Homework 4

Karan Ashar

3/9/2021

library(cluster)  
library(factoextra)

## Warning: package 'factoextra' was built under R version 4.0.4

## Loading required package: ggplot2

## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa

library(gridExtra)  
library(dendextend)

##   
## ---------------------  
## Welcome to dendextend version 1.14.0  
## Type citation('dendextend') for how to cite the package.  
##   
## Type browseVignettes(package = 'dendextend') for the package vignette.  
## The github page is: https://github.com/talgalili/dendextend/  
##   
## Suggestions and bug-reports can be submitted at: https://github.com/talgalili/dendextend/issues  
## Or contact: <tal.galili@gmail.com>  
##   
## To suppress this message use: suppressPackageStartupMessages(library(dendextend))  
## ---------------------

##   
## Attaching package: 'dendextend'

## The following object is masked from 'package:stats':  
##   
## cutree

## Data Preprocessing

An input dataframe was created, this represented the disputed papers. The data was scaled as well. This is important because this can affect our algorithms performance. For the first two tasks i.e to cluster the disputed papers using HAC and K-Means I removed the rows with author ‘Jay’ and ‘HM’. They were not required for our task.

setwd("C:/Users/karan/Desktop/IST707/data")  
data<-read.csv(file ='HW4-data-fedPapers85.csv')  
  
input=data[data$author=='dispt',]  
  
temp=data[data$author!= 'Jay',]  
temp=temp[temp$author!='dispt',]  
filtered\_data=temp[temp$author!='HM',]  
  
nrow(filtered\_data)

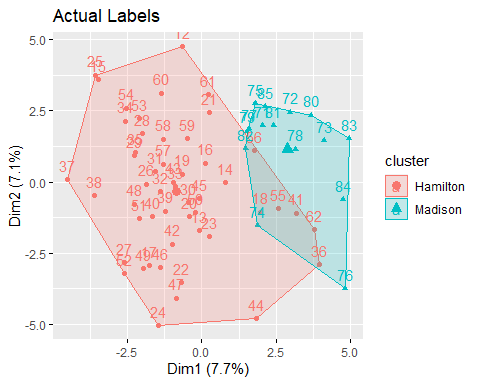
## [1] 66

#str(filtered\_data)  
  
#to\_test  
actual\_labels<-filtered\_data$author  
#actual\_labels  
filtered\_data$author<-NULL  
filtered\_data$filename<-NULL  
  
#Preparing input data   
input\_file\_names=input$filename  
input$author<-NULL  
input$filename<-NULL  
  
  
to\_test<-rbind(input,filtered\_data)  
to\_test<-data.frame(scale(to\_test))  
filtered\_data<-to\_test[-c(1:11),]

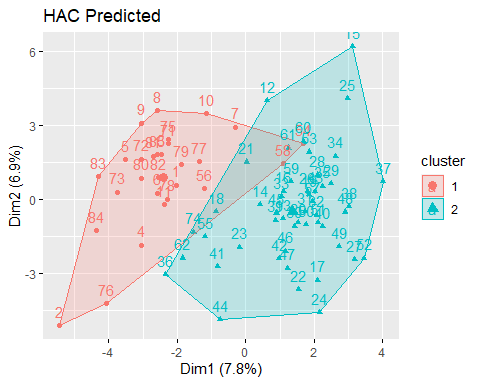
## HAC

First thing I did was to carry out hierarchical clustering. To find our answers to the disputed papers. I tried a few algorithms and various methods. I found this one to be the best. I set k (‘Number of clusters’) = 2. This is because we had two choices Madison or Hamilton. Below we see that all are disputed papers which have a label of 1 to 11 belong to Madison’s cluster. We can also say this because the cluster with lesser number of observations belongs to Madison and the disputed papers are being clustered with the cluster with lesser number of observations.

