Brandon Croarkin

# HW #4 – Use Clustering to Solve a Mystery in History

## Introduction

The Federalist Papers were a series of 85 essays urging citizens of New York to ratify the new United States Constitution. They were written by Alexander Hamilton, James Madison, and John Jay and originally appeared in New York newspapers in 1787 and 1788 under the pseudonym “Publius”. They are considered one of the most important sources for interpreting and understanding the original intent of the Constitution.

It was not until the 1818 edition published by the printer Jacob Gideon that the authors of each essay were identified by name. However, while the authorship of 73 of *The Federalist* essays is fairly certain, 12 of these essays are still disputed over by some scholars. Hamilton initially wrote to claim the authorship before he was killed in a duel by Aaron Burr. While Madison did not immediately dispute Hamilton’s list, he later claimed authorship and claimed the difference in the lists were due to the hurry in which Hamilton’s memorandum was made.

Statistical analysis has been undertaken on several occasions to try to ascertain the authorship question based on word frequencies and writing styles. For instance, in the 1960s, statisticians Mosteller and Wallace analyzed the frequency distribution of common function words in the Federalist Papers and drew their conclusions. This paper will undergo a similar analysis using clustering techniques such as k-Means and HAC to reach a conclusion.

## Analysis

### The Data

This analysis will work with the Federalist Paper data set. The features are a set of “function words”, for example, “upon”. The feature value is the percentage of the word occurrence in an essay. For example, for the essay “Hamilton\_fed\_31.txt”, if the function word “upon” appeared 3 times, and the total number of words in this essay is 1000, the feature value is 3/1000=0.3%.

In the author column, there are 74 essays with identified authors: 51 essays written by Hamilton, 15 by Madison, 3 by Hamilton and Madison, 5 by Jay. The remaining 11 essays, however, are listed as “dispt”. These are the famous essays with disputed authorship.

Below are a series of wordclouds that examine the most frequent “function words” in each group.

|  |  |  |
| --- | --- | --- |
| **Disputed:** | **Hamilton** | **Madison** |
|  |  |  |

|  |  |
| --- | --- |
| **Jay** | **HM** |
|  |  |

### Data Preparation

Much of the data preparation needed for this analysis has already been done by the statisticians Mosteller and Wallace. This includes the transformation of the words and their counts in each essay into a variable with the percentage of the word occurrence in the essay. By doing this the words have been normalized so that they are all on the same scale and the length of the essay does not influence the word frequency. The data in this paper also just includes “function words”, such as “upon”. By just looking at the “function words”, the content of the paper is ignored so the focus can be solely on the writing style of the various authors.

No missing data was found in the dataset.

Some additional data preparation steps included creating various new data frames for the analysis. This includes a new data frame with just the attribute values (dataNoLabels), a data frame with just the papers from Hamilton, Madison, and the disputed essays (fedPapersHMD), and lastly a data frame with just the attributes from fedPapersHMD (dataNoLabelsHMD).

Since most historical records suggest that the disputed essays are either from Hamilton or Madison, it made sense to make a new data frame with just their essays and the disputed ones to help focus the cluster analysis on their two writing styles. The data frames with just the attribute values are the ones used for the cluster analysis since the author and filename fields would get in the way of clustering by the attributes.

### Processing

The analysis in this paper will use two clustering algorithms: k-Means and HAC. Cluster analysis groups data objects based only on information found in the data that describes the objects and their relationships. The goal is that the objects within a group are similar to one another and different from objects in other groups. This approach will use the attributes in the dataset (word frequencies) to see which essays have similar writing styles, which can be used to determine the author of the disputed essays.

K-means is a prototype-based, partitional clustering technique that attempts to find a user-specified number of clusters (K), which are represented by their centroids. It works by first selecting a number of clusters (K). Each point is then assigned to the closest centroid, and each collection of points is assigned to a centroid is a cluster. This analysis will use the mean of the points as the centroid. The centroid of each cluster is then updated based on the points assigned to the cluster. This process is repeated until no point changes clusters, the centroids remain the same, or a maximum number of iterations is met.

For the K-means algorithm, the approach taken was to run the algorithm three times with different randomized starting locations on both the full and subset attribute datasets. This was done to mitigate the effects of a bad centroid starting location affecting the algorithm output. For each iteration, the cluster centers were computed, the cluster assignments were made, tables were created to mark the cluster assignment for each essay by each author, and multiple plots were created to visualize the results. For the full dataset, the results were clustered into 4 clusters whereas the subset dataset was clustered into 2 clusters. In each trial, the disputed papers were determined to be part of the cluster that they were most frequently marked into. Whatever author was the strongest in that cluster was determined to be the author of the disputed papers. By aggregating the results of each trial, the disputed papers can be determined to be authored by the author(s) that was most frequently clustered with the disputed papers.

