Tipología y ciclo de vida de los datos

César Fernández García

Óscar Tienda Beteta

PRA1: ¿Cómo podemos capturar los datos de la web?

1. Contexto

El contexto en el que se ha recolectado esta información es en el ámbito del **análisis de datos** y la **industria cinematográfica**. La información de películas o series de televisión recolectada de la web IMDb (**https://www.imdb.com**) proporciona una gran cantidad de datos relevantes para el análisis de tendencias en el mercado cinematográfico, así como para la investigación académica en el campo de la cultura popular y el cine.

IMDb es un sitio web popular y completo dedicado a la información sobre películas y televisión. Es una fuente de información de alta calidad y muy bien organizada, lo que la hace ideal para obtener datos estructurados y consistentes sobre películas. IMDb es utilizado por una amplia variedad de personas, desde fans del cine hasta críticos y profesionales de la industria, y se actualiza constantemente con nuevas películas y datos.

Además, IMDb ofrece una amplia gama de información para cada película o serie de televisión, como su título, género, clasificación, resumen, elenco, producción y detalles técnicos, lo que la hace una fuente valiosa para el análisis de datos. Los datos recolectados de IMDb se pueden utilizar para analizar tendencias en la industria cinematográfica, para identificar géneros o actores populares, o para evaluar la calidad de las películas. En resumen, el sitio web IMDb es una fuente muy útil para recolectar información sobre películas y televisión, y es ampliamente utilizado en el campo de la investigación de datos cinematográficos.

2. Título

Dataset “**Películas y series de IMDb**”.

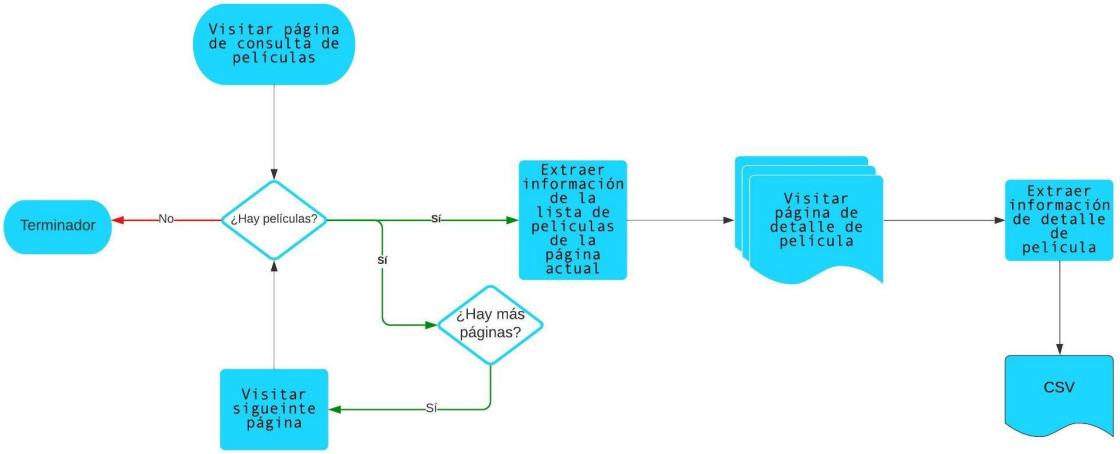
3. Descripción del dataset

El dataset "Películas y series de IMDb" es una colección de datos obtenidos a través de web scraping del sitio web IMDb. Contiene información detallada sobre más de 1,000 películas, incluyendo su título, año de lanzamiento, duración, género, calificación de audiencia, número de votos, sinopsis, director y elenco.

Este conjunto de datos es ideal para analizar tendencias en la industria cinematográfica, identificar patrones en la popularidad de los géneros cinematográficos, comparar la calidad de las películas según la calificación de la audiencia, estudiar el impacto de los actores y directores en el éxito de la película, entre otros usos.

Es importante destacar que el uso de este conjunto de datos debe cumplir con las políticas de uso del sitio web IMDb.

