Joyce Woznica

jlwoznic@syr.edu

November 1, 2019

Joyce Woznica  
Homework #4

Contents

[Introduction 3](#_Toc23543550)

[Analysis 4](#_Toc23543551)

[The Data 4](#_Toc23543552)

[Data Load, Cleanse, Munge and Preparation 4](#_Toc23543553)

[Exploratory Data Analysis 5](#_Toc23543554)

[Visualization 11](#_Toc23543555)

[Results 16](#_Toc23543556)

[Conclusions 16](#_Toc23543557)

List of Figures

[Figure 1: The Federalist Papers 3](#_Toc23543580)

[Figure 2: Example of Federalist Paper – Number 49 4](#_Toc23543581)

[Figure 3: Structure of Corpus created from Federalist Papers 5](#_Toc23543582)

[Figure 4: Setting up Term Frequencies and Removing Stop Words from Federalist Papers 5](#_Toc23543583)

[Figure 5: Inspecting the Document Term Matrix 6](#_Toc23543584)

[Figure 6: Normalizing the Word Matrix 7](#_Toc23543585)

[Figure 7: Wordcloud Code for Each Author 7](#_Toc23543586)

[Figure 8: Wordcloud for Disputed Papers 8](#_Toc23543587)

[Figure 9: Worldcloud for Hamilton Papers 8](#_Toc23543588)

[Figure 10: Wordcloud for Hamilton/Madison Papers 9](#_Toc23543589)

[Figure 11: Worldcloud for Jay Papers 10](#_Toc23543590)

[Figure 12: Wordcloud for Madison Papers 10](#_Toc23543591)

[Figure 13: Distance Matrices 11](#_Toc23543592)

[Figure 14: Plotting Functions for Clustering with Different Distance Matrices 11](#_Toc23543593)

[Figure 15: Clustering Using Euclidean Distance 12](#_Toc23543594)

[Figure 16: Clustering Using Cosine Distance 13](#_Toc23543595)

[Figure 17: Clustering Using Manhattan Distance 13](#_Toc23543596)

[Figure 18: Determining the number of Clusters for k-means 14](#_Toc23543597)

[Figure 19: k-means Clustering R-Code 14](#_Toc23543598)

[Figure 20: k-means Visualization with Manhattan 15](#_Toc23543599)

[Figure 21: k-means Visualization with Euclidean 15](#_Toc23543600)

[Figure 22: k-means Visualization with Spearman 16](#_Toc23543601)

# 

# Introduction

There is a collection of 85 articles and essays written in the second half of 1780 by Alexander Hamilton, James Madison and John Jay under the pseudonym “Publius” called the “Federalist Papers” which were written to promote the ratification of the United States Constitution. The authorship was a closely guarded secret. After Hamilton’s death in 1804, the list of articles attributed to Hamilton became public.

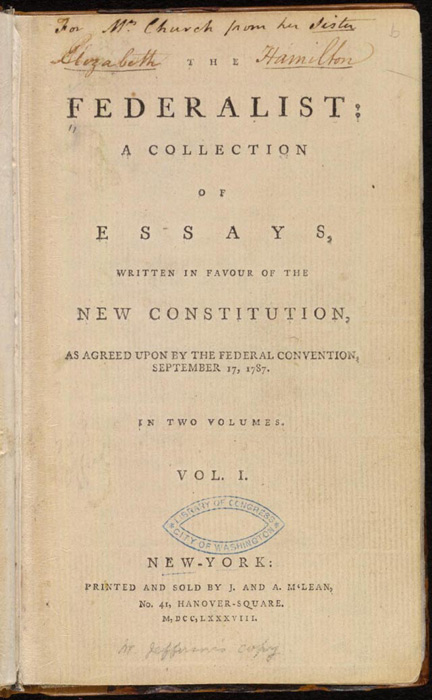


Figure 1: The Federalist Papers

However, who wrote each paper is still mysterious and unfortunately, the truth is has been lost. It has been noted that each author had some unique characteristics to his writing. This could be the use of particular words, phrasing or writing style. Using statistical methods, the goal is to determine which papers can be attributed to which author and how different information in the text draws that conclusion.

This document outlines the data collected, the loading process, any modifications required to the data and why, initial visualizations and information about the data, and the results of various text mining methods and clustering.

# Analysis

## The Data

Each “paper” is a text file (.txt) that is numbered and the content look similar to that represented in the following figure.

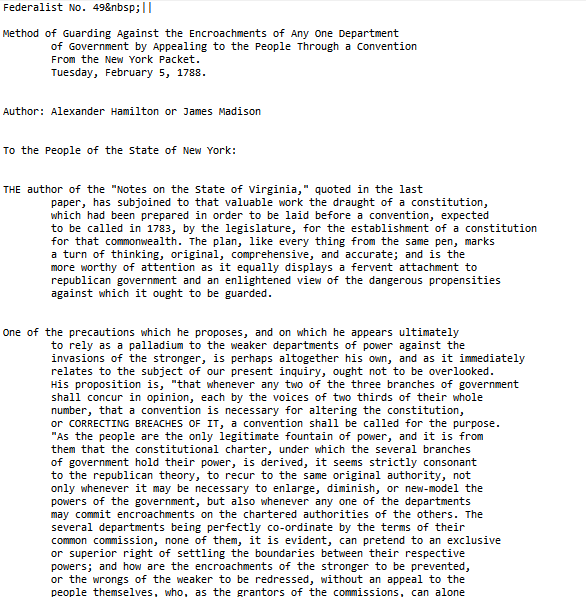


Figure 2: Example of Federalist Paper – Number 49

Each paper is labelled by its contributing author or suspected author. There are 11 disputed papers where authorship is unknown. There are 15 Madison papers, 51 Hamilton papers, 3 Hamilton/Madison (either written by Hamilton or Madison) and 5 Jay papers in the data set.

Before inspecting the articles and attempting to make some determination of the author of the papers, there is some cleaning and work that needs to be done on the data set to prepare it for the statistical functions that will be used.

### Data Load, Cleanse, Munge and Preparation

#### Data Load

To load the federalist papers, we utilize the *corpus* function. This function creates a corpus from the documents which can then be manipulated for statistical analysis. The resulting 85 documents create a structure like that in the following figure. The text of the paper can be found in the “content” section. The metadata contains the paper name, or ID.

