¿Qué libro me recomendarías?

Claudia Quintana Wong

8/11/2020

Introducción

El análisis de datos es un proceso que permite conocer y transformar datos en información valiosa con el fin de descubrir conocimiento y apoyar el desarrollo de toma de decisiones en las diferente aristas de la sociedad.

El objetivo de este proyecto es realizar un análisis exploratorio sobre un determinado conjunto de datos, de manera que sea posible extraer información relevante. El conjunto de datos utilizado en este proyecto fue tomado de Goodreads Book Datasets With User Rating 10M (https://www.kaggle.com/bahramjannesarr/goodreads-book-datasets-10m?select=book1-100k.csv). Goodreads se distingue por tener un sistema de puntuación que permite al usuario otorgar a cada libro una calificación de entre una y cinco estrellas por lo que es muy apropiado para la recomendación de información.

Exploración de los datos

En esta sección se describe el proceso de carga e inspección de los datos. Del conjunto original, debido a su gran tamaño, solo se ha utilizado el fichero book1-100k.csv. El almacenamiento de datos se efectúa en una estructura de tipo dataframe el cual facilita el análisis futuro. A continuación, se listan las diferentes variables presentes en los datos.

```
## 'data.frame':    58292 obs. of  18 variables:
##  $ Id             : int  1 2 3 4 5 6 8 9 10 12 ...
##  $ Name           : chr  "Harry Potter and the Half-Blood Prince (Harry Potter, #6)" "Harr
y Potter and the Order of the Phoenix (Harry Potter, #5)" "Harry Potter and the Sorcerer's St
one (Harry Potter, #1)" "Harry Potter and the Chamber of Secrets (Harry Potter, #2)" ...
##  $ RatingDist1    : chr  "1:9896" "1:12455" "1:108202" "1:11896" ...
##  $ pagesNumber    : int  652 870 309 352 435 734 2690 152 3342 815 ...
##  $ RatingDist4    : chr  "4:556485" "4:604283" "4:1513191" "4:706082" ...
##  $ RatingDistTotal: chr  "total:2298124" "total:2358637" "total:6587388" "total:2560657"
...
##  $ PublishMonth   : int  16 1 1 1 1 28 13 26 12 1 ...
##  $ PublishDay     : int  9 9 11 11 5 9 9 4 9 11 ...
##  $ Publisher      : chr  "Scholastic Inc." "Scholastic Inc." "Scholastic Inc" "Scholastic"
...
##  $ CountsOfReview : int  28062 29770 75911 244 37093 31978 166 1 809 255 ...
##  $ PublishYear    : int  2006 2004 2003 2003 2004 2002 2004 2005 2005 2005 ...
##  $ Language       : chr  "eng" "eng" "eng" "eng" ...
##  $ Authors        : chr  "J.K. Rowling" "J.K. Rowling" "J.K. Rowling" "J.K. Rowling" ...
##  $ Rating         : num  4.57 4.5 4.47 4.42 4.57 4.56 4.78 3.79 4.73 4.37 ...
##  $ RatingDist2    : chr  "2:25317" "2:37005" "2:130310" "2:49353" ...
##  $ RatingDist5    : chr  "5:1546466" "5:1493113" "5:4268227" "5:1504505" ...
##  $ ISBN           : chr  "" "0439358078" "" "0439554896" ...
##  $ RatingDist3    : chr  "3:159960" "3:211781" "3:567458" "3:288821" ...
```

El *dataset* está formado por 58 292 filas y 12 columnas. Cada columna constituye una propiedad de una observación del *dataset*, en este caso, cada observación corresponde a un libro. De los libros se almacena el **Id**, el **Name**, los detalles de su publicación expresados en las variables: **Author**, **pagesNumber**, **Publisher**,

Language, **ISBN**, **PublishDay**, **PublishMonth** y **PublishYear**, así como la evaluación total **RatingDistTotal** y específica **RatingDistn**, que representa la cantidad de *ratings* con valor *n = (1 .. 5)*. Asimismo, se almacena la cantidad de *reviews* en la variable **CountsOfReview** de cada libro en la colección.