4. Representación gráfica



5. Contenido

Campos que se incluyen en el dataset:

- **title**: El título de la película.
- **year**: El año en que se estrenó la película.
- **genre**: Los géneros al que pertenece la película.
- **rating**: El rating de la película en una escala de 0 a 10.
- **duration**: La duración de la película en minutos.
- **summary**: Un resumen breve de la trama de la película.
- **url**: El enlace a la página de la película en IMDb.
- **stars**: Una lista de los nombres de los actores principales que participaron en la película.
- **votes**: El número total de votos que recibió la película en IMDb.
- **num_user_reviews**: El número de reseñas de usuarios de la película en IMDb.
- **num_critic_reviews**: El número de reseñas de críticos de cine de la película en IMDb.
- **director**: El director de la película o serie. En el caso de las series, el director es el de la primera temporada.

Extracto dataset:

director	duration	genre	num_critic_reviews	num_user_revi...	rating	stars	summary	title	url	votes	year
Dennis Dugan	115	Comedy,Romance	139	282	5,70	Adam Sandler,K...	"Two straight, single Brooklyn firefighters pretend to be a g...	I Now Pronounc...	https://www.im...	150.065	2.007
Clark Johnson	117	Action,Adventure...	135	368	5,70	Samuel L. Jacks...	"An imprisoned drug kingpin offers a huge cash reward to a...	S.W.A.T.	https://www.im...	150.407	2.003
Todd Phillips	101	Comedy,Crime	167	344	6,10	Ben Stiller,Owe...	"Two streetwise cops bust criminals in their red and white F...	Starsky & Hutch	https://www.im...	150.408	2.004
Allen Coulter	113	Drama,Romance	152	443	7,10	Robert Pattinso...	"A romantic drama centered on two new lovers: Tyler, whos...	Remember Me	https://www.im...	150.520	2.010
John Erick Do...	80	Horror,Mystery,T...	279	442	6,20	Chris Messina,C...	"A group of people are trapped in an elevator and the Devil ...	Devil	https://www.im...	150.604	2.010
Tom Shadyac	96	Comedy,Family,F...	185	313	1,90	Steve Carell,Mo...	"God contacts Congressman Evan Baxter and tells him to ...	Evan Almighty	https://www.im...	150.623	2.007
David Yates	142	Adventure,Family...	223	1.3K	6,10	Eddie Redmayn...	"Professor Albus Dumbledore must assist Newt Scamande...	Fantastic Beasts...	https://www.im...	150.651	2.022
Larry Charles	81	Comedy	168	433	5,70	Sacha Baron Co...	"Flamboyant and gay Austrian Brüno looks for new fame in ...	Brüno	https://www.im...	150.816	2.009
John Singleton	109	Action,Crime,Dra...	110	395	6,80	Mark Wahlberg,...	"When their adoptive mother is gunned down in a store rob...	Four Brothers	https://www.im...	151.010	2.005
Ian Brennan	44	Comedy,Drama,...	65	291	6,60	Lea Michele,Jan...	"A group of ambitious misfits try to escape the harsh realitie...	Glee	https://www.im...	151.055	2.009

6. Propietario

En esta sección, abordaremos la propiedad y la autoría de los datos, destacando la importancia de IMDb como proveedor de información y sus implicaciones en el uso y manejo del conjunto de datos obtenido a través del web scraping de su sitio web.

IMDb es una empresa subsidiaria de Amazon que proporciona información sobre películas, programas de televisión, actores y más. El conjunto de datos obtenido a través del web scraping de su sitio web se utiliza para análisis y exploración de películas.

La elección de IMDb como sitio web de origen para el web scraping se debe a que es una fuente de información ampliamente utilizada y conocida para películas y programas de televisión.

