Matthew L. Pergolski
IST 707
2/9/2022

Introduction

The Federalist Papers, some would say, are the ultimate precursor to the United States Constitution. They contained many of the same values and initialized principles that would later define the birth of a new nation. Some historians indicate that these works define a larger importance to America than just about any documentation preceding or following it.

The authors of these significant papers to the United States' history are mostly known. In more detail, the main authors are definitively known; however, there lies a mystery in terms of who wrote (i.e., 'authored') each document. In terms of what is known by historians, 85 total papers lie in existence. Alexander Hamilton, close aide and advisor to George Washington, took authorship over at least 51 essays, while two others also contributed. James Madison, who would go on to be the nation's 4th president, was responsible for authoring at least 15. Hamilton and Madison collaborated on an additional 3, and John Jay – who would go on to be the first Chief Justice of the US – wrote 5.

The overall mystery corresponds with the remaining 11 essays, where it was unclear as to who was the true author. Both Hamilton and Madison have allegedly accepted authorship for the documents – which has caused much debate and resulted in a 'disputed' perspective between the congressional (and historical) community.

The basis of the information in the following sections will attempt to utilize clustering analysis to determine an overall author for the remaining 11 'disputed' essays. In the end, the paper will advise/suggest who was the true author – which could be Hamilton alone, Madison alone, or a combination of the two.

Analysis and Models

About the Data / Exploring the Data

The clustering analysis will ultimately consist of matching the writing styles of the two different authors and making a comparison to the 'disputed' essays/documents. This will be done by utilizing a data set with 'function' words used by the various authors. The actual values indicate a percentage of the how often the particular function word appeared in the paper. For example, Hamilton's 31st federalist paper utilized the word "upon" 3 times out of a total of 1,000 words, so the value for this respective observation is 0.3.

In looking at the data set as a whole, a total of 72 attributes exist with 85 observations (one observation for each federalist paper). The 72 attributes consist of various function words such as "should," "shall," "down," "into," etc.

The first two attributes, however, are not function words. Instead, they are a categorization of author and the name of each file/paper/document corresponding to said author(s). These first two variables were imported as 'char' for character, while the rest are considered 'num' for numeric.

In terms of data cleaning and preparation, the first two variable were converted to factors while the remaining were left alone. No 'NAs' existed, so no data was omitted from the analysis. As the report progresses to the next phase, clustering algorithms of K-Means as well as Hierarchical will be conducted and generated.

Results

The overall goal of the 'exploring the data' section will be to generate 5 clusters (one cluster per category in the authors data) utilizing K-Means and HCA. The first approach, K-Means, was generated by first scaling the data set. This way, each variable's mean and standard deviation would equate similarly, 'leveling the playing field' in terms of what analysis is conducted next.

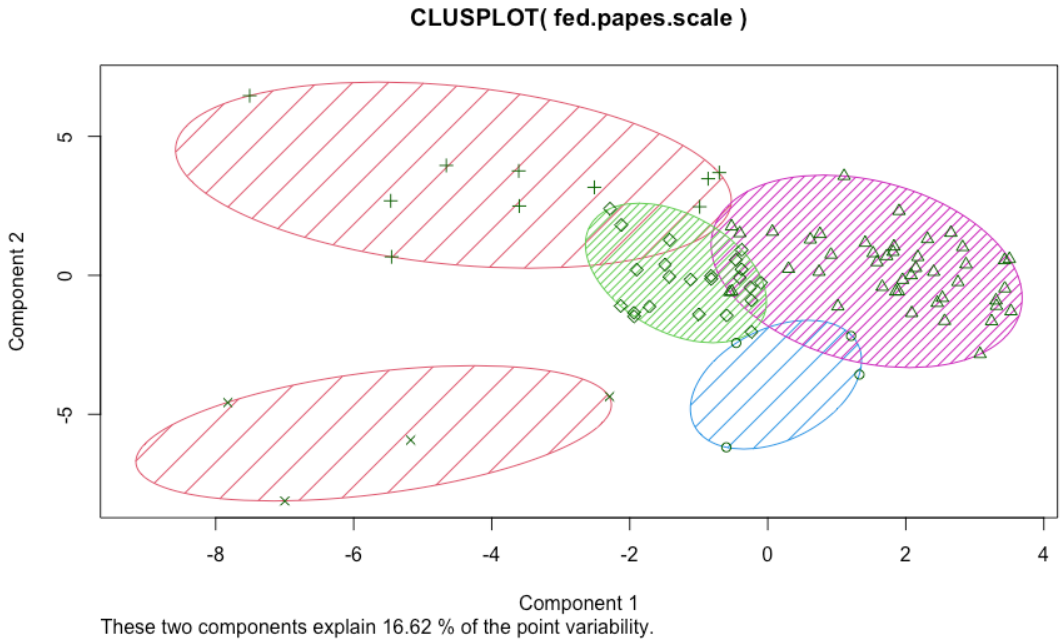
Code developed for the K-Means approach equates to the below:

```
# K MEANS

k.means <- kmeans(fed.papes.scale, 5)
k.means

k.means$centers

assignment.clusters <- data.frame(fed.papes[,1:2], k.means$cluster)
head(assignment.clusters)
View(assignment.clusters)
```



