Lastfm_Clustering

Pranay Singla

1/6/2021

The aim of this code file is to cluster users from lastfm dataset into groups based on listening preferences. This analysis can be used to identify users with similar listerning preferences as well as their artist/genre preferences.

```
# loading libraries
library(forcats)
```

```
## Warning: package 'forcats' was built under R version 3.6.3
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.6.3

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
# reading data
lastfm = read.csv("lastfm1.csv")
```

Most popular artist

```
which.max(colSums(lastfm[,-1]))
```

```
## the.beatles
##         474
```

Artist most correlated with Beatles

```
beatles_index = grep('the.beatles', colnames(lastfm))
temp.cors = cor(lastfm$the.beatles, lastfm[,-c(1,beatles_index)])

# See which artist this is:
names(lastfm[,-c(1,465)])[which.max(temp.cors)]
```

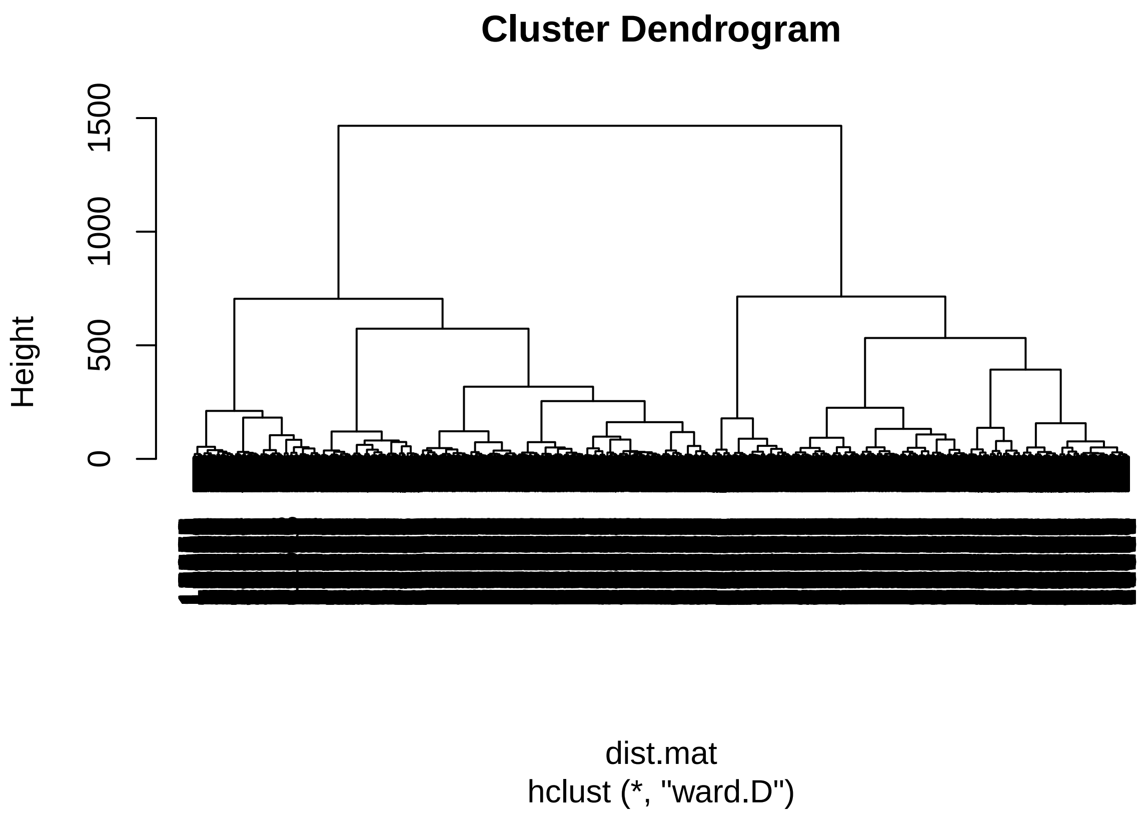
```
## [1] "bob.dylan"
```

Running hierarchical clustering to group users

```
dist.mat = dist(lastfm[,-1], method = "euclidean")
lastfm.clust = hclust(dist.mat, method = "ward.D")
```

Plotting the dendogram

```
# Dendrogram
plot(lastfm.clust)
```



Creating 5 clusters based on dendogram and listing number of users in each cluster

