Práctica 1: Web scraping

Jose Cano Agüero y Raül Villalba Rodríguez

12 de Abril 2021

1 Contexto

El cine siempre ha sido uno de los principales modos de entretenimiento en nuestra sociedad. Desde la aparición de las plataformas de VOD, la oferta ha crecido exponencialmente. Es común que llegado el momento de escoger que película ver, las dudas debidas a la gran cantidad de opciones hagan que el tiempo previo a la visualización de la película se dilate. Es por ello, que se han vuelto especialmente populares las paginas web de valoraciones y criticas de películas. Una de las más populares es la pagina filmaffinity.com.

Nuestro proyecto, por tanto, trata la extracción de datos a partir de esta pagina web, poniendo especial énfasis en el apartado de las críticas. Este tipo de datos se utilizan para entrenar algoritmos NPL que luego permiten detectar críticas negativas en foros y páginas web. Así pues, nuestros dataset final podría utilizarse con este fin, y en definitiva, para facilitar la búsqueda de opciones cinematográficas.

2 Título dataset: filmaffinity

Debido a que no solo recogemos las críticas, hemos considerado que el nombre más adecuado es uno genérico, como bien podría ser filmaffinity.

3 Descripción del dataset

El dataset incluye la información que se encuentra disponible en la pagina web, para cada película.

4 Contenido

Cada fila del dataset corresponde a una película.

Atributos:

- titulo: título de la película.
- referencia: url con la información de la película.
- duracion: duración en minutos.
- imagen: imagen de cartelera de la película.
- descripcion: Sinopsis de la película.
- calificacion: Calificación general atribuida a la película.
- ListaPemios: Lista con los premios obtenidos por la película.
- listaCriticas: Lista con las distintas críticas a la película.