HW#4

Alex Hyman

10/26/2018

## Data Configuration

The csv of feature words were read into the R environment as a data frame. Because euclidean distance is being used in the for the clustering the word frequencies were further scaled to make sure that stopwords included in the word matrix were on the same scale as lesser used words. This will give the words an equal weight. Three different

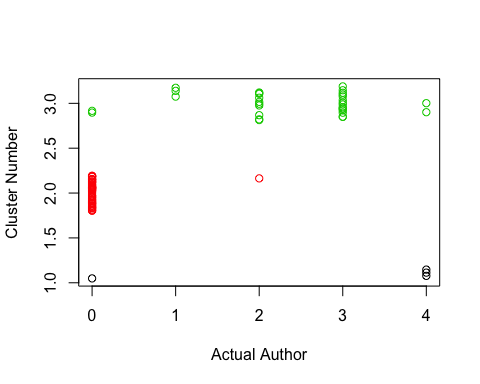
library(factoextra)

library(tidyr)  
  
#Reading the frequencies for the words  
words <- read.csv("fedPapers85.csv")  
#Scaling the words  
words[,3:ncol(words)] <- scale(words[, 3:ncol(words)])

## K-Means

K-Means was first used to cluster the the documents by the scaled frequency. Three centroids were used in the algorithm, one for each of Hamilton, Madison and Jay. The documents written by both Hamilton and Madison will be grouped with one of the clusters. There will be 100 different random sets chosen for the center.

The Authors were also provided unique numbers to distinguish them in a plot with Hamiltion being 0, Hamilton and Madison are 1, Disputed Authors are 2, Madison is 3, and Jay is 4. After the clusters were determined, a plot was created showing the actual author vs. the determined cluster.



## filename cluster  
## 1 dispt\_fed\_49.txt 3  
## 2 dispt\_fed\_50.txt 3  
## 3 dispt\_fed\_51.txt 3  
## 4 dispt\_fed\_52.txt 3  
## 5 dispt\_fed\_53.txt 3  
## 6 dispt\_fed\_54.txt 3  
## 7 dispt\_fed\_55.txt 2  
## 8 dispt\_fed\_56.txt 3  
## 9 dispt\_fed\_57.txt 3  
## 10 dispt\_fed\_62.txt 3  
## 11 dispt\_fed\_63.txt 3

The plot shows a relatively decent classification of known papers. Only three papers by Hamilton were placed in a different cluster and all of the Madison papers were classified correctly; however, two of the five papers by Jay were misclassified. The misclassification of papers by Jay could be due in pary of the small sample size.

The three papers written by both Hamilton and Madison were grouped to be more closely related to works by Madison. Additionally, 9 of the 10 disputed papers were classified as written by Madison. The only federal paper that was classified to be written by Alexander Hamilton, the rest of the papers were clustered to be written by James Madison. However, because all of the papers that were written by both James Madison and Alexander Hamilton were classified as being written by James Madison, it could be possible tha both men wrote the papers.

#table of the clusters vs   
auth.table <- table(words$cluster[words$author %in% c("Hamilton","Jay","Madison")],  
 words$author[words$author %in% c("Hamilton","Jay","Madison")])  
km.known.table <- auth.table[,c("Jay", "Hamilton", "Madison")]  
km.acc <- sum(diag(km.known.table)) / sum(km.known.table) \* 100  
cat("Accuracy of KM on known documents: ", round(km.acc, 2), "%", sep = "")

## Accuracy of KM on known documents: 92.96%

knitr::kable(km.known.table)

|  |  |  |
| --- | --- | --- |
| Jay | Hamilton | Madison |
| 3 | 1 | 0 |
| 0 | 48 | 0 |
| 2 | 2 | 15 |

Further analysis of known authors and their clusters show that if a papers if going to be missclassified, it will likely be grouped in with James Madison’s papers. Only one paper that had been missclassified as being written by John Jay. Overall, the accuracy of the clustering on known papers was determined to be 92.96% accurate.

The most unique words by each author were evaluated by the finding which words in a centroid have the highest positive value. This technique will provide the words that are most uniquely and frequently used by the author. The most frequently used words used by words by Jay include: “one”, “more”, “into”, “her”, and “than”. The most unique, frequently used words in the Hamilton cluster include: “upon”, “to”, “there”, “an”, and “in”. The most unique, frequently used words by Madison include “on”, “by”, “some”, “and”, and “was”.

