Movie-Recommendation.R

Carl Bebli

2021-05-04

```
#Introduction

#The goal of this project is to find out similarities within groups of people in order to build a movie

#GETTING AND CLEANING DATA
movies <- read.table("movielens.txt",header=FALSE,sep="|",quote = "\"")
str(movies)
```

```
## 'data.frame':    1682 obs. of  24 variables:
##  $ V1 : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ V2 : chr  "Toy Story (1995)" "GoldenEye (1995)" "Four Rooms (1995)" "Get Shorty (1995)" ...
##  $ V3 : chr  "01-Jan-1995" "01-Jan-1995" "01-Jan-1995" "01-Jan-1995" ...
##  $ V4 : logi  NA NA NA NA NA NA ...
##  $ V5 : chr  "http://us.imdb.com/M/title-exact?Toy%20Story%20(1995)" "http://us.imdb.com/M/title-exa
##  $ V6 : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ V7 : int  0 1 0 1 0 0 0 0 0 0 ...
##  $ V8 : int  0 1 0 0 0 0 0 0 0 0 ...
##  $ V9 : int  1 0 0 0 0 0 0 0 0 0 ...
##  $ V10: int  1 0 0 0 0 0 0 1 0 0 ...
##  $ V11: int  1 0 0 1 0 0 0 1 0 0 ...
##  $ V12: int  0 0 0 0 1 0 0 0 0 0 ...
##  $ V13: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ V14: int  0 0 0 1 1 1 1 1 1 1 ...
##  $ V15: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ V16: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ V17: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ V18: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ V19: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ V20: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ V21: int  0 0 0 0 0 0 1 0 0 0 ...
##  $ V22: int  0 1 1 0 1 0 0 0 0 0 ...
##  $ V23: int  0 0 0 0 0 0 0 0 0 1 ...
##  $ V24: int  0 0 0 0 0 0 0 0 0 0 ...
```

```
colnames(movies) <- c("ID","Title","ReleaseDate","VideoReleasedate","IMDB","Unknown","Action","Adventure
str(movies)
```

```
## 'data.frame':    1682 obs. of  24 variables:
##  $ ID              : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Title           : chr  "Toy Story (1995)" "GoldenEye (1995)" "Four Rooms (1995)" "Get Shorty (199
##  $ ReleaseDate     : chr  "01-Jan-1995" "01-Jan-1995" "01-Jan-1995" "01-Jan-1995" ...
```