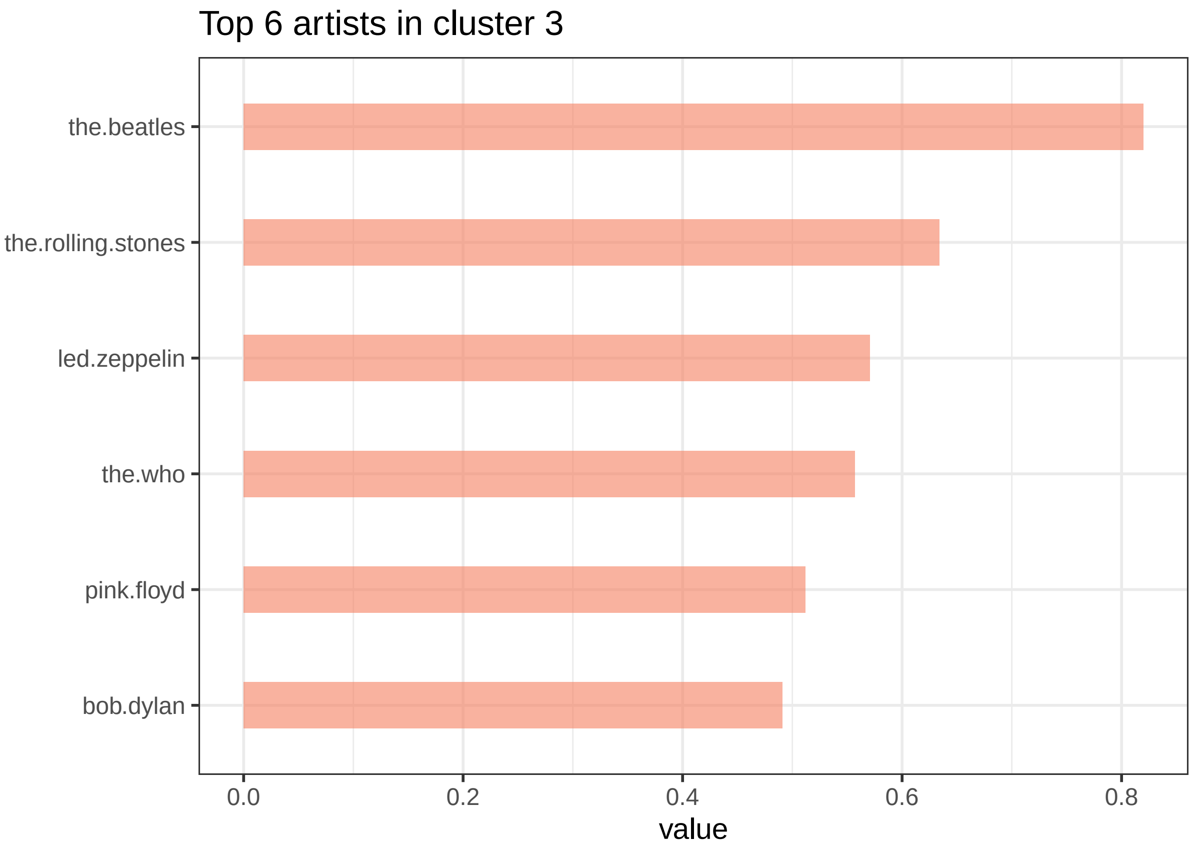
```
# Create the clusters:
clust.groups = cutree(lastfm.clust, k = 5)
table(clust.groups)
```

```
## clust.groups
##    1    2    3    4    5
## 6719 5787 1971 1599 2544
```

Number of users in each cluster:
1 -> 3798
2 -> 3262
3 -> 6037
4 -> 1973
5 -> 1089