From here, the data is ready to generate an assignment data frame containing the author variable, file name, and associated k-means cluster group – which will be the basis for comparing to see which author was responsible for the writing the 'disputed' papers.

For the HCA method, two sub-methods were experimented with. The 'hclust' function as well as the 'agnes' function. For the 'hclust' method, the Euclidean distance was first determined for the (scaled) data set. A custom function was then created to determine the 'best' technique (i.e., average, single, complete, and ward). After running the custom function, it was determined that the ward method yielded the highest value, at ~0.67. Because of this discovery within the data, the ward method was used for for both 'hclust' and 'agnes.'

```
>   hc.methods <- c( "average", "single", "complete", "ward")
>   names(hc.methods) <- c( "average", "single", "complete", "ward")
>   hc.function <- function(x) {
+     agnes(fed.papes.scale, method = x)$ac
+   }
>   map_dbl(hc.methods, hc.function)
  average    single  complete      ward
0.4152046 0.3821415 0.5198525 0.6726487
```

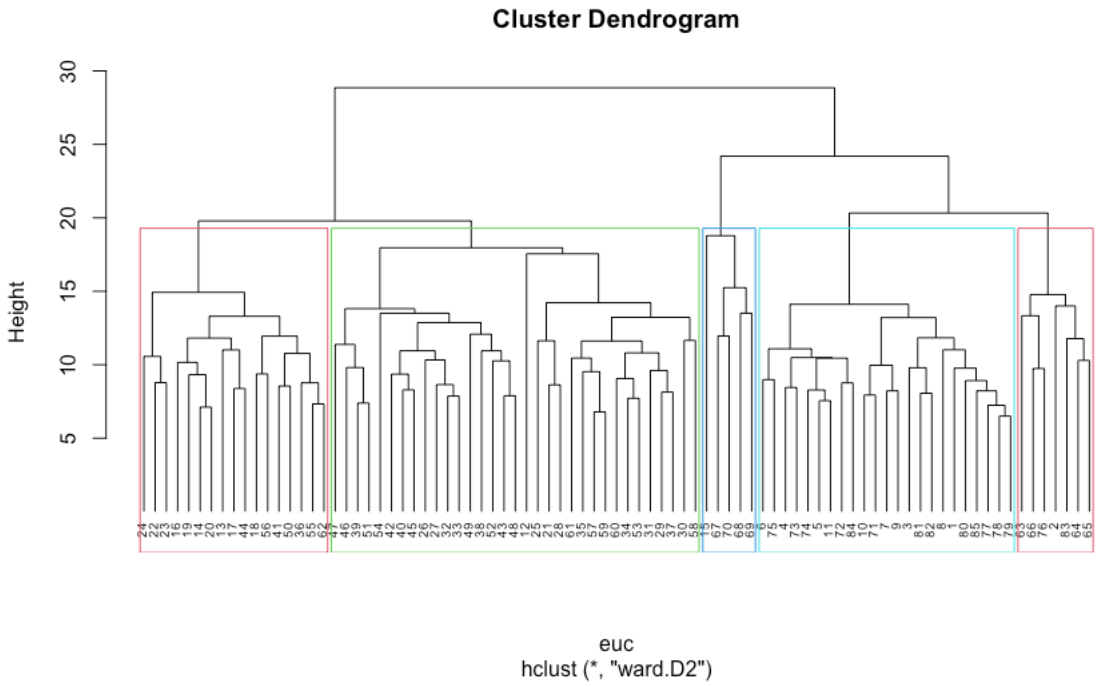
The following was generated for the final visualization shown below:

```r
hc.cluster <- hclust(euc, method = 'ward.D2')
hc.cluster

cluster.groups <- cutree(hc.cluster, k = 5)
cluster.groups

############

plot(hc.cluster, cex = .6, hang = -1)
rect.hclust(hc.cluster, k = 5, border = 2:5)
```



