PRAC1 Tipología - Filmaffinity scraping

Jaime Pérez & Ana Hubel

08/11/2020

Contents

Contexto

Filmaffinity es un sitio web español dedicado al cine que cuenta con una gran cantidad de películas, documentales, cortometrajes, mediometrajes y series de televisión. El funcionamiento del sitio web se basa en una puntuación calculada como la media de las puntuaciones recibidas por los usuarios, que además pueden publicar sus críticas. El objetivo de la creación de nuestra base de datos es obtener información relacionada con las películas más destacabales de cada categoria que se encuentra en la página.

Link repositorio: https://github.com/JaimePerez89/WScrapingR

Título dataset

La base de datos recibe el nombre de **BD_filmaffinity_top30** y engloba una serie de variables relacionadas con las películas que se encuentran en el top 30 de cada género de la página Filmaffinity.

Descripción dataset y contenidos

La base de datos se ha creado empleando la técnica *web scrapping*. Esta técnica se ha aplicado a fecha de 5 de noviembre del 2020 (última actualización), por lo que las películas incluidas son las que se hallaban en la página hasta esa fecha.

La base de datos engloba películas desde el año 1902 hasta el año 2020, incluye las siguientes variables que describen características de las mismas:

- *id*: número de identificación de las películas incluidas en la base de datos.
- *genero_top30*: género principal al que pertenece la película.
- *posicion_top30*: posición en la que se encuentra cada película en relación con el género principal.
- *titulo*: nombre de la pelicula traducido al castellano.
- *titulo_VO*: título original de la película.
- *duracion*: duración de la película en minutos.
- *pais*: pais en el que se ha producido la película.
- *puntuacion_media*: puntuación media de las votaciones recibidas por parte de la película.
- *votos*: número de votos que ha recibido la película.
- *num_criticas*: número de críticas que ha recibido la película.
- *mejor_puntuacion*: puntuación más alta que recibe la película.
- *peor_puntuacion*: puntuación más baja que recibe la película.
- *direccion*: directores de la película.
- *guion*: guionistas de la película.
- *musica*: productores de la música de la película.
- *fotografia*: encargados de las fotografías de la película.
- *reparto*: reparto de la película.
- *productora*: estudio encargado de la producción de la película.
- *sinopsis*: resumen de la película.
- *url*: enlace en el que se encuentra la película.
- *tag1* a *tag13*: categorías secundarias asociadas a las películas.

Representación gráfica

Ver Figure 1.

Análisis de contenidos

En primer lugar, vamos a obtener los países en los que se obtienen las mejores puntuaciones según los usuarios.

