My own project submission
Book recommendation system

Viriya Thach
2022-11-22

Coding a recommendation system for Books

- Content-based recommendation system: this system is based on using the features of the books in order to offer similar products. For example, if I'm a huge fan of John Verdon, the system will recommend other books of the same author or books within the same category.

Loading data to build a book recommendation system

- In order to code a recommendation system in R, the first thing that we need is data. I will use the following book dataset from **Book-Crossing Dataset from the Institut für Informatik, Universität Freiburg.** Collected by Cai-Nicolas Ziegler in a 4-week crawl (August / September 2004) from the Book-Crossing community with kind permission from Ron Hornbaker, CTO of Humankind Systems. Contains 278,858 users (anonymized but with demographic information) providing 1,149,780 ratings (explicit / implicit) about 271,379 books.

```
url = "http://www2.informatik.uni-freiburg.de/~cziegler/BX/BX-CSV-Dump.zip"
download.file(url, destfile = "data.zip")
dir.create("data")
unzip("data.zip",exdir = "data")

files = paste0("data/",list.files("data"))

ratings = read.csv(files[1], sep = ";")
books = read.csv(files[2], sep = ";")
users = read.csv(files[3], sep = ";")

rm(files, url)
```

Understanding book data

First, we will see what type of data we have in each dataset because the type of analysis we should do will depend on the data that we have.

library(dplyr)

```
> str(books)
```

```
'data.frame':	115253 obs. of  8 variables:

 $ ISBN               : chr  "0195153448" "0002005018" "0060973129" "0374157065"
...

 $ Book.Title         : chr  "Classical Mythology" "Clara Callan" "Decision in
Normandy" "Flu: The Story of the Great Influenza Pandemic of 1918 and the Search
for the Virus That Caused It" ...

 $ Book.Author        : chr  "Mark P. O. Morford" "Richard Bruce Wright" "Carlo
D'Este" "Gina Bari Kolata" ...

 $ Year.Of.Publication: chr  "2002" "2001" "1991" "1999" ...

 $ Publisher          : chr  "Oxford University Press" "HarperFlamingo Canada"
"HarperPerennial" "Farrar Straus Giroux" ...

 $ Image.URL.S        : chr
"http://images.amazon.com/images/P/0195153448.01.THUMBZZZ.jpg"
"http://images.amazon.com/images/P/0002005018.01.THUMBZZZ.jpg"
"http://images.amazon.com/images/P/0060973129.01.THUMBZZZ.jpg"
"http://images.amazon.com/images/P/0374157065.01.THUMBZZZ.jpg" ...

 $ Image.URL.M        : chr
"http://images.amazon.com/images/P/0195153448.01.MZZZZZZZ.jpg"
"http://images.amazon.com/images/P/0002005018.01.MZZZZZZZ.jpg"
"http://images.amazon.com/images/P/0060973129.01.MZZZZZZZ.jpg"
"http://images.amazon.com/images/P/0374157065.01.MZZZZZZZ.jpg" ...

 $ Image.URL.L        : chr
"http://images.amazon.com/images/P/0195153448.01.LZZZZZZZ.jpg"
```

```
"http://images.amazon.com/images/P/0002005018.01.LZZZZZZZ.jpg"
"http://images.amazon.com/images/P/0060973129.01.LZZZZZZZ.jpg"
"http://images.amazon.com/images/P/0374157065.01.LZZZZZZZ.jpg" ...
```

```
> summary(books)
```

```
     ISBN            Book.Title        Book.Author        Year.Of.Publication
 Length:115253      Length:115253      Length:115253      Length:115253
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character
  Publisher         Image.URL.S        Image.URL.M        Image.URL.L
 Length:115253      Length:115253      Length:115253      Length:115253
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character
```

Rows: 115,253

Columns: 8

$ ISBN <fct> 0195153448, 0002005018, 0060973129, 0374157065,...

$ Book.Title <fct> Classical Mythology, Clara Callan, Decision in ...

$ Book.Author <fct> Mark P. O. Morford, Richard Bruce Wright, Carlo...

$ Year.Of.Publication <fct> 2002, 2001, 1991, 1999, 1999, 1991, 2000, 1993,...

$ Publisher <fct> Oxford University Press, HarperFlamingo Canada,...

$ Image.URL.S <fct> http://images.amazon.com/images/P/0195153448.01...

$ Image.URL.M <fct> http://images.amazon.com/images/P/0195153448.01...

$ Image.URL.L <fct> http://images.amazon.com/images/P/0195153448.01...

Categories

We have 4 variables referring to some features of the book (Title, Author, Year of Publication and Publisher). We will use these variable to code the content based recommendation system. In order to generate more realistic data, we will include a new variable called 'Category'. This variable will indicate if the book belongs to any of the following categories:

- Action and Adventure.
- Classic.
- Detective and Mystery.
- Fantasy.

```
set.seed(1234)

categories = c("Action and Adventure","Classic","Detective and Mystery","Fantasy")

books$category = sample( categories, nrow(books), replace=TRUE, prob=c(0.25, 0.3, 0.25, 0.20))

books$category = as.factor(books$category)
```

```
rm(categories)
```

We will add the caracters ‘Id’ to all the ISBNs and User-Ids because at some point we will construct matrixes data have ISBNs or User-Ids as column or row names. As they all begin with a number, R would include an X before the column/row name. Adding this ‘Id’ strings will avoid this to happen.