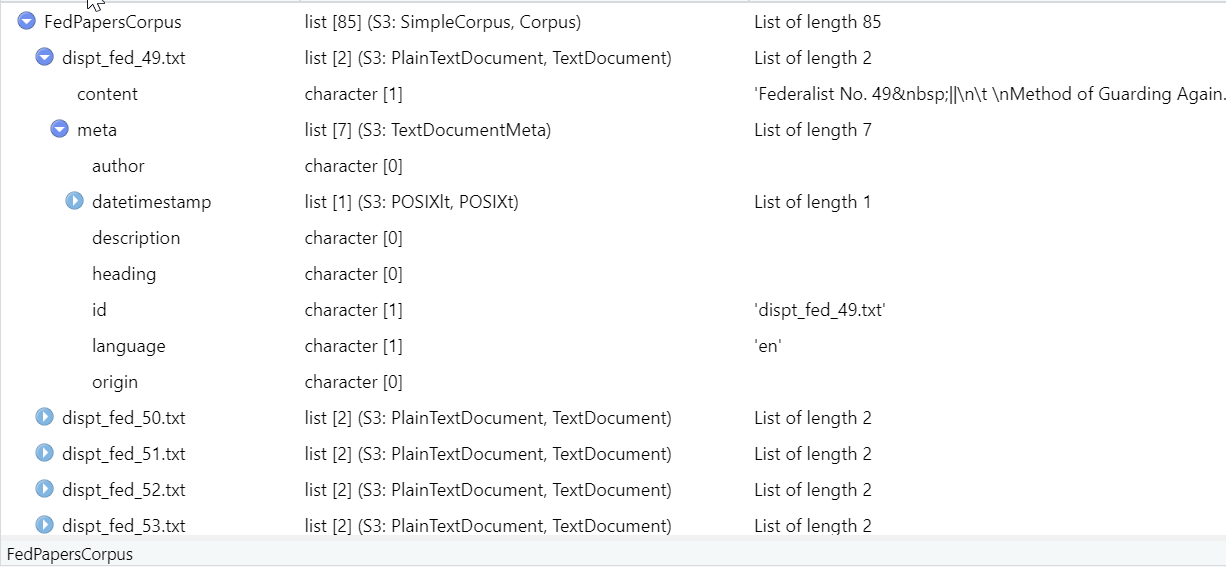


Figure 3: Structure of Corpus created from Federalist Papers

#### Data Cleanse

Some initial cleanup needs to be done on the papers to remove very rare and very common words as well as removing stop words such as common conjunctions like “and”. In addition, all punctuation, numbers and separators are removed as shown in the figure below.

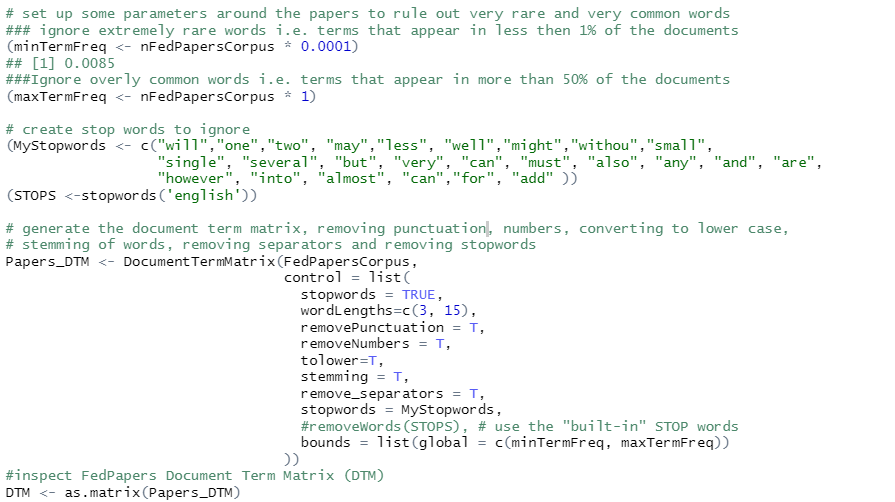
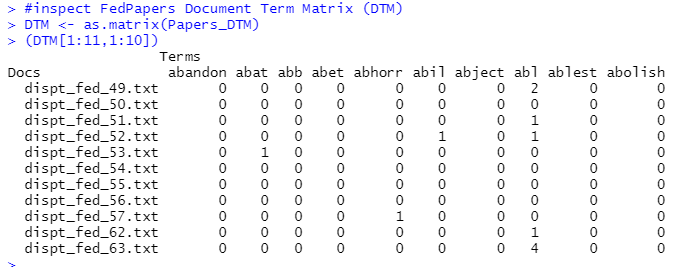


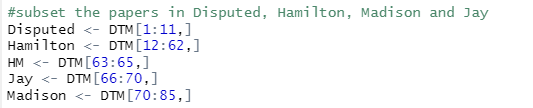
Figure 4: Setting up Term Frequencies and Removing Stop Words from Federalist Papers

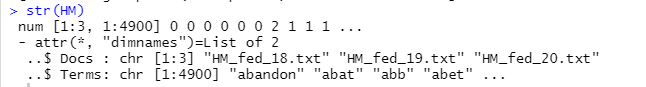
## Exploratory Data Analysis

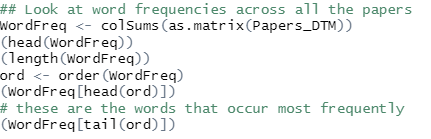
Now that a Document Term Matrix has been created, the data can be explored and various statistical methods can be applied to the words to begin to review patterns in the data set. First, the document term matrix can be explored as well as looking at word frequencies and other details.











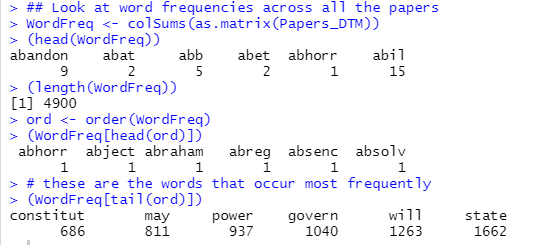


Figure 5: Inspecting the Document Term Matrix

These initial inspections show that there are 4900 words that appear most frequently in the papers and the most frequent words are “constitu” (which is likely words like constitution), “may”, “power”, “govern”, “will” and “state”.

Next, a normalized version of the word matrix is created as show in the following figure.