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 6 x 4
##   pais          mean_puntuacion mean_criticas mean_votos
##   <chr>                   <dbl>         <dbl>      <dbl>
## 1 Italia                   8.11         209.       57.3
## 2 Australia                8.1          105        22.5
## 3 Alemania                 8.07          99.8      18.0
## 4 Nueva Zelanda            8.07         241       178.
## 5 México                   8.03          54         6.99
## 6 Portugal                 8              1       157
```

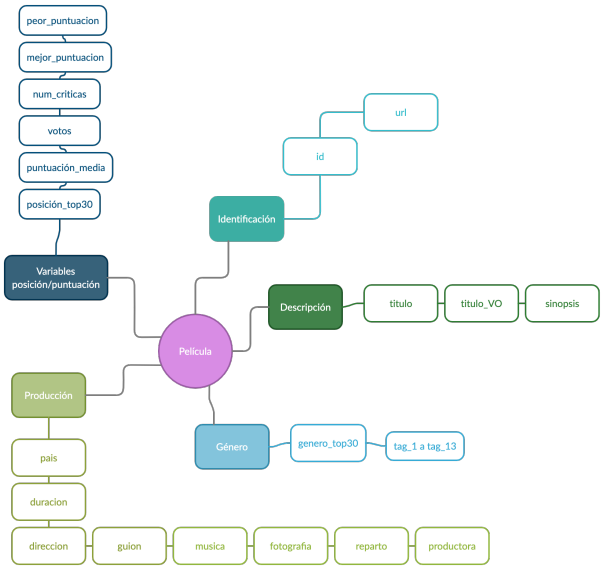
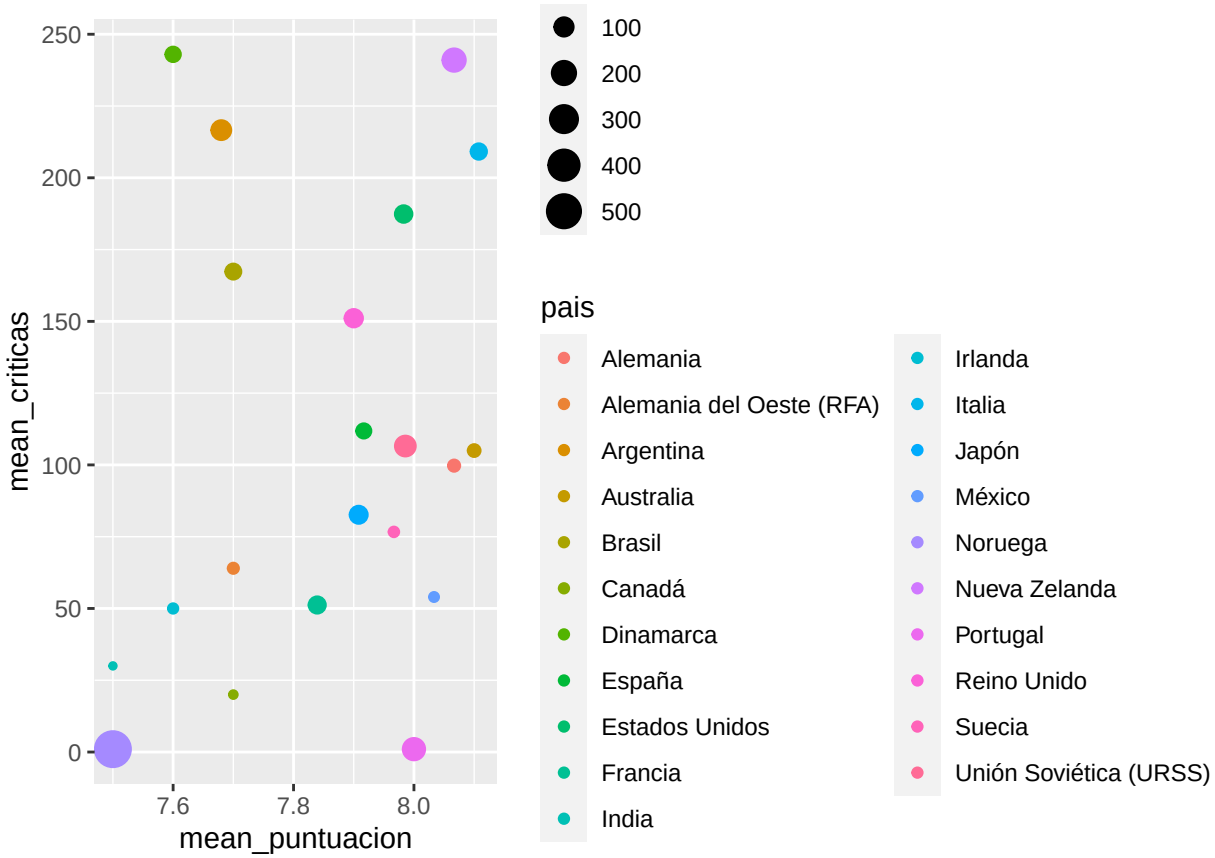


Figure 1: Representación gráfica de la base de datos

En la tabla se puede observar que el país que obtiene las mejores puntuaciones es Italia, seguida de Australia, Alemania, Nueva Zelanda, Mexico y Portugal. Sin embargo, este dato puede estar sesgado por el número de votos o el número de críticas. Ya que, como se observa, los números de críticas y votos medios están muy descompensados en los distintos países.