En cuanto a los principios éticos y legales, se siguieron los siguientes pasos:

1. Se revisaron los términos de uso de IMDb y se respetaron todas las políticas y restricciones de redistribución y propiedad a los datos.
2. Se utilizó un enfoque de web scraping ético, en el que se limitó la velocidad de la solicitud y se evitó sobrecargar el sitio web.
3. Se aseguró de que la recopilación de datos fuera legal y no infringiera ninguna ley de privacidad de datos.
4. No se recopilan datos personales sensibles y cualquier información que pudiera identificar a un usuario específico.
5. El conjunto de datos obtenido a través de web scraping se utilizará solo con **fines educativos y de investigación**, y no se utilizará para ningún otro propósito sin el consentimiento explícito de IMDb. Tampoco se publicará públicamente el dataset generado por no incumplir con su licencia.

En cuanto a las citas a análisis similares, la amplía mayoría de proyectos de scraping en IMDB son para fines puramente educaciones. En algunos casos también hemos encontrado datases públicos de web scraping en IMDb pero como mencionamos en el apartado Licencia y Dataset estos no cumplen con sus términos y condiciones.

En general, en este caso el data scraping no sería una opción realmente viable para fines comerciales debido a su política de datos. Sin embargo, para obtener los datos con obtenidos individuales o educativos, hay que mencionar que IMDb actualiza diariamente sus datasets al público bajo una serie de estrictas condiciones, por lo que aplicar web scraping en este caso solo tiene sentido para practicar el scraping en sí mismo. Si la actualización diaria no es suficiente, IMDb también pone disponible una API mediante AWS Marketplace.

7. Inspiración

El conjunto de datos obtenido a través del web scraping de IMDb ofrece múltiples oportunidades para diversas aplicaciones y análisis, como por ejemplo:

- **Análisis de tendencias en el cine**: El conjunto de datos incluye información sobre el título, año, género, rating, duración, resumen y número de votos de cada película. Estos datos pueden utilizarse para analizar las tendencias en el cine a lo largo del tiempo, como por ejemplo, cuáles son los géneros más populares en una determinada década, cuáles son las películas con mayor rating, etc.
- **Predicción de éxito de una película**: La información sobre el número de votos, el rating, el resumen y el reparto de una película pueden utilizarse para predecir su éxito en taquilla o en plataformas de streaming. Por ejemplo, se podría usar un modelo de aprendizaje automático para predecir el rating de una película antes de su lanzamiento, en función de los datos históricos de películas similares. Además, se podrían identificar factores clave que influyen en el éxito de una película, como la presencia de actores o directores reconocidos, el género, el presupuesto, entre otros.
- **Análisis de la industria del cine**: El conjunto de datos también incluye información sobre el número de críticas de usuarios y críticos profesionales para cada película. Estos datos pueden emplearse para analizar la industria del cine, como por ejemplo, cuáles son las películas más populares entre los críticos profesionales, cómo se comparan las críticas de usuarios y críticos, la relación entre el número de votos y el rating, etc.
- **Estudio de la diversidad en el cine:** Con la información del elenco y el género de las películas, se pueden analizar aspectos relacionados con la diversidad en la industria cinematográfica, como la representación de género y etnia en las películas, la evolución de la diversidad a lo largo del tiempo y la relación entre la diversidad en el elenco y la percepción de la audiencia.
- **Análisis de la relación entre las características técnicas y el éxito de una película:** Al incluir datos de producción y detalles técnicos en el dataset, se podría explorar la relación entre las características técnicas de las películas, como la calidad de la imagen, el sonido y la dirección de arte, y el éxito de una película en términos de rating o taquilla.

Estos son solo algunos ejemplos de las posibles aplicaciones del conjunto de datos obtenido a través del web scraping de IMDb. Este dataset ofrece una gran cantidad de información valiosa para investigadores, analistas y profesionales del cine interesados en comprender las dinámicas y tendencias de la industria cinematográfica.

8. Licencia

Nuestra suposición de que IMDb utiliza la licencia Attribution NonCommercial-NoDerivs (CC BY-NC-ND) se basa en las características y restricciones de esta licencia, que parecen alinearse con las condiciones de uso y políticas de datos de IMDb. La licencia CC BY-NC-ND es la más restrictiva entre las seis licencias principales ofrecidas por Creative Commons y aborda las preocupaciones de integridad y exclusividad que IMDb podría tener en relación con sus datos.