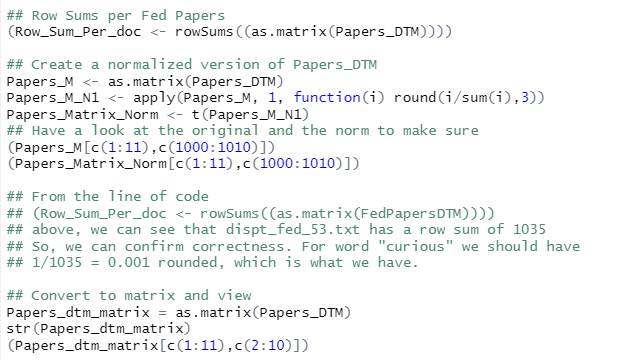
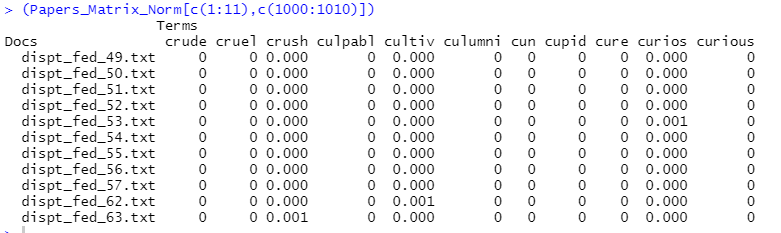
 

Figure 6: Normalizing the Word Matrix

To visualize some of the information obtained about these words by the specific authors. This was done by running “word clouds” for each known author as shown in the following figures.