Por ello, a continuación vamos a realizar un análisis más exhaustivo para tratar de determinar la influencia de estas variables sobre la mejor puntuación.



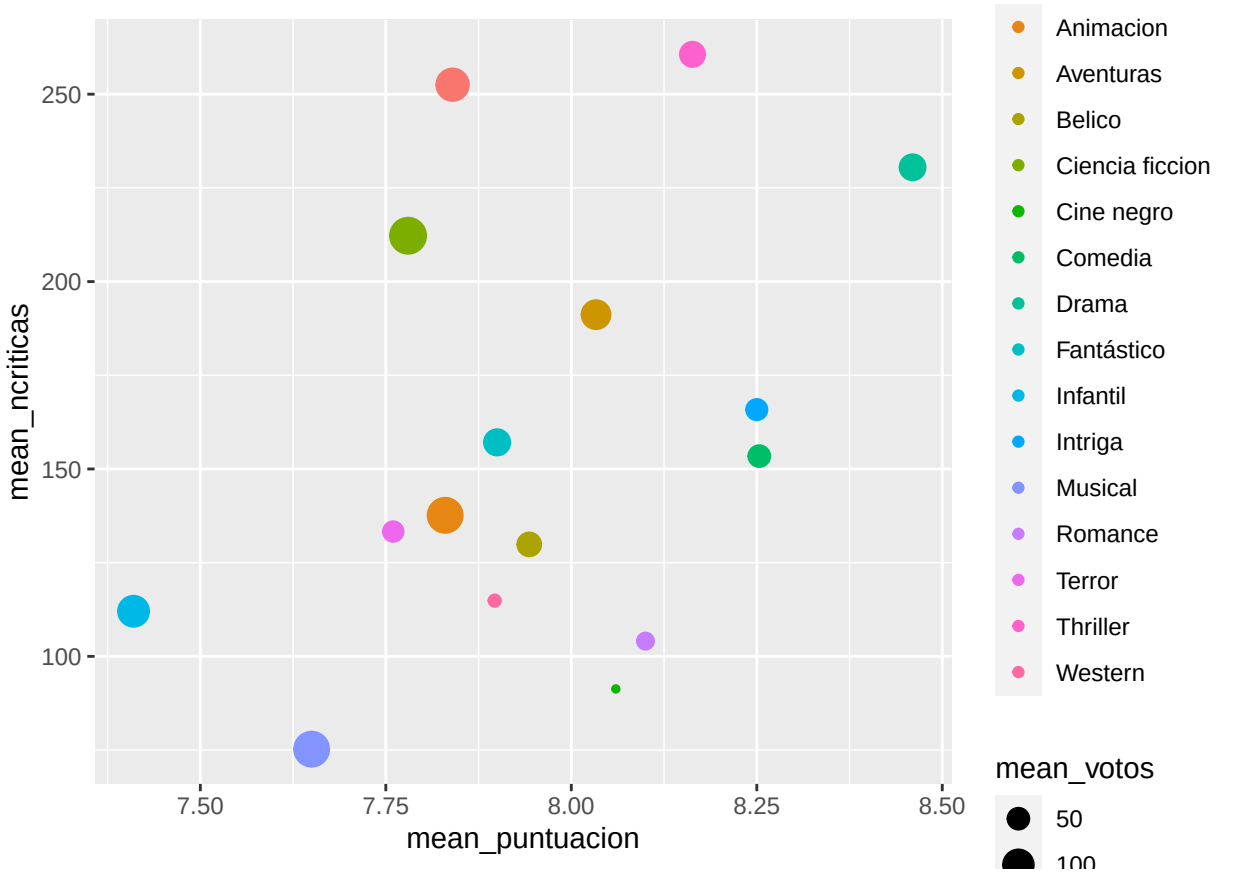
A priori, en la gráfica no se puede observar ningún sesgo o relación entre las variables; no hay un patrón definido. El único hecho que puede ser destacable es que las películas con puntuaciones más altas reciben un mayor número de críticas y votos.

A continuación, vamos a mostrar las 6 mejores categorias en función de la puntuación obtenida.

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 6 x 4
##   genero_top30 mean_puntuacion mean_ncriticas mean_votos
##   <chr>                  <dbl>          <dbl>      <dbl>
## 1 Drama                   8.46          230.        70.5
## 2 Comedia                 8.25          153.        50.1
## 3 Intriga                 8.25          166.        47.5
## 4 Thriller                8.16          261.        65.9
## 5 Romance                 8.1           104.        33.1
## 6 Cine negro              8.06           91.3       19.4
```