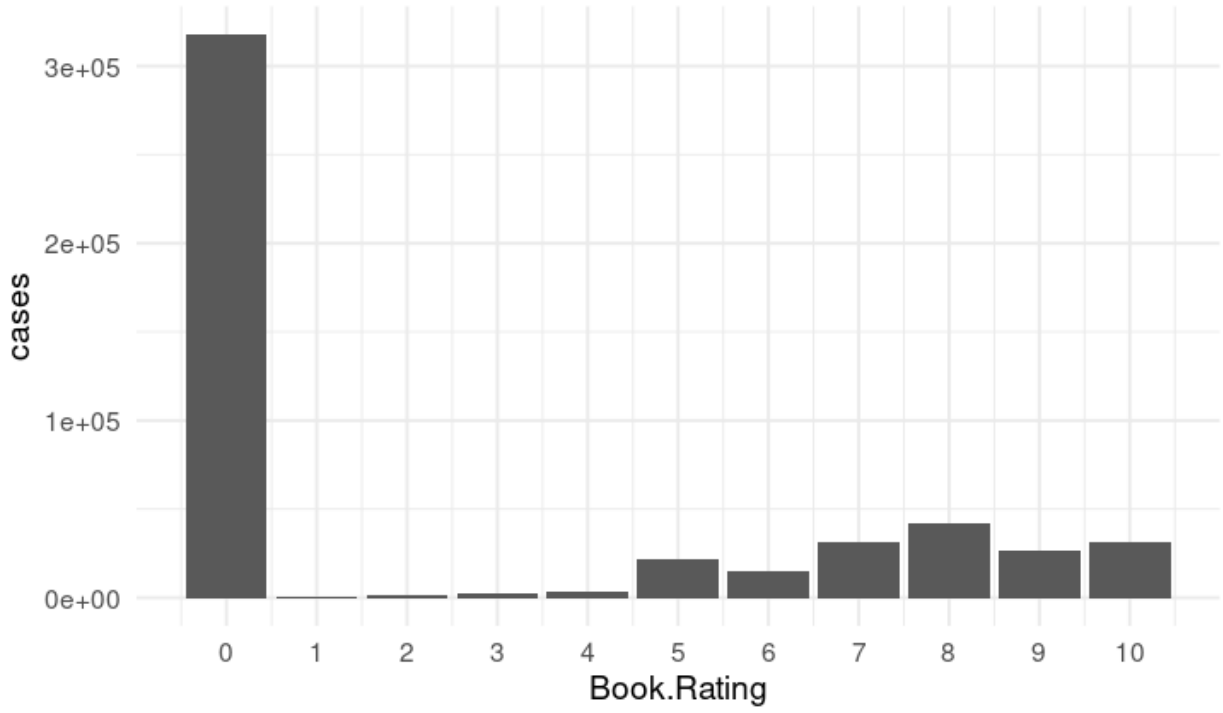
```
books$ISBN = paste0("Isbn.",books$ISBN)

users$User.ID = paste0("User.",users$User.ID)

ratings$ISBN = paste0("Isbn.",ratings$ISBN)

ratings$User.ID = paste0("User.",ratings$User.ID)
```

Rating data

We will see how the ratings of the books are distributed.

```
library(ggplot2)

ratings %>%

  group_by(Book.Rating) %>%

  summarize(cases = n()) %>%
```

```
  ggplot(aes(Book.Rating, cases)) + geom_col() +

  theme_minimal() + scale_x_continuous(breaks = 0:10)
```