#Top Words in papers by Jay  
jay.top <- order(km$centers[1, ], decreasing = T)  
print("Top words in Jay Cluster:")

## [1] "Top words in Jay Cluster:"

print(km$centers[1, jay.top][1:5])

## one more into her than   
## 2.954857 2.436481 2.245474 2.231048 2.157657

#Top words for Hamilton Cluster  
hamilton.top <- order(km$centers[2, ], decreasing = T)  
print("Top words in Hamilton Cluster:")

## [1] "Top words in Hamilton Cluster:"

print(km$centers[2, hamilton.top][1:5])

## upon to there an in.   
## 0.6654067 0.5737830 0.4382765 0.4215049 0.3729280

#Top words by Madison  
madison.top <- order(km$centers[3, ], decreasing = T)  
print("Top words in Madison Cluster:")

## [1] "Top words in Madison Cluster:"

print(km$centers[3, madison.top][1:5])

## on by some and was   
## 0.8687738 0.6738394 0.4312776 0.3755817 0.3735544

Additionally, words that are not used by an author can provide just as much insight into the word choice of an author. The words that were the furthest away from zero with a negative value are considered to be common words that the author uses less than the other authors. The least frequently used words in the Jay cluster include: “the”, “of”, “this”, “an”, and “a”. The least frequently used words in the Hamilton cluster compared to the other clusters include: “on”, “by”, “and”, “also”, and “one”. The least commonly words in the Madison cluster include: “upon”, “to”, “there”, “would”, and “an”.

#Bottom Words in papers by Jay  
jay.bottom <- order(km$centers[1, ])  
print("Least Frequently use words in Jay Cluster:")

## [1] "Least Frequently use words in Jay Cluster:"

print(km$centers[1, jay.bottom][1:5])

## the of this an a   
## -2.393384 -1.933004 -1.783312 -1.396821 -1.370269

#Bottom Words in papers by Hamilton  
hamilton.bottom <- order(km$centers[2, ])  
print("Least Frequently use words in Hamilton Cluster:")

## [1] "Least Frequently use words in Hamilton Cluster:"

print(km$centers[2, hamilton.bottom][1:5])

## on by and also one   
## -0.5567261 -0.4311125 -0.4088216 -0.3550381 -0.3036163

#Bottom Words in papers by Hamilton  
madison.bottom <- order(km$centers[3, ])  
print("Least Frequently use words in Madison Cluster:")

## [1] "Least Frequently use words in Madison Cluster:"

print(km$centers[3, madison.bottom][1:5])

## upon to there would an   
## -0.9198748 -0.8605697 -0.7228298 -0.5766132 -0.4708267

The overall most important words can be determined by looking at the absolute vlaue for each of the words. Because all the words are centered around zero, the words with the greatest absolute value will be the most significant words, either due to usage or lack of usage. The most important words in the classification of the Madison and Hamilton clusters are “upon”, “on”, and “to”. either the frequent use of these words, or the lack of use in the papers can be helpful in choosing the clusters between the two authors. The most important words in the Jay cluster are “one”, “more”, and “the”.

#Jay important words  
print("Most important Jay words")

## [1] "Most important Jay words"

jay.important <- order(abs(km$centers[1,]), decreasing = T)  
print(km$centers[1, jay.important[1:3]])

## one more the   
## 2.954857 2.436481 -2.393384

#Hamilton important words  
print("Most important Hamilton words")

## [1] "Most important Hamilton words"

hamilton.important <- order(abs(km$centers[2,]), decreasing = T)  
print(km$centers[2, hamilton.important[1:3]])

## upon to on   
## 0.6654067 0.5737830 -0.5567261

#Madison important words  
print("Most important Hamilton words")

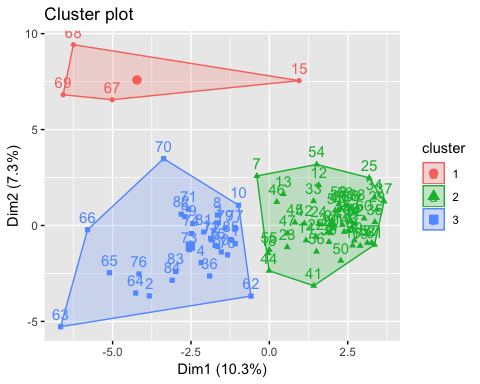
## [1] "Most important Hamilton words"

madison.important <- order(abs(km$centers[3,]), decreasing = T)  
print(km$centers[3, madison.important[1:3]])

## upon on to   
## -0.9198748 0.8687738 -0.8605697

A stripped visualization of the clusters can be provided with a scatter plot of the principal components of two dimensions. The plot shows that the Jay cluster is further in distance from the other two clusters.

