David Doman

IST 707 – HW4

July 29, 2020

**Introduction**

The Federalist Papers are some of the most important and crucial documents when it comes to the origins of the United States of America. They were a series of 85 essays urging the citizens of New York to ratify the United States Constitution. Although many of the essays have a known author there are some that to this day are still a mystery to whom the author is. The known authors of the essays include James Madison, Alexander Hamilton & John Jay.

The focus of this analysis will be to analyze the text of the essays to determine patterns in writing styles, word use, length, etc. to crack the case of the who the true authors are. Although the actual truth lies within the true author certain analytical methods can be used to gain a better understanding of who probably or may have authored the essays.

The remainder of this document will provide information about the data, how the data was cleaned or manipulated, techniques used to analyze the data and visualizations of results.

**Analysis and Models**

**About the Data**

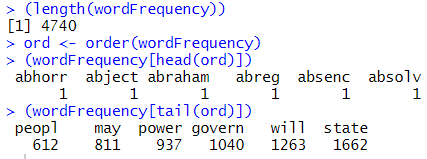
Of the 85 essays, 74 of them have a known author. Hamilton authored an overwhelmingly proportion of the essays laying claim to 51 of them. Madison authored 15 essays followed by Jay authoring 5. Hamilton and Madison teamed up to together to author 3 essays. This leaves 11 essays that are still shrouded in mystery as to whom the true author is. Each of the papers is in a text file format.

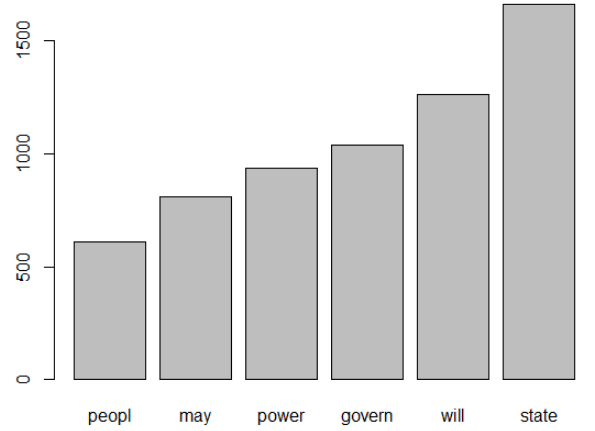
To utilize these 85 individual text files for analysis some cleaning up and restructuring to the files must be performed.

Since the cleaned-up version only contains some words contained in the essay it is beneficial to look at other potential words that could be of use. With 85 individual text files the “corpus” function needs to be used in R to make the data workable for analysis. After the “corpus” function has been applied the next step is to remove anything that can impede our analysis such as punctuation marks.

**Models**

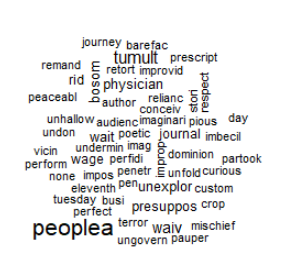
Now that the data has been imported into R and cleaned up, the next step is to run analysis using K-Means and HAC analyses. The starting point will be to view various statistics about the data and to determine word frequencies within the dataset. 4,740 words were shown to occur frequently. The most frequent words include “state”, “will”, “govern”, “power”, “may” & “people”, respectively.



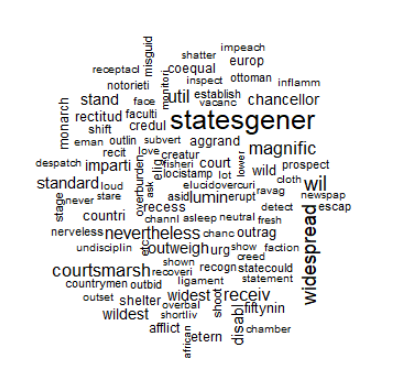


Word Cloud was then used to visualize the most frequent words used by Hamilton and Madison. Bigger words describe those terms that are used most frequent and smaller terms describe those terms that are used less frequent. These word clouds are shown below:

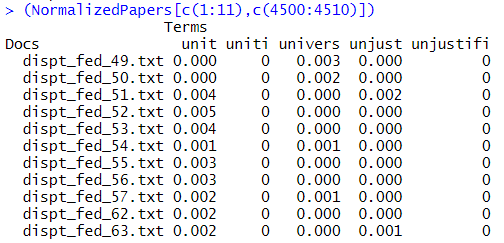
**Hamilton**



**Madison**

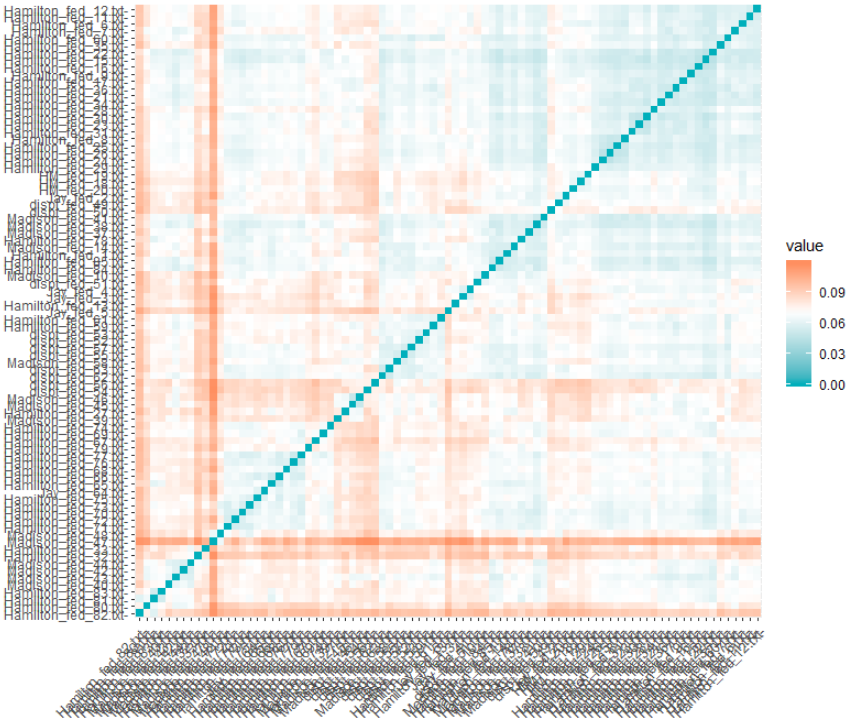


Since the essays vary in length and number of words total in each essay it made sense to normalize the data to create a level playing field.

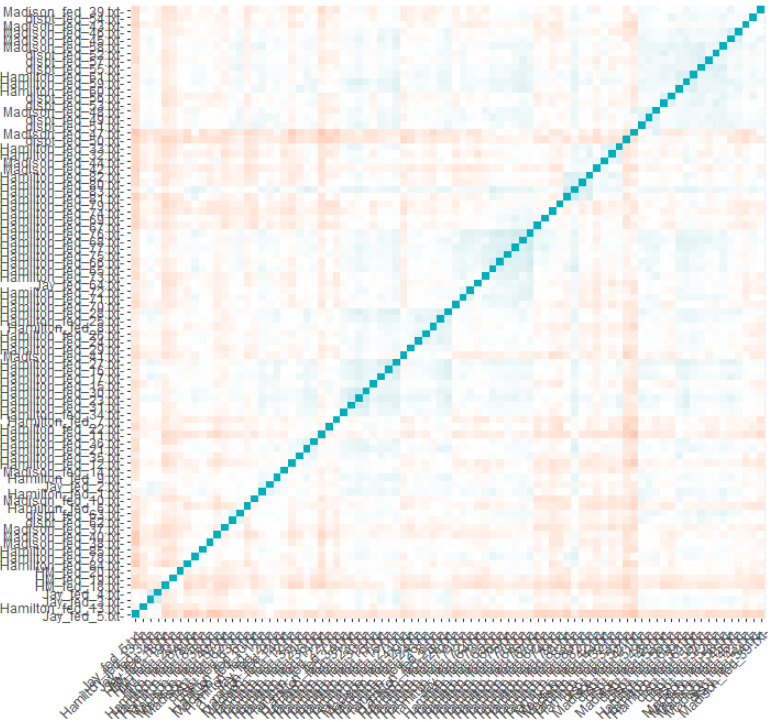


The next step is to run the HAC clustering analysis to visualize the clusters. Before the HAC plots were run the distances using Euclidean, Manhattan and Cosine were calculated. To visually see distances between the various essays we can create the following charts:

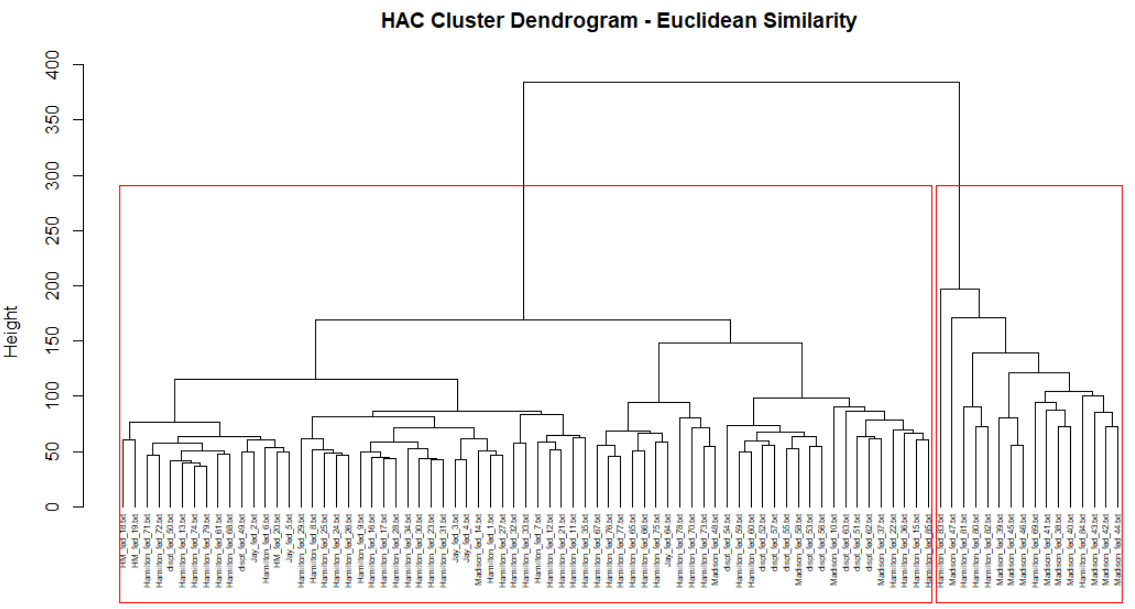
**Euclidean**

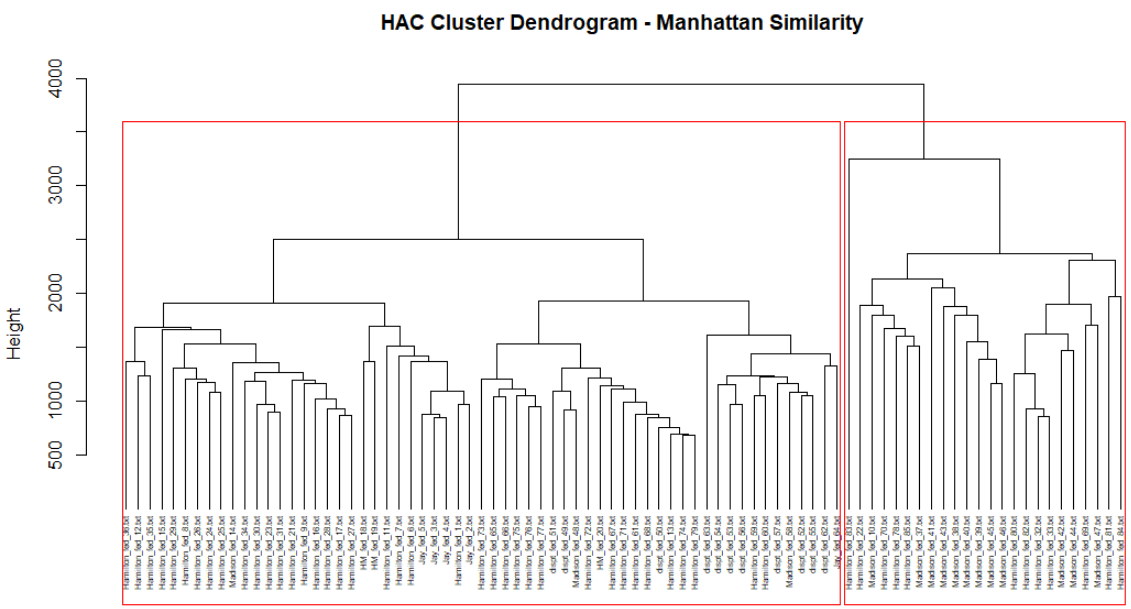


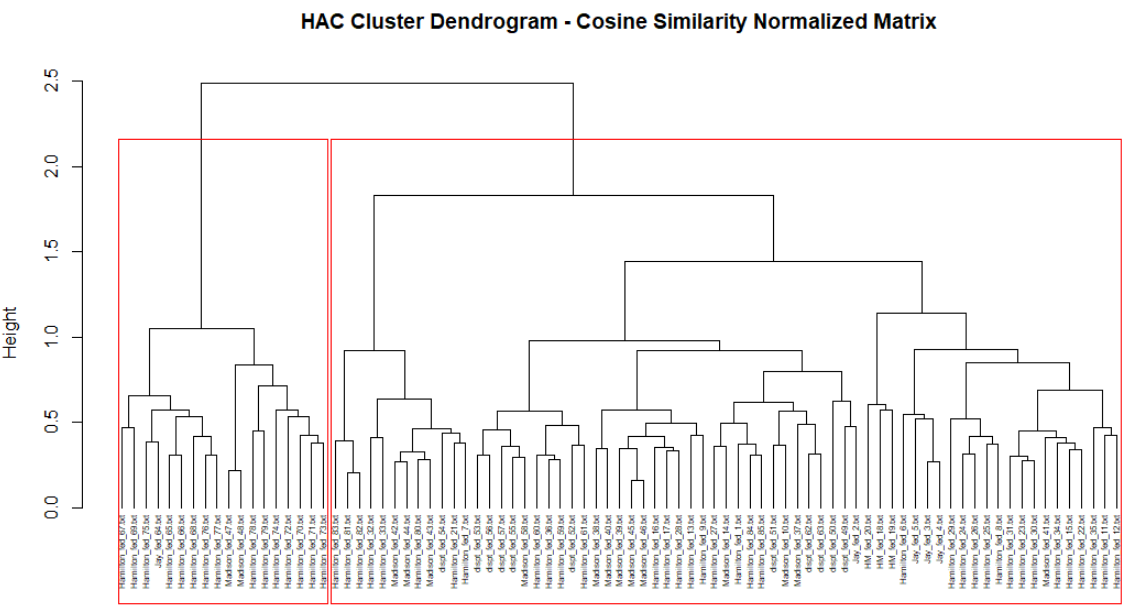
**Manhattan**



These charts tell us how similar 2 essays are to one another. The blue line in the middle is zero since that is where the same 2 essays are being compared. The blue shades represent a smaller distance between 2 essays, while the red spots represent a larger difference between 2 essays. For each distance measurement method 2, 3, and 4 clusters were then tested to determine the optimal number of clusters.

