Zane Alderfer

Dr. Bolton

IST 707

November 3, 2023

Homework 4

Introduction

Believe it or not, not every signer of the Constitution was completely convinced of the idea initially. A group of men under the collective pseudonym “Publius” set out to promote the Constitution in order to get it ratified. By way of 85 essays, these men delivered what they believe is the best argument and interpretation for the intent of the Constitution. About 30 years later, the identity of these men became known. They were, of course, James Madison, Alexander Hamilton, and Jon Jay.

With contrasting beliefs, Hamilton was known as one of the heads of the Federalist party and Madison, a major name for the Democratic-Republicans. Hamilton has been discovered to have written the bulk of the 85 essays as 51 of these articles have been attributed to him, Madison taking ownership of 15, Jon Jay only with 5 attached to his name, and then 3 have a joint ownership of Madison and Hamilton, respectively. This, however, leaves 11 essays under speculation as to who wrote them. Of course, it doesn’t help that both Hamilton and Madison have claimed authorship.

While the true author of these essays may never be known, that doesn’t mean the evidence available can’t be used to develop educated guesses. The Federalist Papers are some of the most important documents in American history so using the essays with known authorship and comparing them to the already distinguished essays could be vital in determining who wrote them. This paper will look into potential ways the truth can be uncovered and possibly unveil the names behind these essays.

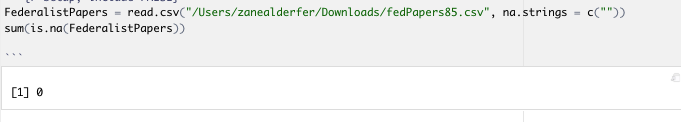
Analysis and Models

**About the Data**

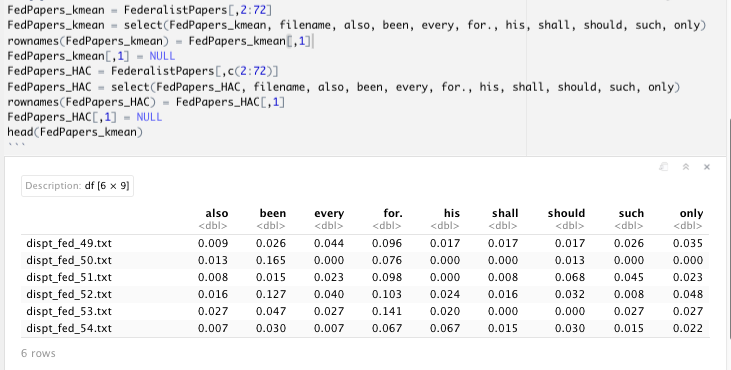
This dataset contains all 85 essays within it. With each essay, is the author attributed to that essay including “dispt” for the essays that are up for debate. Then the entirety of the dataset is the frequency of each word used in each essay. This will be vital in drawing conclusions about who the disputed authors are.

**Reading in the Data**

A simple read\_csv function was used to read in the dataset and then to remove NA values from the string fields, they were replaced with blanks. To be sure all the NA’s were removed, a sum function was used as seen below.

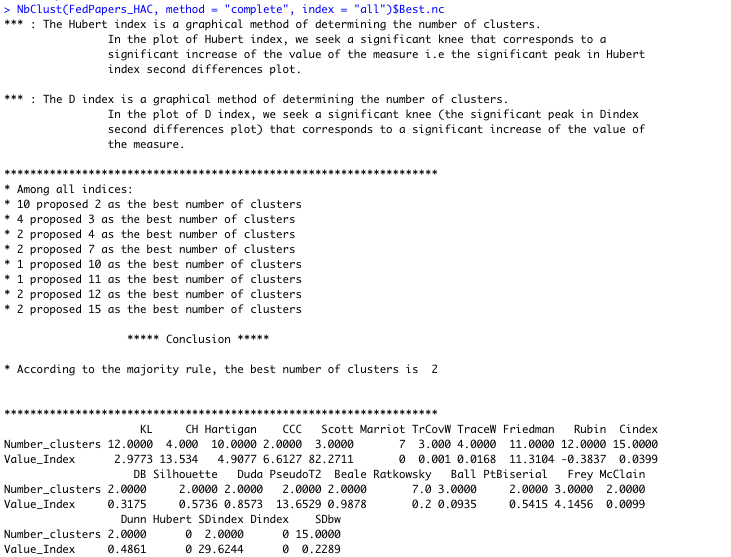
**Transforming the Data**

In order to run the models desired, the dataset needs to be in an all numerical form. In order to do this, the author names must be removed from the data set and then making the row names the names of the essays followed by making all the row names all Null to keep the dataset numerical. These new variables can be used for kmean analysis and an identical set of variables can be made for the HAC analysis. Focusing on specific words to reduce the dimensionality is important so a few nonrandom words were picked in the hopes that this would develop a narrow tunnel for who the disputed authors are. This can be seen below.