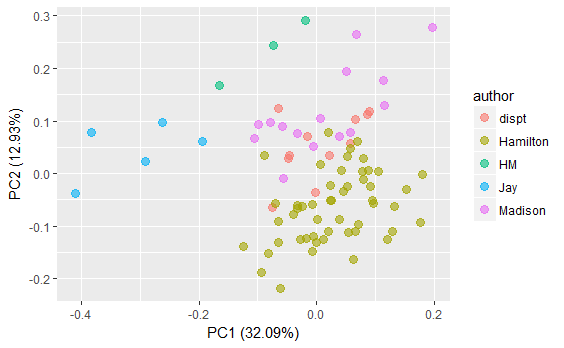
Hierarchical agglomerative clustering (HAC) refers to a collection of clustering techniques that produce a hierarchical clustering by starting with each point as a singleton cluster and then repeatedly merging the two closest clusters until a single, all-encompassing cluster remains. The proximity between two clusters can be defined in a few ways: MIN, MAX, and group average. MIN defines cluster proximity as the proximity between the closest two points that are in different clusters. MAX takes the proximity between the farther two points in different clusters to be the cluster proximity. The group average approach defines cluster proximity to be the average pairwise proximities.

There are also multiple ways to compute the distance between two points. In this analysis, three distance measures are used: Euclidean, Manhattan, and Canberra distance. When points are plotted in a vector space representation, the Euclidean distance measures the straight line between points. This measure tends to perform worse with more dimensions since the distance between any two points is large in high dimensional data. The Manhattan distance takes the sum of the absolute values of the difference in coordinates. The Canberra distance is a weighted version of the Manhattan distance. It examines the sum of series of a fraction differences between coordinates of a pair of objects. It is meant to be a measure of the similarity and dissimilarity between groups.

For the HAC algorithm, the approach taken was to create three different distance matrixes (Euclidean, Manhattan, and Canberra) for both the full and subset dataset. The algorithm was then run using each on these distance matrixes with complete linkages, average linkages, and single linkages. For the full dataset, the dendrogram was cut to create 4 clusters, whereas for the subset dataset, the dendrogram was cut to create two clusters.

## Results

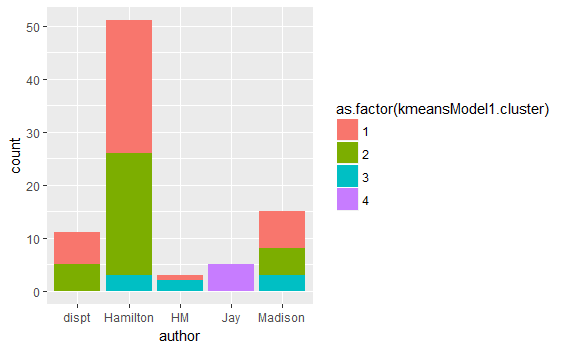
An initial plotting of the points (shown below) using principal component analysis appears to show that the disputed papers lie closest to the Madison data points. The Jay papers seem to be distinct in their own region towards the left-side of the plot. The joint Hamilton and Madison papers are clustered in the middle towards the top of the plot.



**K-Means (Full Dataset) – Trial 1 🡪 Hamilton**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Author** | **Cluster 1** | **Cluster 2** | **Cluster 3** | **Cluster 4** |
| **Disputed** | 0.55 | 0.45 | 0.00 | 0.00 |
| **Hamilton** | 0.49 | 0.45 | 0.06 | 0.00 |
| **HM** | 0.33 | 0.00 | 0.67 | 0.00 |
| **Jay** | 0.00 | 0.00 | 0.00 | 1.00 |
| **Madison** | 0.47 | 0.33 | 0.20 | 0.00 |

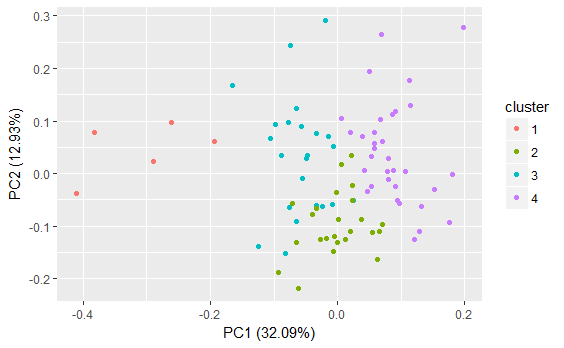
The first K-means model has the disputed papers mostly clustered into cluster 1, where both Hamilton and Madison papers seem to be clustered. Hamilton has the stronger presence in cluster 1, so the disputed papers seem most strongly linked to Hamilton.



