IST 707: Homework 4

**Introduction**

Although the United States (US) famously declared its independence in 1776, at that time it was merely a collection of 13 colonies. These colonies did not officially become part of the US until they chose to adopt the proposed US Constitution, which was written over a decade later in 1787 (National Constitution Center). The Constitution is one of the most famous written documents in history, however, it was highly controversial when it was first proposed.

Those in favor of ratifying the Constitution were called Federalist and those in opposition were the Anti-Federalist (National Constitution Center). Proponents from each party often plead their case with their respective colony by publishing essays in local newspapers (About the Federalist Papers). Perhaps the most famous collection of such essays is the Federalist Papers—a collection of 85 essays published in New York’s newspapers under the pen name “Publius” persuading New Yorkers to ratify the Constitution (About the Federalist Papers).

Today, “Publius” is believed to have been the literary persona of a collection of famed Federalist supporters (About the Federalist Papers). Historians attribute 74 of the essays’ authorship to a combination of Alexander Hamilton (51 sole-author, 3 co-author), John Jay (5 sole-author), and James Madison (15 sole-author, 3 co-author) (The Federalist Papers). However, the exact authorship of the remaining 11 essays is still widely disputed. The Library of Congress has elected to sponsor a study in order to determine the most likely author of these disputed essays.

**Analysis and Models**

**About the Data**

The original 85 Federalist Paper essays were used for this study. Each essay includes the body of text in addition to the author (if known). For the essays with a disputed author, “Disputed” was used as the author.

These documents were compiled as a corpus of text files with the primary indicator of authorship being the filenames. The nomenclature for the original filenames was as follows:

*[ Author ]\_[ fed] \_[ paper # ].txt :* (i.e. Hamilton\_fed\_1.txt)

In order to improve code readability and future visualizations, these filenames were modified to reflect only the author and paper # (i.e. Hamilton\_1).

The corpus of Federalist Papers was compiled into a document term matrix for further analysis. To eliminate “noise” within the data, stopwords were removed along with all punctuation, numbers, and separators. Additionally, all words were made into lower case, and limited to only those between 4 and 15 characters in length. The remaining data consisted of rows with filename labels and the frequencies of words occurring in the respective Federalist Paper.

Figure 1 (below) is a word cloud of the top 30 most frequent words among the 11 disputed Federalist Papers.

**Figure 1**

A screenshot of a cell phone

Description automatically generated

Initial observation of the data showed that the essays varied in length, so to support comparison of the document data within the document term matrix, each document’s word frequency was divided by the document’s total word count. This produced normalized frequencies of words across documents to eliminate document length bias in future analysis.

**Clustering Models**

**Hierarchical Clustering Model with Euclidean Distance**

Initially, a hierarchical clustering model using Euclidean distance was applied to the document term matrix. This model attempts to minimize the Euclidean distance between points in a cluster. This model will group documents in the term matrix in clusters with the shortest “straight-line” distance between each other when graphing the document term similarity. This model requires a predetermined number of clusters to bin the data to, so for the purpose of this analysis 4 clusters were chosen.

**Hierarchical Clustering Model with Cosine Similarity**

In addition to Euclidean distance, cosine similarity was used within a Hierarchical Clustering Model. This model attempts to minimize the cosine similarity distance measure between points in a cluster. This model will group documents in the term matrix in clusters who have the smallest angular distance from each other relative to the origin when graphically comparing the document term similarity. This model also requires a predetermined cluster parameter—5 clusters were chosen.

