunicar al profesor el dataset real de forma privada (p. ej., utilizando un repositorio privado).

El dataset final se ha publicado en Zenodo, a continuación se incluye el enlace del DOI y la URL de acceso al dataset final de la práctica.

- DOI: [10.5281/zenodo.7311723](https://doi.org/10.5281/zenodo.7311723)
- URL: <https://zenodo.org/record/7311723#.Y21G9C0w1gl>

## Características de películas y programas de televisión más populares en la base de datos de Imdb

Joan Arnengual; Lucas Gómez

Se ha realizado una extracción de datos a través de técnicas de web scraping en la web Imdb, con las películas y programas de televisión más populares distribuidos por género.

El dataset cuenta con información referente a las películas y programas de televisión más populares según la comunidad cinéfila de Imdb. Esta información se puede utilizar para clasificar estas películas entre las más votadas, las mejores valoradas, las que más actores aparecen, las se pueden enmarcar en más tipos de géneros o incluso saber el género que presenta las películas peor valoradas. Además, el dataset se ha construido con solo las primeras cincuenta películas más populares de cada género ya que la web contiene más de 2 millones de títulos y no nos interesa tener un dataset tan grande para su posterior tratamiento.

Toda la información que se ha recogido se presenta en un fichero CSV para facilitar su posterior limpieza y análisis en la siguiente práctica.

Preview								
Movie Title	Movie Year	Movie Duration	Movie Genres	Movie Rating	Our Rating	Movie Description	Movie Stars	Movie Votes
The White Lotus	2021	60 min	Comedy, Drama	7.6	7/10	Set in a tropical resort, it follows the exploits of various guests and employees over the span of a week.	Jennifer Coolidge, Eleonora Romandini, Federico Ferrante, Murray Bartlett	75,393
Weird: The Al Yankovic Story	2022	108 min	Biography, Comedy, Music	7.2	6/10	Explores every facet of Yankovic's life, from his meteoric rise to fame with early hits like 'Eat It' and 'Like a Surgeon' to his torrid celebrity love affairs and	Diedrich Bader, Daniel Radcliffe, Lin-Manuel Miranda, Richard Aaron Anderson	9,692

New version

1
1

views
downloads

See more details...

Indexed in



Publication date:  
November 10, 2022

DOI:  
[10.5281/zenodo.7311723](https://doi.org/10.5281/zenodo.7311723)

License (for files):  
[Creative Commons Attribution 4.0 International](#)

Versions

<a href="#">Version V1</a>	Nov 10, 2022
----------------------------	--------------

Cite all versions? You can cite all versions by using the DOI [10.5281/zenodo.7311722](https://doi.org/10.5281/zenodo.7311722). This DOI represents all versions, and will always resolve to the latest one. [Read more](#).

## Vídeo

El vídeo realizado para detallar el proyecto realizado se encuentra en los siguientes enlaces, tanto de Google Drive como Youtube:

### 1. Vídeo en Google Drive de la UOC:

[https://drive.google.com/file/d/118TOtHRW4\\_i5SiX\\_9hYWC1nL-1uVGddU/view?usp=sharing](https://drive.google.com/file/d/118TOtHRW4_i5SiX_9hYWC1nL-1uVGddU/view?usp=sharing)

### 2. Vídeo en Youtube: <https://youtu.be/d8-5t6D9Gtw>

## Complementos realizados

Para llevar un paso más allá esta práctica se han decidido realizar varios complementos que puedan potenciar la ejecución y la extracción de los datos decididos.

## Log In y User Agent

1. Una de las técnicas que hemos utilizado para evitar ser bloqueados realizando web scraping es la **modificación** de la cabecera del **User-Agent**, estableciendo el encabezado que envía nuestro navegador a la hora de hacer solicitudes http.
2. Se ha creado un **login** con el cual se van a establecer dos sesiones abiertas con las mismas cookies en todo momento. La primera es creada con la librería **Requests** y la segunda es creada con la librería **Selenium** a través del navegador Firefox.

## Uso de Docker

3. **Ejecución mediante el uso de Docker:** Docker es una tecnología de contenedores que permite a los usuarios desplegar el código en entornos totalmente controlados y asegurando en todo momento que la ejecución se dará con éxito en cualquier entorno posible. Por ese motivo se ha realizado la implementación del despliegue de nuestro código mediante contenedores de software de Docker.

Para realizar dicha implementación se ha realizado lo siguiente:

- **Creación de un Dockerfile que permita crear la imagen necesaria de Docker.** En este caso, ha sido necesario partir de una imagen base, la cual ha sido Python para poder ejecutar así nuestro código sin problema alguno.
- **Crear un fichero de ejecución que permita ejecutar el contenedor de software partiendo de la imagen de Docker creada previamente.** Las sentencias que se han definido en dicho documento son:

```
- # Create a docker image  
- docker build -t web_scraping:v1 .  
- # Execute docker container  
- docker run web_scraping:v1
```

Es necesario destacar que la implementación realizada de Docker puede usarse siempre y cuando nuestra administración de añadir nuevas valoraciones mediante una interacción dinámica con el navegador se encuentre desactivada, de otro lado la ejecución no será exitosa ya que no se ha realizado la gestión del navegador Firefox internamente en Docker.

## Contribuciones de los integrantes

A continuación se presentan las contribuciones del proyecto firmadas por los integrantes del grupo.

Contribuciones	Firma
Investigación previa	Lucas Gómez, Joan Amengual
Redacción de las respuestas	Lucas Gómez, Joan Amengual
Desarrollo del código	Lucas Gómez, Joan Amengual
Participación en el vídeo	Lucas Gómez, Joan Amengual

El trabajo ha sido dividido de forma modular para que cada integrante pueda recoger parte de las tareas a resolver y enfocarse de manera individual sacar el trabajo adelante. Continuamente se han ido realizando reuniones de grupo (mediante Zoom) para comentar todos los puntos importantes y tomar las decisiones necesarias en cuanto a las tareas que se han necesitado realizar.