# Homework 4

By Shivani S Mahaddalkar

# The Mystery of The Disputed Essays:

## Who wrote them, Hamilton or Madison?

### Introduction:

[The Federalist Papers](http://thomas.loc.gov/home/histdox/fedpapers.html) were a series of eighty-five essays urging the citizens of New York to ratify the new United States Constitution. Written by Alexander Hamilton, James Madison, and John Jay, the essays originally appeared anonymously in New York newspapers in 1787 and 1788 under the pen name "Publius." A bound edition of the essays was first published in 1788, but it was not until the 1818 edition published by the printer Jacob Gideon that the authors of each essay were identified by name. The Federalist Papers are considered one of the most important sources for interpreting and understanding the original intent of the Constitution. (Federalist Papers: Primary Documents in American History, n.d.)

### Objective:

Out of the total 85 essays, 74 essays were identified with authors: 51 essays written by Hamilton, 15 by Madison, 3 by Hamilton and Madison, 5 by Jay. Hamilton claimed the authorship for the remaining 11 essays before he was killed in a duel. Later Madison also claimed authorship of the remaining 11 essays. This analysis aims to cluster similar essays together and provide a solution to the dispute of the authorship.

### About the Data:

### The data has 85 entries with 72 attributes. The first attribute is the author’s name. In case of the disputed essays, the record says *‘dispt’;* and in the case of joint authorship by Hamilton and Madison, the record says *‘HM’.* The second attribute states the name of the document. The remaining 70 attributes contains frequently occurring words in the essays. In 1960s, statistician Mosteller and Wallace analyzed the frequency distributions of common function words in the Federalist Papers, and drew their conclusions. <<http://www.stat.cmu.edu/~vlachos/courses/724/final/mosteller.pdf>>

### The records in these columns is the percentage of the frequency of these words in the essay.

The unsupervised algorithm of k means clustering and hierarchical agglomerative clustering was applied to assess whether the disputed essays were closer to the essays authored by Hamilton or Madison.

### Technical Analysis and Observations:

#Calling the appropriate packages and Loading the data frame  
df <- read.csv("HW4-data-fedPapers85.csv")  
head(df)

The data frame shows that there are 85 entries with 72 attributes. The first attribute is the author’s name. In case of the disputed essays, the record says ‘dispt’; and in the case of joint authorship by Hamilton and Madison, the record says ‘HM’. The second attribute states the name of the document. The remaining 70 attributes contains frequently occurring words in the essays. The records in these columns is the percentage of the frequency of these words in the essay.

#Dropping all the rows who have Jay as their author.  
df1 <- df[df$author != "Jay",]

Since the disputed authorship is between Hamilton or Madison, we take out the entries with Jay as the author.

As we do not want the clustering algorithm to depend to an arbitrary variable unit, we start by scaling/standardizing the data using the R function scale:

df1[,3:72] <- scale(df1[,3:72])

#Checking how many papers each author has published  
table(df1$author)

##   
## dispt Hamilton HM Madison   
## 11 51 3 15

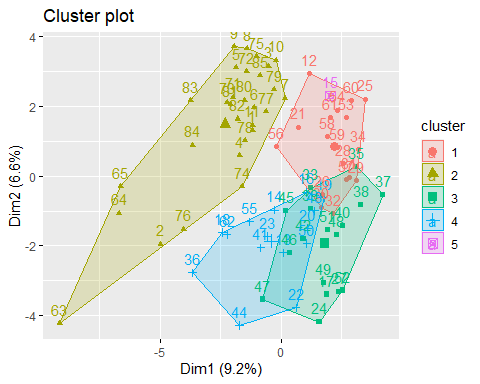
Starting with 5 different clusters, we conduct k means clustering:

set.seed(123)  
model1 <- kmeans(df1[,3:72], centers = 5)  
(table(df1$author,model1$cluster))

##   
## 1 2 3 4 5  
## dispt 0 11 0 0 0  
## Hamilton 18 0 17 15 1  
## HM 0 3 0 0 0  
## Madison 0 15 0 0 0

In this clustering it can be seen that the papers authored by Madison and the disputed papers along with the co-authored papers lie in a single cluster. We visualize this clustering to see how the clusters overlap. However, from this single analysis, we can make a claim that the similarity of the disputed papers is very high to that of the papers by Madison.

fviz\_cluster(model1, data = df1[,3:72])



The visualization shows that cluster two does not overlap with the other clusters. This cluster contains the essays by Madison, co-authored papers and the disputed papers. The other 4 clusters namely 1,3,4 and 5 overlap with each other(which gives us confidence that our analysis is moving in the correct direction). As the essays are authored by Hamilton, it was expected that these clusters overlap.

Now, we need to check how many clusters are optimal. We model 5 different models with different numbers of clusters.