**K-Means Clustering Model**

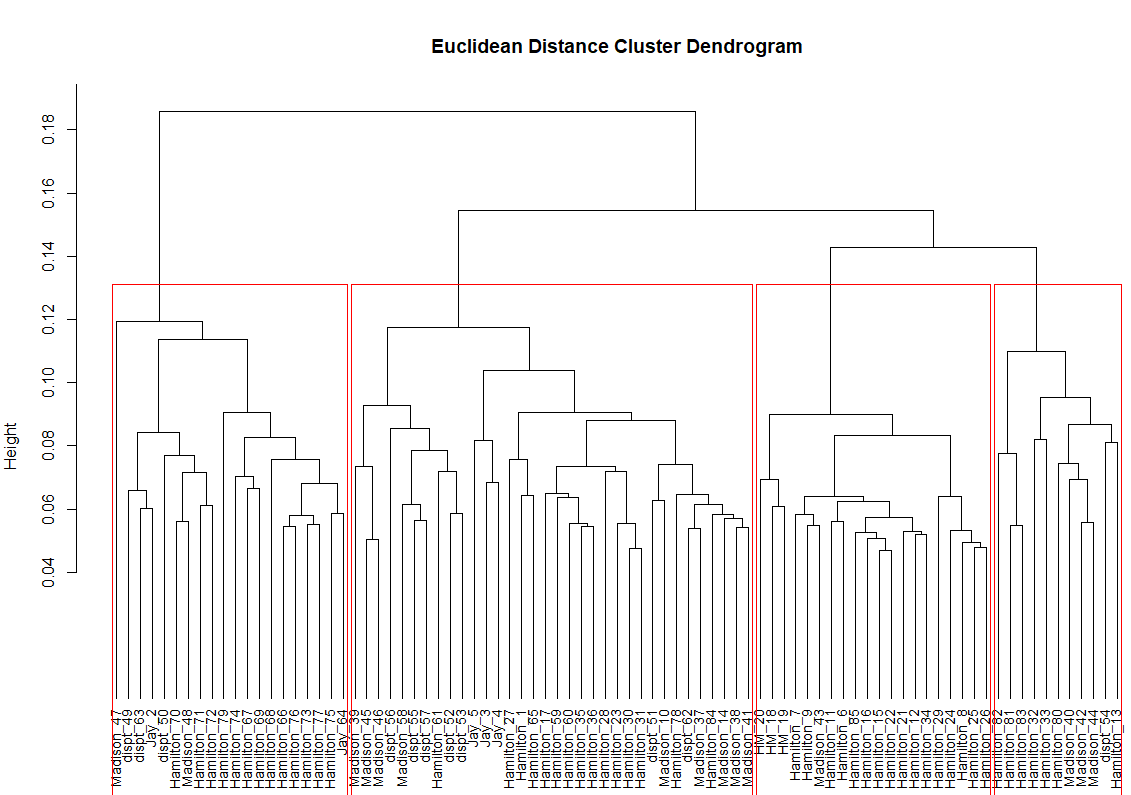
Furthermore, a K-Means clustering model was also applied to the Federalist Paper document term matrix. This model uses predetermined cluster numbers (k) to affix k center points among the data. The model then assigns data points to a cluster based on which k center it is closest to and tries to reduce the mean of all data points within a cluster. This model will group documents whose mean distance from each other is the least. K was set to 3 for the purpose of this study.

**Results**

**Hierarchical Clustering Model with Euclidean Distance Results**

The results of the Hierarchical Clustering Model using Euclidean distance can be seen in Figure 2 (below). Examination of the dendrogram does not reveal any distinctive document groupings that would help predict the author of the disputed papers. All the disputed essays are clustered with essays authored by multiple known Federalist Paper authors. The best prediction that can be made from these results is that the disputed essays were either co-authored by a combination of Hamilton, Madison, and Jay, or the author used the works of the other known authors as inspiration when writing the disputed Federalist Papers.

