Martin\_Alonso\_Hw4

Martin Alonso

October 29, 2018

# Hw4 Instructions

You are going to try solving this mystery using clustering algorithms k-Means, EM, and HAC. Document your analysis process and draw your conclusion on who wrote the disputed essays. Provide evidence for each method to demonstrate what patterns had been learned to predict the disputed papers, for example, visualize the clustering results and show where the disputed papers are located in relation to Hamilton and Madison’s papers. By the way, where are the papers with joint authorship located? For k-Means and EM, analyze the centroids to explain which attributes are most useful for clustering. Hint: the centroid values on these dimensions should be far apart from each other to be able to distinguish the clusters.

require(stats)  
require(cluster)

## Loading required package: cluster

require(dplyr)

## Loading required package: dplyr

##   
## 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

require(ggvis)

## Loading required package: ggvis

require(tidyr)

## Loading required package: tidyr

## Loading the data

Let’s load the data and do an initial analysis of the contents of the data set.

feds <- read.csv("fedPapers85.csv")  
  
# Exploring the data set.  
dim(feds)

## [1] 85 72

# We can see that the dataset has 85 observations - one for each paper - and  
# 72 variables.  
  
# If we check the six first rows, we can see that the first two columns  
# refer to the author of the text and the text itself. Columns three and so  
# forth refer to the frequency of letters occurring in the text.  
head(feds)

## author filename a all also an and any are as  
## 1 dispt dispt\_fed\_49.txt 0.280 0.052 0.009 0.096 0.358 0.026 0.131 0.122  
## 2 dispt dispt\_fed\_50.txt 0.177 0.063 0.013 0.038 0.393 0.063 0.051 0.139  
## 3 dispt dispt\_fed\_51.txt 0.339 0.090 0.008 0.030 0.301 0.008 0.068 0.203  
## 4 dispt dispt\_fed\_52.txt 0.270 0.024 0.016 0.024 0.262 0.056 0.064 0.111  
## 5 dispt dispt\_fed\_53.txt 0.303 0.054 0.027 0.034 0.404 0.040 0.128 0.148  
## 6 dispt dispt\_fed\_54.txt 0.245 0.059 0.007 0.067 0.282 0.052 0.111 0.252  
## at be been but by can do down even every for. from  
## 1 0.017 0.411 0.026 0.009 0.140 0.035 0.026 0.000 0.009 0.044 0.096 0.044  
## 2 0.114 0.393 0.165 0.000 0.139 0.000 0.013 0.000 0.025 0.000 0.076 0.101  
## 3 0.023 0.474 0.015 0.038 0.173 0.023 0.000 0.008 0.015 0.023 0.098 0.053  
## 4 0.056 0.365 0.127 0.032 0.167 0.056 0.000 0.000 0.024 0.040 0.103 0.079  
## 5 0.013 0.344 0.047 0.061 0.209 0.088 0.000 0.000 0.020 0.027 0.141 0.074  
## 6 0.015 0.297 0.030 0.037 0.186 0.000 0.000 0.007 0.007 0.007 0.067 0.096  
## had has have her his if. in. into is it its may  
## 1 0.035 0.017 0.044 0 0.017 0.000 0.262 0.009 0.157 0.175 0.070 0.035  
## 2 0.101 0.013 0.152 0 0.000 0.025 0.291 0.025 0.038 0.127 0.038 0.038  
## 3 0.008 0.015 0.023 0 0.000 0.023 0.308 0.038 0.150 0.173 0.030 0.120  
## 4 0.016 0.024 0.143 0 0.024 0.040 0.238 0.008 0.151 0.222 0.048 0.056  
## 5 0.000 0.054 0.047 0 0.020 0.034 0.263 0.013 0.189 0.108 0.013 0.047  
## 6 0.022 0.015 0.119 0 0.067 0.030 0.401 0.037 0.260 0.156 0.015 0.074  
## more must my no not now of on one only or our shall  
## 1 0.026 0.026 0 0.035 0.114 0 0.900 0.140 0.026 0.035 0.096 0.017 0.017  
## 2 0.000 0.013 0 0.000 0.127 0 0.747 0.139 0.025 0.000 0.114 0.000 0.000  
## 3 0.038 0.083 0 0.030 0.068 0 0.858 0.150 0.030 0.023 0.060 0.000 0.008  
## 4 0.056 0.071 0 0.032 0.087 0 0.802 0.143 0.032 0.048 0.064 0.016 0.016  
## 5 0.067 0.013 0 0.047 0.128 0 0.869 0.054 0.047 0.027 0.081 0.027 0.000  
## 6 0.045 0.015 0 0.059 0.134 0 0.876 0.141 0.052 0.022 0.074 0.030 0.015  
## should so some such than that the their then there things  
## 1 0.017 0.035 0.009 0.026 0.009 0.184 1.425 0.114 0.000 0.009 0.009  
## 2 0.013 0.013 0.063 0.000 0.000 0.152 1.254 0.165 0.000 0.000 0.000  
## 3 0.068 0.038 0.030 0.045 0.023 0.188 1.490 0.053 0.015 0.015 0.000  
## 4 0.032 0.040 0.024 0.008 0.000 0.238 1.326 0.071 0.008 0.000 0.000  
## 5 0.000 0.027 0.067 0.027 0.047 0.162 1.193 0.027 0.007 0.007 0.000  
## 6 0.030 0.007 0.045 0.015 0.030 0.208 1.469 0.089 0.007 0.007 0.000  
## this to up upon was were what when which who will with  
## 1 0.044 0.507 0 0.000 0.009 0.017 0.000 0.009 0.175 0.044 0.009 0.087  
## 2 0.051 0.355 0 0.013 0.051 0.000 0.000 0.000 0.114 0.038 0.089 0.063  
## 3 0.075 0.361 0 0.000 0.008 0.015 0.008 0.000 0.105 0.008 0.173 0.045  
## 4 0.103 0.532 0 0.000 0.087 0.079 0.008 0.024 0.167 0.000 0.079 0.079  
## 5 0.094 0.485 0 0.000 0.027 0.020 0.020 0.007 0.155 0.027 0.168 0.074  
## 6 0.126 0.445 0 0.000 0.007 0.030 0.015 0.037 0.186 0.045 0.111 0.089  
## would your  
## 1 0.192 0  
## 2 0.139 0  
## 3 0.068 0  
## 4 0.064 0  
## 5 0.040 0  
## 6 0.037 0