Como se puede observar, la categoria mejor valorada es el Drama. A continuación de esta podemos encontrar la Comedia, la Intriga, Thriller, Romance y el Cine Negro. De igual manera que ocurría en el caso anterior, en el que realizábamos un análisis por países, la media de número de críticas y votos, está muy descompensado entre las películas asociadas a cada categoria.



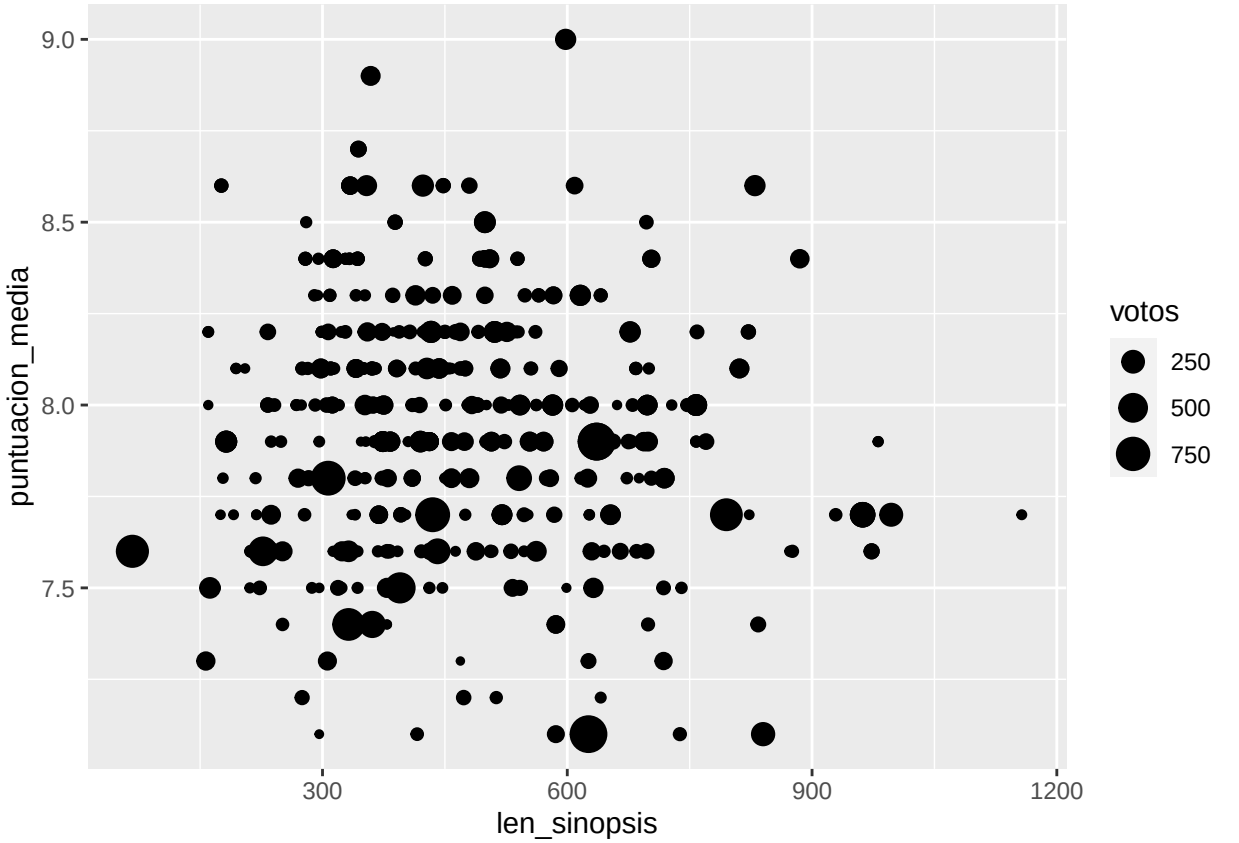
En la gráfica podemos observar que la mayor parte de las películas tienen un alto número de votos en función del género. La mayor parte de las películas con un bajo número de votos, también tienen menor número de críticas; sin embargo, no destacan por una puntuación más baja.

