Homework4

Noah Mott

2/14/2021

The purpose of this assignment was to solve a mystery. The Federalist Papers, written so very long ago have quite a bit of intrigue behind them because several of the papers were published anonymously. James Madison and Alexander Hamilton, both contributors to the Federalist Papers, had each claimed to be the authors of the unclaimed papers. My job was to use clustering analysis to determine which, if any, of the two was the actual author.

**STEP1**

##Loading the necessary libraries etc.

#install.packages('data.table')  
library(wordcloud2)

library(wordcloud)

library(slam)

library(quanteda)

library(SnowballC)

library(arules)

library(proxy)

library(cluster)

library(stringi)  
library(Matrix)  
library(tidytext)

library(plyr)

library(ggplot2)

library(factoextra)

library(mclust)

library(dplyr)

library(tm)

library(ggplot2)  
library(dplyr)  
library(gridExtra)

library(ggforce)

library(grid)  
library(data.table)

**STEP2**

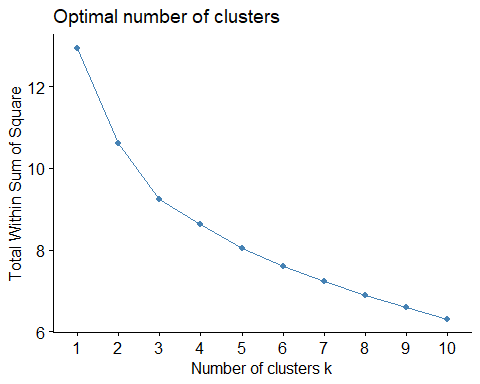
papers<-read.csv('C:/Users/noahm/OneDrive/Desktop/fedPapers85.csv')  
paperscopy<-papers

#Format Data for use in kmeans clustering

paperskmeans<-paperscopy[,2:72]  
rownames(paperskmeans)<-paperskmeans[,1]  
paperskmeans[,1]<-NULL

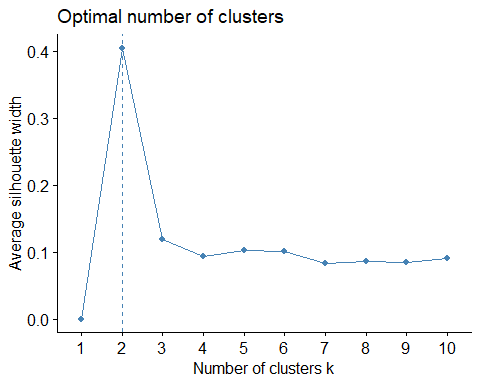
#Conduct initial analysis of the data to determine ideal number of clusters

fviz\_nbclust(papers, FUN=hcut, method = 'wss')



fviz\_nbclust(papers, FUN=hcut, method='silhouette')

**STEP3**

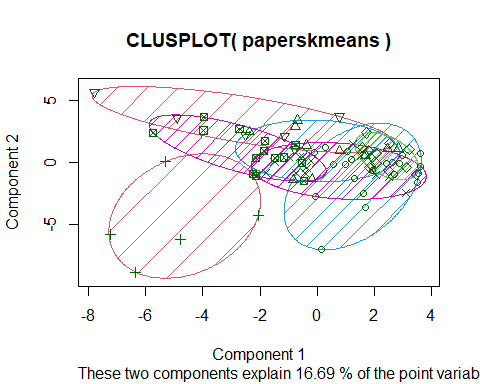


The initial value fell short of the necessary outcome. While a K of 2 might be most optimal, It wouldn’t have addressed our problem directly. After some tweaking of the parameters, I decided on 7 clusters for future use based on assessment of the data.

clusters<-kmeans(paperskmeans, 7)  
paperskmeans$Clusters<-as.factor(clusters$cluster)

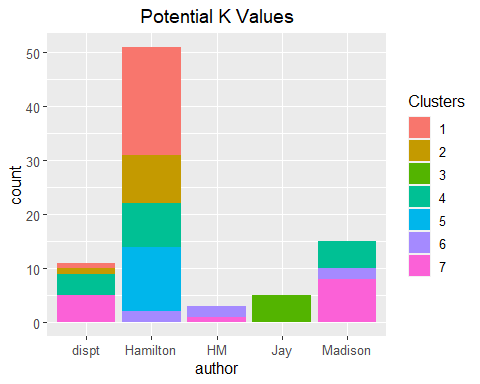
Conducted additional visual exploration of the Data

papers3<-papers  
papers3$Clusters<-as.factor(clusters$cluster)  
clusplot(paperskmeans, paperskmeans$Clusters, color = T, shade = T, labels = 0, lines = 0)



#Additional visual exploration made a K of 7 seem like the ideal choice.

library(ggplot2)  
ggplot(data=papers3, aes(x=author, fill=Clusters))+geom\_bar(stat='count')+labs(title="Potential K Values")+theme(plot.title=element\_text(hjust = 0.5), text=element\_text(size=12))