En primer lugar, al considerar que IMDb es una de las fuentes de información cinematográfica más importantes y confiables en la actualidad, es comprensible que quieran proteger la integridad y calidad de sus datos. La licencia CC BY-NC-ND no permite la creación de obras derivadas, lo que significa que los datos no pueden ser alterados ni tergiversados de ninguna manera. Esta restricción garantiza que los usuarios de IMDb tengan acceso a información precisa y coherente en todo momento.

En segundo lugar, la restricción de uso no comercial en la licencia CC BY-NC-ND coincide con la política de IMDb de limitar el uso de sus datos a fines no comerciales. Es posible que IMDb desee mantener el control sobre la distribución y monetización de sus datos para proteger su modelo de negocio y la exclusividad de su información. La licencia CC BY-NC-ND asegura que otros no obtengan beneficios económicos directos a partir del contenido de IMDb, a menos que se obtenga una autorización expresa y se establezca un acuerdo de licencia específico.

Por estos motivos, creemos que la licencia de IMDb se ajusta perfectamente a la licencia Creative Commons Attribution NonCommercial-NoDerivs (CC BY-NC-ND).

9. Código

Repositorio github: https://github.com/OscarTienda/imdb-web-scraping

El objetivo de este proyecto es recolectar información de películas de la página web de IMDb.

Scrapy

Scrapy es un framework de Python utilizado para realizar web scraping y procesamiento de datos de manera estructurada y escalable. Ofrece una

arquitectura flexible y modular que permite construir y personalizar arañas de web (spiders) para extraer datos de sitios web de manera eficiente.

Entre las principales ventajas de Scrapy se encuentran la capacidad de realizar solicitudes HTTP de manera automatizada, la posibilidad de seguir enlaces en las páginas web y la capacidad de procesar y almacenar datos de manera estructurada en diferentes formatos (JSON, CSV, XML, entre otros). Además, cuenta con una amplia comunidad de desarrolladores y una documentación completa y actualizada.

Por estas razones, Scrapy es una excelente opción para realizar web scraping en proyectos de diferentes escalas y complejidades.

Descripción del proceso

En primer lugar, creamos un nuevo proyecto de Scrapy utilizando el comando:

scrapy startproject IMDb

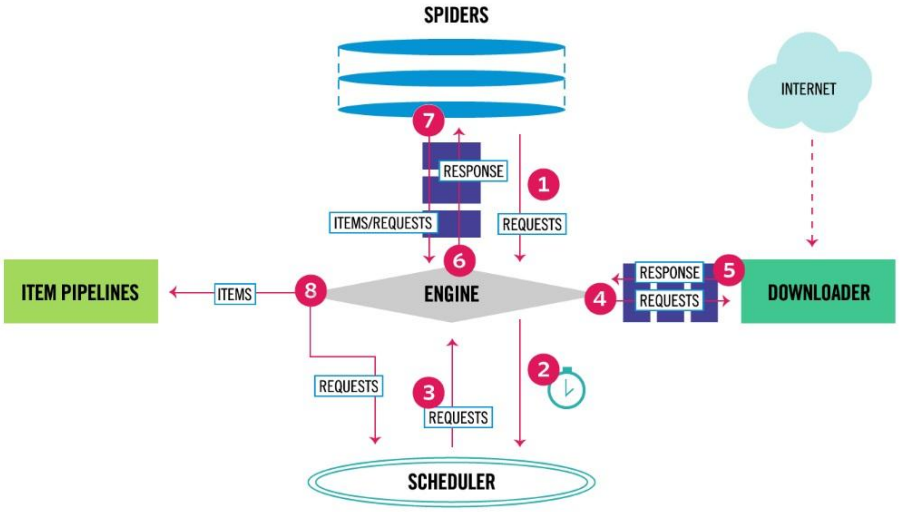
A continuación, creamos el spider de Scrapy con el comando:

scrapy genspider IMDbspider

El flujo de datos que sigue Scrapy para extraer la información de la lista de películas de IMDb es el siguiente:

