Top Filmaffinity Web Scraper

1. Contexto

Nos hallamos ante un proyecto en el que se ha implementado una web crawler o araña, basado en la técnica de web scraping, para la obtención de un número de películas o series introducido por el usuario que ejecuta el programa, de la sección Top Filmaffinity de la página web homónima. Este proyecto ha sido realizado por **Iñigo Arregui** dentro del marco de la asignatura **Tipologia i cicle de vida de les dades**.

Filmaffinity es una conocida página web en el mundo hispanohablante que reúne información sobre el mundo audiovisual relacionado con el cine y las series. Entre los diferentes servicios que ofrece existe el comentado **Top Filmaffinity** donde las películas o series son clasificadas en función de la valoración que los usuarios de la página hacen de ellas al calificar su calidad con un numero comprendido entre 0 y 10. De esta manera este Top nos da la oportunidad de acceder a un gran número de películas ordenadas en función de su calidad de acuerdo con la valoración de los usuarios de la página, y para así poder decidir contenidos audiovisuales que visualizar y disfrutar.

2. Titulo del Dataset

El dataset que se obtiene fruto del proceso de web scraping se guarda en la carpeta *filmaffinity/Output Scraper Csv* con el siguiente formato de nombre **{número de peliculas/series scrapeadas}_ TopFilmaffinity _{fecha}.csv**. De esta manera se pretende identificar el número de productos audiovisuales de los que dispone información cada fichero, así como la fecha y hora en la que se ha realizado el scrapeo. Esto se debe al hecho de que el programa desarrollado permite determinar el número de películas cuya información obtener, de manera que se puede disponer de diferentes ficheros csv con distinto número de películas/series.

3. Descripción del dataset

Tal y como se ha mencionado previamente, el dataset lo conforma un número de instancias (películas o series contenidas en el ranking Top Filmaffinity) que el usuario que ejecuta el programa determina en primera instancia antes de que se realice la extracción de información. Tal y como puede suponerse, el número de películas cuya información se pretende extraer obedece a la lógica de ordenamiento del ranking, de modo que si se solicita la extracción de información de 100 películas, éstas corresponderán a las 100 primeras películas/series del ranking. Por otro lado, cada instancia extraída se caracterizará por diferentes campos que describirán diferentes atributos de la misma como titulo, duración, género, año de estreno, diferentes profesionales encargados de los diferentes apartados técnicos, así como la productora del proyecto y la puntación y número de valoraciones que recibe la película/serie por parte de los usuarios de Filmaffinity.

4. Representación gráfica/ esquemática del dataset

 # Instancia 1

atributo 1: titulo
atributo 2: genero
atributo 3: anyo
atributo 4: duración
atributo 5: país
atributo 6: direccion
atributo 7: guion
atributo 8: musica
atributo 9: fotografia

atributo 10: reparto
atributo 11: productor
atributo 12: puntuacion
atributo 13: numVal

 # Instancia 2

atributo 1: titulo
atributo 2: genero
atributo 3: anyo
atributo 4: duración
atributo 5: país
atributo 6: direccion
atributo 7: guion
atributo 8: musica
atributo 9: fotografia
atributo 10: reparto
atributo 11: productor
atributo 12: puntuacion
atributo 13: numVal

{…}

 # Instancia n

5. Contenido

En cuanto a la información que se extraer para cada película o serie, cada instancia del dataset se caracteriza por los siguientes atributos:

- **titulo** *(string)*: título del producto audiovisual.
- **genero** *(string)*: género del producto audiovisual
- **anyo** *(integer)*: año de publicación del producto audiovisual.
- **duracion** *(string)*: duración de la película o de cada capitulo de serie en minutos.
- **pais** *(string)*: país de realización del producto audiovisual.
- **direccion** *(string)*: director del producto audiovisual.
- **guion** *(string)*: responsable(s) de la elaboración del guion del producto audiovisual.
- **musica** *(string)*: responsable(s) de la composición de la banda sonora del producto audiovisual.
- **fotografia** *(string)*: responsable(s) del apartado fotográfico del producto audiovisual.
- **reparto** *(string)*: actores y actrices que participan en el producto audiovisual.
- **productor** *(string)*: compañía(s) responsable de la producción del producto audiovisual.
- **puntuacion** *(integer)*: nota media del conjunto de valoraciones realizadas por los usuarios.
- **numVal** *(integer)*: número de valoraciones que dispone cada producto audiovisual presente en la clasificación.

Como puede observarse se ha evitado el empleo de acentos y caracteres propios del lenguaje castellano como la letra ñ para evitar errores en la codificación de nombres de los atributos.

6. Agradecimientos

A través de la siguiente se pretende agradecer a los responsables y personas que hay detrás de la existencia de un proyecto como Filmaffinity, dado el servicio que dan a todos los enamorados del séptimo arte y de las series en nuestro eterno proceso de descubrimiento de joyas audiovisuales que disfrutar, así como por ser la plataforma que he empleado para la puesta en práctica y desarrollo de un proyecto de web scraping.

7. Inspiración

Más allá de una primera funcionalidad como es el hecho de disponer de un archivo con un número determinado de películas con la que poder descubrir diferentes productos audiovisuales introduciendo filtros, otra razón que **ha inspirado** la realización de este proyecto es disponer de un conjunto de datos a través del cual poder analizar diferentes tendencias en el cine a partir de su evolución a través del tiempo, como puede ser la variación en duración y genero en función del año de producción, u otros fenómenos como la vinculación de diferentes productoras a determinado tipo de cine, directores y demás elenco, así como el origen nacional de las diferentes películas y la importancia de cada país en cuanto a aportación al cine de calidad, entre otros.

8. Licencia

9. Código

Se ha cargado un carpeta con el nombre de filmaffinity al **respositorio Github filmaffinity_scraper** de @iarregis donde se hayan las diferentes partes del proyecto. Así, el directorio **filmaffinity/filmaffinity/spiders** incluye las dos arañas empleadas en este proyecto.

10. Dataset

El archivo csv con que contiene el dataset se localiza en el directorio filmaffinity/'Output scraper csv' con el nombre **4500_TopFilmaffinity_11-11-2019.csv.**

Contribuciones	Firma
Investigación Previa	Iñigo Arregui
Redacción de Respuestas	Iñigo Arregui
Desarrollo de código	Iñigo Arregui