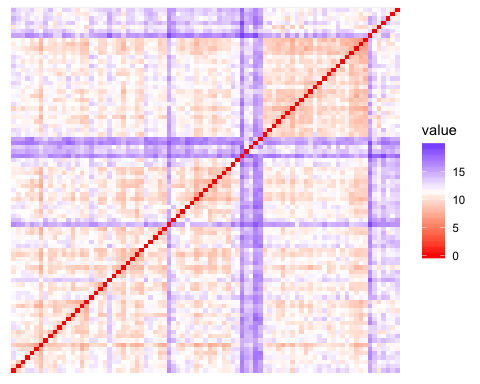
#looking at the kmeans clusters  
p <- fviz\_cluster(km, words[,3:ncol(words)])  
print(p)



## HAC

Hierarchical clustering was also used to see if a better classification can be obtained via hierarchical clustering. Hierarchical clustering will be conducted using the complete, sigle, and average methods. First, a distance matrix will need to be created. This distance matrix wiil be the basis of the hierarchical clustering for each of the methods.

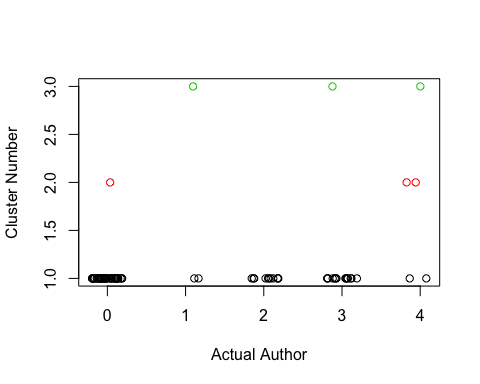
#distance of the scaled words  
word.dist <- words %>% select(c(-"author", -"filename", -"authorNum")) %>% dist(method = "euclidean")  
#heatmap of distance  
fviz\_dist(word.dist, show\_labels = F)



### Complete

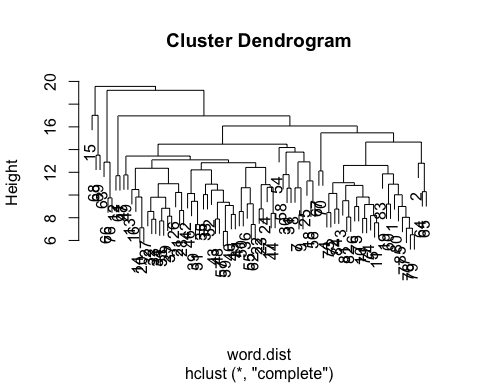
The first hierarchical clustering method used was complete clustering. After the dendogram was created, the tree was cut to have three clusters. This method groups clusters based on the distance of every item in the cluster. This clustering has a tendency to group a majority of the papers into cluster 1, with only three papers being classified into each cluster 2 and cluster 3. This clustering cannot even be properly evaluated due to the large grouping in cluster 1. This large grouping that includes a majority of each of the authors can not be distinguished as a cluster for a particular artist.

#hierarchical cluster of word distance  
hc <- hclust(d = word.dist, method = "complete")  
#cutting the hierarchical cluster at k = 3  
complete\_hac <- cutree(hc, k = 3)  
#combining hierarchical cluster to actual id  
complete\_hac <- data.frame("Author" = words$authorNum, "cluster" = complete\_hac)  
plot(complete\_hac$cluster ~ jitter(complete\_hac$Author, 1), pch = 21,  
 col = as.factor(complete\_hac$cluster), xlab = "Actual Author", ylab = "Cluster Number")

The

dendrogram of the clusters shows the problem. Because these two branches are very high up in the dendogram and only contain a few papers, it is likely that these papers are outliers. If the numbe rof authors was not already known, I would likely set the number of clusters to a higher number.

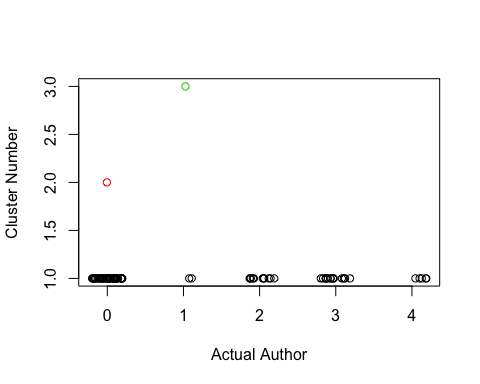
plot(hc)



### Average

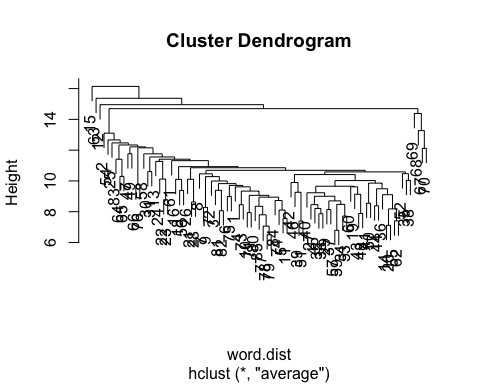
The average method was used to find the hierarchical clusters by grouping clusters with the smallest average distance between the clusters. Once again, the clustering grouped a majority of the papers in cluster 1 and only 1 paper was grouped in each of clusters 2 and 3.