- El Engine de Scrapy envía una petición al Scheduler para que comience el scraping.
- El Scheduler encola la petición y la envía al Downloader.
- El Downloader descarga los datos de la página solicitada y los envía al Spider.
- El Spider (***IMDbspider***) extrae los elementos necesarios de la página (campos definidos en el dataset: title, year, rating, …) y los empaqueta en un ítem (***ImdbMovie***), que son contenedores para la información extraída. Para cada uno de estos elementos se define un selector (xpath/css) para extraer la información correspondiente de la página web y guardarlos en el item (***ImdbMovie***). En nuestro caso, ***IMDbspider*** navega por la lista de películas y la página de detalle de cada película para extraer esa información que hemos definido relevante para el estudio.
- El ítem (***ImdbMovie***) se envía a través de los pipelines (***MyCsvItemExporter***) para su procesamiento, en nuestro caso, los almacenamos en un archivo CSV.

El siguiente diagrama muestra una descripción general de la arquitectura Scrapy con su componentes y un esquema del flujo de datos que tiene lugar dentro del sistema:



Dificultades

En cuanto a las dificultades encontradas en el sitio web, IMDb es un sitio web bien estructurado que sigue buenas prácticas web, lo que hace que la extracción de datos sea relativamente sencilla. No obstante, hemos encontrado algunas dificultades:

- **Paginación**: Para obtener más resultados de búsqueda, es necesario navegar por varias páginas. IMDb utiliza un esquema de paginación estándar, por lo que el spider debe ser capaz de navegar por múltiples páginas y extraer información de cada una de ellas.
- **Estructura del HTML**: Aunque el sitio web de IMDb sigue buenas prácticas web, la estructura del HTML puede ser compleja en algunos lugares. En nuestro caso, la estructura de la página o el formato de campos a extraer (year) cambia dependiendo de si el item encontrado es una película o una serie, para resolverlo, se han desarrollado funciones para tratar estos datos (*input_processor*). También hay distintos tipos de campos, atómicos (título, year) y no atómicos(genre, stars) cuya información se trata de manera distinta (*output_processor*). El spider debe ser capaz de navegar por el HTML de forma inteligente para extraer la información adecuada, encontrar y extraer los elementos relevantes en una página muy densa, sin ser engañado por los anuncios o enlaces adicionales. Para resolver este problema, se utilizaron selectores XPath/CSS específicos para identificar los elementos relevantes.

- **Politeness**: IMDb es un sitio web popular, lo que significa que tiene muchas visitas. Para evitar sobrecargar el sitio web y ser considerados como un ataque de denegación de servicio (DoS), debemos respetar los límites establecidos en el archivo robots.txt y limitar la frecuencia de las solicitudes para no sobrecargar el servidor del sitio web y afectar su rendimiento. Scrapy nos proporciona herramientas para manejarlo automáticamente en el fichero de configuración del Engine *settings.py:*
    - *USER_AGENT = 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/93.0.4577.63 Safari/537.36'*
    - *ROBOTSTXT_OBEY = True #Obedecer las reglas definidas en robots.txt*
    - *DOWNLOAD_DELAY = 3 # En segundos*

Archivo robots.txt

Debido a la presencia de un archivo robots.txt en el sitio web en el cual se ha desarrollado el web scraping, consideramos interesante profundizar en el contenido de este y, por supuesto, hablar primero de lo que es un archivo robots.txt.

El archivo robots.txt es un archivo utilizado por los sitios web para comunicar a los web crawlers o scrapers cuáles partes del sitio no deben ser visitadas o indexadas. Este archivo permite a los propietarios de sitios web controlar el acceso a ciertas áreas del sitio y proporciona información sobre las prácticas de web scraping permitidas en el sitio.

