**Introduction:**

**General:**

The Federalist Papers were written and circulated in order to persuade New Yorkers to ratify the proposed United States Constitution, which had been drafted in Philadelphia during the summer of 1787. The essays discuss specific aspects of the Constitution in depth in order to advocate for its adoption over the current Articles of Confederation. Originally, intended to consist of 25 documents, The Federalist Papers eventually totaled to 85. Between October 1787 and August 1788, seventy-seven of the essays were published in a series in The Independent Journal, The New York Packet, and The Daily Advertiser[[1]](#footnote-0). Afterwards, the final eight were added. The "Federalist Papers" were initially referred to as "The Federalist.". Because of this, and due to Hamilton, Jay and Madison, the founding fathers of the Constitutional Convention, the Federalists are often consulted today to interpret the purposes of those who drafted the Constitution.

The essays were written by Alexander Hamilton, James Madison, and John Jay the three men collaborated under the name Publius. The authorship of the papers was a closely guarded secret at the time of publication. A list naming Hamilton as one of the writers was not made public until after his death. Hamilton received almost two-thirds of the essays. Madison would later refute many of these claims, claiming that he had written a number of the papers credited to Hamilton. The matters of authorship of the individual essays remained a matter of controversy. That is, not until two mathematicians named Frederick Mosteller and David Wallace decided to spend 3 years combing through Madison's and Hamilton's written works for subtle patterns in word choice—linguistic fingerprints that could be used to pinpoint the contested papers' roots back to their original author.

Mosteller and Wallace, statisticians, studied the frequency distributions of different function terms in the Federalist Papers. Was Hamilton, the sole author of these essays? Was there another author who was the sole author? Or were the essays the product of the three men working together? When Mosteller and Wallace published their findings, they not only solved a long-standing mystery, but they also developed a new mathematical method for analyzing written text. Suddenly, statisticians and computer scientists can be called upon to act as textual detectives. Authorship attribution is now a common problem in data mining.

**Analysis:**

**About the Data:**

First, the libraries for Clustering Analysis were added to the script such as:

library(ggplot2)

library(factoextra)

library(stringr)

library(tidyr)

library(gridExtra)

library(FunCluster)

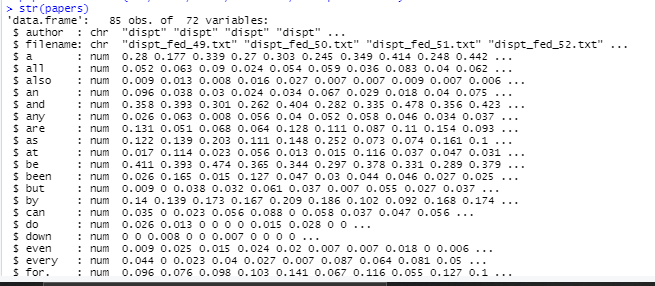
library(rpart)

library(caret)

library(rattle)

Then, the data was read in from the csv file provided and examined using the structure function. Initially, the data looked like Figure 1 below.

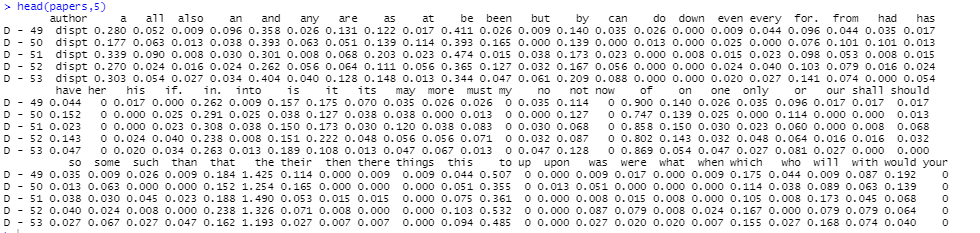
Figure 1: Original Structure of the dataframe



Data Manipulation

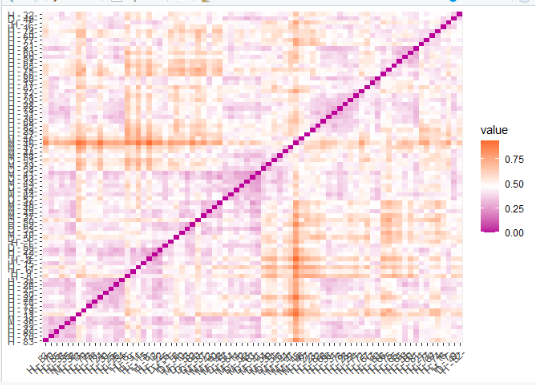
Once the Federalist dataset is loaded some pre-processing steps are used to clean the data. To start, a new column was created with an abbreviation of the author name. Secondly, the file name and file number was split. A new column was thus created to combine the author name along with the file number. Dropped unused levels, unwanted columns and only concerned on the authorship of the disputed articles and only among Hamilton and Madison. Lastly, since there were a few changes to the dataset, reviewed the first top 5 rows as shown in Figure 2.