**Developing Optimal # of Clusters**

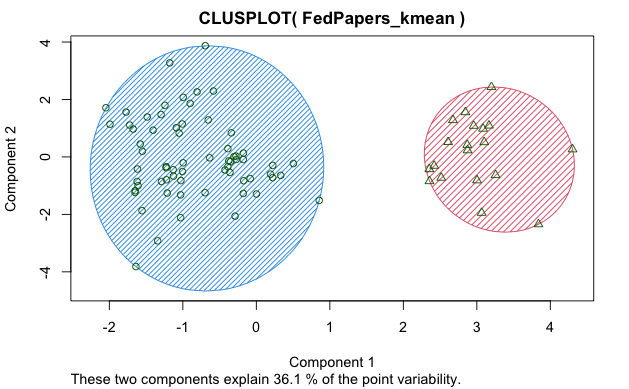
There are a few ways to figure the optimal number of clusters such as the fviz\_nbclust function but optimally the NbClust function from the NbClust library was used. The results can be seen below.

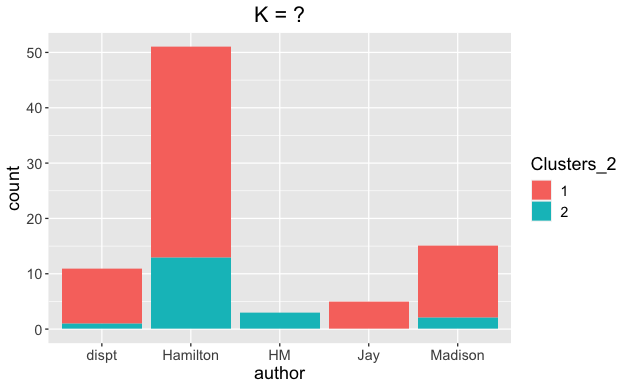


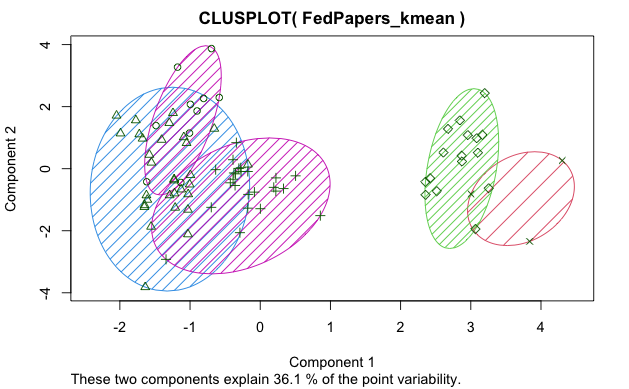
The result is interesting here as only 2 clusters is the “majority rule.” However, to get a more complete analysis, 3 of the suggested clusters will be examined. The three being 2, 5, and 15.

