Name: Jie Wang

Email: [JWANG326@Syr.edu](mailto:JWANG326@Syr.edu)

Assignment: HW4

Topic: Identify the author of disputed essays

Models/methods: Clustering

Date: Aug, 01,2022

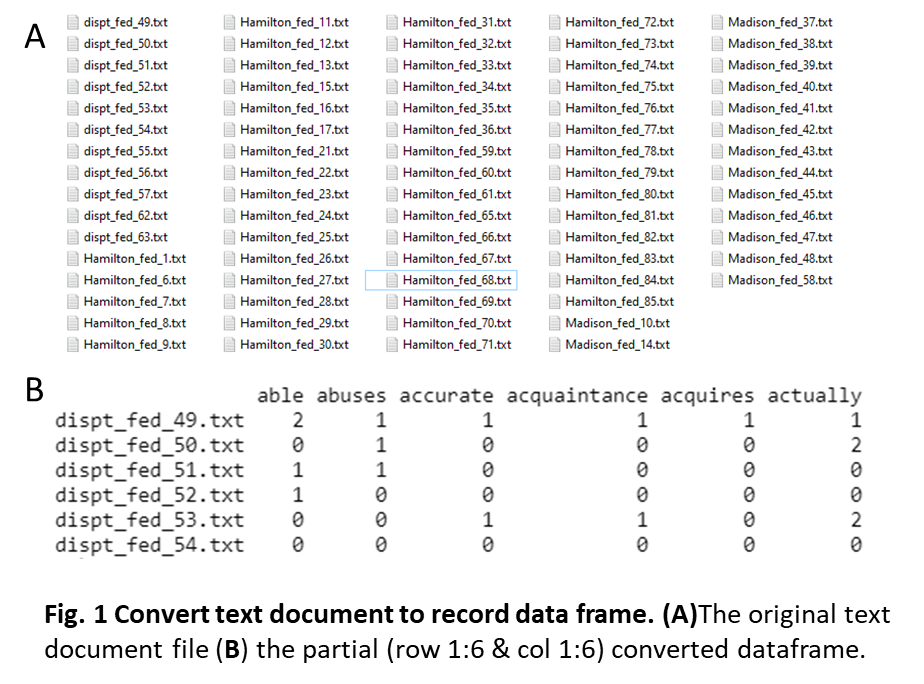
# Identify the author of disputed essays

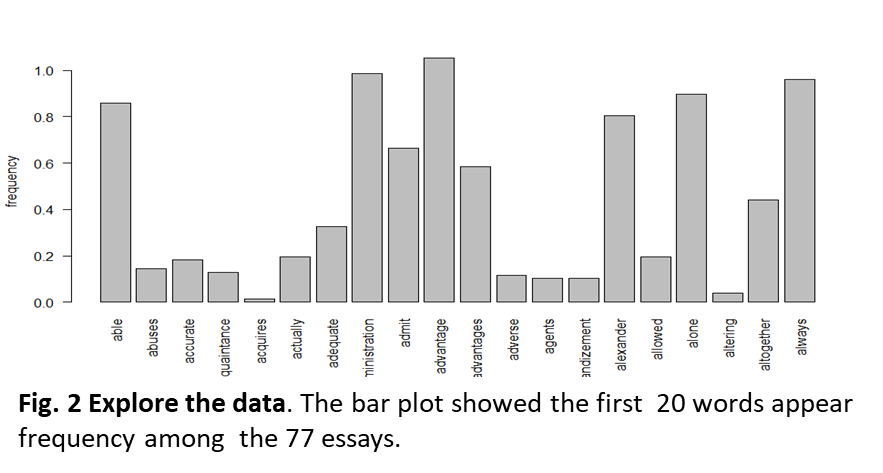
The Federalist Papers were a series of eighty-five essays urging the ratification of the Constitution, which were written by Alexander Hamilton, James Madison, and John Jay for publication in New York newspapers, the essays originally appeared anonymously. The first essay was published in 1787, edited by John Mclean. The Federalist Papers are considered one of the most important sources for interpreting and understanding the original intent of the Constitution.