Preparación y limpieza

Con el fin de facilitar el tratamiento y manipulación de los datos en pasos posteriores son necesarios los siguientes cambios:

- En las columnas de tipo **character** aparece la cadena vacía para representar que se desconoce el valor de esa propiedad para una observación (Ejemplo: columna **ISBN**, observación 1), para un mayor control se reemplaza la cadena vacía por el valor NA.

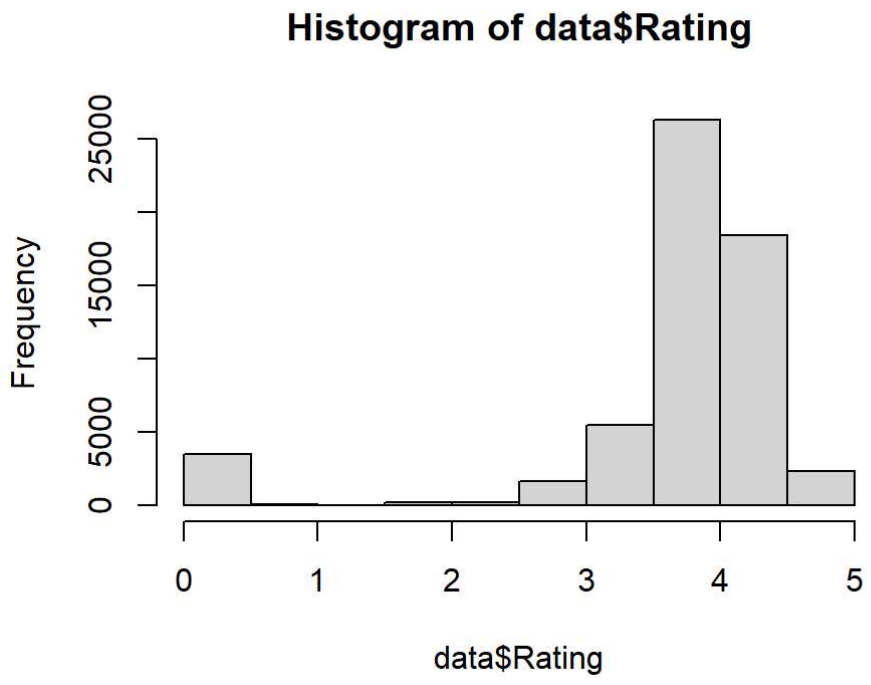
```
data[data==""]<-NA
```

- Las columnas **RatingDistn** de tipo **character** siguen el formato *n:m*, donde *m* representa la cantidad de evaluaciones con valor *n*. Asimismo, la columna **RatingDistTotal** con formato *total:m* representa el total de evaluaciones que ha recibido un libro. Con el fin de facilitar la comparación entre valores numéricos es apropiado contar solamente con el valor numérico *m*. A continuación se muestra cómo se lleva a cabo dicha transformación.

```
library(stringr)
data$RatingDist1 <- as.numeric(sub("\\d:", "", data$RatingDist1))
data$RatingDistTotal <- as.numeric(sub("(total):", "", data$RatingDistTotal))
```

Manipulación de datos

Una vez hechos los cambios necesarios, es posible pasar a la manipulación y descubrimiento de conocimiento a partir de los datos. A continuación, se expone un histograma que muestra la distribución de los *ratings* numéricos.



Es posible observar que la mayor parte de los valores de los *ratings* oscila entre los valores 3.5 y 4.5. Analizando el gráfico desde el valor 2 hasta el 5 se puede notar que la curva se asimila a una distribución normal, lo que significa que existe mayor probabilidad de que los usuarios den valores cercanos a la media. Asimismo, se puede constatar que es mayor la cantidad de personas que no emiten ninguna evaluación que los que evalúan con valores cercanos a 1 estrella.

A continuación se listan un conjunto de requerimientos informacionales y las respuestas sobre el conjunto de datos en cuestión. Nótese que en todos los casos se ha truncado el dataframe respuesta para una mejor visualización.

1. Determinar cuántos libros tienen la máxima evaluación

```
##   count
## 1  1015
```

De los 58 292 libros existentes en el *dataset* sólo 1 015, lo que representa el 1,7 %, tienen una calificación media de 5 estrellas.