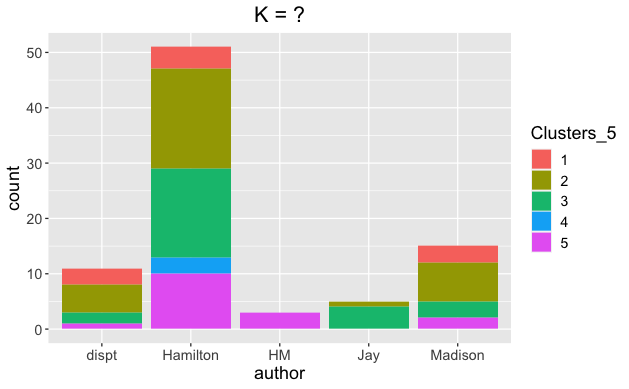
**Analysis of the Models**

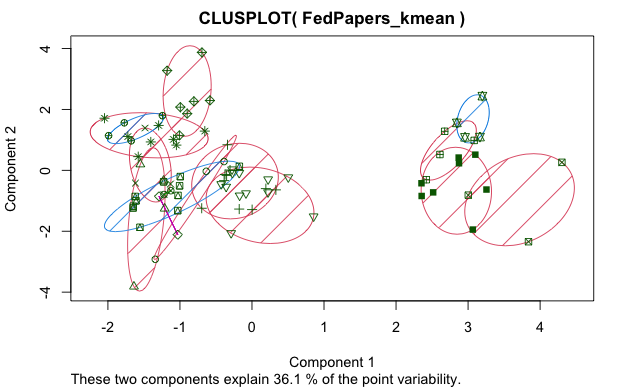
After developing variables that contained the number of clusters for each evaluation, those variables were then put up against the “cluspoint” function followed by being plotted in a “ggplot” based on the number of clusters in each variable. Before being able to put the data into ggplot graphs, the data had to be factorized to actually develop the clusters that needed to be visualized. The results are as shown below.

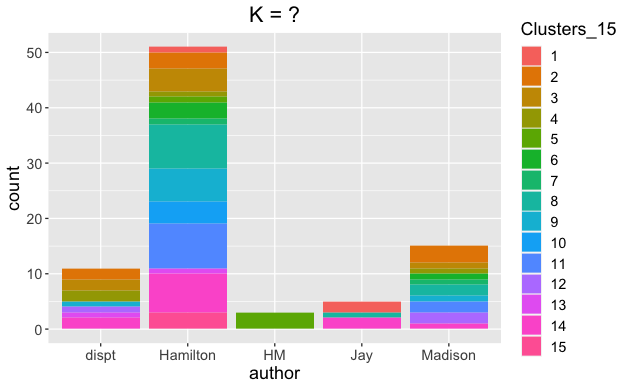
**Graph 1 (Cluster Plot) - 2 Clusters**

**Graph 2 (GGplot) - 2 Clusters**

**Graph 3 (Cluster Plot) - 5 Clusters**

**Graph 4 (GGplot) - 5 Clusters**

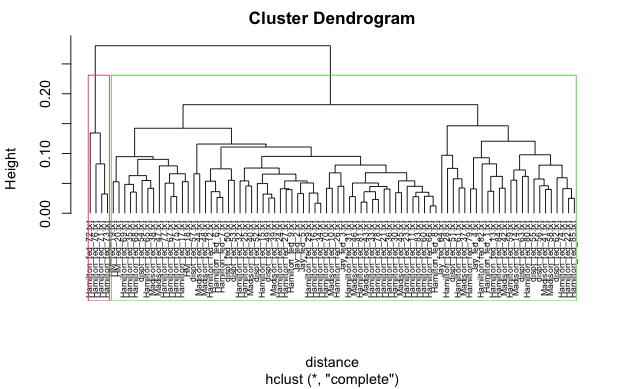
**Graph 5 (Cluster Plot) - 15 Clusters**

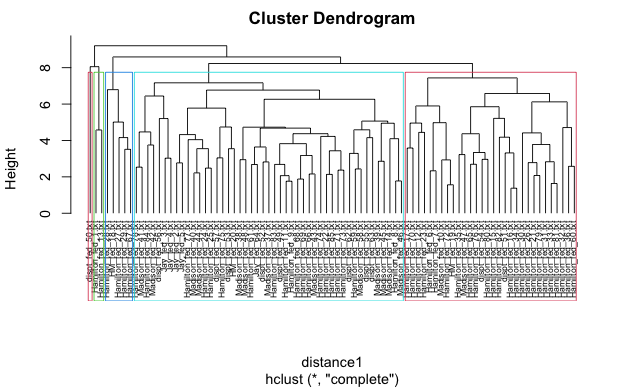
**Graph 6 (GGplot) - 12 Clusters**

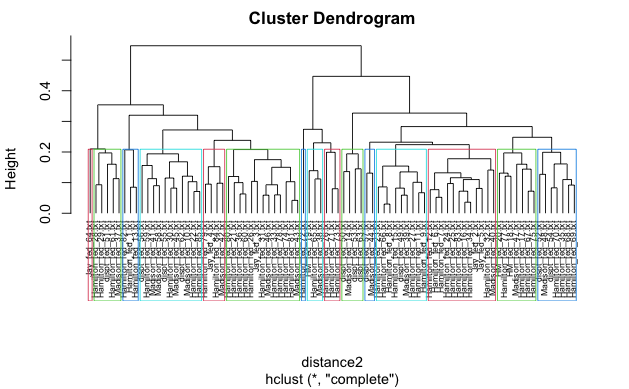
That concludes the model development for kmeans. To further analyze the data, hierarchical clustering algorithms will be created for the dataset using three separate methods: euclidean, cosine, and manhattan. To do this, more variables will be created using the “dist” function and then defining the type of preferred method. This can be seen below.