Within those 85 essays, 74 essays with identified authors, including 51 essays written by Hamilton, 15 by Madison, 3 essays were co-authored by both of Hamilton and Madison, and 5 essays written by Jay. The remaining 11 essays’ author are still under debate, which could either written by Hamilton or by Madison. To understand the original intent of the Constitution, historians had been trying to identify the real author of the remaining 11 essays.

The statisticians have been trying to analysis the words distribution in the essay and drew their conclusion on the authorship of the remaining 11 disputed essays, however, more evidence is needed to solve this mystery in history. This project is focusing on identifying the author of essays, which could be either Hamilton or Madison.

# Analysis

Data preparation and cleaning: original dataset is a total of 85 .txt files essays, in which 51 essays written by Hamilton, 15 by Madison, 3 by Hamilton and Madison, 5 by Jay, and 11 papers probably written by written by Hamilton or Madison. The papers written by Madison only, Hamilton only, and 11 undefined papers (**Fig. 1A**) were load into document, the corpus. **1)** Corpus(DirSource(‘folder’)) was use to load in the documents (the corpus). **2**) tm\_map() function is used to remove unnecessary white space, to convert the text to lower case, to remove common stopwords like 'the', “we”, “and”. “like”, “very”, “can”, “I”, “also”, and “lot”. **3)** use DocumentTermMatrix() function to convert text data to record data (**Fig. 1B**).