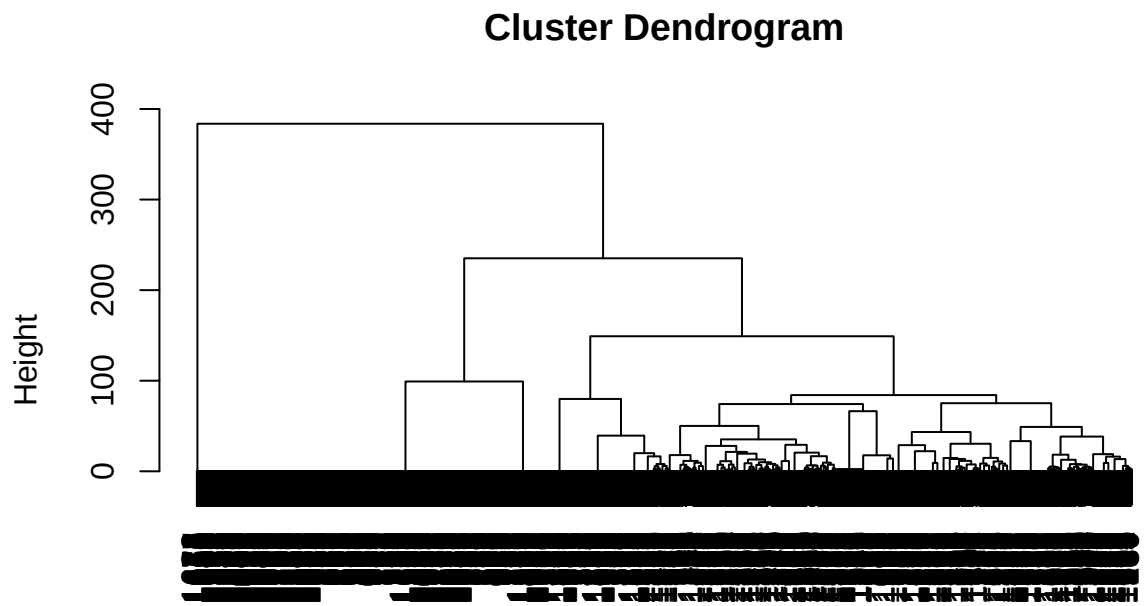
```
##  $ VideoReleasedate: logi  NA NA NA NA NA NA ...
##  $ IMDB            : chr  "http://us.imdb.com/M/title-exact?Toy%20Story%20(1995)" "http://us.imdb.co
##  $ Unknown         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Action          : int  0 1 0 1 0 0 0 0 0 0 ...
##  $ Adventure       : int  0 1 0 0 0 0 0 0 0 0 ...
##  $ Animation       : int  1 0 0 0 0 0 0 0 0 0 ...
##  $ Childrens       : int  1 0 0 0 0 0 0 1 0 0 ...
##  $ Comedy          : int  1 0 0 1 0 0 0 1 0 0 ...
##  $ Crime           : int  0 0 0 0 1 0 0 0 0 0 ...
##  $ Documentary     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Drama           : int  0 0 0 1 1 1 1 1 1 1 ...
##  $ Fantasy         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ FilmNoir        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Horror          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Musical         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Mystery         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Romance         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ SciFi           : int  0 0 0 0 0 0 1 0 0 0 ...
##  $ Thriller        : int  0 1 1 0 1 0 0 0 0 0 ...
##  $ War             : int  0 0 0 0 0 0 0 0 0 1 ...
##  $ Western         : int  0 0 0 0 0 0 0 0 0 0 ...
```

```
movies$ID=NULL
movies$ReleaseDate=NULL
movies$VideoReleasedate=NULL
movies$IMDB=NULL
movies <- unique(movies)
str(movies)
```

```
## 'data.frame':    1664 obs. of  20 variables:
##  $ Title      : chr  "Toy Story (1995)" "GoldenEye (1995)" "Four Rooms (1995)" "Get Shorty (1995)" .
##  $ Unknown    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Action     : int  0 1 0 1 0 0 0 0 0 0 ...
##  $ Adventure  : int  0 1 0 0 0 0 0 0 0 0 ...
##  $ Animation  : int  1 0 0 0 0 0 0 0 0 0 ...
##  $ Childrens  : int  1 0 0 0 0 0 0 1 0 0 ...
##  $ Comedy     : int  1 0 0 1 0 0 0 1 0 0 ...
##  $ Crime      : int  0 0 0 0 1 0 0 0 0 0 ...
##  $ Documentary: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Drama      : int  0 0 0 1 1 1 1 1 1 1 ...
##  $ Fantasy    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ FilmNoir   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Horror     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Musical    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Mystery    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Romance    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ SciFi      : int  0 0 0 0 0 0 1 0 0 0 ...
##  $ Thriller   : int  0 1 1 0 1 0 0 0 0 0 ...
##  $ War        : int  0 0 0 0 0 0 0 0 0 1 ...
##  $ Western    : int  0 0 0 0 0 0 0 0 0 0 ...
```

```
#HIERARCHICAL CLUSTERING
```

```r
distances <- dist(movies[2:20],method="euclidean")
clustermovies <-hclust(distances,method = "ward.D")
plot(clustermovies)
```




```r
clustergroups <- cutree(clustermovies,k = 10)

tapply(movies$Action, clustergroups, mean)
```

```
##         1         2         3         4         5         6         7         8
## 0.1784512 0.7839196 0.1238532 0.0000000 0.0000000 0.1015625 0.0000000 0.0000000
##         9        10
## 0.0000000 0.0000000
```

```r
tapply(movies$Animation, clustergroups, mean)
```

```
##          1          2          3          4          5          6          7
## 0.13468013 0.01005025 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
##          8          9         10
## 0.00000000 0.00000000 0.00000000
```

```r
# When the user selects a random movie we use clustering to predict movies the user is likely to enjoy
subset(movies, Title == "G.I. Jane (1997)")
```

```
##                Title Unknown Action Adventure Animation Childrens Comedy Crime
## 326 G.I. Jane (1997)       0      1         0         0         0      0     0
##     Documentary Drama Fantasy FilmNoir Horror Musical Mystery Romance SciFi
## 326           0     1       0        0      0       0       0       0     0
##     Thriller War Western
## 326        0   1       0
```

```
clustergroups[326]
```

```
## 328
##   3
```

```
cluster2 <- subset(movies, clustergroups == 3)
cluster2$Title[1:20]
```

```
##  [1] "Four Rooms (1995)"                "Copycat (1995)"
##  [3] "Seven (Se7en) (1995)"             "Usual Suspects, The (1995)"
##  [5] "From Dusk Till Dawn (1996)"       "Taxi Driver (1976)"
##  [7] "Rumble in the Bronx (1995)"       "Batman Forever (1995)"
##  [9] "Strange Days (1995)"              "Disclosure (1994)"
## [11] "Dolores Claiborne (1994)"         "Pulp Fiction (1994)"
## [13] "Carlito's Way (1993)"             "Firm, The (1993)"
## [15] "Blade Runner (1982)"              "Silence of the Lambs, The (1991)"
## [17] "Fargo (1996)"                     "Diabolique (1996)"
## [19] "Lone Star (1996)"                 "Godfather, The (1972)"
```

```
#Based on the analysis, the user is likely to enjoy the following movies
```