A continuación, vamos a realizar un análisis inferencial para determinar si dichas variables tienen una relación estadísticamente significativa entre ellas. Para ello, vamos a realizar un ANCOVA con el objetivo de hallar una posible relación entre el número de críticas, el número de votos, el género y el país respecto a la puntuación media obtenida.

```
##                Df Sum Sq Mean Sq F value   Pr(>F)
## votos           1  0.744   0.744  21.924 3.78e-06 ***
## num_criticas    1  3.633   3.633 107.058  < 2e-16 ***
## genero_top30   15 27.060   1.804  53.168  < 2e-16 ***
## pais           20  1.155   0.058   1.701   0.0302 *
## Residuals     442 14.997   0.034
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

En este caso, la modelización de las variables nos indica que todas las variables resultan significativas en el modelo. Es decir, la puntuación media depende tanto de los votos recibidos, el número de críticas, el género y el país del que procede de forma estadísticamente significativa.

Otro análisis que parece interesante puede ser la relación que puede haber entre la longitud de la sinopsis de cada película y el número de votos o la puntuación que recibe. Es decir, ¿puede ser que las personas estén viendo o puntuando, de forma inconsciente, películas que tengan unos resumenes más elaborados?



En vista de la gráfica, podemos deducir que la mayor parte de los resumenes se encuentran entre una longitud de 300 y 900 caracteres. No parece haber una relación clara con el número de votos, ni con la puntuación media. De todas formas, vamos a realizar un análisis de ANCOVA para ver si la puntuación media se ve influenciada de forma estadísticamente significativa por la longitud de la sinopsis, incluyendo las variables mencionadas anteriormente para una modelización más completa..

En primer lugar, vamos a modelizar el número de votos en función de la longitud de la sinopsis.

```
##                   Df  Sum Sq Mean Sq F value   Pr(>F)
## len_sinopsis       1  124434  124434   9.678 0.001986 **
## num_criticas       1  108475  108475   8.437 0.003861 **
## puntuacion_media   1  159565  159565  12.411 0.000471 ***
## genero_top30      15  455078   30339   2.360 0.002855 **
## pais              20  414195   20710   1.611 0.046361 *
## Residuals        441 5669978   12857
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El análisis del número de votos, en presencia del resto de variables, nos indica que la longitud del resumen tiene un efecto estadísticamente significativo sobre el número de votos que recibe cada película.

A continuación, vamos a proceder a realizar el mismo tipo de modelización con la puntuación media de cada película como variable respuesta.

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## len_sinopsis   1  0.468   0.468  13.899 0.000218 ***
## votos          1  0.605   0.605  17.987 2.71e-05 ***
## num_criticas   1  3.872   3.872 115.104  < 2e-16 ***
## genero_top30  15 26.615   1.774  52.741  < 2e-16 ***
```

```
## pais          20  1.193   0.060   1.773 0.021268 *
## Residuals    441 14.836   0.034
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Una vez más, todas las variables resultan significativas en el análisis. Es decir, adicionalmente a lo anteriormente comentado, la longitud del resumen parece tener un efecto sobre la puntuación recibida.

Por último, vamos a realizar una serie de representaciones en las que mostraremos los tags asociados a cada género de película. Posteriormente, vamos a centrarnos de forma más detallada en el género *thriller*.

En primer lugar, vamos a mostrar el primer tag en función de todos los géneros, para ver cuáles son los más frecuentes.

```
## `summarise()` regrouping output by 'genero_top30' (override with `.groups` argument)
```

```
## # A tibble: 10 x 4
## # Groups:   genero_top30 [10]
##    genero_top30    tag_1              n  freq
##    <chr>           <chr>          <int> <dbl>
##  1 Animacion       Animación         29 0.967
##  2 Infantil        Animación         28 0.933
##  3 Cine negro      Cine negro        26 0.867
##  4 Western         Western           26 0.867
##  5 Comedia         Comedia           23 0.767
##  6 Ciencia ficcion Ciencia ficción   18 0.6
##  7 Belico          Bélico            17 0.567
##  8 Drama           Drama             17 0.567
##  9 Terror          Terror            16 0.533
## 10 Musical         Musical           14 0.467
```

De los resultados obtenidos, se puede observar que prácticamente todas las categorías destacan por poseer un tag igual a su género, excepto el género infantil que osee un tag de animación.

A continuación, vamos a mostrar el segundo tag.

```
## `summarise()` regrouping output by 'genero_top30' (override with `.groups` argument)
```

```
## # A tibble: 10 x 4
## # Groups:   genero_top30 [10]
##    genero_top30 tag_2          n  freq
##    <chr>        <chr>      <int> <dbl>
##  1 Romance      Romance       19 0.633
##  2 Intriga      Intriga       17 0.567
##  3 Fantástico   Fantástico    16 0.533
##  4 Belico       Drama         14 0.467
##  5 Cine negro   Intriga       14 0.467
##  6 Aventuras    Aventuras     12 0.4
##  7 Thriller     Thriller      12 0.4
##  8 Accion       Acción        11 0.367
##  9 Animacion    Drama         11 0.367
## 10 Infantil     Fantástico    11 0.367
```

En este caso, vemos algo más de diversidad. El género bélico tiene como tag más repetido el drama, el cine negro la intrifa, la animación el tag de drama y el género infantil el tag fantástico.