Data exploration: to explore the record data set, appy str() function on the conversed dataframe, it shows that there are total 77 observations and 8081 variables. All the variables are numeric values. **Fig. 2** shows the frequency of the first 20 words from the dataset.

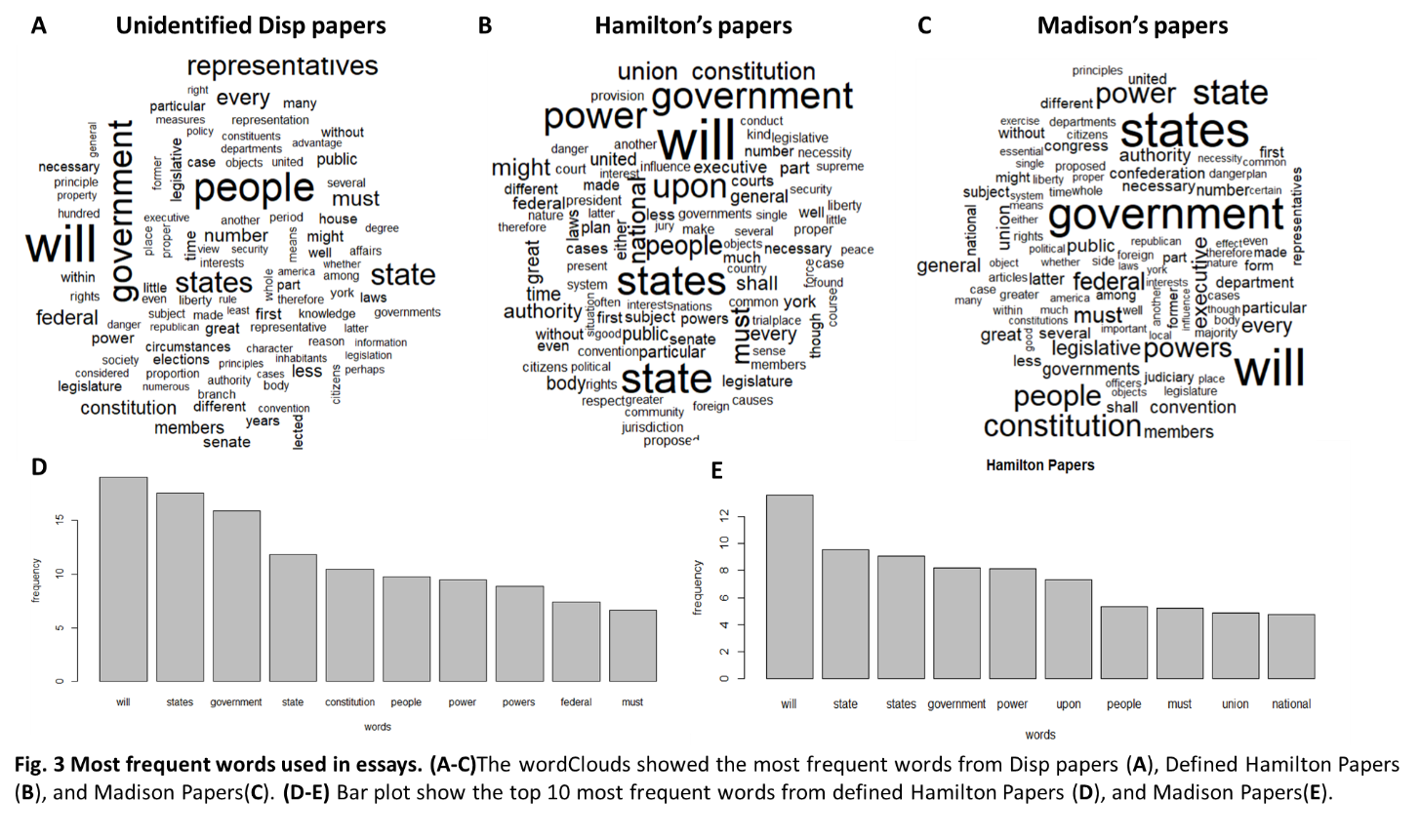
Methods: “tm”, “stringr”, “Wordcloud”, “slam”, “quanteda”, “SnowballC”, “arules”, “proxy”, “cluster”, “stringi”, “proxy”, “Matrix”, “tidytext”, “plyr”, “ggplot2”, “factoextra”, “mclust”, “textstem”, “amap” and “ClusterR” libraries and base R were used for analysis and data visualization. With the convert record data, **1)** use wordcloud() function to present the most frequent words used by author Hamilton and author Madison. **2**) apply K-Means approach to cluster essays into two clusters. Different algorithms including “Manhattan” (shown in the result), “Euclidean”, “Pearson”, “Canberra”, and “Spearman” were used to map and visualize the distance. **3**) apply Hierarchical approach with cosine method to cluster the essays, based on the hierarchical cluster tree, 10 clusters were identified to further check the similarities between each essay, and define the author for the remaining 11 disputed essays.