#HAC   
hc3 <- agnes(to\_test, method = "ward")  
cluster\_label <- cutree(hc3, k = 2)  
  
  
fviz\_cluster(list(data = filtered\_data, cluster = actual\_labels))+ggtitle('Actual Labels')



fviz\_cluster(list(data = to\_test, cluster = cluster\_label))+ ggtitle('HAC Predicted')



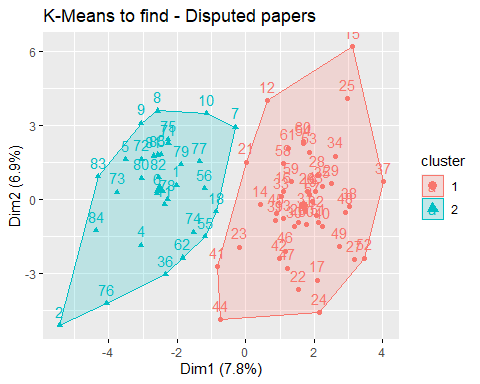
table(head(cluster\_label,11))

##   
## 1   
## 11

## K-Means

Next I tried the K-means algorithm with two clusters as well. I got the same results as before. The first 11 rows i.e the disputed ones belong to Madison. We can view the same using the plot and the label numbers.

#K-means  
  
model\_r = kmeans(to\_test, centers = 2, nstart = 25)  
  
fviz\_cluster(model\_r, data = to\_test) + ggtitle('K-Means to find - Disputed papers')



head(model\_r$cluster,11)

## 1 2 3 4 5 6 7 8 9 10 11   
## 2 2 2 2 2 2 2 2 2 2 2

## Identify important attributes

Below we can see the important words helping us to cluster.

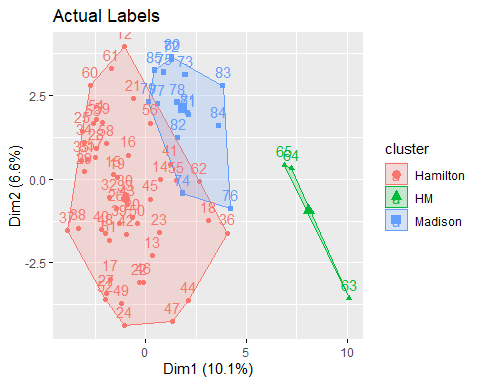
df<-data.frame(cbind(model\_r$centers[1,],model\_r$centers[2,]))  
colnames(df)<-c('one','two')  
df$abs\_distance<-abs(df$one-df$two)  
rownames(head(df[order(-df$abs\_distance),],10))

## [1] "upon" "to" "on" "there" "in." "by" "and" "an" "would"  
## [10] "any"

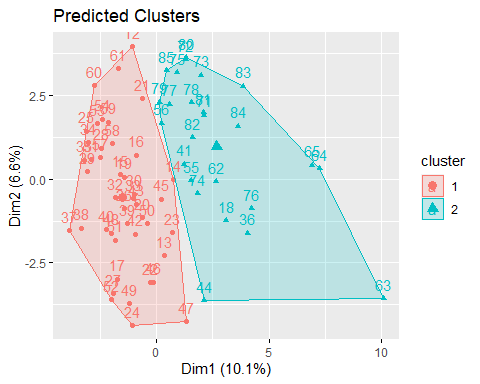
## Which side to the joint authorship papers go?

For this question I kept only the rows containing Madison, HM and Hamilton and made 2 clusters. This would help us identify which way did the joint authorship papers go. First I plotted a visualization on where the 3 were located before clustering. Then I carried out k-means clustering after removing the ‘author’ variable. Note : The HM papers had a label of ‘63’,‘64’,‘65’. We notice that the HM papers got clustered with the Maddison’s papers.

temp=data[data$author!='Jay',]  
temp=temp[temp$author!='dispt',]  
hm=temp[temp$author=='HM',]  
temp$filename<-NULL  
actual\_lables\_part3<-temp$author  
temp$author<-NULL  
scaled\_temp<-data.frame(scale(temp))  
model\_r = kmeans(scaled\_temp, centers = 2, nstart = 25)  
  
  
fviz\_cluster(list(data = temp, cluster = actual\_lables\_part3)) +ggtitle('Actual Labels')



fviz\_cluster(model\_r, data = scaled\_temp) + ggtitle('Predicted Clusters')



#Labels representing HM.  
rownames(hm)

## [1] "63" "64" "65"

## Conclusion

I can conclude based on my experiments that all the ‘disputed’ belonged to Madison and not Hamilton. Also we can see a strong input of Madison in the joint authorship papers as well.