Finding top 6 artists in different clusters

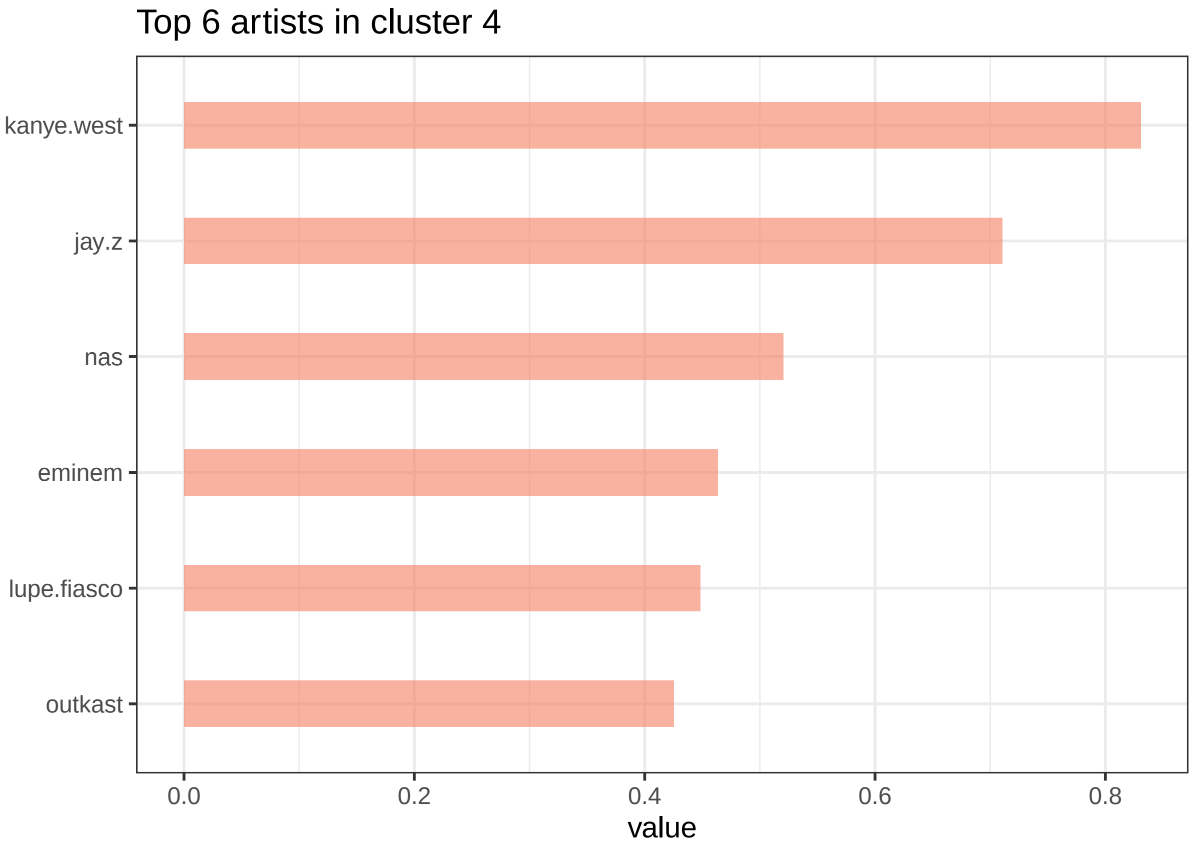
```r
# Display top 6 artists in cluster 3:
top_6 <- tail(sort( colMeans(lastfm[clust.groups == 3,-1]), decreasing = F ), 6)
top_6 <- data.frame(artist = names(top_6), value = top_6, row.names = NULL)
top_6 <- top_6[order(-top_6$value),]
top_6 %>%
  mutate(artist = fct_reorder(artist, value)) %>%
  ggplot( aes(x=artist, y=value)) +
    ggtitle('Top 6 artists in cluster 3') +
    geom_bar(stat="identity", fill="#f68060", alpha=.6, width=.4) +
    coord_flip() +
    xlab("") +
    theme_bw()
```



Users in cluster 3 prefer to listen to classic rock artists

```r
# Display top 6 artists in cluster 4:

top_6 <- tail(sort( colMeans(lastfm[clust.groups == 4,-1]), decreasing = F ), 6)
top_6 <- data.frame(artist = names(top_6), value = top_6, row.names = NULL)
top_6 <- top_6[order(-top_6$value),]
top_6 %>%
  mutate(artist = fct_reorder(artist, value)) %>%
  ggplot( aes(x=artist, y=value)) +
    ggtitle('Top 6 artists in cluster 4') +
    geom_bar(stat="identity", fill="#f68060", alpha=.6, width=.4) +
    coord_flip() +
    xlab("") +
    theme_bw()
```



Users in cluster 4 prefer to listen to rap artists