

Contexto General.

IMDb, o la Base de Datos de Películas en Internet, es una plataforma en línea que proporciona información detallada sobre películas, programas de televisión, series, videojuegos y personas relacionadas con la industria del entretenimiento. Fue fundada en 1990 por el programador británico Col Needham, y desde entonces se ha convertido en una de las fuentes de información más importantes y populares para los amantes del cine y la televisión.

IMDb es una base de datos en constante expansión, que recopila información de una variedad de fuentes, incluyendo estudios cinematográficos, redes de televisión, festivales de cine, distribuidores de medios y, en algunos casos, los propios usuarios. La plataforma tiene como objetivo proporcionar información detallada y precisa sobre cada producción, incluyendo su título, fecha de lanzamiento, duración, género, país de origen, elenco y equipo, sinopsis y calificaciones de los usuarios. Además, IMDb ofrece una variedad de características adicionales, como noticias de la industria del entretenimiento, trailers, reseñas y listas de las películas y programas de televisión más populares.

La importancia de IMDb radica en su capacidad para proporcionar información detallada y precisa sobre la industria del entretenimiento. Al recopilar información sobre producciones y personas relacionadas con la industria, IMDb ha creado una base de datos exhaustiva que puede ser utilizada por fans, periodistas, profesionales de la industria y analistas de datos. La plataforma ha permitido a los fans descubrir nuevas películas y programas de televisión, a los periodistas acceder a información precisa para sus reseñas y entrevistas, y a los profesionales de la industria acceder a información relevante para sus trabajos.

Además de su valor informativo, IMDb también es importante para la industria del entretenimiento en términos de promoción y evaluación del éxito financiero y crítico de una producción. La plataforma proporciona una plataforma en línea para que las producciones promuevan su trabajo y se conecten con los fans. Los estudios cinematográficos y las redes de televisión pueden publicar trailers, imágenes y noticias en sus páginas de IMDb para generar interés en sus producciones. La plataforma también es utilizada por la industria para evaluar el éxito financiero y crítico de una producción, lo que puede tener un impacto en la financiación y la distribución de futuras producciones.

IMDb también ha desarrollado una comunidad activa de usuarios, que utilizan la plataforma para discutir y debatir sobre películas, programas de televisión y la industria en general. Los usuarios pueden calificar y revisar las producciones, lo que puede influir en la percepción del éxito de una producción. La plataforma también ofrece una variedad de características interactivas, como listas de reproducción

personalizadas y opciones de recomendación, que permiten a los usuarios descubrir nuevas producciones.

En conclusión, IMDb es una plataforma en línea esencial para cualquier persona interesada en la industria del entretenimiento. Al proporcionar información detallada sobre películas, programas de televisión y personas relacionadas con la industria, la plataforma se ha convertido en una herramienta indispensable para fans, periodistas, profesionales de la industria y analistas de datos. Su importancia radica en su capacidad para proporcionar información precisa y exhaustiva, y en su papel en la promoción y evaluación del éxito.

Conjuntos de datos de IMDb

Los subconjuntos de datos de IMDb están disponibles para el acceso de los clientes para uso personal y no comercial. Es posible contar con copias locales de estos datos, y están sujetos a ciertos términos y condiciones que son explicitados al obtener dicha información.

Detalles del conjunto de datos de IMDb

Cada conjunto de datos está contenido en un archivo con formato de valores separados por tabuladores (TSV) en el conjunto de caracteres UTF-8. La primera línea de cada archivo contiene encabezados que describen lo que hay en cada columna. Un '\N' se utiliza para indicar que un campo en particular falta o es nulo para ese título/nombre.

Los conjuntos de datos disponibles son los siguientes:

title.akas.tsv: contiene la siguiente información para los títulos:

- titleId (cadena) - un tconst, un identificador único alfanumérico del título
- ordering (entero) – un número para identificar filas de forma única para un titleId dado
- título (cadena) – el título localizado
- región (cadena) - la región para esta versión del título
- language (cadena) - el idioma del título
- tipos (matriz): conjunto enumerado de atributos para este título alternativo. Uno o más de los siguientes: "alternativo", "dvd", "festival", "tv", "video", "en funcionamiento", "original", "imdbDisplay". Es posible que se agreguen nuevos valores en el futuro sin previo aviso

- atributos (matriz): términos adicionales para describir este título alternativo, no enumerados
- isOriginalTitle (booleano) – 0: no es el título original; 1: título original

title.basics.tsv: contiene la siguiente información para los títulos:

- tconst (cadena) - identificador único alfanumérico del título
- titleType (cadena): el tipo/formato del título (p. ej., película, corto, serie de televisión, episodio de televisión, video, etc.)
- PrimaryTitle (cadena): el título más popular / el título utilizado por los cineastas en los materiales promocionales en el momento del lanzamiento
- originalTitle (cadena) - título original, en el idioma original
- isAdult (booleano) - 0: título no adulto; 1: título de adulto
- startYear (YYYY): representa el año de lanzamiento de un título. En el caso de Series de TV, es el año de inicio de la serie
- endYear (YYYY) – Año de finalización de la serie de TV. 'N' para todos los demás tipos de títulos
- runtimeMinutes: tiempo de ejecución principal del título, en minutos
- géneros (matriz de cadenas): incluye hasta tres géneros asociados con el título

title.crew.tsv: contiene la información del director y del escritor para todos los títulos en IMDb. Los campos incluyen:

- tconst (cadena) - identificador único alfanumérico del título
- directores (matriz de nconsts) - director(es) del título dado
- escritores (matriz de nconsts) – escritor(es) del título dado
- title.episode.tsv.gz : contiene la información del episodio de televisión. Los campos incluyen:
 - tconst (cadena) - identificador alfanumérico del episodio
 - parentTconst (cadena) - identificador alfanumérico de la serie de TV principal
 - seasonNumber (entero) – número de temporada al que pertenece el episodio
 - episodioNumber (entero) – número de episodio del tconst en la serie de TV

title.principals.tsv: contiene el reparto/equipo principal de los títulos

- tconst (cadena) - identificador único alfanumérico del título
- ordering (entero) – un número para identificar filas de forma única para un titleId dado
- nconst (cadena) - identificador único alfanumérico del nombre/persona
- categoría (cadena) - la categoría de trabajo en la que estaba esa persona
- trabajo (cadena) - el título de trabajo específico si corresponde, de lo contrario '\N'
- caracteres (cadena) - el nombre del personaje reproducido si corresponde, de lo contrario '\N'

title.ratings.tsv: contiene la calificación de IMDb y la información de votos para los títulos

- tconst (cadena) - identificador único alfanumérico del título
- AverageRating: promedio ponderado de todas las calificaciones de los usuarios individuales.
- numVotes - número de votos que ha recibido el título
- name.basics.tsv.gz : contiene la siguiente información para los nombres:
- nconst (cadena) - identificador único alfanumérico del nombre/persona
- PrimaryName (cadena): nombre por el que se acredita a la persona con mayor frecuencia
- año de nacimiento: en formato AAAA
- deathYear - en formato AAAA si corresponde, de lo contrario '\N'
- PrimaryProfession (matriz de cadenas): las 3 profesiones principales de la persona
- knownForTitles (matriz de tconsts): títulos por los que la persona es conocida

Descarga:

<https://we.tl/t-bmskzZ0LGa>

Utilización del Conjunto de Datos.

A continuación, hay una lista de datos, comportamientos, relaciones o información relevante que se pueden obtener utilizando los datasets públicos de IMDb y algunas otras fuentes de información:

- Identificar las películas más populares de cada año, género y país.
- Analizar las tendencias de taquilla de películas durante ciertos períodos de tiempo.
- Identificar las películas con la calificación más alta en IMDb.
- Analizar las calificaciones de IMDb para encontrar patrones de género y edad.
- Analizar la relación entre la calificación de IMDb y la taquilla de la película.
- Analizar la distribución geográfica de las producciones de cine y televisión.
- Identificar las películas con las críticas más positivas y negativas.
- Identificar las películas más exitosas dirigidas por mujeres.
- Identificar las películas más populares basadas en libros o cómics.
- Identificar las actrices y actores con mayor éxito en la industria del entretenimiento en cada época, género y país.
- Identificar las actrices y actores que trabajan más frecuentemente en producciones.
- Analizar las relaciones entre los actores, actrices y directores en la industria.
- Identificar las películas y programas de televisión más vistos en diferentes países y regiones.
- Analizar la popularidad de los diferentes géneros de películas y programas de televisión en diferentes países y regiones.
- Identificar las películas y programas de televisión más influyentes en la cultura pop.
- Identificar los directores con mayor éxito financiero en la industria.
- Analizar la relación entre la duración de la película y el éxito financiero.
- Identificar los personajes más populares en las películas y programas de televisión.
- Analizar la evolución de los roles de género en las películas y programas de televisión.
- Identificar los géneros y temas de películas más populares entre diferentes grupos demográficos en función del tiempo.
- Identificar los géneros de películas y programas de televisión más rentables.
- Identificar la duración promedio de una producción según su género y analizar la relación de dicha información con el éxito de cada producción.

- Analizar la evolución de la popularidad de un actor, actriz o director a lo largo del tiempo.
- Identificar los actores, actrices o directores que han trabajado juntos con más frecuencia y analizar su relación con el éxito de cada producción, según su género.

Tareas a realizar por Grupo.

Cada grupo debe escoger y desarrollar 10 (diez) actividades y desarrollarlas de acuerdo a las instrucciones enviadas con anterioridad.