**K-Means (Full Dataset) – Trial 2 🡪 Hamilton and Madison**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Author** | **Cluster 1** | **Cluster 2** | **Cluster 3** | **Cluster 4** |
| **Disputed** | 0 | 0.18 | 0.45 | 0.36 |
| **Hamilton** | 0 | 0.43 | 0.16 | 0.41 |
| **HM** | 0 | 0 | 1 | 0 |
| **Jay** | 1 | 0 | 0 | 0 |
| **Madison** | 0 | 0 | 0.47 | 0.53 |

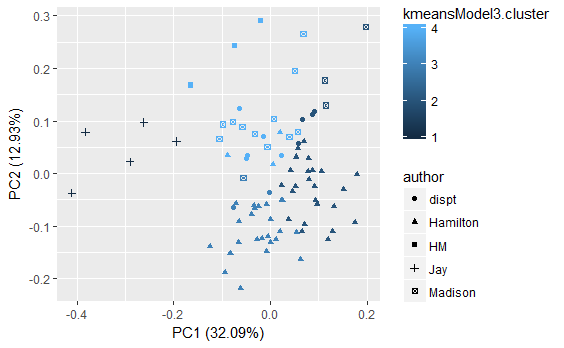
The second K-means model has the disputed papers mostly clustered into cluster 3, where the joint Hamilton and Madison papers seem to be clustered. This trial seems to suggest that the disputed papers were a joint writing from Hamilton and Madison.



**K-Means (Full Dataset) – Trial 3 🡪 Madison**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Author** | **Cluster 1** | **Cluster 2** | **Cluster 3** | **Cluster 4** |
| **Disputed** | 0 | 0.36 | 0.18 | 0.45 |
| **Hamilton** | 0 | 0.47 | 0.47 | 0.06 |
| **HM** | 0 | 0 | 0 | 1 |
| **Jay** | 1 | 0 | 0 | 0 |
| **Madison** | 0 | 0.2 | 0.07 | 0.73 |

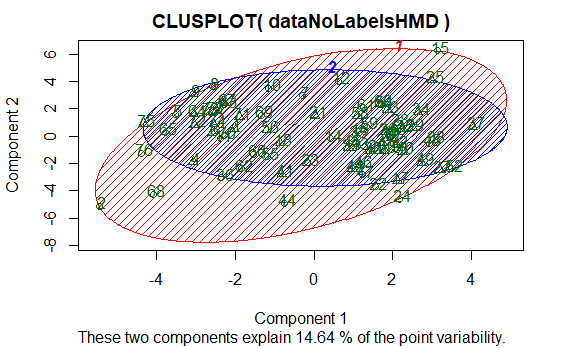
The third K-means model has the disputed papers mostly clustered into cluster 4, where the Madison papers seem to be clustered. The disputed papers thus seem most strongly linked to Madison.



**K-Means (Subset Dataset) – Trial 1 🡪 Unknown**

|  |  |  |
| --- | --- | --- |
| **Author** | **Cluster 1** | **Cluster 2** |
| **Disputed** | 0.55 | 0.45 |
| **Hamilton** | 0.49 | 0.51 |
| **Madison** | 0.47 | 0.53 |

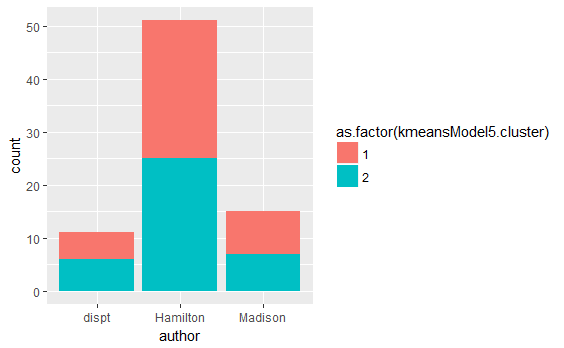
The first K-means model with the subset data has the disputed papers mostly clustered into cluster 1, where both Hamilton and Madison papers have high presence, but they are still in cluster 2. There is no clear result gained from this model.