Figure 2: First 5 rows post manipulation



The Euclidean distance was used to calculate the distance between the vectors and in this analysis the similarity between files was measured.

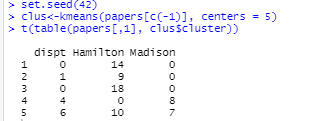
Figure 3: Euclidean plot

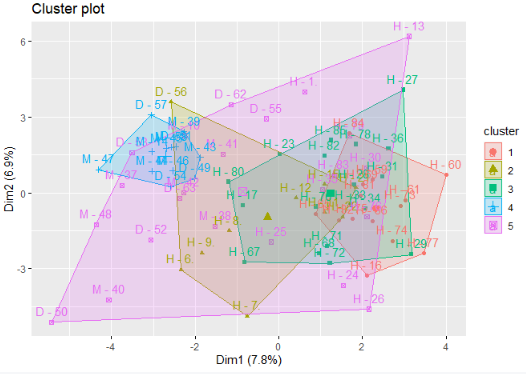


K-Means

Clustering is a technique that groups together objects together in such a way that objects in the same cluster are more similar than objects in other clusters. When solving for the k-means, the aim is to partition the points into k groups such that the sum of squares from points to the assigned cluster centres is minimized. The next step, would be to find the optimal number of clusters to gain a better insight.

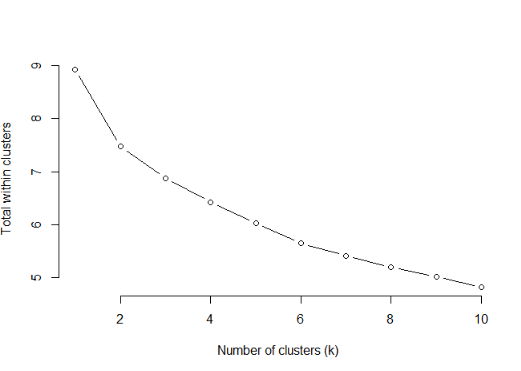
Figure 4: Cluster table and plot of the clusters





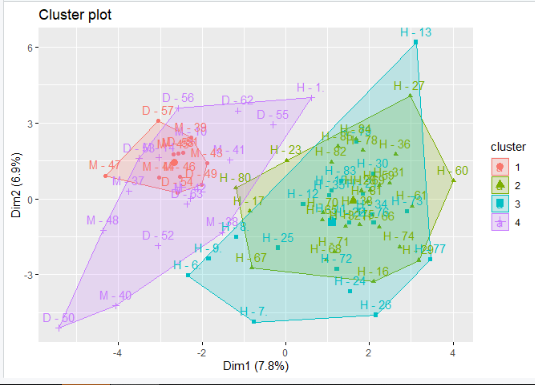
In order, to find the right number of clusters. A function was created to assist in finding the optimal number of clusters by comparing the Number of clusters (k) and the total within clusters-sum of squares as shown in Figure 5.

Figure 5: Number of clusters vs Sum of squares



Next step is to plot the 4 clusters to assess who wrote the disputed articles as shown in Figure 6. The document names were condensed to be more readable while still being unique. The letter “H” represents Hamilton, “M” represents Madison, and “D” is disputed.

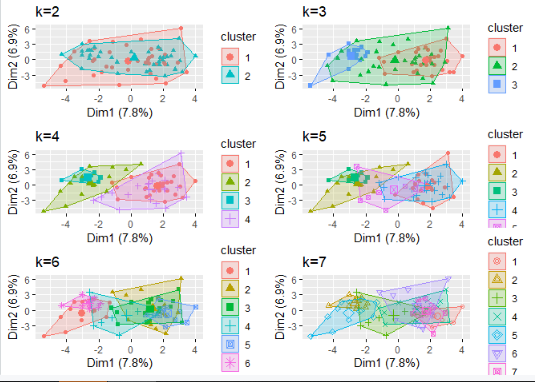
Figure 6:Cluster plot of 4 clusters



Cluster Growth

The next section of the code shows cluster growth and how the cluster formation varies with gradual increase in the number of clusters (Figure 7)

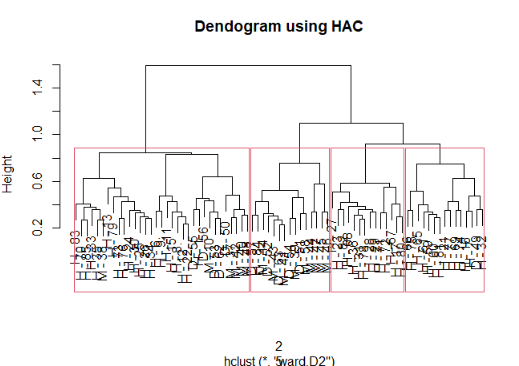
Figure 7: Cluster Growth



Hierarchical clustering

In Hierarchical clustering, the endpoint is a set of clusters, each of which is distinct from the others, and the objects within each cluster are broadly identical to one another. In this cluster analysis, the disputed articles are clustered together with the articles authored by Madison as shown in Figure 8.