Conducted additional cleaning of the data for kmeans evaluation

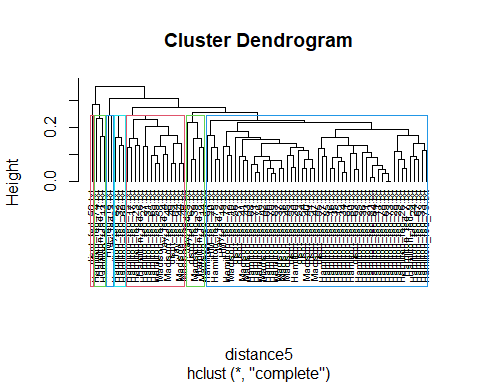
FedPapers\_HAC <- papers[,c(2:72)]  
rownames(FedPapers\_HAC) <- FedPapers\_HAC[,1]  
FedPapers\_HAC[,1] <- NULL

#Developed various distance formulas between variables using several different methods

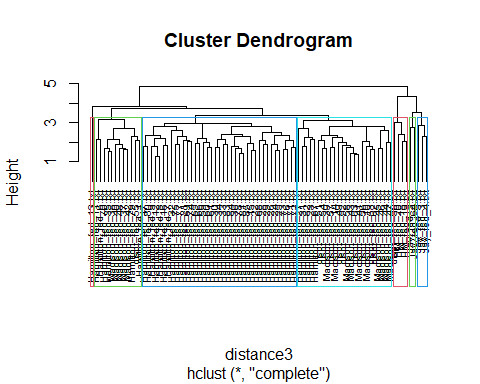
distance <- dist(FedPapers\_HAC, method = "euclidean")  
distance2 <- dist(FedPapers\_HAC, method = "maximum")  
distance3 <- dist(FedPapers\_HAC, method = "manhattan")  
distance4 <- dist(FedPapers\_HAC, method = "canberra")  
distance5 <- dist(FedPapers\_HAC, method = "binary")

#Developed Hierarchical Dendograms for each distance method before settling on manhattan measurement

HAC <- hclust(distance5, method="complete")  
plot(HAC, cex=0.6, hang=-1)  
rect.hclust(HAC, k =7, border=2:5)

 #Manhattan measurement with a k of 7 provided the most distinct order to the various papers and a patern began to emerge. You’ll note that the disputed papers largely fall in with Madison’s work within the Hierarchy. Seen below.

HAC2 <- hclust(distance3, method="complete")  
plot(HAC2, cex=0.6, hang=-1)  
rect.hclust(HAC2, k =7, border=2:5)



I conducted additional cleaning of the data and separated subsets of the data for visual comparison of where they fell within the clusters.

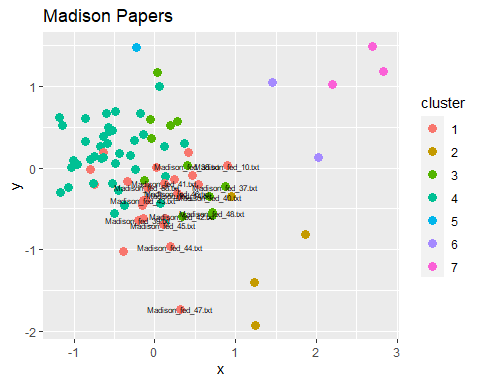
hc <- hclust(distance3)  
cluster <- cutree(hc, k=7)  
xy<-data.frame(cmdscale(distance3), factor(cluster))  
names(xy) <- c("x", "y", "cluster")  
xy$model <- rownames(xy)  
  
#note the split of the data for comparison purposes.  
m<-xy[xy$model %like% 'Madison',]  
peac<-xy[xy$model %like% 'dispt',]  
hamton<-xy[xy$model %like% 'Ham',]  
joint<-xy[xy$model %like% 'HM',]  
  
n<-as.data.frame(bind\_rows(m,peac))  
n<-as.data.frame(bind\_rows(n,hamton))

**STEP 4**

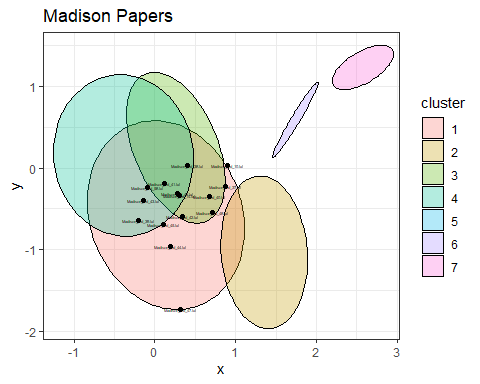
The following charts depict the location of each of the papers produced by Madison, Hamilton, and our mystery writer with relation to one another corresponding with Manhattan Distance with a K of 7.