**K-Means (Subset Dataset) – Trial 2 🡪 Unknown**

|  |  |  |
| --- | --- | --- |
| **Author** | **Cluster 1** | **Cluster 2** |
| **Disputed** | 0.45 | 0.55 |
| **Hamilton** | 0.51 | 0.49 |
| **Madison** | 0.53 | 0.47 |

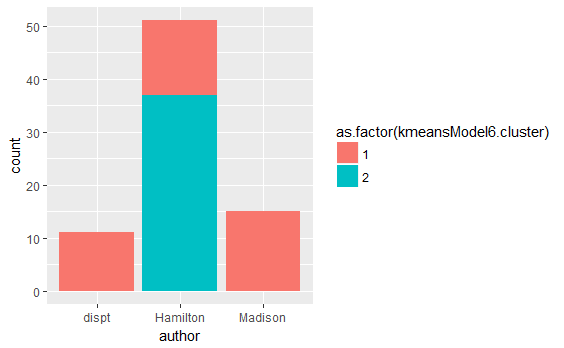
The second K-means model with the subset data has the same results of the first K-means model, but with the cluster assignments reversed. Again, there is no clear result gained from this model.



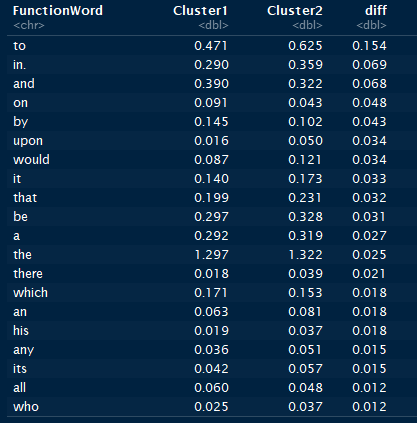
**K-Means (Subset Dataset) – Trial 3 🡪 Madison**

|  |  |  |
| --- | --- | --- |
| **Author** | **Cluster 1** | **Cluster 2** |
| **Disputed** | 1 | 0 |
| **Hamilton** | 0.27 | 0.73 |
| **Madison** | 1 | 0 |

The third K-means model has the disputed papers mostly clustered into cluster 1, where the Madison papers are strongly clustered here. Hamilton is strongly in cluster 2. This is the strongest K-means model result and has the disputed papers linked with Madison.



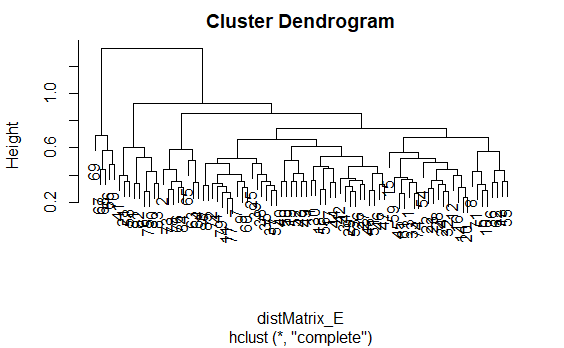
By examining the cluster centers for these two clusters, we can see the words that most affected the clustering into each cluster. For each function word, the average word occurrence was aggregated for both clusters. By comparing the absolute difference in this number between the two clusters and then examining the top 20 words by size difference, we can find the words that had the biggest impact in creating the cluster. From the table below, it appears that “to”, “in”, and “and” were the three words that had the biggest impact in creating the clusters with the rest of the list visible below.



**HAC (Full Dataset) – Euclidean Distance – Complete Linkage 🡪 Unknown (edge Hamilton)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Author** | **Disputed** | **Hamilton** | **HM** | **Jay** | **Madison** |
| Cluster 1 | 9 | 48 | 0 | 0 | 6 |
| Cluster 2 | 1 | 0 | 3 | 0 | 4 |
| Cluster 3 | 1 | 3 | 0 | 0 | 5 |
| Cluster 4 | 0 | 0 | 0 | 5 | 0 |

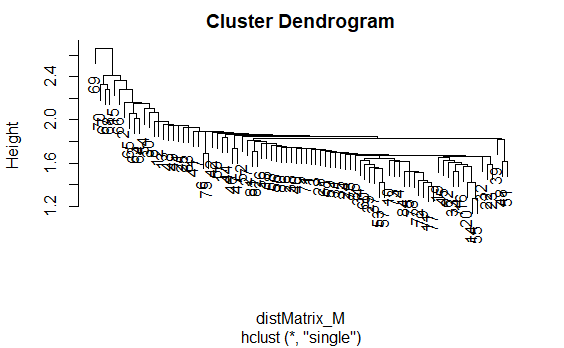
The first HAC model with Euclidean distance and complete linkage has the disputed papers mostly clustered into cluster 1, where both the Hamilton and Madison papers are clustered. Hamilton is strongly in cluster 1, whereas the Madison papers are in cluster 1, but with a weak association. This model seems to lean towards Hamilton but cannot make a clean association.



