Práctica 1: Web Scraping

Las 1.000 Películas Mejor Valoradas en FilmAffinity (2013–2023)

Juan Antonio Tora Cánovas y *Tim Thorp*

Noviembre 2023

Contents

1 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.

Los datos se han recolectado con el propósito de analizar tendencias y preferencias en la industria cinematográfica a lo largo de la última década. FilmAffinity es un reconocido agregador de reseñas y calificaciones de películas que cuenta con una vasta base de usuarios activos. Las calificaciones en FilmAffinity se generan por millones de usuarios, lo que proporciona una muestra representativa de la recepción de las películas. La dirección web desde donde se han extraído los datos es www.filmaffinity.com.

2 Título

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

Aunque ya se ha presentado en la portada, se incluye aquí también para mantener la enumeración del enunciado: **Las 1.000 Películas Mejor Valoradas en FilmAffinity (2013–2023)**

3 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 conjunto de datos extraído para este estudio comprende las 1.000 películas mejor valoradas en FilmAffinity desde el año 2013 hasta 2023. Este periodo de tiempo se seleccionó para capturar las tendencias contemporáneas en la valoración de películas y para reflejar los cambios en las preferencias de la audiencia a lo largo de los últimos diez años.

4 Representación gráfica

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

(Escribir respuesta aquí)

5 Contenido

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

El dataset abarca el período de 2013 a 2023 y los datos se extrajeron de FilmAffinity en Noviembre de 2023. A continuación, se detallan los campos incluidos:

5.1 Título

El `Título` representa el nombre de la película tal como se lanzó en España. Este campo es de tipo cadena de caracteres.

5.2 Título Original

El `Título Original` refiere al nombre de la película en su país de origen, el cual puede coincidir o no con el nombre en España. Este campo es de tipo cadena de caracteres.

5.3 Año

El campo `Año` indica el año de lanzamiento de la película. Todos los valores son enteros y se encuentran en el rango de 2013 a 2023.

5.4 Duración

La `Duración` se registra en minutos y representa la duración total de la película. Este campo es de tipo entero.

5.5 Género

`Género` describe los géneros cinematográficos de la película, pudiendo incluir múltiples géneros separados por comas. Este campo es de tipo cadena de caracteres.

5.6 País

El campo `País` detalla el país o países de origen de la producción de la película. Este campo es de tipo cadena de caracteres.

5.7 Puntuación Media

La `Puntuación Media` es un valor decimal que refleja la calificación promedio otorgada por los usuarios, en una escala de 1,0 a 10,0, usando la coma como separador decimal.

5.8 Número de Puntuaciones

`Número de Puntuaciones` indica cuántas valoraciones ha recibido la película. Este campo es de tipo entero.

5.9 Director

`Director` lista el nombre del director o directores de la película. En caso de múltiples directores, sus nombres se separan por comas. Este campo es de tipo cadena de caracteres.

5.10 Reparto

El `Reparto` incluye los nombres de los actores principales, separados por comas. Este campo es de tipo cadena de caracteres.

5.11 Sinopsis

La `Sinopsis` ofrece un breve resumen del argumento de la película. Este campo es de tipo cadena de caracteres y usualmente consiste en dos o tres frases.

5.12 Enlace

El `Enlace` es la URL a la página específica de FilmAffinity para la película en cuestión. Este campo es de tipo cadena de caracteres.

6 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.

(Escribir respuesta aquí)

7 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.

(Escribir respuesta aqui)

8 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).*

(Escribir respuesta aquí)

9 Código

Código implementado para la obtención del 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 la memoria en 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 se han resuelto.*

(Escribir respuesta aquí)

10 Dataset

Publicar el dataset obtenido en formato CSV en Zenodo, incluyendo una breve descripción del mismo. 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:*

a. *Comentar esta circunstancia y justificar el motivo.*

b. *Generar un dataset simulado y publicarlo en Zenodo, obteniendo el enlace del DOI*
c. *Comunicar al profesor el dataset real de forma privada (p. ej., en el repositorio privado o en una carpeta de Google Drive privada)*

(Escribir respuesta aquí)

11 Video

Realizar un breve vídeo explicativo de la práctica ***(máximo 10 minutos)****, que deberá contar con la participación de los dos integrantes del grupo. En el vídeo se deberá realizar una presentación del proyecto, destacando los puntos más relevantes, tanto de las respuestas a los apartados como del código utilizado para extraer los datos. Indicar el enlace del vídeo (https://drive.google.com/), que deberá ubicarse en el Google Drive de la UOC.*

(Escribir respuesta aquí)

Bibliografía Utilizada

1. Subirats, L., Calvo, M. (2018). *Web Scraping.* Editorial UOC.
2. Lawson, R. (2015). Scraping the Data. En *Web Scraping with Python* (Capítulo 2). Packt Publishing Ltd.
3. Selenium. (2023). *Getting started.* Recuperado de https://www.selenium.dev/documentation/webdriver/getting_started/