Práctica 1

Contexto

El conjunto de datos ha sido extraído de la información disponible en www.imbd.com, aprovechando la funcionalidad de búsqueda avanzada. Después de descubrir esta potente herramienta de búsqueda, surge la idea de elaborar un *dataset* recompilando datos sobre las películas y series de TV y servicios de steraming modernos (Netflix, Amazon Prime, HBO…) cuyo país de origen fuera España. Además, para reducir ligeramente el tamaño del *dataset* de cara a su tratamiento, se acotó la búsqueda a las series producidas después del año 2000.

Título

Series y películas de TV y plataformas de streaming españolas lanzadas entre el año 2000 y 2022, valoradas por IMBD.

Descripción

El conjunto de datos resultante de la ejecución del script elaborado consiste en un fichero en formato .csv en el que cada una de las entradas se corresponde con una película o serie española lanzada entre 2000 y 2022. Para cada una de las entradas, se recogen los siguientes datos:

- Nombre de la película o serie
- Url de su página de análisis en IMDB
- Año o años de emisión
- Edad recomendada
- Duración, duración por capítulo para las series
- Valoración de la plataforma
- Número de votos (de la valoración)

Representación gráfica

Contenido

Se procede con la explicación del formato y contenido de cada uno de los campos del *dataset.*

- Name (String): Nombre de la película o serie, en español.
- url (String): La dirección en internet de la página de detalle de la película o serie. Si se accede a ella, se puede obtener más información sobre la película y sus valoraciones (sólo se almacena la ruta, el dominio es siempre www.imdb.com).
- Year (String): Año o años de emisión. Para las series obtendremos dos años separados por un guión (“2015-2022”), y para las películas sólo un guión, que indica el año de emisión(“2021-”).

- Certificate (Integer): es la edad recomendada para cada película o serie.
- Duración (String): La duración en minutos de la película o serie. Para las series, obtendremos la duración media de cada capítulo
- Rating (Double): Es la valoración asignada por la plataforma para cada serie o película.
- Votes (int): Es el total de votos recogidos para calcular la valoración media

Proceso de la recolección de datos

Para la recolección de los datos, se crea un script en Python que realiza las siguientes tareas:

1. Petición HTTP a la dirección de la página suministrada, para obtener el HTML raw
2. Se instancia la librería BeautifulSoup para procesar el raw HTML
3. Se separan todas las ocurrencias de etiquetas 'div' con el nombre de clase 'lister-item', que gracias a la fase de análisis previa se ha identificado como el contenedor de la información de cada una de las entradas de la tabla.
4. Se itera a través del conjunto de divs extraídos del HTML, y dentro de ellos se extrae (gracias también a una previa fase de exploración manual) los datos deseados de las tags HTML en las que se encuentran.
5. Una vez se han transferido estos datos exitosamente a un objeto en Python, se procede a iterar el objeto para generar un fichero .csv que contenga toda la información extraída.

** Nota: Se realizó la implementación para ampliar los datos recogidos para cada una de las entradas accediendo a la URL de detalle de cada una de ellas, pero el tiempo de ejecución se veía incrementado en niveles inadmisibles (del orden de horas).

Agradecimientos

Naturalmente, los datos empleados para la elaboración del conjunto de datos son propiedad de IMBD. Para actuar de acuerdo con los principios éticos y legales en el contexto del proyecto, se han emprendido dos acciones principales. La primera ha sido la visualización del fichero *robots.txt* (https://www.imdb.com/robots.txt). Del análisis de este fichero, se concluye que la ruta en concreto utilizada para la extracción de los datos (/search/title/) no se encuentra dentro de la lista de rutas restringidas por este fichero. Además , el bot empleado para realizar el Scrapping ('Python-urllib/3.9') no se encuentra entre los 'User agents' con acceso restringido.

La segunda acción fue la lectura del fichero que contiene la información sobre la licencia de los servicios de datos producidos por IMDB (https://help.imdb.com/article/imdb/general-information/can-i-use-imdb-data-in-my-software/G5JTRESSHJBBHTGX?ref_=helpart_nav_26#), para conocer bajo qué licencia se publicará el dataset.

Inspiración

Siendo yo un aficionado a las películas y series, especialmente de las plataformas de streaming principales como Netflix o HBO, siempre suelo consultar este tipo de páginas donde usuarios y críticos valoran y clasifican las películas y series. Siempre

he pensado que en este tipo de webs se albergan datos muy interesantes de cara a un análisis multivariable de las valoraciones de las películas, donde se extraigan conclusiones interesantes como cuáles son las plataformas con las películas mejor valoradas, o si existe algún criterio para que las películas tengan mejor valoración (su duración, la edad recomendada…)

Licencia

La licencia bajo la que se publicará este conjunto de datos será CC BY 4.0 License.

Código

El código desarrollado para completar la especificación de la práctica se puede encontrar en el siguiente repositorio: https://github.com/javidiazdom/pra1-tcvd.

Dataset

El *dataset* está publicado en el siguiente enlace.
https://zenodo.org/record/6449802#.YlR4lcjMKPo

Contribuciones

Contribuciones	**Firma**
Investigación previa	Javier Díaz Domínguez
Redacción de las respuestas	Javier Díaz Domínguez
Desarrollo del código	Javier Díaz Domínguez