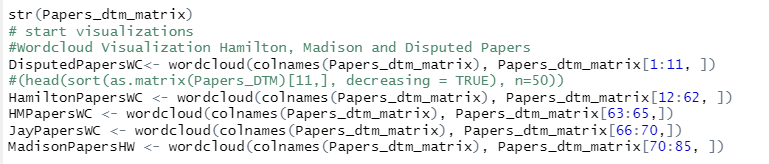


Figure 7: Wordcloud Code for Each Author

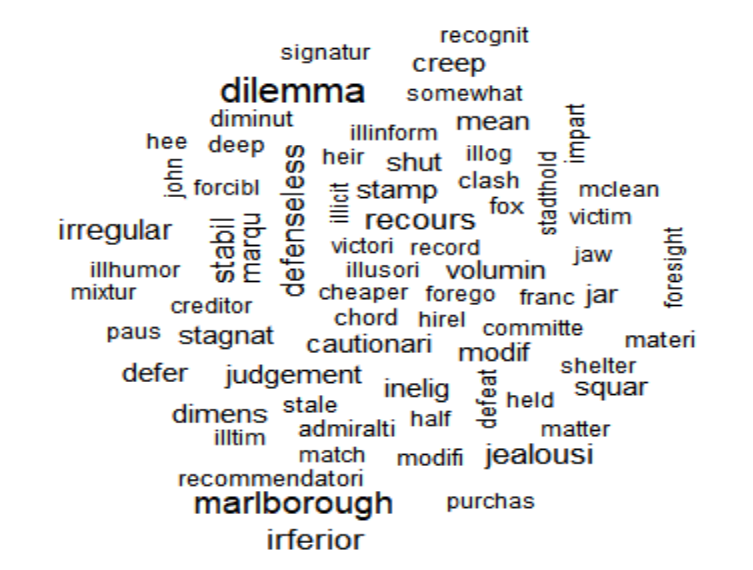


Figure 8: Wordcloud for Disputed Papers

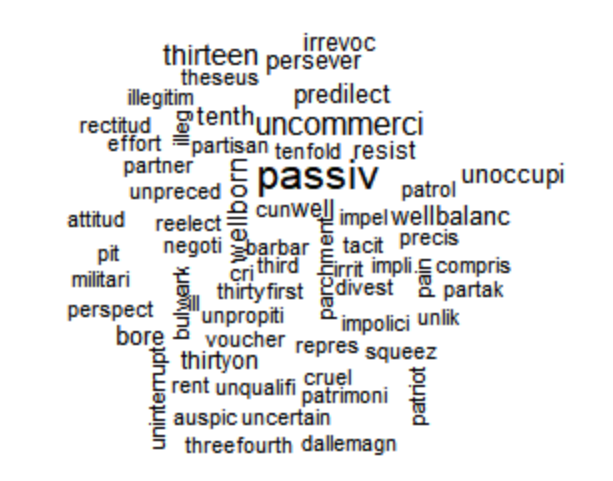


Figure 9: Worldcloud for Hamilton Papers

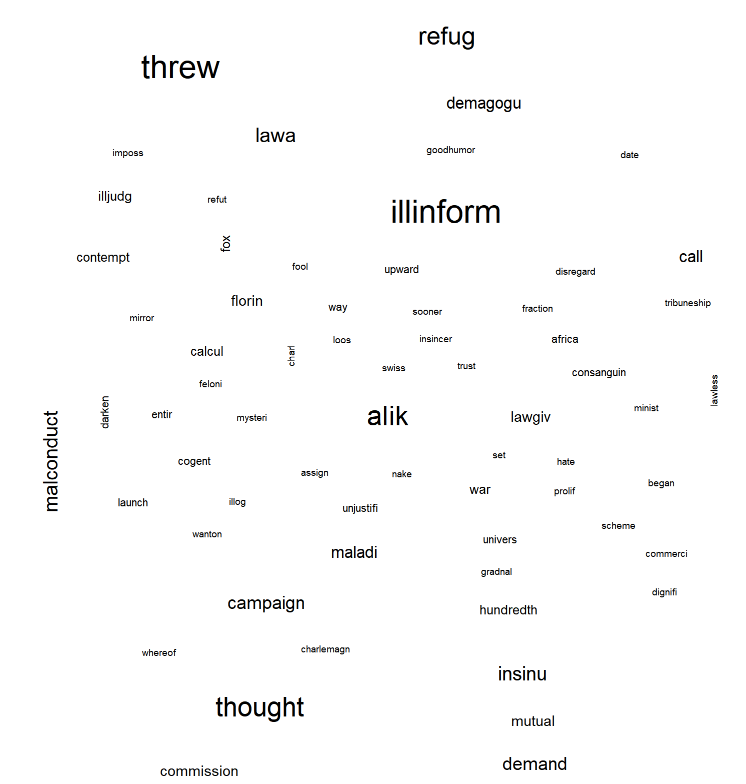


Figure 10: Wordcloud for Hamilton/Madison Papers

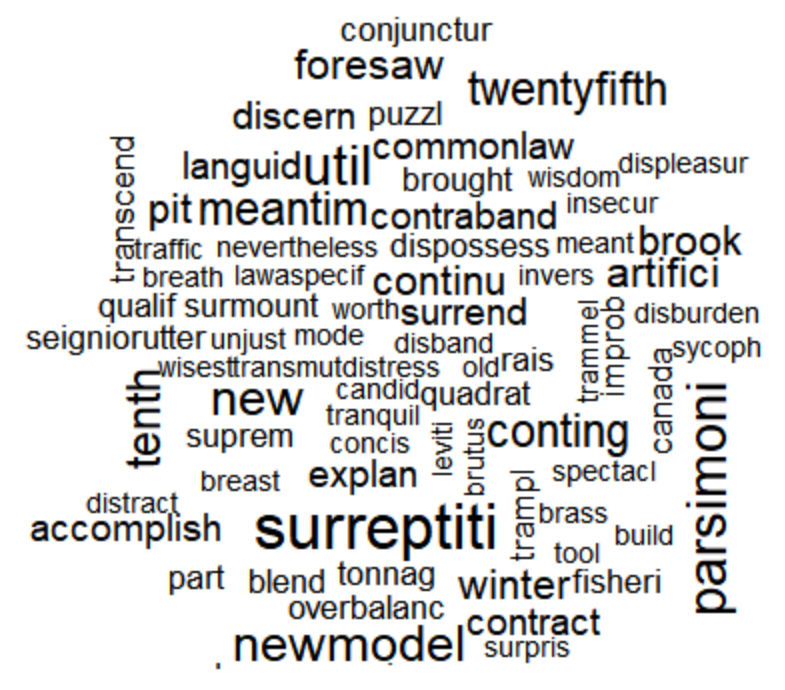


Figure 11: Worldcloud for Jay Papers

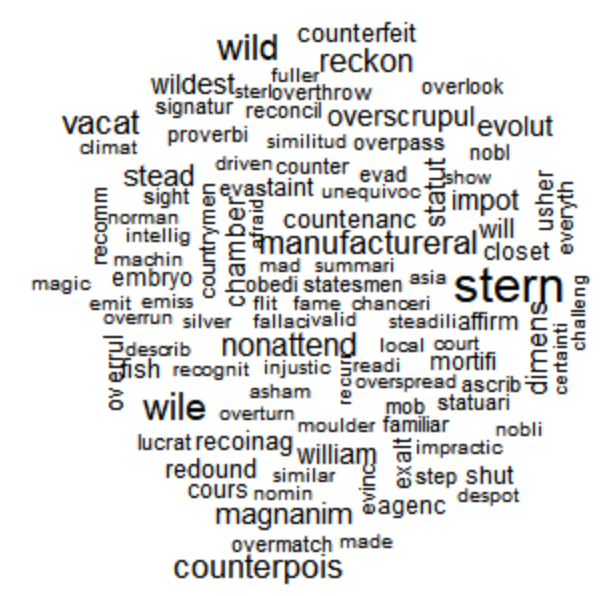


Figure 12: Wordcloud for Madison Papers

The larger words represented in these wordclouds are the ones that appear more often. This type of visualization helps to see what words each author uses most often which will help narrow down the authorship.

### Visualization

In order to study the papers and assist in determining the actual authorship, clustering was used. For this exercise, varying distance types were calculated for the use in the clustering algorithms. The code used can be found in the following figure.

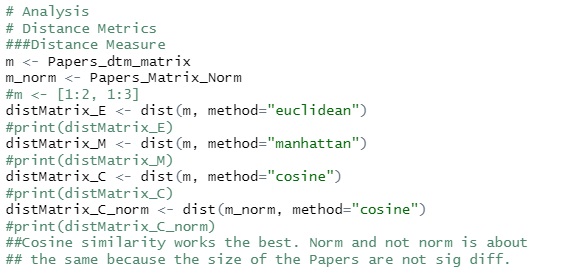


Figure 13: Distance Matrices

This information is then used to plot hierarchical clusters as shown in the code in Figure 14.

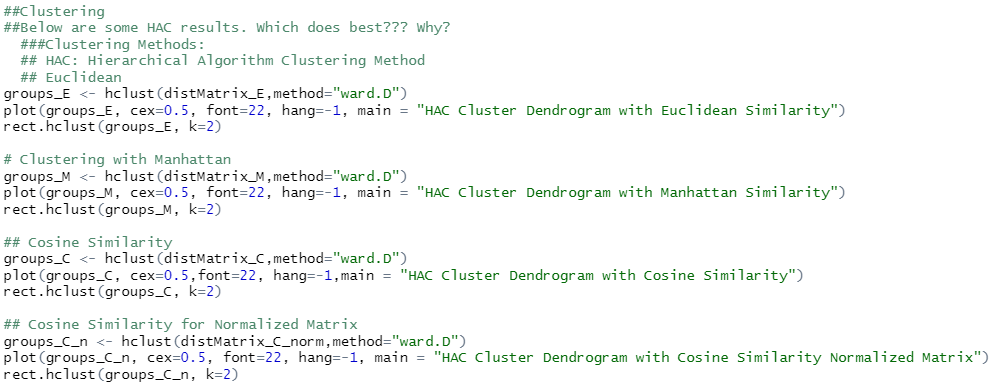


Figure 14: Plotting Functions for Clustering with Different Distance Matrices

As shown, the Euclidean, Manhattan and Cosine distance matrices are obtained. Calculating these matrices provide the input to plots for clustering to show how related texts or articles can be grouped together.