**HAC (Full Dataset) – Manhattan Distance – Single Linkage 🡪 Unknown**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Author** | **Disputed** | **Hamilton** | **HM** | **Jay** | **Madison** |
| Cluster 1 | 10 | 50 | 0 | 0 | 15 |
| Cluster 2 | 1 | 0 | 3 | 0 | 0 |
| Cluster 3 | 0 | 1 | 0 | 0 | 0 |
| Cluster 4 | 0 | 0 | 0 | 5 | 0 |

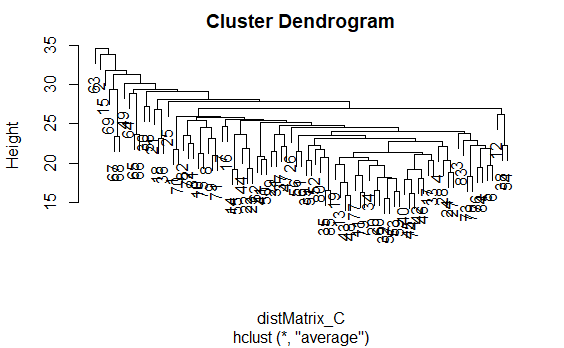
No clear authorship can be determined from the HAC model with Manhattan distance and single linkage as the disputed, Hamilton, and Madison essays are all grouped in the same cluster.



**HAC (Full Dataset) – Canberra Distance – Average Linkage 🡪 Unknown**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Author** | **Disputed** | **Hamilton** | **HM** | **Jay** | **Madison** |
| Cluster 1 | 10 | 50 | 2 | 5 | 15 |
| Cluster 2 | 1 | 0 | 0 | 0 | 0 |
| Cluster 3 | 0 | 1 | 0 | 0 | 0 |
| Cluster 4 | 0 | 0 | 1 | 0 | 0 |

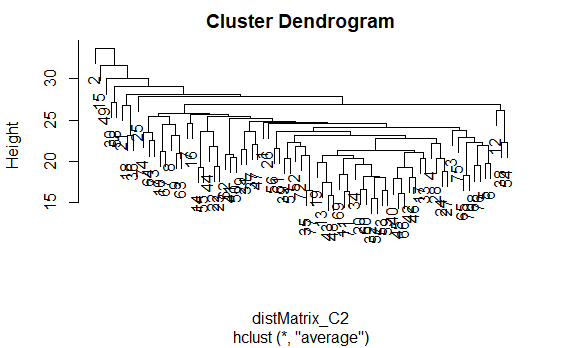
No clear authorship can be determined from the HAC model with Canberra distance and average linkage as the all the authors are grouped in the same cluster.



**HAC (Subset Dataset) – Euclidean Distance – Complete Linkage 🡪 Unknown**

|  |  |  |  |
| --- | --- | --- | --- |
| **Author** | **Disputed** | **Hamilton** | **Madison** |
| Cluster 1 | 10 | 48 | 10 |
| Cluster 2 | 1 | 3 | 5 |

No clear authorship can be determined from the HAC model on the subset data with Euclidean distance and complete linkage as both Hamilton and Madison are grouped in the same cluster.



The additional HAC algorithms run are not shown since they did not give any meaningful results.

## Conclusions

From the analysis shown above, no clear author of the disputed papers can be determined, although the results tend to lean towards Madison being the author. The initial PCA plot shows the disputed papers being most closely associated with Madison. Additionally, there were 2 models that clustered Madison with the disputed papers, compared to 1 that associated Hamilton with the disputed papers, with the remaining models often clustering the two authors together and reaching an inconclusive result.

Additional analysis of the Federalist Papers would be needed to help reach a stronger result. Some additional approaches to be considered include a classification analysis and inclusion of all words in the papers, not just the “function words”. The classification analysis would appear to be better suited to this analysis since the supervised learning approach could use the prior knowledge of the essay authors to reach the conclusion. Using all the words in the essay, in either a classification or clustering approach, could also help give additional information that would help to cluster/classify the authors. For instance, if different authors discussed different aspects of the U.S. Constitution that would be able to help cluster/classify the results.