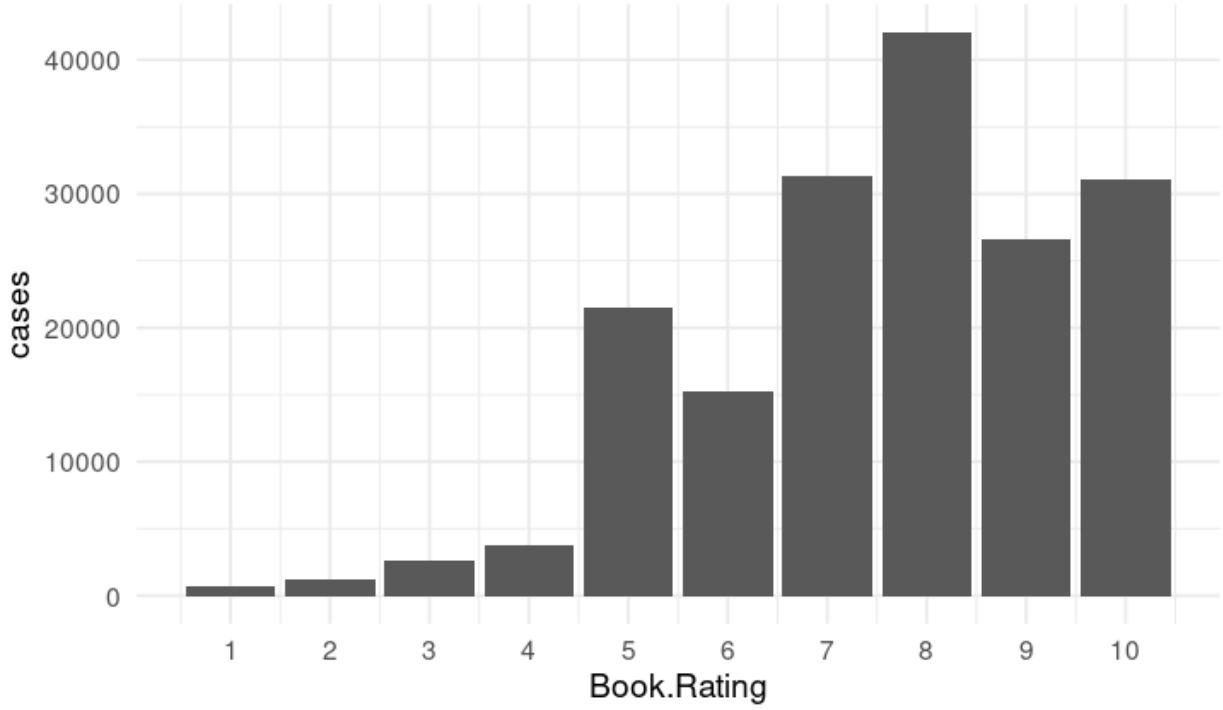
There are many zeros and it might indicate that a person has read the book but has not rate it. Therefore, we will just get with the recommendations that are non-zero.

```
ratings = ratings[ratings$Book.Rating!= 0, ]
```

How the ratings are distributed

We can redo the same graph and better see how the ratings are distributed:

```
ratings %>%
  group_by(Book.Rating) %>%
  summarize(cases = n()) %>%
  ggplot(aes(Book.Rating, cases)) + geom_col() +
  theme_minimal() + scale_x_continuous(breaks = 0:10)
```



How much each person scores

```
ratings_sum = ratings %>%
  group_by(User.ID) %>%
  count()

summary(ratings_sum$n)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    1.000    1.000    1.000    5.319    3.000 1906.000
```

Keeping only significant users

75% of users have given 3 recommendations or less. We are going to remove these people to keep only more significant users.

```
user_index = ratings_sum$User.ID[ratings_sum$n>4]

users = users[users$User.ID %in% user_index, ]
ratings = ratings[ratings$User.ID %in% user_index, ]
books = books[books$ISBN %in% ratings$ISBN,]

rm(ratings_sum, user_index)
```

Coding a content based recommendation system in R

Using Gower distance

```
library(cluster)

books_distance = books[,c("ISBN","Book.Author","Publisher")]

# Convert variables to factors
books_distance[,1] <- as.factor(books_distance[,1])
books_distance[,2] <- as.factor(books_distance[,2])
books_distance[,3] <- as.factor(books_distance[,3])

# Calculate Gower Distance
dissimilarity = daisy(books_distance, metric = "gower")


library(dplyr)

book_feature = books[1:10000,c("Book.Author","Publisher","category")]

# convert to factors
book_feature[,1] <- as.factor(book_feature[,1])
book_feature[,2] <- as.factor(book_feature[,2])
book_feature[,3] <- as.factor(book_feature[,3])

dissimilarity = daisy(book_feature, metric = "gower", weights = c(2,0.5,1))
```

```
dissimilarity = as.matrix(dissimilarity)

row.names(dissimilarity)<-  books$ISBN[1:10000]
colnames(dissimilarity)<- books$ISBN[1:10000]

Dissimilarity[15:20,15:20]
```

```
##                 Isbn.0971880107 Isbn.0345402871
Isbn.0345417623 Isbn.0684823802
## Isbn.0971880107       0.0000000       1.0000000
1.0000000       1.0000000
## Isbn.0345402871       1.0000000       0.0000000
0.5714286       1.0000000
## Isbn.0345417623       1.0000000       0.5714286
0.0000000       1.0000000
## Isbn.0684823802       1.0000000       1.0000000
1.0000000       0.0000000
## Isbn.0375759778       0.7142857       1.0000000
1.0000000       1.0000000
## Isbn.0375406328       1.0000000       1.0000000
1.0000000       0.7142857
##                 Isbn.0375759778 Isbn.0375406328
## Isbn.0971880107       0.7142857       1.0000000
## Isbn.0345402871       1.0000000       1.0000000
## Isbn.0345417623       1.0000000       1.0000000
## Isbn.0684823802       1.0000000       0.7142857
## Isbn.0375759778       0.0000000       1.0000000
## Isbn.0375406328       1.0000000       0.0000000
```

Conclusion

This recommendation systems recommends books that are:

- Written by the same author, publisher and category.
- Written by the same author and category.
- That they are only written by the same author.
- They are on the same category and are of the same publisher.
- That they are only of the same category.
- That they are only from the same publisher.