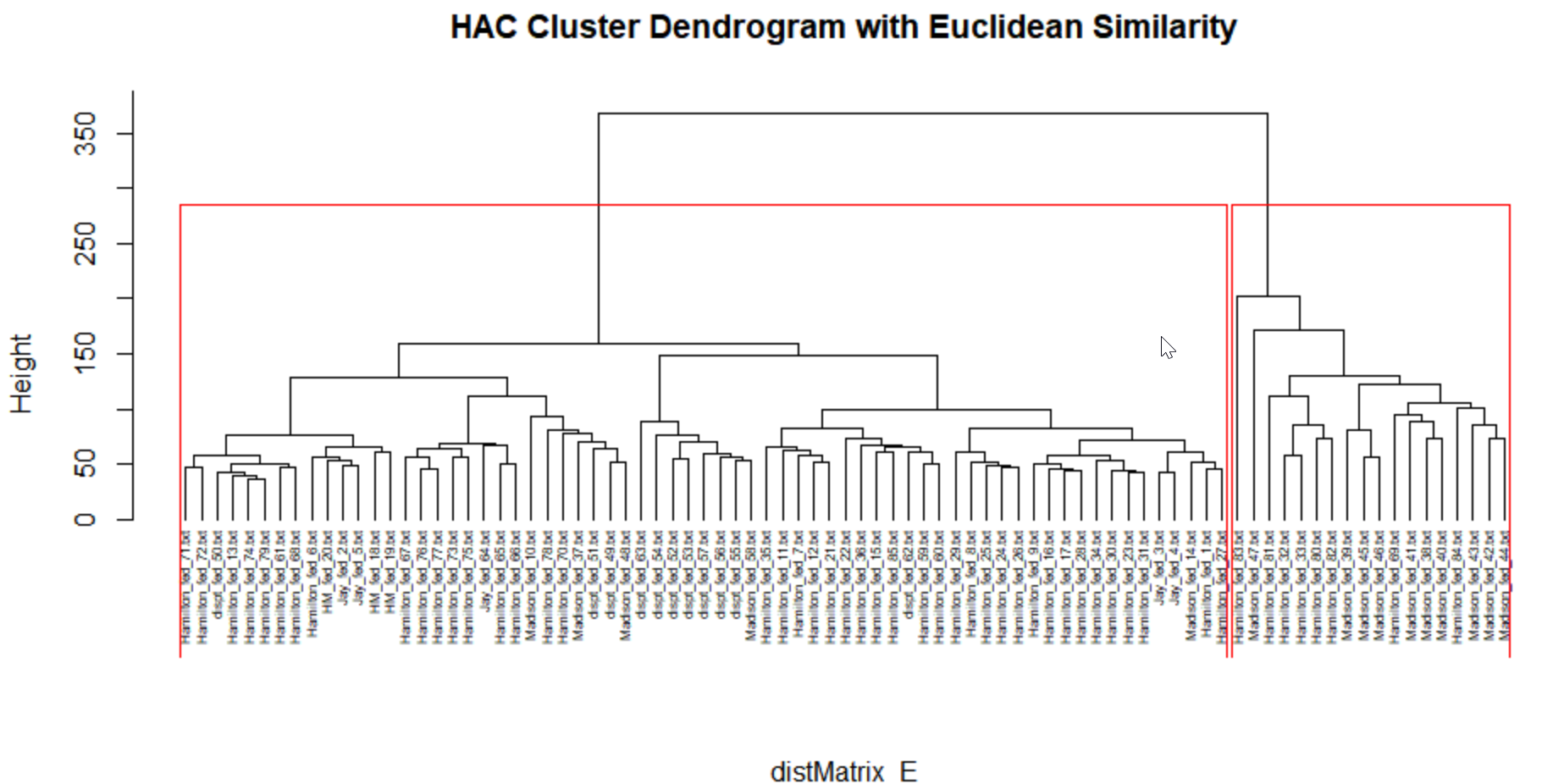


Figure 15: Clustering Using Euclidean Distance

As shown above, there are two major clusters; however, these do not clearly delineate the authors Madison and Hamilton. In fact, the articles are not well separated and many Hamilton and Madison articles are mixed together leaving no direction to the author of the disputed papers even though these fall in the first major cluster.

Unfortunately, using the Cosine distance for hierarchical clustering, as reflected in Figure 16, provides limited benefit. Again, the papers are grouped together and there is no definitive grouping by author, so there cannot be any strong conclusions drawn related to the disputed papers.

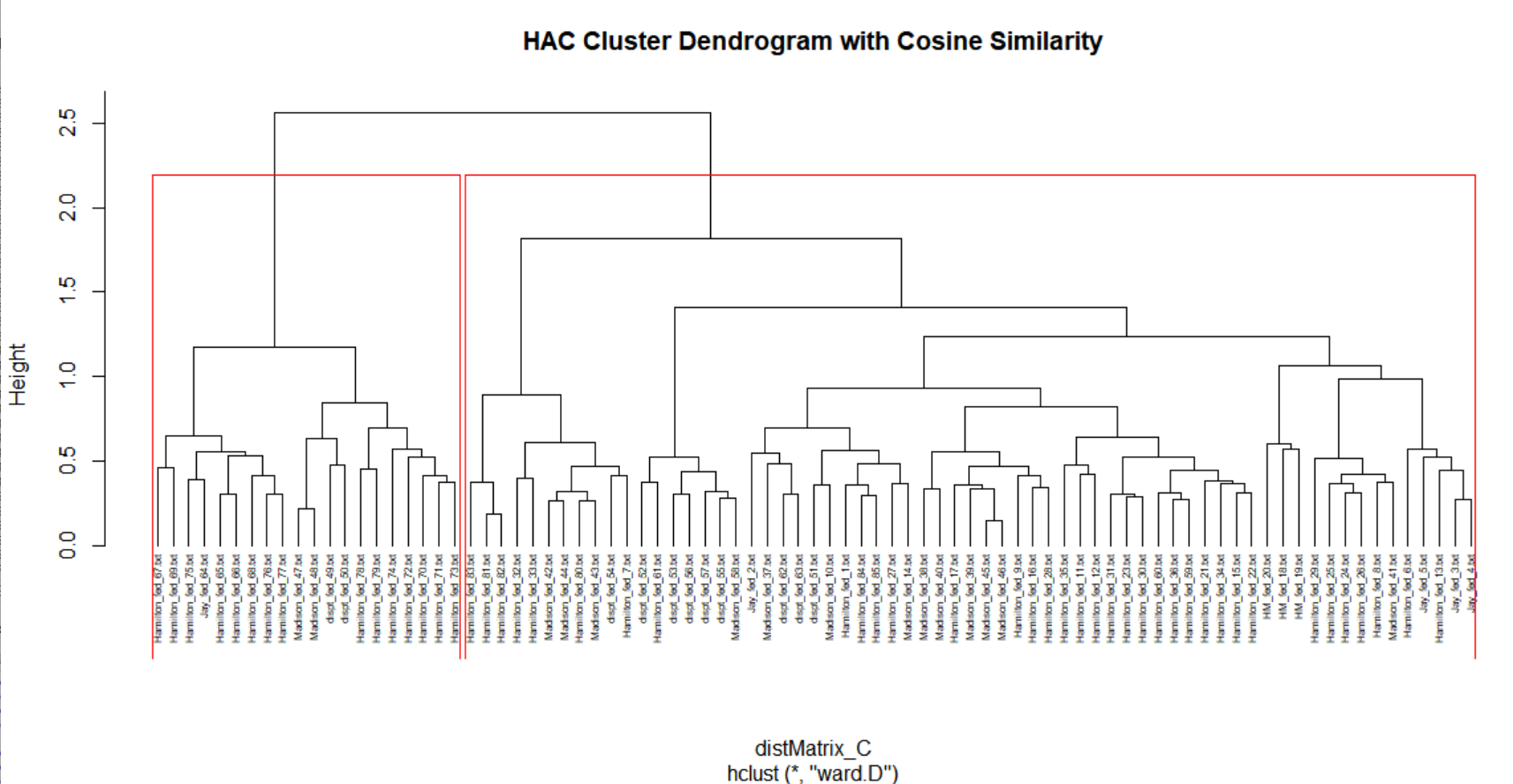


Figure 16: Clustering Using Cosine Distance

However, running the same code with the Manhattan distance, an entirely different result can be seen. As shown above in Figure 17, this approach shows a much more definitive distinction in the clusters with the majority of Madison clusters appearing on the right cluster and the majority of Hamilton and HM (joint authored between Hamilton and Madison) appearing on the left hand cluster. Using this method, the disputed papers show strongly that they were authored by Hamilton.