De igual manero, se muestra el tercer tag.

```
## `summarise()` regrouping output by 'genero_top30' (override with `.groups` argument)
```

```
## # A tibble: 10 x 4
## # Groups:   genero_top30 [9]
##    genero_top30 tag_3                   n  freq
##    <chr>        <chr>               <int> <dbl>
##  1 Cine negro   "Crimen"               11 0.367
##  2 Thriller     "Crimen"               10 0.333
##  3 Accion       "Acción"                9 0.3
##  4 Western      ""                      9 0.3
##  5 Aventuras    "Aventuras"             8 0.267
##  6 Infantil     "Aventuras"             8 0.267
##  7 Intriga      "Crimen"                7 0.233
##  8 Accion       "Drama"                 6 0.2
##  9 Animacion    "Aventuras"             6 0.2
## 10 Belico       "II Guerra Mundial"     6 0.2
```

De cara a este tag, cabe destacar que el número de tags desciende de forma notable respecto a los dos anteriores: cada vez encontramos más tags vacíos para las películas. También cabe destacar una mayor diversidad: el género ya no coincide con el valor del tag.

De igual manera, se muestran los resultados del cuarto y quinto tag.

```
## `summarise()` regrouping output by 'genero_top30' (override with `.groups` argument)
```

```
## # A tibble: 10 x 4
## # Groups:   genero_top30 [10]
##    genero_top30 tag_4           n  freq
##    <chr>        <chr>       <int> <dbl>
##  1 Infantil     "Infantil"     15 0.5
##  2 Western      ""             15 0.5
##  3 Cine negro   "Crimen"        8 0.267
##  4 Comedia      ""              7 0.233
##  5 Intriga      ""              7 0.233
##  6 Musical      ""              7 0.233
##  7 Romance      "Melodrama"     6 0.2
##  8 Thriller     "Crimen"        6 0.2
##  9 Belico       ""              5 0.167
## 10 Drama        ""              5 0.167
```