#distance of the scaled words  
word.dist <- words %>% select(c(-"author", -"filename", -"authorNum")) %>% dist(method = "euclidean")  
#hierarchical cluster of word distance  
hc <- hclust(d = word.dist, method = "average")  
#cutting the hierarchical cluster at k = 3  
average\_hac <- cutree(hc, k = 3)  
#combining hierarchical cluster to actual id  
average\_hac <- data.frame("Author" = words$authorNum, "cluster" = average\_hac)  
plot(average\_hac$cluster ~ jitter(average\_hac$Author, 1), pch = 21,  
 col = as.factor(average\_hac$cluster), xlab = "Actual Author", ylab = "Cluster Number")



The dendogram plot shows that once again, outliers are having a significant effect on the hierarchical clustering. Federal papers 15 and 63 are such a far distance from the other papers, that there really is not a grat way to use hierarchical clustering with these papers in the included in the dataset.

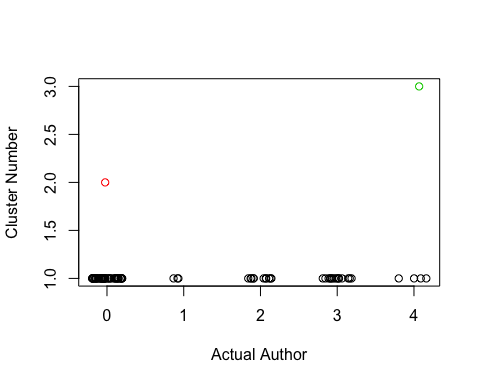
plot(hc)



### Single

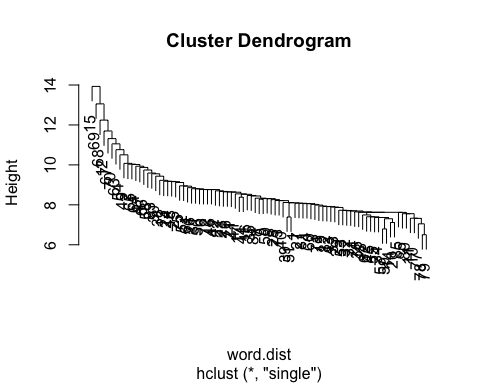
Finally, hierarchical clustering with grouping by the single closest paper was conducted, and once again, the outliers greatly affected the ability of the clustering to take place. All of the papers but two were grouped into cluster 1, with clusters 2 and 3 each containing only a single paper.

#distance of the scaled words  
word.dist <- words %>% select(c(-"author", -"filename", -"authorNum")) %>% dist(method = "euclidean")  
#hierarchical cluster of word distance  
hc <- hclust(d = word.dist, method = "single")  
#cutting the hierarchical cluster at k = 3  
single\_hac <- cutree(hc, k = 3)  
#combining hierarchical cluster to actual id  
single\_hac <- data.frame("Author" = words$authorNum, "cluster" = single\_hac)  
plot(single\_hac$cluster ~ jitter(single\_hac$Author, 1), pch = 21,  
 col = as.factor(single\_hac$cluster), xlab = "Actual Author", ylab = "Cluster Number")



The dendogram plot shows that federalist paper 15 was once again an outlier, but federalist paper 69 was also an outlier. The single method would be a decent comparison of which two papers are most similarly related, however, it is once again a terrible classifer of the authors.

plot(hc)



## Conclusion

The best method to classify the disputed papers would be with the k-means algorithm. The k-means algorithm had a 92% accuracy on known authors, and grouped the papers jointly written by Hamilton and Madison in the Madison cluster.

This k-means classifier grouped all but one of the disputed papers in the cluster along with known Madison papers. This could mean that the disputed papers were written solely by Madison, or that these papers could have been jointly written by Madison and Hamilton.

The words that most distinguished the papers by Madison and Hamilton were “upon”, “to”, and “on”. Hamilton was likely to use “upon” and “to” more often than the other authors, and to use “on” less than the other authors. John Jay tended to use the words “one” and “more” more often than the other authors, and he used the word “the” significantly less than the other authors.

The analysis of the hierarchical clustering showed that the distance between some of the papers was too large to be accurate then “cutting the tree”. This resulted in branches that were either too small, or way too large. Using the complete method could have produced a decent result if the number of clusters was not already known, as it would have likely branched off the outliers and eventually had decent grouping of the papers. The single method and average method were too affected by the outlier (papers 15, 69, 63) to have produced any decent results.