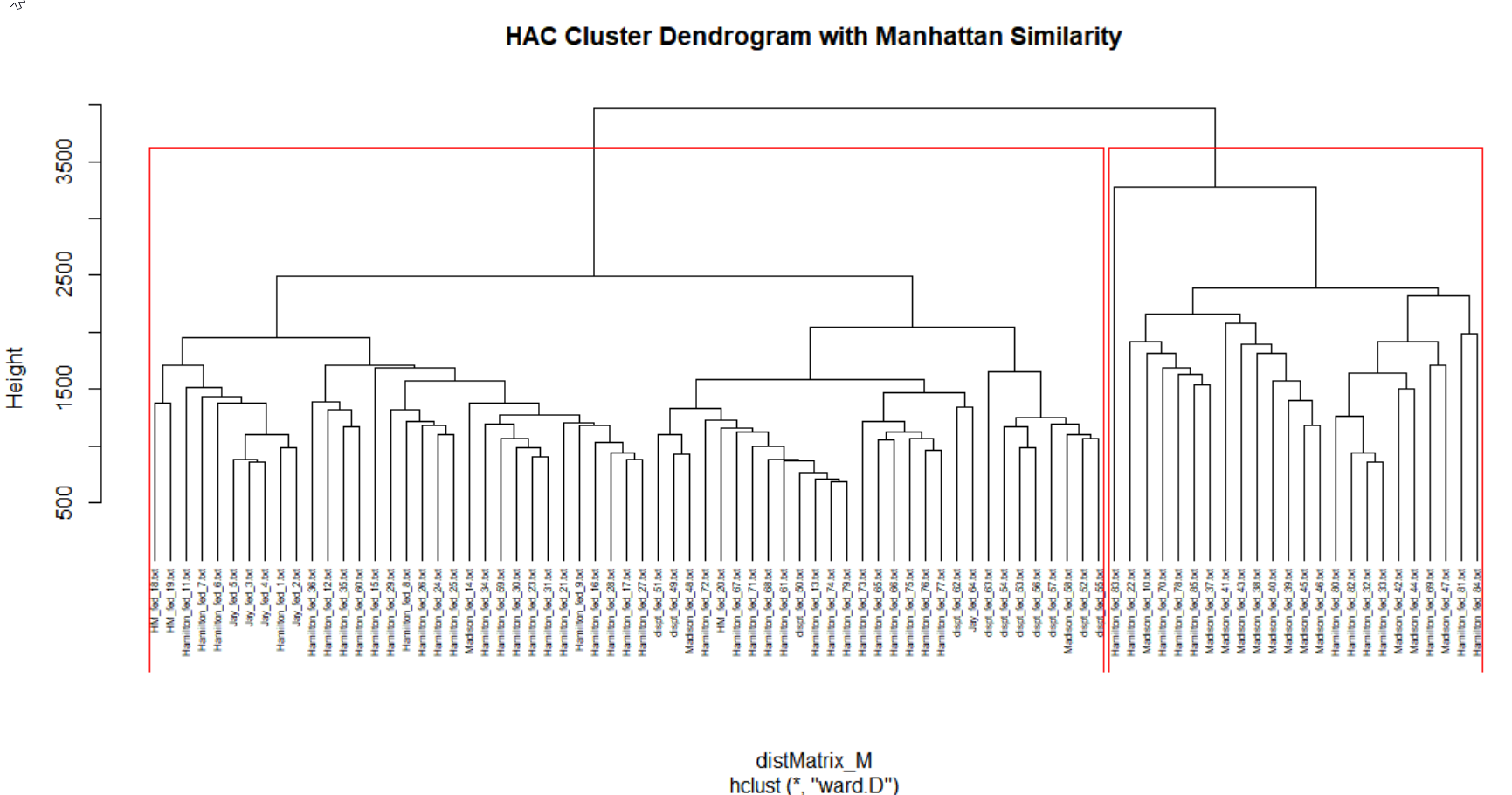
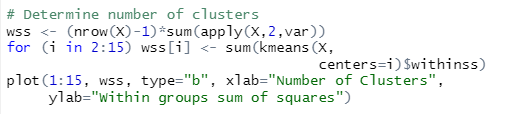


Figure 17: Clustering Using Manhattan Distance

The next step was to review clustering with k-means. One of the most important things to consider when working with k-means is the number of clusters that will be used for the analysis. For this exercise, a plot of the within group sums of squares by number of clusters extracted can help determine the appropriate number of clusters.



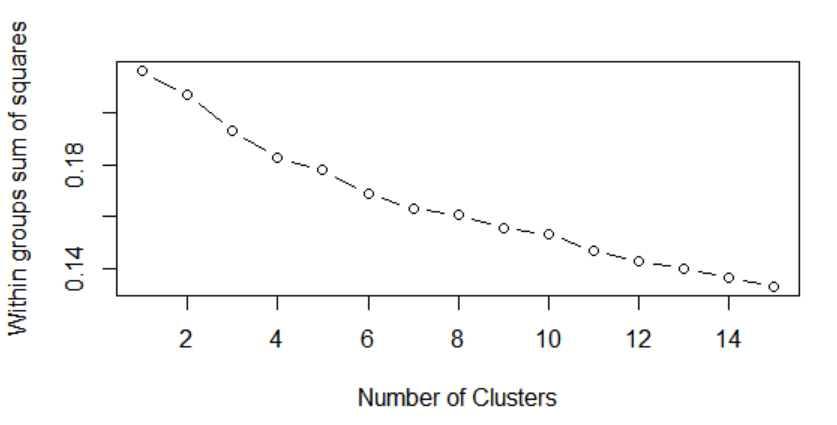


Figure 18: Determining the number of Clusters for k-means

To run the k-means best fit clusters, the k-means function is used within R. The code for k-means can be found in the following figure. Based on the information above, 8 clusters were used for this analysis.