Analysis goals and Parameters: The goal of this project is to identify the potential authorship of remaining 11 essays, to this end, some parameters were set up and adjusted to get a better cluster solution, including cluster number K value and clustering algorithm methods. Cluster number k value was adjusted to gain meaningful clusters. **1)** For the K Means clustering, k was set as 2, which was decided based on the fviz\_nbclust() function result, as the result showed a significant peak at cluster 2. **2)** For hierarchical cluster analysis, the items were mapped into 10 groups based on the cluster tree for better exploration. **3)** Different clustering algorithm methods were tested when map the distance, including “Manhattan”, “Euclidean”, “Pearson”, “Canberra”, and “spearman” for better a understanding data and visualizing the data.

# Results

## The most frequent words used in the Hamilton’s paper and Madison’s paper

To get a first overview of the words-distribution of 11 disputed papers, Hamilton’ s paper, and Madison’s paper. WordClouds (**Fig. 3A-B**) were generated for each catalogue with the max. words = 100. **Fig. 3A** represent the most frequent words have been used in 11 remaining disputed papers, including ‘will’, ‘people’, ’government’, ‘state’, ‘representatives’, and ‘states’. **Fig. 3B** shows the most frequent words that Hamilton used in the papers, such as ‘will’, ‘people’, ’government’, ‘power’, and ‘upon’, the top 10 of the most frequent words were also showed with frequency decreasing order in **Fig.** **3D**. **Fig. 3C** and **E**. represent the most frequent words, such as ‘will’, ‘state’, and ‘government’, used by Madison. Taken together, both authors prefer using ‘will’, ’state’, ’states’, and ‘government’ in their papers, which also highly showed in 11 remaining unidentified author papers. In addition, as compared to Madison, Hamilton would more likely use ‘upon’, ‘union’, and ‘national’ in the essays, while Madison would more likely use ‘constitution’, ’powers’, and ‘federal’ in the essays.