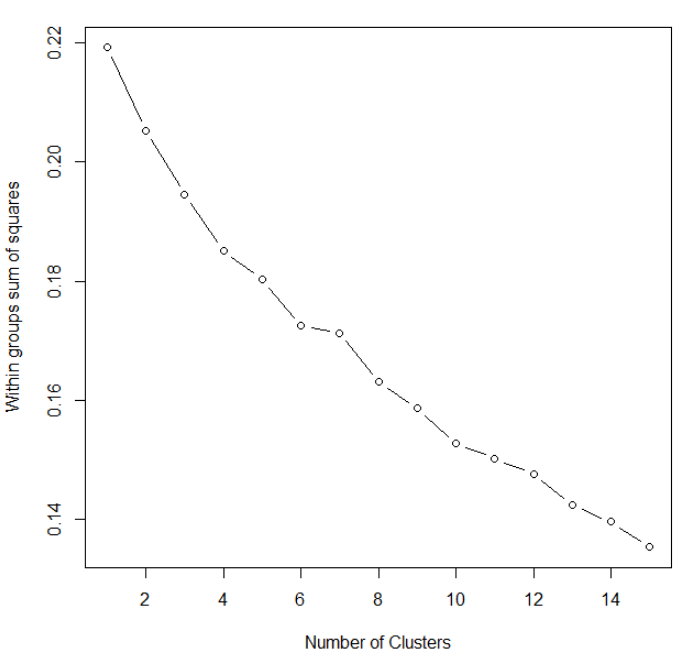




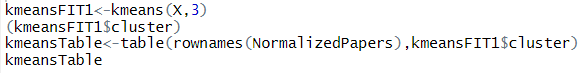


The differences in the results from each of the distancing methods is why it is important to analyze clusters using various options and techniques. Each technique works differently depending on the type of data being analyzed and with unsupervised analysis the more options to view the data the better. The objective is to see if there is a cluster that contains most of the unknown papers along with one author. It was determined that 2 clusters provided the most useful results. From the above cluster graphs, Euclidean and Cosine did not offer any such type of cluster. The clusters using these methods contained various artists and in multiple clusters. On the other hand, looking at the Manhattan cluster chart it is clear to see that one cluster contains mainly Madison’s work, while in the other cluster contains Hamilton’s and the unknown author essays. This tells us that Hamilton is the author responsible for writing the unknown essays.

Next, the K-Means technique was utilized to analyze the dataset. The chart below was created to help determine what the optimal number of clusters should be for K-Means analysis. K-Means relies heavily on picking the correct number of clusters for the results to be sufficient.



From the graph above it seems as though 3 clusters is a good number to use for K-Means. The following code was executed to get K-Means results:



This yielded the following results:

|  |  |  |  |
| --- | --- | --- | --- |
| **Author** | **Cluster 1** | **Cluster 2** | **Cluster 3** |
| Unknown | 0 | 3 | 8 |
| Hamilton | 15 | 26 | 10 |
| HM | 0 | 3 | 0 |
| Jay | 0 | 4 | 1 |
| Madison | 5 | 5 | 5 |
| **Grand Total** | **20** | **41** | **24** |

Based on frequency in each cluster some assumptions can be made:

|  |  |  |
| --- | --- | --- |
| **Disputed Paper** | **Cluster** | **Most Likely Author** |
| Fed Paper 49 | 3 | Hamilton |
| Fed Paper 50 | 3 | Hamilton |
| Fed Paper 51 | 2 | Undetermined |
| Fed Paper 52 | 2 | Undetermined |
| Fed Paper 53 | 2 | Undetermined |
| Fed Paper 54 | 1 | Undetermined |
| Fed Paper 55 | 2 | Undetermined |
| Fed Paper 56 | 2 | Undetermined |
| Fed Paper 57 | 2 | Undetermined |
| Fed Paper 62 | 2 | Undetermined |
| Fed Paper 63 | 3 | Hamilton |

**Results**

Results showed that HAC analysis turned out to be a better predictor as to who authored the disputed essays. More specifically, the Manhattan method was overwhelmingly more effective than the Euclidean and Cosine methods. It is clear to see from the Manhattan model that a good number of the disputed essays most likely belonged to Hamilton. Although it is not a certainty this model provided evidence that this most likely is the case. On the other hand, the K-Means technique did not provide as effective results. Authors were dispersed throughout multiple clusters providing no clear conclusion as to who the mystery author is for the 11 disputed essays. Some assumptions were made based on the results even though they were not as strong as the HAC model.

**Conclusions**

The Federalist Papers are some of the most important and recognized papers when it comes to the origination of the United States of America. These essays and the individuals who wrote them had a lasting impact on the country that we can even see today. Alexander Hamilton, James Madison, and Jon Jay are known to be the authors of what was published under the name “Publius” in New York in support of ratification. Of the 85 essays written, 11 still to this day do not have an author assigned to it. Obviously knowing who authored these 11 disputed essays would be fantastic to know for historical reasons but from the results of this analysis it is tough to determine who the true authors are.

Reviewing results from both the HAC and K-Means analysis, the only essays that could be possibly attributed to an author would be Federalist Papers 49, 50, and 63. Both analysis point in the direction of Alexander Hamilton being the author. Beyond that, further analysis would need to be performed to assign an author to the other 8 disputed essays.