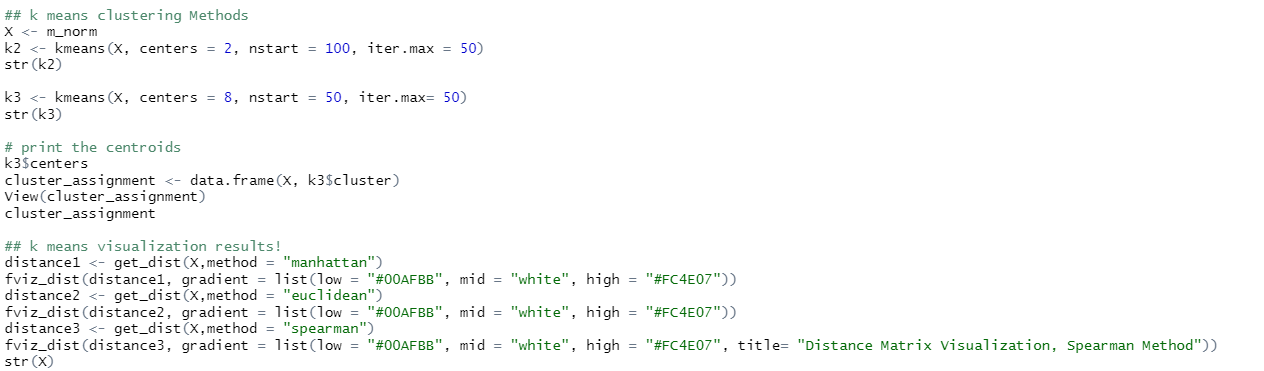


Figure 19: k-means Clustering R-Code

The visualizations of the data as per this code can be found in the following figures.

