MÁSTER U. DE CIENCIA DE DATOS

Tipología y Ciclo de Vida de los Datos. Aula 1

(M2.851)

PRÁCTICA 1

Alumnos:

Olast Arrizibita Iriarte - oarrizibita@uoc.edu

Enrique Pérez Balbuena - eperezbal@uoc.edu

1.- Contexto

Debido al estado de alarma que sufre nuestro país y ante la expectativa de tenernos que quedar en casa por la cuarentena decretada por el gobierno, hemos pensado que podríamos recolectar información sobre las películas más 'top' de la web IMDb, para tener diferentes referencias y opiniones a la hora de elegir qué película nos gustaría ver durante este tiempo de confinamiento para así, estar más entretenidos.

IMDb (según sus siglas en inglés 'Internet Movie Database') es una de las web (y también una aplicación móvil) con la mayor información relacionada con películas, programas de televisión, videojuegos, premios, eventos y celebridades. Cuenta en la actualidad con unos 83 millones de usuarios registrados y es de las más populares cuando se trata de consultar información sobre actores, productores, directores. Su versión en castellano se estrenó en 2009, aunque en la actualidad sólo existe la web en inglés. Desde 1998 pasó a ser subsidiaria de Amazon [1].

2.- Título de dataset

Escoge tu película gracias a IMDb ¡Y a disfrutar!

3.- Describir el dataset

El dataset contiene básicamente los datos esenciales de una película, como su título, fecha de estreno, edad recomendada, duración o género, pero además y esto es lo interesante, para poder decantarnos por una u otra película, hemos recopilado los votos, puntuaciones junto con las valoraciones de las películas más 'top' por rango de edad de los usuarios registrados en el portal de IMDb.

4.- Imagen que representa el dataset

5.- Contenido

Para crear el dataset, primero estuvimos estudiando y analizando qué tipo de datos queríamos incluir. Qué variables nos parecía interesantes a la hora de poder entender qué películas nos podría gustar más en función de las puntuaciones y votos que había recibido. Las variables que hemos incluido son:

- **Puesto**: la posición dentro del ranking que tiene la película en los más 'top'.

- **Título**: hemos recogido los títulos de cada una de las películas que corresponden con la sección 'Top Rated Movies'.
- **Año**: Fecha de estreno de la película.
- **Rating**: Puntuación sobre 10, que los usuarios de IMDb otorga a la película.
- **Votos**: total de votos contabilizados de la película.
- **Edad**: clasificación de la película para la edad recomendada.
- **Duración**: tiempo que dura la película en horas y minutos.
- **Género**: estilo, género de la película.
- **<18**: puntuación media de todos los usuarios registrados de IMDB con una edad inferior a 18 años.
- **18-29**: puntuación media de todos los usuarios registrados de IMDb con una edad comprendida entre 18 y 29 años.
- **30-44**: puntuación media de todos los usuarios registrados de IMDb con una edad comprendida entre 30 y 44 años.
- **45+**: puntuación media de todos los usuarios registrados de IMDb con una edad superior a los 45 años.

6.- Agradecimientos

Agradecer a 'IMDb.com, Inc' (http://www.imdb.com) que es el portal que aporta los diferentes datos y a sus usuarios registrados, que con su voto y puntuación hace posible el que podamos realizar estudios de este tipo.

7.- Inspiración

Nos parece que este conjunto de datos pueden ayudar a otras personas a elegir de forma más acertada que tipo de película puede ver durante este periodo de confinamiento en casa. No solo sirve para los adultos, también para todos aquellos que tengan hijos y quieran ver qué tipo de películas podrían ponerles.

8.- Licencia

9.- Código

Adjuntamos en este archivo, el fichero de Notebook 'top_peliculas.ipynb' en código python

10.- Dataset

Adjuntamos en este archivo, el fichero 'top_peliculas.csv' con los datos

Contribuciones	Firma
Investigación previa	OA, EP
Redacción de las respuestas	OA, EP
Desarrollo código	OA, EP

Bibliografía

[1] Luís Castro (2019). IMDB: la base de datos del mundo del cine y la televisión. [En línea]. Actualizado: 1 noviembre 2019. Disponible en: https://www.aboutespanol.com/imdb-la-base-de-datos-del-mundo-del-cine-y-television-157980 [Fecha acceso: 18 marzo 2020]

[2] Lawson, R. (2015). *Web Scraping with Python*. Packt Publising Ltd. Chapter 2. Scraping the Data.

[3] Mitchel, R. (2015). *Web Scraping with Python: Collecting Data from the Modern Web*. O'Reilly Media, Inc. Chapter 1. Your Frist Web Scraper.