2. Determinar los 10 libros que han logrado recoger el mayor número de reviews y *ratings* de los lectores.

```
##                                                    Name CountsOfReview
## 1                                 Twilight (Twilight, #1)          94850
## 2                                           The Book Thief          87685
## 3   Harry Potter and the Sorcerer's Stone (Harry Potter, #1)          75911
## 4                                The Giver (The Giver, #1)          57034
## 5                                      Water for Elephants          52918
## 6  The Lightning Thief (Percy Jackson and the Olympians, #1)          48630
## 7                                        Looking for Alaska          48042
## 8                                           Eat, Pray, Love          47852
## 9                                         The Glass Castle          46551
## 10                               The Catcher in the Rye          44046
##    RatingDistTotal Rating
## 1          4734773   3.59
## 2          1727186   4.37
## 3          6587388   4.47
## 4          1681531   4.13
## 5          1327834   4.09
## 6          1864960   4.25
## 7          1060299   4.04
## 8          1414495   3.56
## 9           855213   4.27
## 10         2610840   3.80
```

Los libros anteriores son los 10 que más han provocado que los usuarios expresen su opinión, tanto positiva como negativa.

3. Mostrar los 5 libros que mayor cantidad de ratings con valor 5 tienen y sus autores.

```
##                                                      Name      Authors
## 1                   Harry Potter and the Sorcerer's Stone J.K. Rowling
## 2                          Harry Potter agus an Ã“rchloch J.K. Rowling
## 3      Harri Potter a Maen yr Athronydd (Harry Potter, #1) J.K. Rowling
## 4                        Harry Potter e la Pietra Filosofale J.K. Rowling
## 5 Harry Potter und der Stein der Weisen (Harry Potter, #1) J.K. Rowling
##   RatingDist5
## 1     4292138
## 2     4277616
## 3     4276245
## 4     4276245
## 5     4276127
```

De la consulta anterior se puede llegar a la conclusión que J.K. Rowling es uno de los autores más populares entre los lectores.

4. Determinar la cantidad de libros publicados por editorial.

```
## # A tibble: 5 x 2
##   Publisher                     books_count
##   <chr>                               <int>
## 1 Vintage                               829
## 2 Oxford University Press, USA          723
## 3 Penguin Books                         692
## 4 Routledge                             604
## 5 Cambridge University Press            528
```

5. Listar la media de ratings obtenida por autor para aquellos autores que han publicado al menos 50 títulos.

```
## # A tibble: 5 x 3
##   Authors           mean count
##   <chr>            <dbl> <int>
## 1 J.K. Rowling      4.52    60
## 2 J.R.R. Tolkien    4.19    91
## 3 Rumiko Takahashi  4.17    83
## 4 C.S. Lewis        4.13    77
## 5 P.G. Wodehouse    4.13    77
```

6. Cantidad de libros publicados por año.

```
## # A tibble: 105 x 2
##    PublishYear count
##          <int> <int>
##  1        2006  7739
##  2        2005  6060
##  3        2004  4917
##  4        2003  4560
##  5        2002  3964
##  6        2001  3378
##  7        2000  2960
##  8        2007  2541
##  9        1999  2523
## 10        1998  2222
## # ... with 95 more rows
```

Conclusiones

En este proyecto se ha realizado un análisis que ha permitido a partir de un conjunto de libros evaluar la aceptación por parte del público de los diferentes títulos. Como resultado, es posible afirmar que uno de los libros que mayor éxito ha alcanzado desde su lanzamiento en 2006 es “**Harry Potter and the Sorcerer’s Stone**” y su autora J.K. Rowling, una de las más populares entre el público lector. Asimismo, destacan las editoriales inglesas y los libros escritos en inglés por encima del resto de los idiomas. Se puede determinar que entre los años 2000 y 2007 se han publicado la mayor cantidad de libros de acuerdo a los datos recogidos en el *dataset*. Más detalles sobre el proyecto pueden ser encontrados en https://github.com/claudiaqw/programacion-r (https://github.com/claudiaqw/programacion-r)