Tipología y ciclo de vida de los datos

Práctica 1

¿Cómo podemos capturar los datos de la web?
25% nota final

Fecha de entrega
12 de noviembre de 2024

Realizada por:

Jorge Moreno Fuentes

Jorge Martín Rota

Índice

Contexto:

Explicar en qué contexto específico se han recolectado los datos y argumentar por qué el sitio web seleccionado es una fuente pertinente y fiable de esa información. Indicar la dirección del sitio web.

Se ha elegido IMDb como fuente de datos para crear un conjunto de datos basado en las 250 mejores películas según su clasificación en el sitio web. IMDb (Internet Movie Database) es una plataforma reconocida y confiable en la industria del cine, que ofrece información detallada sobre películas, series de televisión, actores, directores, y otros elementos relacionados con la industria cinematográfica. La página utilizada para este proyecto es IMDb Top 250. Esta lista define las 250 películas mejor valoradas según los usuarios de IMDb.

IMDb es una buena fuente para este tipo de práctica de scraping porque ofrece datos públicos bien organizados y de fácil acceso. Se ha elegido IMDb por la fiabilidad de su información, por su popularidad en la industria y la gran cantidad de datos que proporciona. Esta fuente de datos permite explorar aspectos variados de las películas como su popularidad, los directores, los géneros y el reparto. Además, su estructura clara facilita el uso de técnicas de scraping sin requerir autenticación o el uso de APIs.

Título:

Definir un título conciso y que sea descriptivo para el dataset.

Análisis de las 250 mejores películas de IMDb: Extracción y exploración de datos del cine.

Descripción del dataset:

Desarrollar una breve descripción del conjunto de datos que se ha extraído. Es necesario que esta descripción sea coherente con el título elegido.

El dataset extraído contiene información detallada sobre las 250 películas mejor valoradas en IMDb. El objetivo es analizar diversos aspectos del cine y su impacto en la audiencia. Para cada película, se han recopilado datos como el título, la calificación, el género, la fecha de lanzamiento, el director, los escritores y los actores principales.

Estos datos permiten explorar patrones de popularidad, analizar el perfil de los directores y actores más recurrentes en el cine de alta valoración y entender las características que comparten las películas mejor clasificadas. Además, el dataset incluye enlaces directos a las páginas individuales de IMDb de cada película, lo que permite realizar futuras exploraciones

o análisis adicionales. Este conjunto de datos proporciona una visión completa y organizada de las películas más influyentes según la comunidad de IMDb.

Representación gráfica:

Dibujar un esquema o diagrama que refleje visualmente el dataset y el proyecto elegido

Contenido:

Explicar los campos que se incluyen en el dataset y el período de tiempo al que pertenecen los datos.

Este dataset obtiene información sobre las 250 películas mejor valoradas en IMDb. Está pensado para explorar lo que hace que estas películas sean tan apreciadas por el público. Con estos datos, se puede investigar desde los aspectos narrativos y de producción hasta las tendencias en las preferencias de la audiencia.

Cada película se describe en una fila y está organizada en las siguientes columnas:

- **Title**: El título de la película, que la identifica de manera única.
- **Rating**: La calificación promedio que le han otorgado los usuarios de IMDb. Es una señal del impacto que ha tenido en la audiencia.
- **Genre**: Los géneros que definen la película, como Drama, Comedia o Acción, lo que ayuda a ver qué tipos de historias capturan más la atención del público.
- **Web**: Un enlace directo a la página de IMDb de la película. Puede ser útil para consultar detalles adicionales o actualizados.
- **Date**: La fecha de estreno, útil para explorar cómo ha cambiado la popularidad del cine a lo largo del tiempo y descubrir qué décadas han dado lugar a las películas más icónicas.
- **Director**: El director o directores que dieron vida a cada película. Esto permite descubrir quiénes están detrás de estas grandes obras y observar patrones en su estilo o en la evolución de sus carreras.
- **Writers**: Los guionistas que crearon la historia. Es una oportunidad para ver quiénes están detrás de las mejores películas y qué temas o enfoques suelen abordar.
- **Actors**: Los actores principales del elenco. Este dato ayuda a identificar las estrellas más recurrentes y explorar cómo la presencia de ciertos actores puede influir en la popularidad de una película

Este dataset proporciona un vistazo a las películas favoritas del público y también permite estudiar patrones: **¿hay géneros más comunes en las películas mejor valoradas? ¿Algunos directores o actores aparecen más de una vez? ¿Qué tendencias han cambiado con el tiempo?** Son algunas de las preguntas que el dataset puede resolver y que se pueden estudiar en futuras prácticas.

Propietario:

Presentar al propietario del conjunto de datos. Es necesario incluir citas de análisis anteriores o, en su defecto, justificar esta búsqueda con análisis similares. Indicar qué pasos se han seguido para actuar de acuerdo con los principios éticos y legales en el contexto del proyecto elegido.