A continuación, se presenta primero el contenido del archivo robots.txt de IMDb y después un análisis de este:

```
# robots.txt for https://www.imdb.com properties
User-agent: *
Disallow: /OnThisDay
Disallow: /ads/
Disallow: /ap/
Disallow: /mymovies/
Disallow: /r/
Disallow: /register
Disallow: /registration/
Disallow: /search/name-text
Disallow: /search/title-text
Disallow: /find
Disallow: /find$
```

```
Disallow: /find/
Disallow: /tvschedule
Disallow: /updates
Disallow: /watch/_ajax/option
Disallow: /_json/video/mon
Disallow: /_json/getAdsForMediaViewer/
Disallow: /list/ls*/_ajax
Disallow: /list/ls*/export
Disallow: /*/*/rg*/mediaviewer/rm*/tr
Disallow: /*/rg*/mediaviewer/rm*/tr
Disallow: /*/mediaviewer/*/tr
Disallow: /title/tt*/mediaviewer/rm*/tr
Disallow: /name/nm*/mediaviewer/rm*/tr
Disallow: /gallery/rg*/mediaviewer/rm*/tr
Disallow: /tr/
Disallow: /title/tt*/watchoptions
Disallow:
/search/title/?title_type=feature,tv_movie,tv_miniseries,documentary,short,video,tv
_short&release_date=,2020-12-31&lists=%21ls538187658,%21ls539867036,%21l
s538186228&view=simple&sort=num_votes,asc&aft
Disallow: /name/nm*/filmotype/*
Disallow: /user/ur*/ratings
Disallow: /user/ur*/checkins
Disallow: /_json/*

User-agent: Baiduspider
Disallow: /list/*
Disallow: /user/*
```

El archivo robots.txt de iMDb especifica dos bloques de reglas: el primero es aplicable a todos los web crawlers o scrapers (User_agent: *) y el segundo es específico para Baiduspider, un web crawler de Baidu (User-agent: Baiduspider).

La directivas “Disallow” enumeradas en el primer bloque indican las áreas del sitio que están restringidas para todos los web crawlers. Entre estas áreas se encuentran la mayoría de las secciones del sitio web: páginas de registro, búsqueda, programas de televisión, actualizaciones, opciones de visualización, listas de películas, etcétera.

El segundo bloque de reglas, el específico para el web crawler chino, restringe el acceso a las listas de películas y a las páginas de usuarios.

Librerías y versiones utilizadas

A continuación, se adjunta el contenido del archivo requirements.txt también alojado en el repositorio cuyo contenido fue definido mediante pip freeze en un entorno creado específicamente para aplicar el proceso de scraping y el proceso de sanitización del dataset público. La gran mayoría de librerías son dependencias de Scrapy, Numpy o Pandas.

```
asttokens==2.2.1
attrs==23.1.0
Automat==22.10.0
backcall==0.2.0
certifi==2022.12.7
cffi==1.15.1
charset-normalizer==3.1.0
colorama==0.4.6
comm==0.1.3
constantly==15.1.0
cryptography==40.0.2
cssselect==1.2.0
debugpy==1.6.7
decorator==5.1.1
executing==1.2.0
filelock==3.11.0
hyperlink==21.0.0
idna==3.4
incremental==22.10.0
ipykernel==6.22.0
ipython==8.12.0
itemadapter==0.8.0
itemloaders==1.0.6
jedi==0.18.2
jmespath==1.0.1
jupyter_client==8.2.0
jupyter_core==5.3.0
lxml==4.9.2
matplotlib-inline==0.1.6
nest-asyncio==1.5.6
numpy==1.24.2
```

```
packaging==23.1
pandas==2.0.0
parsel==1.7.0
parso==0.8.3
pickleshare==0.7.5
platformdirs==3.2.0
prompt-toolkit==3.0.38
Protego==0.2.1
psutil==5.9.4
pure-eval==0.2.2
pyasn1==0.4.8
pyasn1-modules==0.2.8
pycparser==2.21
PyDispatcher==2.0.7
Pygments==2.15.0
pyOpenSSL==23.1.1
python-dateutil==2.8.2
pytz==2023.3
pywin32==306
pyzmq==25.0.2
queuelib==1.6.2
requests==2.28.2
requests-file==1.5.1
Scrapy==2.8.0
service-identity==21.1.0
six==1.16.0
stack-data==0.6.2
tldextract==3.4.0
tornado==6.2
traitlets==5.9.0
Twisted==22.10.0
twisted-iocpsupport==1.0.3
typing_extensions==4.5.0
tzdata==2023.3
urllib3==1.26.15
w3lib==2.1.1
wcwidth==0.2.6
zope.interface==6.0
```

10. Dataset

Dado que las condiciones de IMDb no permite la publicación de un dataset, hemos desarrollado un conjunto de datos sanitizado que se desvincula de IMDb, permitiéndonos así subirlo a Zenodo y completar la práctica.