# Having said that, the first two columns are nominal, while the remaining  
# columns are numeric.  
str(feds[, 1:6])

## 'data.frame': 85 obs. of 6 variables:  
## $ author : Factor w/ 5 levels "dispt","Hamilton",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ filename: Factor w/ 85 levels "dispt\_fed\_49.txt",..: 1 2 3 4 5 6 7 8 9 10 ...  
## $ a : num 0.28 0.177 0.339 0.27 0.303 0.245 0.349 0.414 0.248 0.442 ...  
## $ all : num 0.052 0.063 0.09 0.024 0.054 0.059 0.036 0.083 0.04 0.062 ...  
## $ also : num 0.009 0.013 0.008 0.016 0.027 0.007 0.007 0.009 0.007 0.006 ...  
## $ an : num 0.096 0.038 0.03 0.024 0.034 0.067 0.029 0.018 0.04 0.075 ...

table(sapply(feds, class))

##   
## factor numeric   
## 2 70

Having done an initial assessment of the data, we can proceed to test different clustering methods that might give as a clue as to who is (or are) the trre author of the disputed texts.

## Cluster analysis.

We shall do three types of clustering analysis: k-means, EM, and HAC.

### 1. k-means

We’ll start using the k-means algorithm by selecting three centroids. The reason why we choose three centroids is because there are four unique authors: Hamilton, Madison, and Jay. We want to check to which of these groups the 11 disputed papers belongs to.