El propietario de los datos originales es IMDb (Internet Movie Database). Es una plataforma ampliamente utilizada y reconocida en la industria cinematográfica. Es propiedad de Amazon. IMDb recopila y organiza datos sobre películas, series, actores, directores y otros aspectos del cine y la televisión. Toda la información presentada en IMDb es accesible públicamente. Esto permite su uso para análisis y estudios siempre que se respeten las políticas de uso de la plataforma.

Para esta práctica de scraping, hemos seguido un enfoque respetuoso y ético. En primer lugar, el acceso a los datos se realiza mediante técnicas de web scraping. Se incluyen medidas para evitar saturar los servidores de IMDb como incluir pausas aleatorias entre las solicitudes. Además, se ha utilizado un *User-Agent* válido para identificar la naturaleza de la solicitud y cumplir con los términos de uso del sitio.

Este proyecto se inspira en estudios y análisis similares de datos de IMDb que han sido utilizados en el pasado para explorar tendencias cinematográficas, identificar patrones de popularidad o evaluar la influencia de ciertos actores y directores en la industria. La naturaleza educativa y analítica de este trabajo está en línea con el uso responsable y permitido de los datos de IMDb, ya que no se involucra ningún uso comercial ni modificación de la información original. Además, no se emplearon APIs externas como fuente de los datos. Esto sigue las indicaciones de la práctica y asegura que la extracción se haga directamente desde el sitio de IMDb de manera controlada y respetuosa.

Inspiración:

Explicar por qué puede ser interesante este conjunto de datos y qué preguntas se pretenden responder con ellos. Es necesario comparar con los análisis anteriores o análisis similares presentados en el apartado 6.

Este conjunto de datos es valioso porque permite explorar qué hace que ciertas películas sean consideradas "las mejores" por la comunidad de IMDb. Esta comunidad tiene una audiencia de millones de usuarios. Al analizar factores como la calificación, el género, el equipo creativo y los actores principales, es posible investigar las características que tienen en común las películas más apreciadas y cómo han cambiado las preferencias a lo largo del tiempo.

Este tipo de análisis puede ayudar a responder preguntas como:

- ¿Qué géneros son más comunes entre las películas mejor valoradas?

  Esto permite observar si ciertos temas o estilos tienen una aceptación universal o varían según las épocas.

- ¿Existen directores o actores recurrentes en las películas de alta valoración?

  Identificar patrones en la elección de directores o actores puede revelar el impacto de ciertos individuos en el éxito cinematográfico.

- ¿Cómo han evolucionado las tendencias de popularidad a lo largo de las décadas?

  Al analizar la fecha de estreno, se puede ver cómo las preferencias han cambiado y si hay épocas que se consideran más influyentes en la historia del cine.

- ¿Qué impacto tiene el reparto o el equipo creativo en la valoración de una película?

  Esto permite investigar si la presencia de ciertos nombres en el reparto o en el equipo creativo aumenta la popularidad de una película.

Este tipo de análisis no es nuevo; estudios previos han utilizado datos de IMDb para explorar patrones de popularidad, identificar los géneros favoritos de cada década, o estudiar la influencia de ciertos actores y directores en la cultura popular. Sin embargo, cada análisis aporta nuevas perspectivas y puede centrarse en preguntas específicas de interés actual. Por ejemplo, investigaciones similares han encontrado que ciertos géneros, como el drama y el crimen, son recurrentes entre las películas mejor valoradas, y que algunos directores tienen una frecuencia notable en la lista de películas favoritas. Este trabajo pretende sumarse a estos estudios explorando el dataset desde un enfoque único que considera múltiples variables, como género, directores, escritores y actores, para entender mejor los factores detrás del éxito cinematográfico.

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)

Para este proyecto, la **licencia CC BY-NC-SA 4.0 (Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International)** es una opción adecuada para el dataset resultante. Esta licencia permite que otros usuarios compartan, adopten y utilicen el dataset siempre que:

- Den el crédito adecuado al autor.
- No lo usen con fines comerciales.
- Compartan el resultado bajo los mismos términos de esta licencia.

La elección se basa en los siguientes puntos:

1. Reconocimiento del Propietario Original: El dataset se basa en datos públicos de IMDb, propiedad de Amazon. Al aplicar esta licencia, permitimos su uso y reutilización mientras se reconoce a IMDb como fuente original de los datos. Con esta licencia se cumplen con los principios éticos de atribución.
2. No Comercial: La licencia restringe el uso comercial del dataset, este punto es importante ya que los datos originales de IMDb no fueron creados para fines comerciales sin autorización. Esto asegura que el dataset mantenga un uso educativo y de investigación sin fines de lucro.
3. Compatibilidad con el Propósito Educativo: Al requerir que cualquier trabajo derivado utilice la misma licencia se fomenta un uso responsable y colaborativo del dataset en la comunidad educativa. Con este punto se asegura que futuras versiones del conjunto de datos sigan siendo accesibles para fines no comerciales.