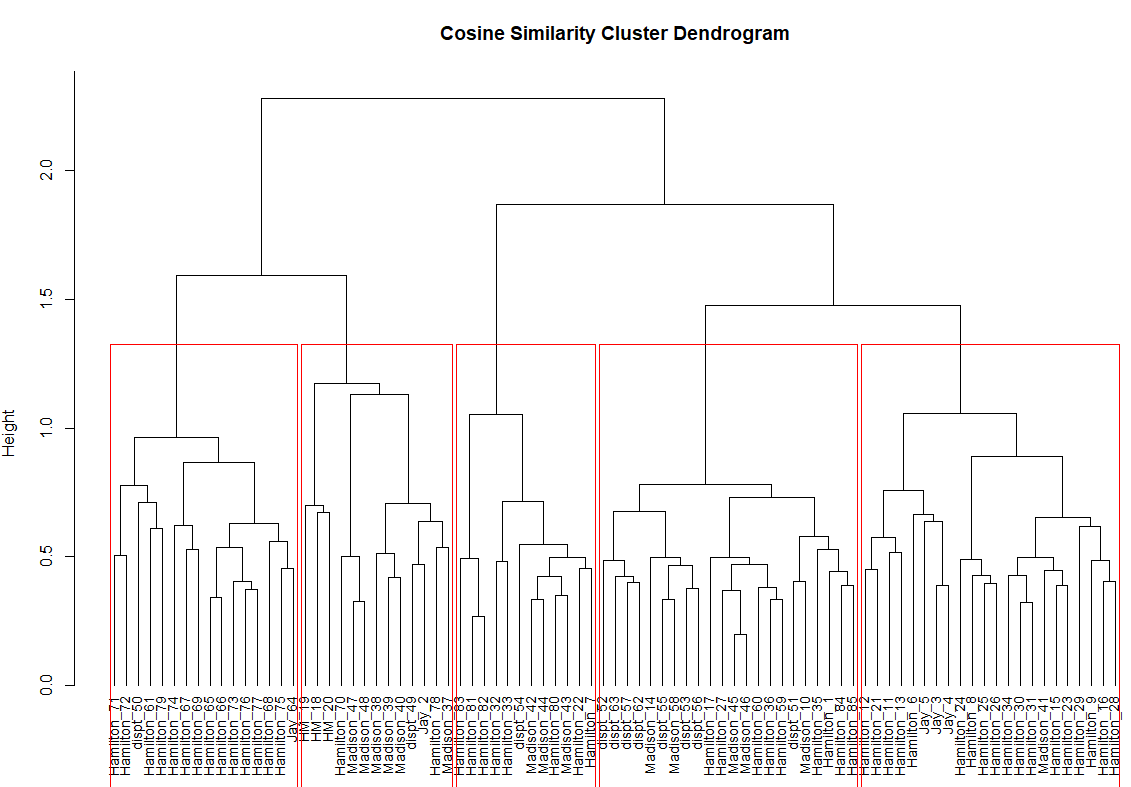
**Figure 2:**



**Hierarchical Clustering Model with Cosine Similarity Results**

The results of the cosine similarity variant Hierarchical Clustering Model can be seen in Figure 3. Contrary to the Euclidean distance-based model, the dendrogram does identify some helpful clustering with respect to the disputed essays. Federalist Paper #50 is clustered heavily with Alexander Hamilton authored essays which supports that he also authored this essay. Additionally, the remaining disputed papers appear to be clustered with Hamilton and Madison authorships, so this model provides somewhat less ambiguity than the Euclidean-based model. Therefore, the results of this model identify disputed essay #50 as Hamilton written and the remaining 10 essays as co-authored by Hamilton and Madison.

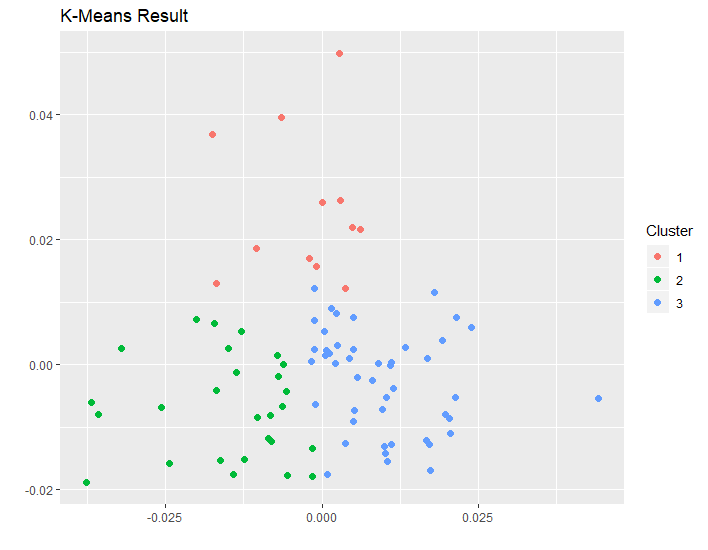
**Figure 3:**



**K-Means Clustering Model Results**

The results of the K-Means clustering model are shown in Figure 4 (below). The resulting clusters from this model are relatively spread out with respect to intra-cluster distance and rather close together in terms of inter-cluster distance. This may mean that some of the documents were mis-assigned if they were on the periphery of one of the 3 clusters.

**Figure 4:**

****

The summary of papers authored grouped by cluster is as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| Author(s) | Cluster 1 | Cluster 2 | Cluster 3 |
| Disputed | 1 | 7 | 3 |
| Hamilton | 7 | 12 | 32 |
| Hamilton & Madison | 0 | 0 | 3 |
| Jay | 0 | 2 | 3 |
| Madison | 4 | 6 | 5 |

Using these frequencies as a basis, the most likely authors of the disputed papers are:

|  |  |  |
| --- | --- | --- |
| 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 |

Of the 3 models, this one predicts the authorship of the most disputed essays; however, it still leaves 8 essays in question. Due to the mixed similarity in writing style, it may not be possible to distinctively determine a singular author of the remaining Federalist Papers.

**Conclusions**

In summation, before the United States officially came to be, the original states (then colonies) had to elect to ratify the proposed Constitution. In order to persuade or dissuade residents to ratify the Constitution, politicians often wrote anonymous essays in local newspapers. Perhaps the most famous collection of these essays is the Federalist Papers, published under the pen name “Publius” in New York in support of ratification (About the Federalist Papers). Publius is believed to have been used by Alexander Hamilton, James Madison, and Jon Jay who later claimed some of the essays. However, of the 85 essays comprising the Federalist Papers, 11 have disputed authorship. To help clarify, the Library of Congress initiated a study of the documents (About the Federalist Papers).

After analyzing the data, it appears that Federalist Paper #50 was authored by Alexander Hamilton, as confirmed by multiple models. Furthermore, Hamilton is likely to have authored #49 and #63. As for the remaining disputed Federalist Papers, it is best predicted that they were either co-authored by a combination of Hamilton and Madison, or that the authors were inspired by the work of the other when writing these essays.

If the Library of Congress wishes to act on the results of this study, it could only alter the authorship of Federalist Papers #49, #50, and #63 to be Alexander Hamilton. The remaining essays would be best left as either disputed or co-authored by Alexander Hamilton and James Madison.

**Citations**

About the Federalist Papers. (n.d.). Retrieved February 8, 2020, from www.congress.gov/resources/display/content/About+the+Federalist+Papers

The Federalist Papers. (n.d.). Retrieved February 8, 2020, from https://www.congress.gov/resources/display/content/The+Federalist+Papers

National Constitution Center. (n.d.). *The Constitution of the United States: Fast facts*. Retrieved February 8, 2020, from constitutioncenter.org/learn/educational-resources/historical-documents/constitution-fast-facts