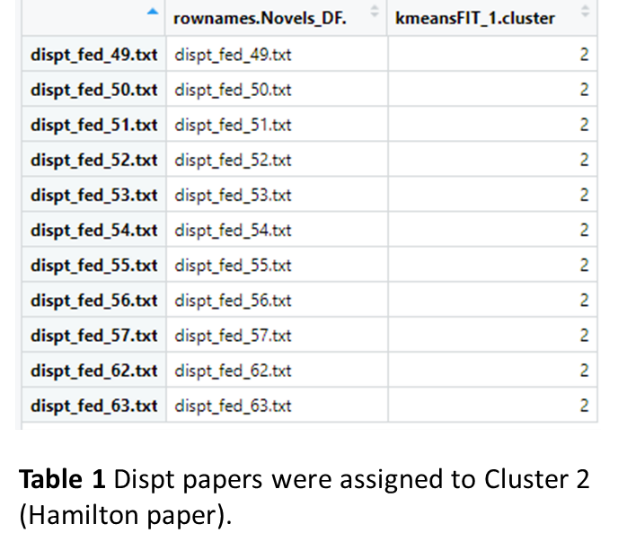


## Manhattan based distance map for 77 essays

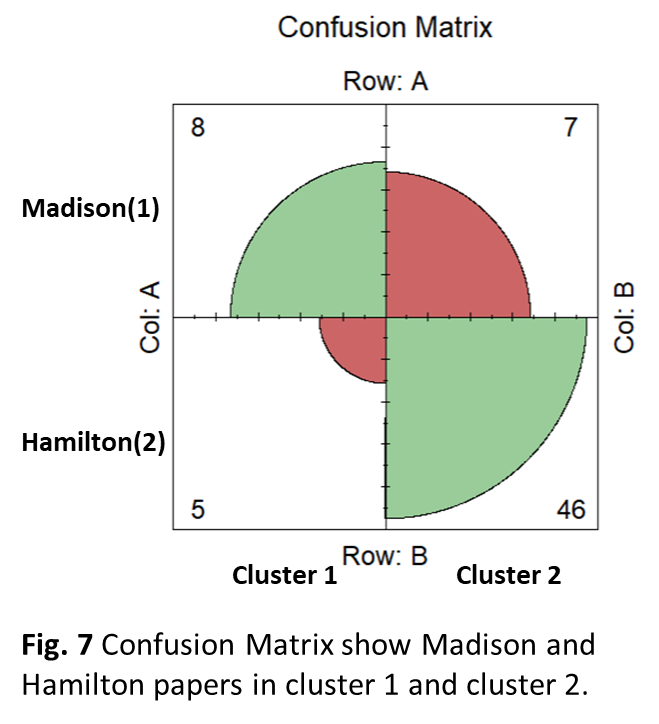
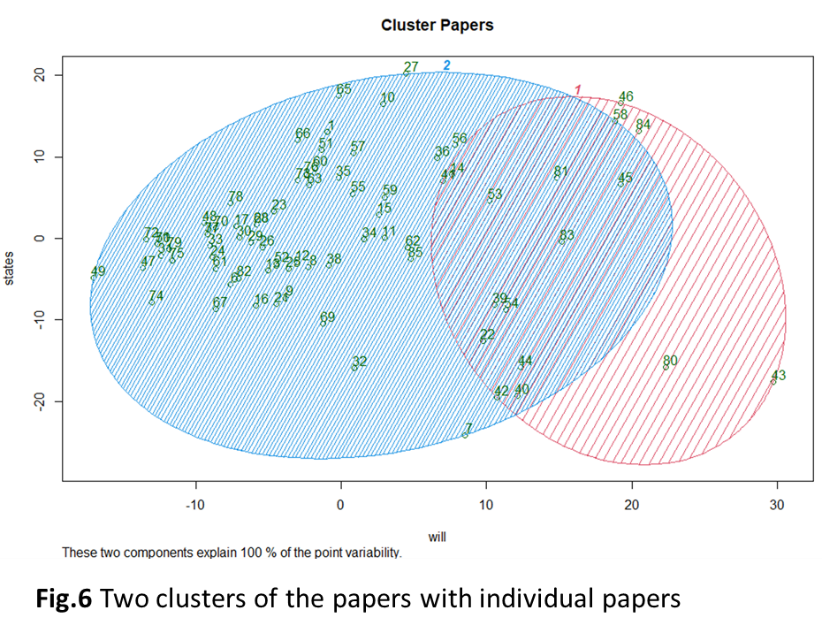
To further measure how each essay close or distance to each other based on the word distribution, the distance within the essays were applied with Manhattan method and further visualized in heatmap as shown in **Fig. 4**,each square colored between blue and orange are the distance value between indicated two essays. blue is low (more similar), while orange is high (less similar). The squares showed along diagonal line are distance equal to 0, which means the essay is 0 distance between itself. From this heatmap, essays from Hamilton, including “Hamilton\_fed\_83”, “Hamilton\_fed\_81”, “Hamilton\_fed\_84”, “Hamilton\_fed\_22”, “Hamilton\_fed\_15”, and “Hamilton\_fed\_69” showed a great distance from other essays; Essays from Madison, such as “Madison\_fed\_40”, “Madison\_fed\_43”, “Madison\_fed\_42”, “Madison\_fed\_44”, “Madison\_fed\_37”, “Madison\_fed\_10”, and “Madison\_fed\_38” showed a great distance from other essays. Essay “dispt\_fed\_53”, “Madison\_fed\_37”, and “Madison\_fed\_10”. This data suggested that essays written by Madison and Hamilton are significant different based on the words use.