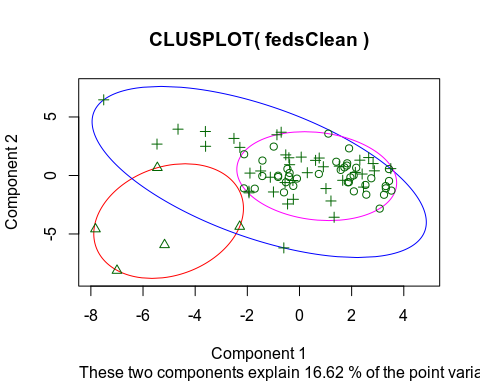
# Check how many unique authors there are. We disregard disputed papers as  
# an author as we want to identify who the real author might be.  
unique(feds$author)

## [1] dispt Hamilton HM Jay Madison   
## Levels: dispt Hamilton HM Jay Madison

# Since there are three authors, we will work with three centroids.  
  
# Now, we'll create a new data set in which the author and text columns are  
# removed.  
fedsClean <- feds[, 3:72]  
  
set.seed(825)  
k1 <- kmeans(fedsClean, 3)  
k1$size

## [1] 39 5 41

# We seem to have run into an unexpected grouping as the sizes of the  
# clusters are not what was expected. Let's plot the clusters to see how  
# they compare.  
clusplot(fedsClean, k1$cluster, color = T, shade = F, labels = 0, lines = 0)



# We can clearly see that there is overlap between the four clusters.  
# Perhaps a better approach would be to do an elbow chart and determine the  
# ideal number of centroids to use.  
wss <- kmeans(fedsClean, centers = 1)$tot.withinss  
for (i in 3:72) {  
 wss[i] <- kmeans(fedsClean, centers = i)$tot.withinss  
}  
sse <- data.frame(c(1:72), c(wss))  
colnames(sse) <- c("Clusters", "SSE")  
sse %>% ggvis(~Clusters, ~SSE) %>% layer\_points(`:=`(fill, "blue")) %>% layer\_lines()

Renderer: SVG | Canvas

Download

# There is no clear point where ideal sets of clusters would converge. In  
# any case, if we stick to the four original clusters we chose, then the  
# resulting papers would be grouped the following way:  
fedsCluster <- cbind(feds, k1$cluster)  
colnames(fedsCluster)[73] <- "cluster"  
fedsClusterClean <- fedsCluster %>% group\_by(author, cluster) %>% summarise(number = n())  
spread(fedsClusterClean, key = cluster, value = number)

## # A tibble: 5 x 4  
## # Groups: author [5]  
## author `1` `2` `3`  
## <fct> <int> <int> <int>  
## 1 dispt 5 NA 6  
## 2 Hamilton 26 NA 25  
## 3 HM NA NA 3  
## 4 Jay NA 5 NA  
## 5 Madison 8 NA 7

It appears that the clusters have done a better job than what was apparent. K-means successfully grouped all of Jay’s texts into a single cluster without adding any of the other texts. And, though it has missed on the mark for the other texts, there is some value to rescue.  
Cluster 1 houses over half of Hamilton’s and Madison’s texts and under half of the disputed texts. This may mean that the disputed texts are entirely attributable to Hamilton but he uses similar words to those of Madison.  
Cluster 3, however, encompasses the three joint author texts and the remaining texts by both Hamilton and Madison. Having also the majority of the disputed texts, there is no clear distinction as to whether the disputed texts might be further evidence of collaboration, or the usage of similar words by two different authors.

### 2. EM

### 3. HAC

We’ll now look at how the documents cluster through a heirarchical clustering algorithm (HAC).