The first is Madison’s known works:

ggplot(xy, aes(x, y)) + geom\_point(aes(colour=cluster), size=3)+geom\_text(data=m ,aes(label=model), size=2)+ggtitle('Madison Papers')

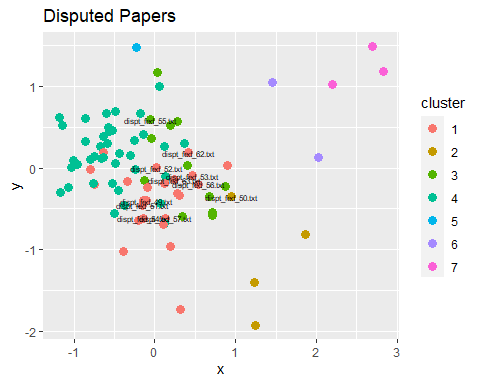


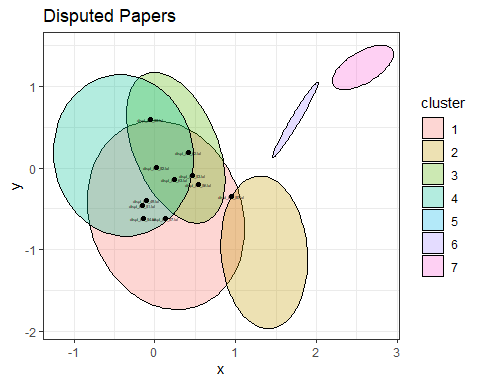
ggplot(xy, aes(x, y)) + geom\_mark\_ellipse(expand=0,aes(fill=cluster))+geom\_point(data=m)+theme\_bw()+geom\_text(data=m,aes(label=model), size=1)+ggtitle('Madison Papers')



The second is our mystery writer

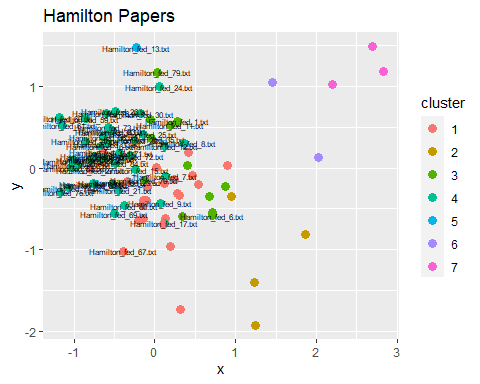
ggplot(xy, aes(x, y)) + geom\_point(aes(colour=cluster), size=3)+geom\_text(data=peac ,aes(label=model), size=2)+ggtitle('Disputed Papers')

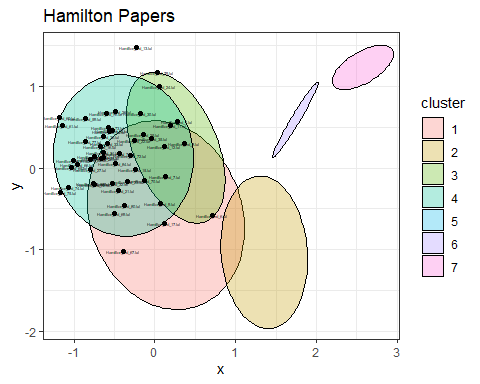
ggplot(xy, aes(x, y)) + geom\_mark\_ellipse(expand=0,aes(fill=cluster))+geom\_point(data=peac)+theme\_bw()+geom\_text(data=peac,aes(label=model), size=1)+ggtitle('Disputed Papers')



The third is Hamilton’s known works

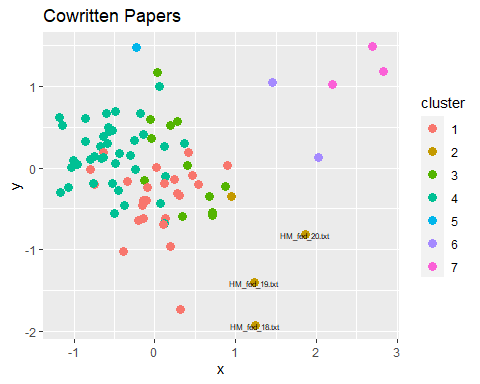
ggplot(xy, aes(x, y)) + geom\_point(aes(colour=cluster), size=3)+geom\_text(data=hamton ,aes(label=model), size=2)+ggtitle('Hamilton Papers')

ggplot(xy, aes(x, y)) + geom\_mark\_ellipse(expand=0,aes(fill=cluster))+geom\_point(data=hamton)+theme\_bw()+geom\_text(data=hamton,aes(label=model), size=1)+ggtitle('Hamilton Papers')



I also plotted the location of papers coauthored by both Hamilton and Madison to demonstrate the clear difference such a paper has with relation to the others.

ggplot(xy, aes(x, y)) + geom\_point(aes(colour=cluster), size=3)+geom\_text(data=joint ,aes(label=model), size=2)+ggtitle('Cowritten Papers')



What I was able to gleen from the various clustering that was conducted on the Federalist Papers was that it is highly likely that the mystery writer for the papers with the unknown author are attributable to Madison. You will note in the plots that while the plots overlap, the majority of Hamilton’s work resides in the upper left of the plot while Madison’s and the mystery author’s works are closer together and occupy more of the middle of the plot. This holds true, not only for the cluster plot but also for the hierarchical dendogram demonstrating a clustering of Madison and the mystery author’s works.