## 

## Identify the remaining 11 disputed authors with K-Means clustering

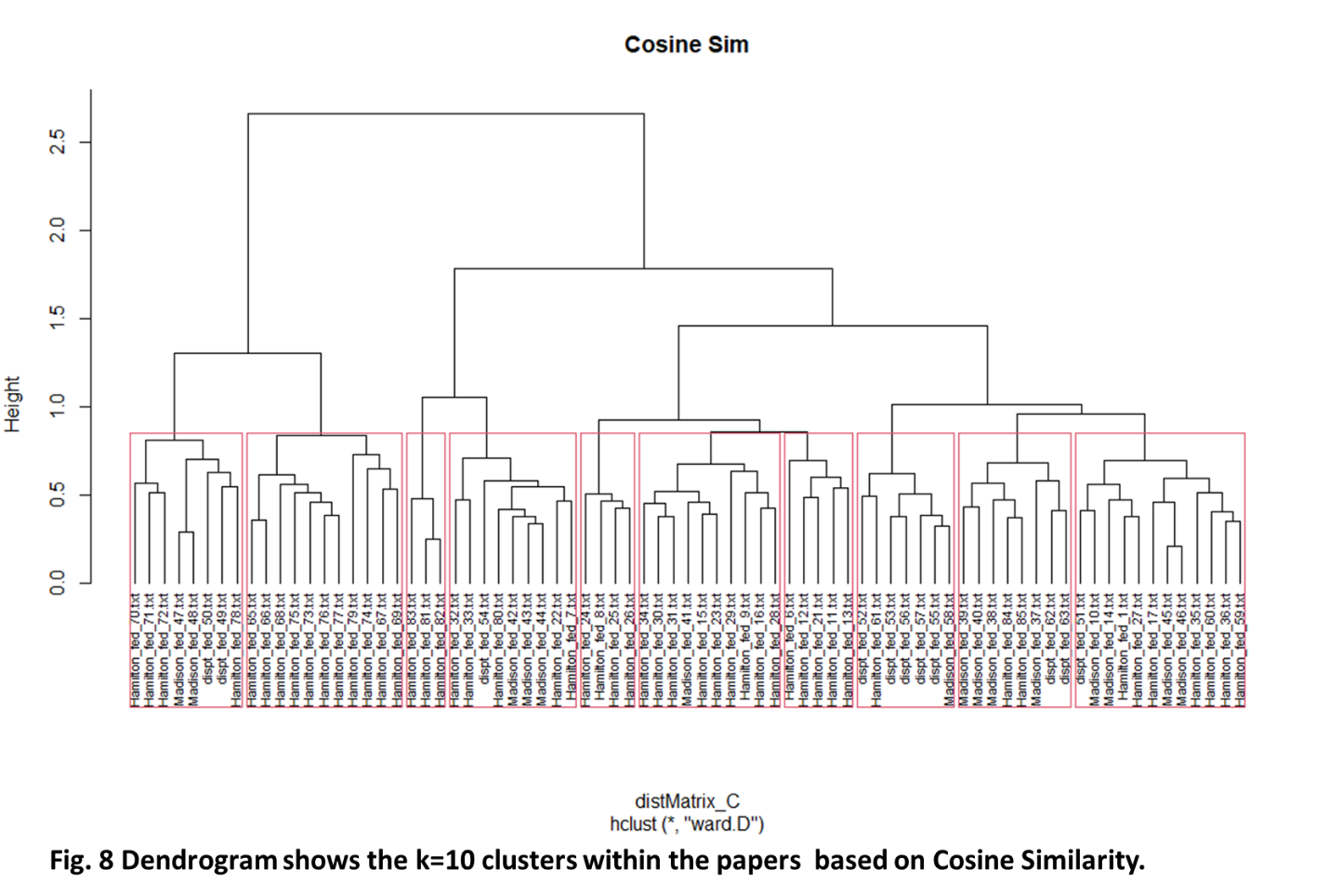
The number of cluster (k) is most important for k- means clustering, Silhouette method was used to identify the optimal cluster numbers, as shown in **Fig. 5A**, the significant peak is at two clusters in the graph. With k=2, the total 77 essays, include 51 from Hamilton, 15 from Madison, and 11 unidentified were grouped into 2 clusters. **Fig.5B** showed two cluster plots with two different colors with most frequent word “states” and “will”, the centroid of each cluster was marked with cross signs with the same color of the clusters, in which cluster 1 showed in grey circle, and cluster 2 showed in red circle. **Fig. 6** represent the individual essays (labelled as essay number only) assignment in the 2 clusters. The clusters formed with varying “will” and “state” are not well separated, since there is a big overlap between cluster 1 and cluster 2. Based on above results, all the 11 remaining unidentified essays were clustered into cluster 2 (**Table 1**). To verify whether k-means clustering work well in this text mining task, a confusion matrix was generated to map the essays from Hamilton and Madison with the clusters (cluster 1 and cluster 2) predicted with K-Means. As shown in **Fig. 7**, among 15 essays written by Madison, 8 were assigned to cluster 1, and 7 were assigned to cluster 2; among 51 essays written by Hamilton, 5 were assigned to cluster 1, and 46 were assigned to cluster 2, clearly, this is not a very good predict model. These data indicated that even though all remaining unidentified essays were grouped into cluster 2, which were assuming written by Hamilton, it has to be notice that based on the confusion matrix data, the K-Means clusters with k =2, works not good as expected. Therefore, alternative method Chart, scatter chart

Description automatically generatedis needed for better exploration.