#defining different models for 3,4,5,6 and 7 number of clusters  
set.seed(123)  
k3 <- kmeans(df1[,3:72], centers = 3)  
set.seed(123)  
k4 <- kmeans(df1[,3:72], centers = 4)  
set.seed(123)  
k5 <- kmeans(df1[,3:72], centers = 5)  
set.seed(123)  
k6 <- kmeans(df1[,3:72], centers = 6)  
set.seed(123)  
k7 <- kmeans(df1[,3:72], centers = 7)

(table(df1$author,k3$cluster))

##   
## 1 2 3  
## dispt 0 10 1  
## Hamilton 21 0 30  
## HM 0 3 0  
## Madison 0 15 0

(table(df1$author,k4$cluster))

##   
## 1 2 3 4  
## dispt 0 10 1 0  
## Hamilton 17 0 18 16  
## HM 0 3 0 0  
## Madison 0 15 0 0

(table(df1$author,k5$cluster))

##   
## 1 2 3 4 5  
## dispt 0 11 0 0 0  
## Hamilton 18 0 17 15 1  
## HM 0 3 0 0 0  
## Madison 0 15 0 0 0

(table(df1$author,k6$cluster))

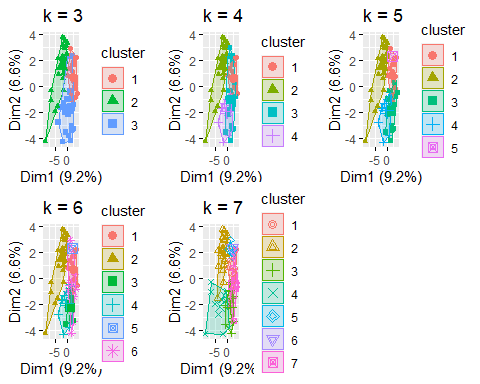
##   
## 1 2 3 4 5 6  
## dispt 0 10 0 0 0 1  
## Hamilton 16 0 8 15 1 11  
## HM 0 3 0 0 0 0  
## Madison 0 15 0 0 0 0

(table(df1$author,k7$cluster))

##   
## 1 2 3 4 5 6 7  
## dispt 0 9 0 1 1 0 0  
## Hamilton 10 0 13 9 3 1 15  
## HM 0 0 0 3 0 0 0  
## Madison 0 15 0 0 0 0 0

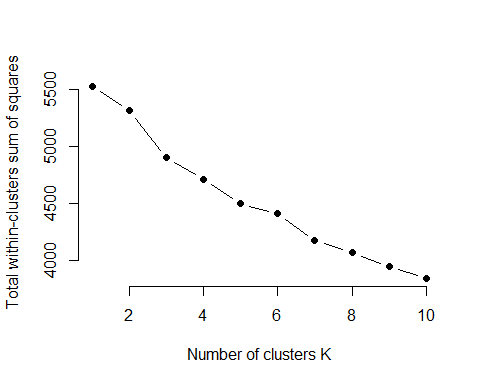
Total number of clusters 5 give us results where all the disputed essays are put together in the same cluster. In the other clustering models, it can be seen that atleast 9 essays are put into the same cluster as that of Madison’s.

# plots to compare  
p1 <- fviz\_cluster(k3, geom = "point", data = df1[,3:72]) + ggtitle("k = 3")  
p2 <- fviz\_cluster(k4, geom = "point", data = df1[,3:72]) + ggtitle("k = 4")  
p3 <- fviz\_cluster(k5, geom = "point", data = df1[,3:72]) + ggtitle("k = 5")  
p5 <- fviz\_cluster(k6, geom = "point", data = df1[,3:72]) + ggtitle("k = 6")  
p4 <- fviz\_cluster(k7, geom = "point", data = df1[,3:72]) + ggtitle("k = 7")  
library(gridExtra)  
grid.arrange(p1, p2, p3, p5, p4, nrow = 2)



The comparison shows that with 5 clusters, the cluster with Madison as the author does not overlap with the others and has some distance between the boundaries as well.

# set up a list that records cluster size from 1 to 10   
k.values <- 1:10  
  
# set up an empty list, which will be used to record within-cluster SSE for each k  
sse.values <- vector(mode = "list", length = length(k.values))  
  
# compute total within-cluster sum of square for each possible k  
for (i in k.values) {  
 sse.values[i] = kmeans(df1[,3:72], i)$tot.withinss  
}  
  
# plot the relationship between k and sse  
plot(k.values, sse.values,  
 type="b", pch = 19, frame = FALSE,   
 xlab="Number of clusters K",  
 ylab="Total within-clusters sum of squares")



From this plot and the observations we have been seeing, we can say that 5 is the optimal number of clusters.

#Hierarchical Agglomerative Clustering Using the agnes function, with the ward method we visualise the clusters of the essays with a dendogram

hc <- agnes(df1[,3:72], method = "ward")  
pltree(hc, cex = 0.6, hang = -1, main = "Dendrogram of agnes")  
rect.hclust(hc, k=5)

Chart, diagram

Description automatically generated

Essays numbered 1-11 are disputed essays, 12-62 are Hamilton’s, 63-65 are HM, and 66-85 are Madison.On visual inspection of the dendogram, we can see that the essays from 12-62 are classified into their own clusters of 3, whereas the other three categories are spread within 2 other clusters. Therefore, a conclusion can be drawn that the disputed essays were written by Madison.

### Conclusion:

From k means clustering and hierarchical agglomerative clustering we can deduce that the disputed papers were authored by Madison. However, that is not a certainty. There is another possibility that we cannot ignore. On analysis, the three papers written by both Hamilton and Madison showed a strong lexical similarity to Madison’s essays. These three papers were never contested and it is safe to assume that both Hamilton and Madison co-authored these essays. Essay writing is not a quantitative task. The process of collaboration ideally should make the co-authored papers in clusters between Madison and Hamilton, however, no such thing happens. Collaborative writing does not follow a mechanical route and therefore, it cannot be quantified and be made to fall in the middle of their clusters. This begs the question whether the disputed essays were also written by both Madison and Hamilton and not just either of the two. What we can say with certainty that these essays were not written by Hamilton alone as they show a very strong similarity with the ones written by Madison and both of them.