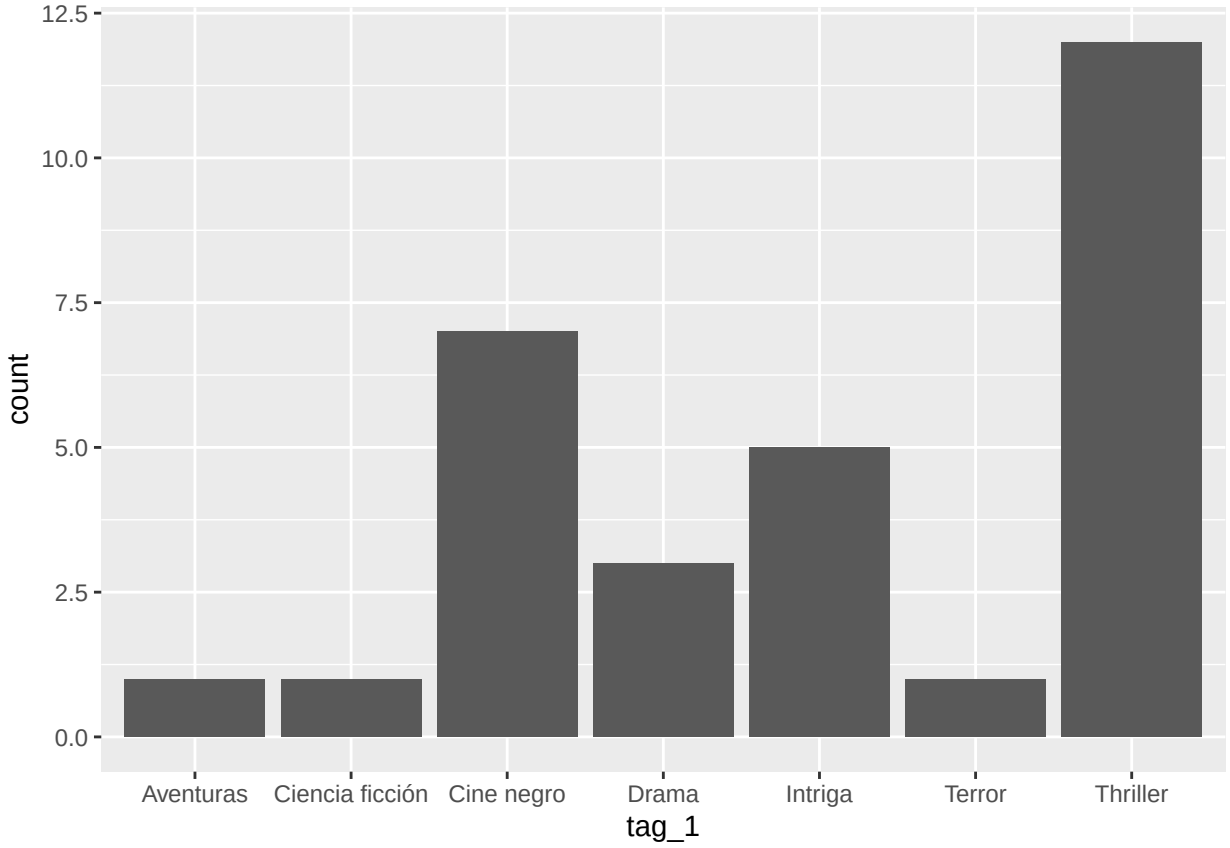
```
## `summarise()` regrouping output by 'genero_top30' (override with `.groups` argument)
```

```
## # A tibble: 10 x 4
## # Groups:   genero_top30 [10]
##    genero_top30 tag_5          n  freq
##    <chr>        <chr>      <int> <dbl>
##  1 Western      ""            24 0.8
##  2 Cine negro   ""            14 0.467
##  3 Intriga      ""            14 0.467
##  4 Comedia      ""            13 0.433
##  5 Romance      ""            12 0.4
##  6 Belico       ""            11 0.367
##  7 Musical      ""            10 0.333
##  8 Drama        ""             8 0.267
##  9 Infantil     "Infantil"     6 0.2
## 10 Thriller     ""             6 0.2
```

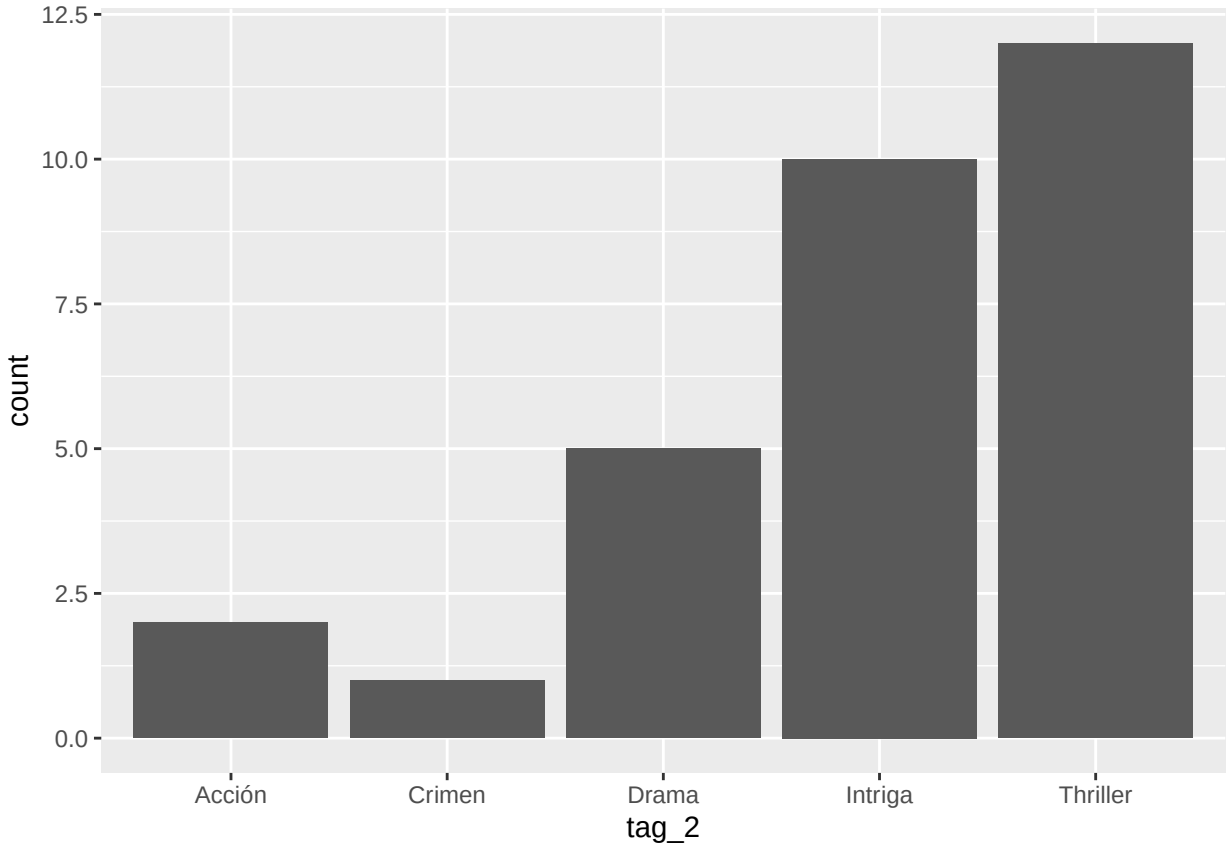
De la información mostrada, podemos extraer que la mayoría de los tags más comunes son los correspondientes al mismo género al que pertence la película. También se puede observar que los principales tags se hallan en