These variables will then be implemented into developing HAC diagrams. To see a diversified output, implementing the differing number of clusters for each HAC method will be done. The euclidean method will use 2 clusters, cosine will use 5 and manhattan will use 15.

The results are as follows.

**Graph 7 - 2 Clusters - Euclidean Method**

**Graph 8 - 5 Clusters - Canberra Method**

**Graph 9 - 15 Clusters - Manhattan Method**

Results

Looking at graphs 1 and 2 collectively, there seems to be a very distinct difference between clusters 1 and 2. There is no crossover whatsoever between the two clusters which would imply a distinct difference between the usage of specific words by the authors. Graph 2 doesn’t give too much value but an interesting take away would be that Jay clearly had nothing to do with the mixed authorship essays as all of Jay’s word usage is within cluster 1 and the mixed essays’ are all within cluster 2. Another interesting takeaway here is that the disputed essays are primarily cluster 1 and slightly cluster 2. Based on this, there’s no telling who wrote the disputed essays, however the one disputed essay as well as the mixed essays in cluster 2 look to be more than likely written by Hamilton as Hamilton seems to be more engaged in cluster 2. Bouncing to graph 7, this tells a different story for the two clusters. This graph suggests that the entirety of cluster 1 is Hamilton. The only significant thing from this graph would be the two mixed essays to the far left of the dendrogram as they’re surrounded by Hamilton which could potentially suggest that these are more Hamilton than Madison authorship.

A consistency developed across the three cluster models. As mentioned above, there is a distinct difference in wording which has led to two separate islands of clusters regardless of how many clusters are being used. In graph 3, there seems to be more of an overlap in the three clusters to the left than the two clusters to the right. Matching with its corresponding ggplot in graph 4, Jay is once again isolated from the mixed essays. Cluster 4 is home only to Hamilton. Other than that, this chart doesn’t present much value. Looking at graph 8’s five cluster dendrogram, one of the disputed essays is completely isolated from the rest in its own cluster. Cluster five has 30 essays within it, 2 being disputed essays. This cluster is heavily in favor of Hamilton which could be suggestive of authorship for those 2 disputed essays.

Graph 3 has 15 clusters in a cluster plot and there’s no deviation from the distinct style of cluster formulating two clear cut groups. Observing graph 6, there’s no commonality between the mixed essays and the disputed essays. Hamilton and Jay are the only two authors who fall into cluster 1. With 15 clusters involved, the mixed HM essays only falling into one cluster is an outlier. 15 clusters makes it particularly hard to distinguish meaningful relationships across clusters in a dataset. The corresponding 15 cluster dendrogram has a cluster with 3 disputed essays and just Madison implying the style of writing in these essays are indigenous to Madison. There’s another cluster with multiple disputed essays as well as multiple Hamilton and Madison essays which could suggest that these disputed essays were more than likely not written by the same person.

Conclusion

It is easy to see why so many historians or proof readers had trouble identifying the authors of these disputed essays. Based on the clustering models created in this report, there’s no definitive answer as to who wrote these essays. In fact, the clustering models actually contradicted themselves at times. Although no opinion can be formulated in terms of identifying the authors of the disputed essays overall, it was interesting to see that specific graphs were able to give a good indication of who the authors were of some of the disrupted essays. Unfortunately, the results weren’t unanimous across the board so no concrete conclusion could be made.

The clustering models could have been developed very differently had the specific words that were chosen to represent the essays been chosen differently. If the models were to be recreated, perhaps using all of the words in the essays versus picking specific words would have been more proficient in identifying better clusters. After changing the specific words a few times for the models, the amount of suggested clusters changed which further suggests that the models that were developed were more than likely not the most accurate ones. Furthermore, the words chosen weren't the only thing that could have had an impact on the outcome as the methods used in the HAC models could have been developed more effectively in that perhaps using a single or ward methodology would have created a better representation of the data.

Some takeaways from the models that can be said confidently would be that Jay evidently had a different style of writing and would have been almost beneficial to remove his essays from the models to get a more thorough comparison between Hamilton and Madison. Although they had differing political views, it’s amazing that their authorship can be hard to distinguish from one another, almost implying that maybe they had more in common than they thought. Last thought would be that considering Hamilton wrote the bulk of the essays, it could be assumed that he had some kind of a hand in the disputed essays. Although the evidence in favor of Madison is also readily apparent. This may never be completely solved but regardless of the true results, these men were paramount in the development of the United States.