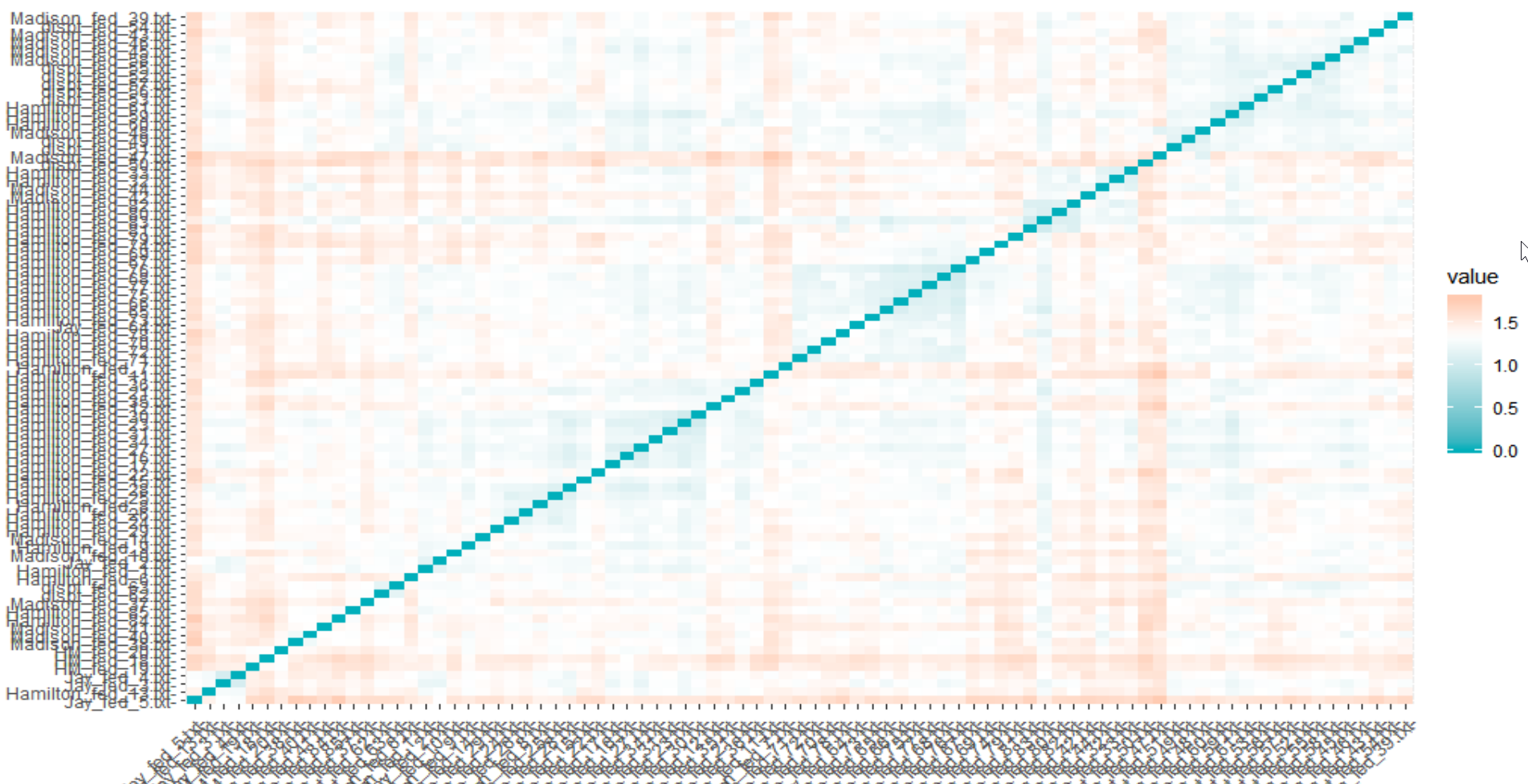


Figure 20: k-means Visualization with Manhattan

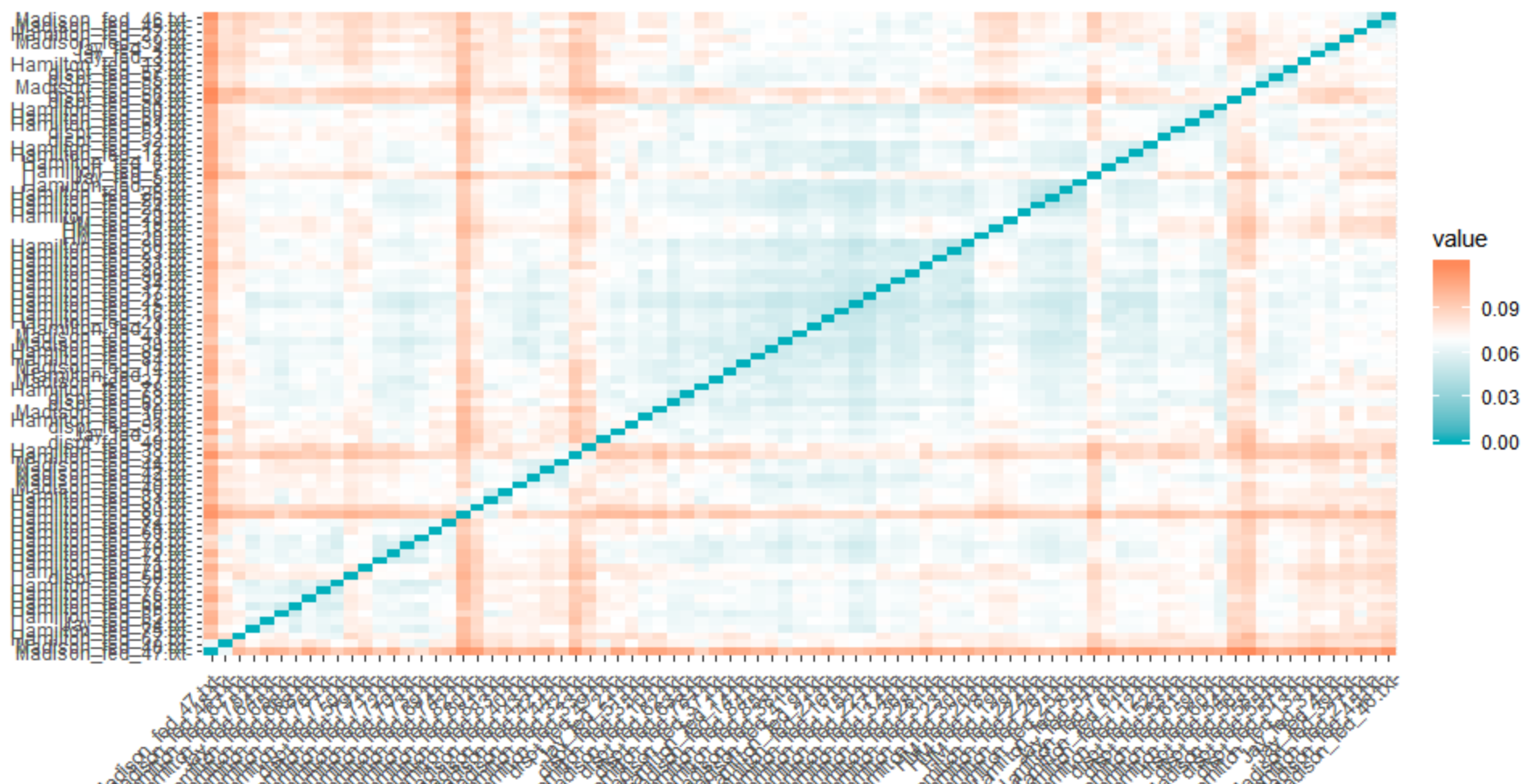


Figure 21: k-means Visualization with Euclidean

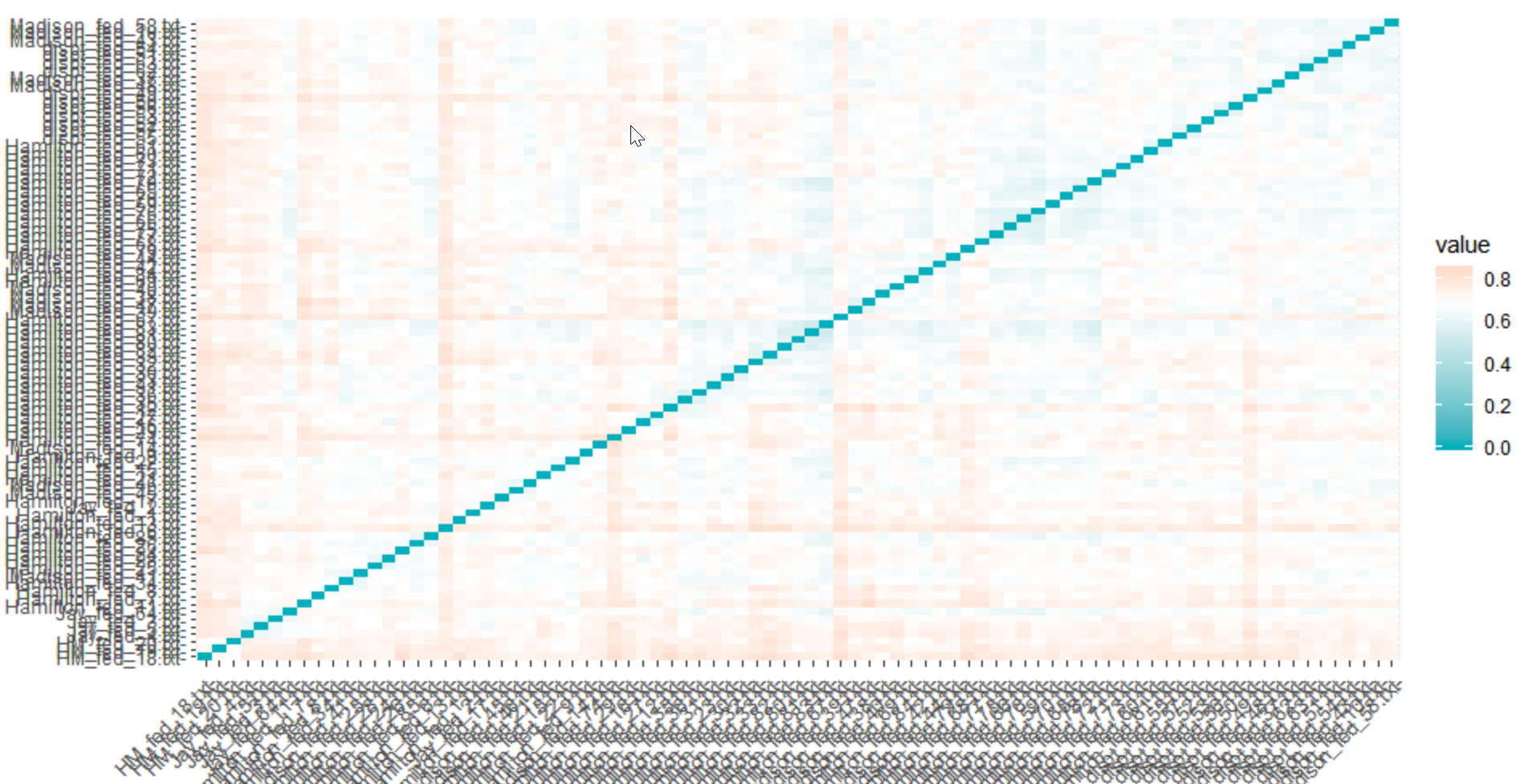


Figure 22: k-means Visualization with Spearman

Unfortunately, due to lack of understanding of these models as well as lack of time, the further analysis of k-means was not completed for this data set.

# Results

As shown in the previous section, the various clustering and other statistical methods helped to determine the authorship of the eleven disputed papers. Clearly, the Manhattan distance was the best distance method for hierarchical clusters.

As per the previous section, the results for the k-means analysis was not completed.

# Conclusions

The only conclusion found in this data set with the limited statistical methods that were understood and run, reflected that using the Manhattan distance method with HAC provided some direction as to the disputed authorship.

Unfortunately, due to lack of time and understanding of this section, all further conclusions were not drawn as the final models and visualizations could not be completed.