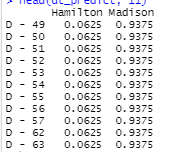
Figure 8:



Decision Tree Algorithm

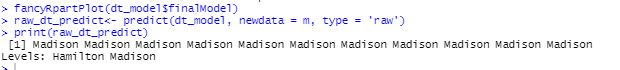
Using the decision tree is to create a training model which can be used to predict class or value of target variables by learning decision rules inferred from prior data. First, the data was split into l (train) and m(test) based on author name. The unused levels were dropped and the decision tree analysis was ran on the l (train) dataset.

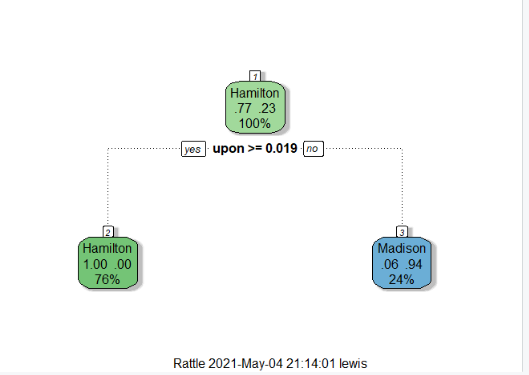
Figure 9: 93.75% probability the disputed articles belonging to Madison



**Model:**

Figure 10: Predicting Model which reconfirms that the disputed articles are authored by Madison

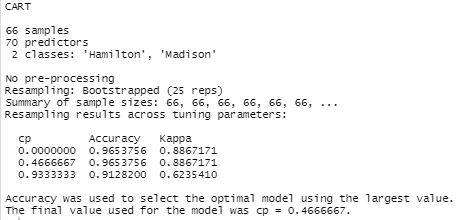




**Results:**

From the Euclidean plot (Figure 3), the files intersecting at the purple point are quite close suggesting similarities and the ones at the red (high) section are not. In Figure 4, the following code below, shows how the disputed articles have been spread across the authors. The reason being, the usage of clusters. Figure 5 concludes that 4 is the right number of clusters for this dataset. The cluster plot shown in Figure 6 suggests that the author of the disputed articles is most likely Madison. The plot above shows Hamilton occupying cluster 2 and 3. Madison occupying cluster 1 and 4. Using the Decision Tree Algorithm, the probability the disputed articles belong to Madison is 93.75%.

Figure 11: Model Prediction



**Conclusion:**

The essays were co-written by Alexander Hamilton, James Madison, and John Jay, who went by the alias Publius. At the time of print, the authorship of the papers was a tightly guarded secret. The matters of authorship of the individual essays remained a matter of controversy. That is, not until two mathematicians named Frederick Mosteller and David Wallace decided to join in on the controversy and try to solve the mystery of who wrote the disputed papers. The frequency distributions of different function words in the Federalist Papers were analyzed by statisticians Mosteller and Wallace. Was Hamilton the only person who wrote these essays? Is there any person who wrote the essays by themselves? Or were the essays the product of collaboration between the three men?

The evidence indicates that Madison was the author of the contested papers after reviewing the findings gathered. Madison’s total essays authored are 29. Hamilton, in turn, wrote 51 essays, with Jay providing the remaining 5. Despite the fact that Hamilton is the most prolific author of the three, this assertion significantly extends Madison's contribution to the papers, making this a significant historical moment. This isn't the first time The Federalist Papers have been studied in order to solve this puzzle. The essays have played a significant role in American history, both in the past and present. Identifying the writers of each of these seminal works will reveal essential truths about the founding fathers' values and achievements.

Some people were concerned that the Constitution would grant the federal government too much control. Not only did people think that the government would have too much power but also restrict individual liberty. To help ease these concerns, Hamilton, Jay, and Madison studied the Constitution in depth. The authors explained the built-in checks and balances intended to balance power. The Federalist Papers are a valuable compilation today. The papers are valuable because they provide insight into the intentions of key individuals who discussed the elements of the Constitution. Thank you Mosteller and Wallace for developing a new mathematical method for analyzing written text.

1. Dan T. Coenen, “Fifteen Curious Facts About the Federalist Papers.” [University of Georgia School of Law](https://digitalcommons.law.uga.edu/cgi/viewcontent.cgi?article=1001&context=fac_pm), April 1, 2007. [↑](#footnote-ref-0)