DAANMO

Versión en castellano.

Diomedes Barbero

DAANMO (*Data Analysis of Movies*) es el primer proyecto encuadrado en el bootcamp de *Data Science*. Es un proyecto centrado en el análisis y la visualización de datos (EDA)

El objetivo del proyecto es la visualización y análisis de *movies dataset*. El archivo zip fue descargado de Kaggle y de acuerdo a la descripción del archivo, contiene 45000 entradas diferentes referentes a películas cuya fecha de estreno llega hasta 2017. El zip contiene los siguientes archivos csv:

- Créditos (*Credits*)
- Palabras clave (*Keywords*)
- Links
- Links reducido (*Links small*)
- Metadatos de películas (*Movies metadata*)
- Puntuación (*Ratings*)
- Puntuación reducida (*Ratings small*)

El primer csv, Créditos, parece hacer referencia a las opciones de casting de la película en cuestión. Lo cual ofrecería datos acerca de los principales roles (protagonista, antagonista e interés romántico), así como de papeles de apoyo (secundarios) o incluso del equipo técnico en sí (directores, directores de fotografía, productores… etc.).

El segundo, Palabras clave, parece ser una recolección de una práctica habitual en Hollywood: la clasificación de películas en torno a la trama central. A tal efecto se espera encontrar múltiples entradas relacionadas con el cine de acción, susense, romance, ciencia ficción… entre otros.

Se desconoce la utilidad del tercer y cuarto csv (Links y Links reducido), aunque se sospecha que pueda tener relación con las fuentes de información.

Metadatos de películas mientras tanto podría contener tanto el *pitch* o premisa de la película, como el título y el género en el que se ubica.

Por último, los dos últimos csv. Puntuación y Puntuación reducida. Prácticamente con toda seguridad estos archivos contendrán una serie de puntuaciones dadas tanto por parte de los usuarios como los críticos profesionales y puede que incluso críticas en formato texto.

Tras un rápido reconocimiento de los datasets contenidos en el dataset, se pueden llegar a las siguientes conclusiones preliminares: Relacionando:

a. Créditos con Palabras Clave
    i. De esta forma podrían determinarse los géneros que los actores recogidos visitan con mayor frecuencia (*Typecasting*).
        1. De gran utilidad tanto para productores como directores de casting o guionistas.
        2. También extremadamente útil para los propios actores y actrices a la hora de determinar a qué papeles pueden acceder en determinado momento de sus vidas profesionales.
b. Créditos con Puntuación
    i. Se podría establecer qué estrellas, directores y productores están habitualmente situados en los mejores puestos frente a la crítica.
        1. Sería de especial utilidad si además se tuviera acceso al coste y beneficios de la película. Sin embargo puede seguir siendo de utilidad a la hora de determinar quién es más rentable.
c. Palabras clave con Puntuación.
    i. Podría utilizarse para establecer qué género cinematográfico y/o qué temática tiende a reunir críticas más favorables e incluso (si se establece una línea temporal) determinar la predisposición del público hacia un tipo determinado de película en determinado momento.

Fuentes

- *The Movies Dataset*, en Kaggle, https://bit.ly/2Jc6BV2
- Snyder, B. (2019). *Save the cat!* (6ª). Barcelona: Alba.

DAANMO

English version

Diomedes Barbero

DAANMO (*Data Analysis of Movies*) is the first of a series of projects situated withing the frame provided by the Data Science Bootcamp currently being attended by the author. DAANMO is at its core an EDA[1] project.

The main aim of DAANMO is to share light onto the "*The movies dataset*" (TMD) available on the website *Kaggle*. According to the dataset description TMD contains 45.000 entries of a varied selection of movies up to and including the releases of 2017. The data is distributed amongst the following dataframes:

- Credits
- Keywords
- Links
- Links Small (*links_small*)
- Movie Metadata (*movie_metadata*)
- Links
- Ratings
- Ratings Small (*ratings_small*)

Brief preliminary description:

The first csv file, credits, appears to contain data related to the casting choices and starring roles of the movies contained. That information may reveal both the protagonist and antagonist's roles, secondary protagonist (also the love interest in most cases) as well as the secondary/support characters. It may also contain information about the crew: producer, director, director of photography, special effects supervisor… etc.

Keywords on the other hand, may refer to the usual practice in Hollywood of labeling films along thematic lines. Different cinematic genres are expected to appear, such as action, drama or noir.

Links is at the time of writing of the present document impossible to decipher from just the title alone. Though it may contain links for the different websites of the sources, movies or studios responsible.

Movie Metadata is again, difficult to ascertain. Nonetheless it is expected to contain general information about the movie, such as budget, title and possibly, the pitch[2].

[1]Author's note: Exploratory Data Analysis

[2] A short and concise description of the movie's: premise, central conflict and main characters.

The last two dataframes, Ratings and Ratings Small are an easy guess: it is expected that their content will be related with various points of view expressed online about the quality of the movies presented. It is possible that those opinions have been recorded in text form.

Therefore, it is necessary to reach the following conclusions about the whole data pack. By combining the data contained within:

a. Credits with Keywords:
    i. We may obtain information relevant to which movies certain actors seem to appear most in. Also known as typecasting.
        1. This is useful for producers casting directors and even writers, as it allows them to "tighten the search" for a determinate role.
        2. Actors and actresses may also find this information useful, as to avoid being typecast in a specific role.

a. Credits with Ratings:
    i. Again, beneficial for those involved in the casting procedure. As it allows the casting of those performers better viewed by the critics and the general public.
        1. If revenue information also exists it may prove of even further use, as it would also show the most profitable stars.

b. Keywords with Ratings:
    i. At last, it is important to any producer to know whether or not the financial undertaking of the movie is worth the risk. One of the ways to inform their decision is to categorize which film genre is more popular in any given time.

Sources

- *The Movies Dataset*, downloaded from Kaggle, https://bit.ly/2Jc6BV2
- Snyder, B. (2019). *Save the cat!* (6th edition). Barcelona: Alba.