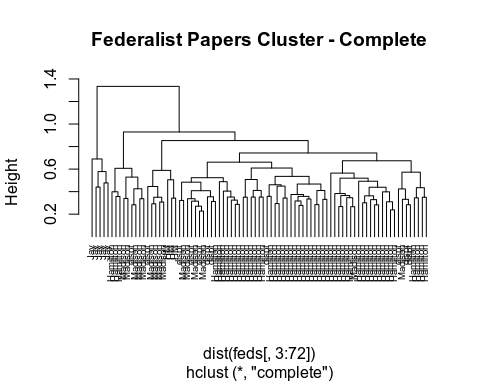
# We'll first build two different groups of clusters: one using the complete  
# method and the other the average, so as to compare how the clusters form.  
clusComp <- hclust(dist(feds[, 3:72]), method = "complete")  
clusAvg <- hclust(dist(feds[, 3:72]), method = "average")  
  
# We can see that both groups use Euclidean distance and present 85  
# observations, but the cluster method varies.  
clusComp

##   
## Call:  
## hclust(d = dist(feds[, 3:72]), method = "complete")  
##   
## Cluster method : complete   
## Distance : euclidean   
## Number of objects: 85

clusAvg

##   
## Call:  
## hclust(d = dist(feds[, 3:72]), method = "average")  
##   
## Cluster method : average   
## Distance : euclidean   
## Number of objects: 85

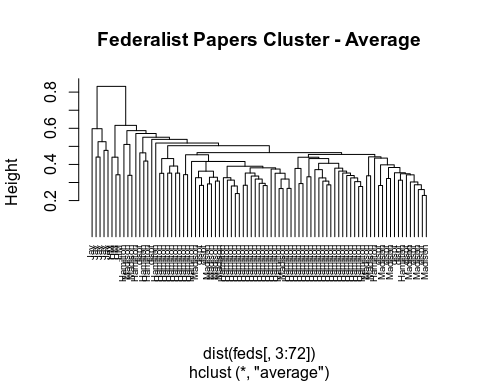
# Time to graph the clusters and see how they compare to the k-means  
# algorithm.  
plot(clusComp, hang = -1, cex = 0.6, main = "Federalist Papers Cluster - Complete",   
 label = feds$author)



From this first graph, we can see that Jay’s papers are all grouped together to the left side of the dendogram. No problem there as it keeps up with the k-means algorithm.  
However, as we start looking to the right, we find that Hamilton and Madison’s papers are intertwined, with their three co-authored papers stacked between Madison’s papers. A conclusion we can draw from this is that, though the papers were co-authored, Madison may have done the actual writing, something that would explain why they are grouped together.  
Also within this first section, we find several of the disputed papers grouped with Madison. These 8 papers seem to be authored by Madison; whilst a group of three papers at the far right of the dendogram seem to be attrubitable to Hamilton.

Let’s see how the clusters form if we use the ‘average’ method.

plot(clusAvg, hang = -1, cex = 0.6, main = "Federalist Papers Cluster - Average",   
 label = feds$author)



Again, from this graph, we can see that the Jay papers are grouped together, followed by the Hamilton-Madison papers. This gives us a different idea that the one we saw in the previous graph; the language use is completely different from a pure Hamilton or Madison verbiage. However, we then find Madison’s papers are interspersed with Hamilton’s papers, with many of the disputed papers crunched in between. There are some papers that can be argued to be authored by Hamilton, and other by Madison; but if we look at the far right of the dendogram, we find that there are seven papers grouped together with a couple of Hamilton papers and five Madison papers. These papers might be tougher to determine true authorship, though one could argue that they could be attributed to Madison.

## Conclusions

Though the initial k-means algorithm gave us a clear idea of how the papers were grouped, in the end the Hamilton, Madison, and disputed papers were so grouped together that we could not descern any true authorship.  
EM  
The HAC algorithm, however, provided more insight into how the papers were being grouped, with the complete method showing that they could be more likely attributed to Madison, while the average method showed that some of the papers were more likely to be authored by Hamilton, some by Madison, and a couple that could yet be determined since the language used is very similar to that by both authors.  
Perhaps, to this effect, these 11 papers could be a mix of co-authored papers or a very realistic attempt of both authors copying each other’s writing style, which in effect makes identifying a true author an even more difficult task.