### Identify the remaining 11 disputed authors with Hierarchical clustering with cosine similarity

Cosine similarity has been widely used for text classification, therefore, a hierarchical clustering approach with cosine method was used here to cluster 77 essays. As showed in **Fig. 8**, the hierarchical clustering is organized as a tree, the leaf nodes of tree are singleton clusters of the individual 77 essays, two essays that are the most similar were grouped together, by cutting the tree in a certain level, 10 clusters were identified as highlighted in red boxes. Within those clusters, the anonymous disputed essays would be able to be identified as either written by Hamilton or by Madison. Following are the details of the clustering: **1)** Within the first cluster (cluster 1) from the left side, total 8 essays were clustered together, with 4 from Hamilton, 2 from Madison, and 2 are disputed essays, given that ‘dispt\_fed\_49’ and ‘Hamilton\_fed\_78‘ are first merged together, indicating that ‘dispt\_fed\_49’ and ‘Hamilton\_fed\_78‘ are most similar, therefore, it is very possible that ‘dispt\_fed\_49’ was written by Hamilton as well, “dispt\_fed\_50” was then joined into the subcluster with ‘dispt\_fed\_49’ and ‘Hamilton\_fed\_78‘, which reflexed the close distance of the tree essays, therefore, “dispt\_fed\_50” might be authored by Hamilton as well. **2)** in the second cluster (cluster 2), total 11 essays are clustered together, which are all authored by Hamilton, which is very reason, as the author show a similar writing pattern in his essays. **3)** the same case was observed in cluster 3, cluster 5, and cluster 7. All the essays within those clusters were authored by Hamilton, the reason they were separated into different cluster is probably due to the essay topic difference. **4)** In cluster 4, there are total 9 essays, 5 by Hamilton, 3 by Madison, and 1 is disputed essay. It shows that “disp\_fed\_54” is joint the node with 3 Madison essays and 3 Hamilton essays, further all those 7 essays were joint the up-level nodes with another two essays from Hamilton, therefore it is possible that “disp\_fed\_54” is written by Hamilton, however, the possibility of “disp\_fed\_54” written by Madison is also big. **5)** Cluster6 showed 9 Hamilton essays were joint with 1 Madison essays, it is possible that Madison ‘s paper (Madison\_fed\_41) in this issue talked similar topic as Hamilton’s essays, specifically article “Hamilton\_fed\_15” and “Hamilton\_fed\_23” . **6)** In cluster 8, 1 Hamilton, 1 Madison, and 5 disputed essays were included, according the cluster rules mentioned above, it is possible that “disp\_fed\_53”, “disp\_fed\_56”, “disp\_fed\_57”, and “disp\_fed\_55” are written by Madison, as they were clustered with “Madison\_fed\_58” in a subcluster, while “dispt\_fed\_52” is more possible written by Hamilton. However, it has to be noted that only one reference (“Madison\_fed\_58”) within each subcluster, and only one reference (“Hamilton\_fed\_61”) within the other subcluster in cluster 7. the essay bias could be existed, therefore the evidence of assignment of disputed essays in this cluster is still not solid. **7)** In cluster 9, there are total 8 essays, 4 by Madison, 2 by Hamilton, and 2 are disputed. the two disputed essays are grouped together at beginning, indicating they could be written by one author, further, those two essays were then merged with the essay by Madison, therefore, it is very possible both two disputed essays (“disp\_fed\_62” and “disp\_fed\_63”) were written by Madison. **8)** At last, in cluster 10, total 12 essays including 7 by Hamilton, 4 by Madison, and only one is disputed, given that “dispt\_fed\_51” is more similar with the essay by Madison (“Madison\_fed\_10), it is very possible that the “dispt\_fed\_51” is written by Madison.



# Conclusions

The aim of this project is to identify the author who wrote the remaining disputed essays urging the ratification of the Constitution, based on the known authorship of the other essays in the Federalist Papers. Different authors have different writing styles, in this case, the essays from Hamilton and Madison are different regarding the words use.

This study showed that the frequent of different words are varying between Hamilton’s essays and Madison’s essays. Even though the overall remaining disputed essays showed a quite similar words distribution as the overall Hamilton’s essays and Madison’s essays, different individual disputed essay showed similarities with either Hamilton’s essays or Madison’s essay. And with that, it seems like that essay “dispt\_fed\_49” , “dispt\_fed\_50”, “dispt\_fed\_52”, and “dispt\_fed\_54” are more likely written by Hamilton, while “dispt\_fed\_51”, “dispt\_fed\_62”, and “dispt\_fed\_63” are more likely written by Madison, however, “dispt\_fed\_53”, “dispt\_fed\_55”, “dispt\_fed\_56”, and “dispt\_fed\_57” could be written either by Madison or Hamilton, which is still not clear.

It has to be noted that writing style (or words use) is not the only factor could distinct the essays from Hamilton and Madison. Some essays from Madison and from Hamilton shared similarities, it is possible that Madison and Hamilton were focusing on a very similar (or same) specific topic, alternatively, the other possibility is that the essays were edited by the same editor. Taken together, even though some progress has been made for identifying the authorship of disputed essays, some of the disputed essays are still under debate.