Nos hemos tomado en serio el proceso de sanitización: no solo hemos eliminado por completo la URL en el conjunto de datos, sino que también hemos modificado de manera integral la columna "summary", que aunque es muy parecida en la mayoría de las webs del estilo, es propiedad de IMDb. Además, hemos sanitizado otros campos, incluyendo "rating", "votes", "num_user_reviews" y "num_critic_reviews" siguiendo diversas estrategias, como la generación de números aleatorios o el cálculo de porcentajes.

Debido a las restricciones impuestas por IMDb, no podemos publicar abiertamente el conjunto de datos real en Zenodo. Por lo tanto, hemos generado un conjunto de datos simulado que se asemeja al original y no infringe las condiciones de uso de IMDb. Hemos publicado este conjunto de datos simulado en Zenodo en formato CSV, incluyendo una breve descripción del mismo. A continuación, se encuentra el enlace de Zenodo del conjunto de datos:
https://zenodo.org/record/7838158#.ZD123XbP1D8

En nuestro repositorio privado de GitHub, hemos incluido tanto la versión completa vinculada a IMDb, que no podemos distribuir, como la versión sanitizada disponible en Zenodo. Asimismo, hemos comunicado al profesor el conjunto de datos real de forma privada, proporcionando acceso al repositorio privado donde se encuentran ambas versiones del conjunto de datos.

11. Video

URL:
https://drive.google.com/file/d/1h57-cnPNXxZw20mE2hBBokDvgoyKNAso/view?usp=sharing

12. Contribución

Contribuciones	Firma
Investigación previa	**César Fernández , Óscar Tienda**
Redacción de las respuestas	**César Fernández , Óscar Tienda**
Desarrollo del código	**César Fernández , Óscar Tienda**
Participación en el vídeo	**César Fernández , Óscar Tienda**

Referencias

[1] –.Amazon Web Services. (s. f.). AWS Marketplace: Search Results. Recuperado de https://aws.amazon.com/marketplace/search/results?FULFILLMENT_OPTION_TYPE=DATA_EXCHANGE&CREATOR=0af153a3-339f-48c2-8b42-3b9fa26d3367&DATA_AVAILABLE_THROUGH=API_GATEWAY_APIS&filters=FULFILLMENT_OPTION_TYPE%2CCREATOR%2CDATA_AVAILABLE_THROUGH

[2] – IMDb. (s. f.). IMDb: Interfaces. Recuperado de https://www.imdb.com/interfaces/

[3] – IMDb Developer. (s. f.). IMDb Developer. Recuperado de https://developer.imdb.com/

[4] – IMDb. (s. f.). ¿Puedo utilizar datos de IMDb en mi software? Recuperado de https://help.imdb.com/article/imdb/general-information/can-i-use-imdb-data-in-my-software/G5JTRESSHJBBHTGX?ref_=helpart_nav_19#

[5] – IMDb. (s. f.). IMDb: Condiciones de uso. Recuperado de https://www.imdb.com/conditions?ref_=helpms_ih_gi_usedata

[6] – freecodecamp.org. (s. f.). Web scraping sci-fi movies from IMDb with Python. Recuperado de https://www.freecodecamp.org/news/web-scraping-sci-fi-movies-from-imdb-with-python/

[7] – Kaggle. (s. f.). Web scraper IMDb movies. Recuperado de https://www.kaggle.com/code/akdagmelih/web-scraper-imdb-movies

[8] – Scrapy. (s. f.). Documentación de Scrapy. Recuperado de https://docs.scrapy.org/en/latest/