Upon calling the assignments variable that was created, a clear structure is seen (in record data format) that makes sense of these cluster groups.

Conclusions

After utilizing the cluster algorithms K-Means and Hierarchical (HCA), a model suggestion has become apparent as to who authored each of the 11 remaining disputed. The record data is provided below.

For federalist papers 49, 51-54, and 56-63, the associated cluster group is five (5). Upon looking at other authors within the data set, cluster five (5) is most commonly associated with Madison. Therefore, it is assumed Madison authored these documents.

For federalist paper 50, the cluster group associated is three (3). Upon looking at other authors within the data set, cluster three (3) is most commonly associated with Hamilton and Madison. Therefore, it is assumed Hamilton and Madison collaborated on this document.

For federalist paper 55, the cluster group associated is one (1). Upon looking at other authors within the data set, the only other appearance of cluster 1 occurs with a Hamilton document. Therefore, it is assumed Hamilton authored this document – although this could be further debated due to the lack of 'density' within cluster one (1).

Alas, at least according to the above cluster models, the centuries-old mystery has a proposed answer to the question of 'who wrote the 11 disputed federalist paper essays?'

author	filename	k.means.cluster
dispt	dispt_fed_49.txt	5
dispt	dispt_fed_50.txt	3
dispt	dispt_fed_51.txt	5
dispt	dispt_fed_52.txt	5
dispt	dispt_fed_53.txt	5
dispt	dispt_fed_54.txt	5
dispt	dispt_fed_55.txt	1
dispt	dispt_fed_56.txt	5
dispt	dispt_fed_57.txt	5
dispt	dispt_fed_62.txt	5
dispt	dispt_fed_63.txt	5
Hamilton	Hamilton_fed_1.txt	1
Hamilton	Hamilton_fed_11.txt	2
Hamilton	Hamilton_fed_12.txt	2
Hamilton	Hamilton_fed_35.txt	2
Hamilton	Hamilton_fed_68.txt	2
Hamilton	Hamilton_fed_69.txt	2
Hamilton	Hamilton_fed_7.txt	3
Hamilton	Hamilton_fed_8.txt	2
Hamilton	Hamilton_fed_80.txt	2
Hamilton	Hamilton_fed_81.txt	2
Hamilton	Hamilton_fed_82.txt	2
Hamilton	Hamilton_fed_83.txt	2
Hamilton	Hamilton_fed_84.txt	2
Hamilton	Hamilton_fed_85.txt	2
Hamilton	Hamilton_fed_9.txt	3
HM	HM_fed_18.txt	3
HM	HM_fed_19.txt	3
HM	HM_fed_20.txt	3
Jay	Jay_fed_2.txt	3
Jay	Jay_fed_3.txt	4
Jay	Jay_fed_4.txt	4
Jay	Jay_fed_5.txt	4
Jay	Jay_fed_64.txt	4
Madison	Madison_fed_10.txt	5
Madison	Madison_fed_14.txt	5
Madison	Madison_fed_37.txt	5
Madison	Madison_fed_38.txt	5
Madison	Madison_fed_39.txt	5
Madison	Madison_fed_40.txt	3
Madison	Madison_fed_41.txt	5
Madison	Madison_fed_42.txt	5
Madison	Madison_fed_43.txt	5
Madison	Madison_fed_44.txt	5
Madison	Madison_fed_45.txt	5
Madison	Madison_fed_46.txt	5
Madison	Madison_fed_47.txt	5
Madison	Madison_fed_48.txt	5
Madison	Madison_fed_58.txt	5