los primeros, conforme aumentamos de número cada vez obtenemos más tags vacíos.

Por último, vamos a centrarnos en el primer y segundo tag asociado al género thriller. La elección de estos tags se hace porque, como acabamos de mencionar, son los tags más completos y con menor número de valores vacíos.



Como hemos comentado anteriormente, se observa que el tag más repetido es el del thriller. A continuación, destacan los tags de cine negro, intriga y drama. por último, de forma irrelevante encontramos los tags de acenturas, terror y los vacíos.



En el segundo tag, y de igual forma que en el caso anterior, el tag más repetido es el de thriller. Posteriormente destacan los tags de intriga y drama. En este caso, los valores vacíos son más numerosos que en el caso anterior. Por último, el tag menos importante es el del crimen.

Agradecimientos

Especial agradecimiento a Pablo Kurt Verdú Schumann y Daniel Nicolás Aldea, creadores de Filmaffinity, por mantener un portal independiente como éste sin vinculación alguna con ningún grupo mediático. Adicionalmente, nos gustaría agradecer a todas aquellas personas que publican paquetes/vignettes/tutoriales/etc en R y que hacen que su comunidad sea cada día más grande y mejor.

Inspiración

La obtención de esta base de datos puede resultar interesante para la gente aficionada al cine, tanto expertos como no expertos. Con ella se pretende satisfacer la curiosidad de aquellas personas que, pese a tener delante el ranking de películas, no se termina de fiar del mismo o pretende indagar un poco más para ver qué características influyen en que una película se halle en una determinada posición o reciba una determinada puntuación. Es decir, se puede utilizar para responder a preguntas como:

¿En qué países se producen las mejores películas?

¿Qué generos reciben unas puntuaciones más elevadas?

¿Cómo se relaciona el número de votos y el de críticas con la puntuación obtenida?

¿Tiene algo que ver la longitud de la sinopsis con el número de votos o la puntuación que recibe una película?

¿Qué tags se encuentran más relacionados con cada género de película?

Es decir, el análisis de dicha base de datos pretende establecer relaciones entre las variables que pueden estar influyendo en la puntuación o posición de las películas. En conclusión, la principal pregunta a la que se pretende responder es

¿Están esas películas en esa posición porque de verdad son muy buenas o hay algún tipo de sesgo que las hace estar ahí?

Por último, además de las variables que se analizan en el presente estudio, se incluyen muchas otras para que aquellos que tengan curiosidad y algo de imaginación intenten buscar relaciones entre las mismas.

Licencia

Dataset

El DOI correspondiente a la base de datos es http://doi.org/10.5281/zenodo.4256706

Tabla de contribuciones al trabajo

Contribuciones	Firma
Investigación previa	J.P, A.H
Redacción de las respuestas	J.P, A.H
Desarrollo código	J.P, A.H