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 2008

Debido a que no solo recogemos las críticas, hemos considerado que el nombre más adecuado es uno genérico, como bien podría ser filmaffinity, y dado que la extracción concreta ha sido realizada con peliculas del año 2008, el nombre escogido finalmente es filmaffinity 2008.

3 Descripción del dataset

El dataset incluye la información que se encuentra disponible en la pagina web, para cada película. Esta información se puede encontrar descrita en el apartado número 5.

4 Representación gráfica

Figure 1: Muestra de la gran oferta de peliculas en plataformas VOD

5 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.

6 Agradecimientos

La extracción se ha realizado desde el sitio web filmaffinity.com. Más en concreto, a partir de su buscador. Escogiendo un año (al inicio de la ejecución del

programa), se hace la extracción de los resultados del buscador para ese año en concreto. El lenguaje de programación utilizado ha sido Python, con la ayuda de la librería BeautifulSoup de bs4.

7 Inspiración

Tal y como hemos indicado en el apartado 1, uno de los usos más interesantes para este conjunto de datos podría ser la aplicación de dicho conjunto para el entrenamiento de redes neuronales las cuales fuesen capaces de detectar críticas positivas, críticas negativas, e incluso la fiabilidad de esas críticas. Debido al origen de los datos, gracias a el atributo que tiene cada película para clasificarla, como es la calificación general de la película, se podría llegar a clasificar el valor de las críticas. Por ejemplo, una critica muy negativa, de una película con una calificación muy alta, o una critica muy positiva para una película con una calificación muy baja, podrían indicar un valor bajo de dicha crítica.

8 Licencia

Nuestra elección de licencia inicial fue CC0: Public Domain License, dado que nuestra intención es la de dar total libertad de uso a nuestro data set. Por lo tanto, dando permiso para copiar, modificar y distribuir el dataset, incluso para fines comerciales, sin necesidad de pedir permiso. Sin embargo, a la hora de crear el DOI, entre las opciones disponibles no se encontraba CC0. Por lo tanto, la licencia final ha sido: Creative Commons Attribution 4.0 Internacional, la cual permite copiar, modificar y distribuir el dataset, incluso para fines comerciales, pero con la obligación de citar la fuente, en este caso, nuestro dataset, ofreciendo link a la licencia e indicando las modificaciones realizadas.

9 Código

El código se puede encontrar en la carpeta src del repositorio de github
https://github.com/raulvillalba/web_scraping

10 Dataset

DOI del dataset: https://doi.org/10.5281/zenodo.4672708

11 Contribuciones

Contribuciones	Firma
Investigación previa	J.C / R.V
Redacción de las respuestas	J.C / R.V
Desarrollo del código	J.C / R.V