**Who Wrote the Federalist Papers?**

Ben Harwood

**Introduction**

During the recent impeachment trial, members of both the House prosecutors and the President’s defense team repeatedly referenced the Federalist Papers. These 85 articles and essays were written as an exhaustive explanation and defense of the proposed Constitution that had been submitted to the states for ratification in 1787. There had been a collection of articles and letters written in criticism of the Constitution, and it was Alexander Hamilton that sought to win over the hearts and minds of the “Anti-Federalists” as they became known.

Hamilton was not alone in this endeavor. He initially enlisted John Jay who wrote four of the first five essays (Jay did write one of the later essays, but this was his last due to illness). Hamilton and Jay also recruited James Madison, who contributed numerous papers and became Hamilton’s primary collaborator. The articles were published in New York newspapers beginning on October 27, 1787. They were published under pseudonyms because the men had attended the Constitutional Convention, though many people were able to determine the authors.

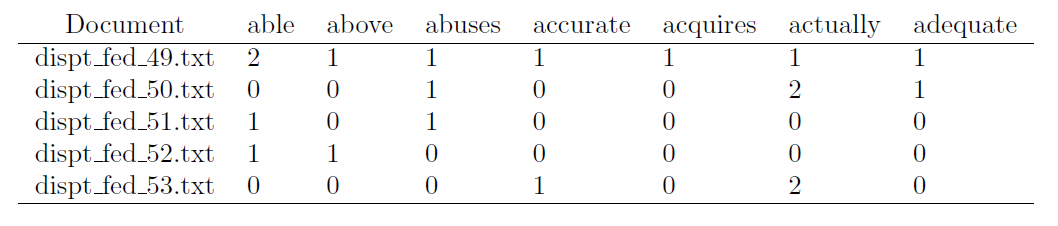
Hamilton provided a list that associated the authors with the paper numbers when demand for a more official publication of the collected papers arose. The authorship of seventy-three of the articles is generally well-established, but there is some debate about the remaining twelve among scholars. This is partly due to Madison providing a different list of authorship that disputed Hamilton’s list[[1]](#footnote-1), indeed it known that Hamilton attributed authorship of paper 54 to John Jay, when in fact Jay wrote paper 64. Much, much more can be said about the Federalist Papers however the purpose of this article is to use data mining techniques to see if the authorship of the 12 disputed papers can be established.

**Section 1: Analysis and Models**

As mentioned in the introduction, there are 85 papers total in the collection. The idea of the study is to look at the word usage in each of the papers to see what similarities in writing styles and vocabulary exist between them. To do this, the documents will be turned into a collection of values measuring the frequency each word appears in each document, and then the “distance” between the different documents will be measured to form clusters of the documents. With luck, the Hamilton and Madison papers will be separate, with some of the disputed papers in each cluster, whereby the authorship can be determined.

**Section 1.1: Data Cleaning and Preparation**

Thankfully, the Federal Papers are widely available. To begin the study, the papers in their raw forms (i.e. text files) need to be converted into something that can be mathematically analyzed. The papers are collected together into a digital corpus[[2]](#footnote-2) and then a so-called “document term matrix” (DTM going forward) is formed with the content of the papers. The DTM is a matrix where each row is a different document, and the columns are the individual words from any of the papers. The elements of the matrix are the frequencies in which the words appear in the documents. The below example shows the word counts for the first seven (alphabetically) words that appear in any of the papers[[3]](#footnote-3).

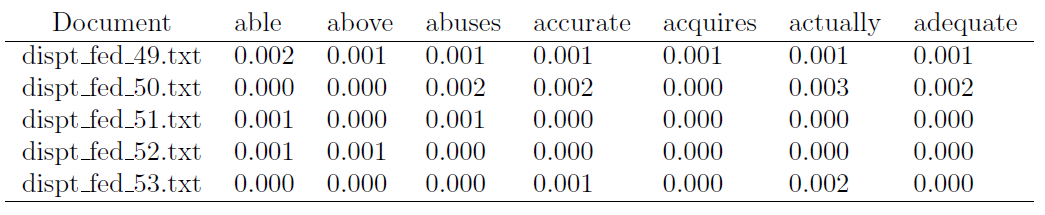


To get an idea for which words are more common than others, here is a “word cloud”, where the size of the word indicates its frequency amongst the papers (bigger is more frequent).

A screenshot of a cell phone

Description automatically generated

In normal circumstances, an important part of data cleaning is dealing with missing values. In this case, however, there are no missing values, because the zeros are word counts, and there is no way for data to be missing. One thing that does need to be done, however, is to normalize the counts in each row, because each document is a different length. To this end, each word count is divided by the total number of words in the associated document, resulting in the following transformed matrix:



Additionally, because it is known that John Jay wrote five of the papers and that three of the papers were cowritten by Hamilton and Madison, those eight documents should not be part of the analysis.

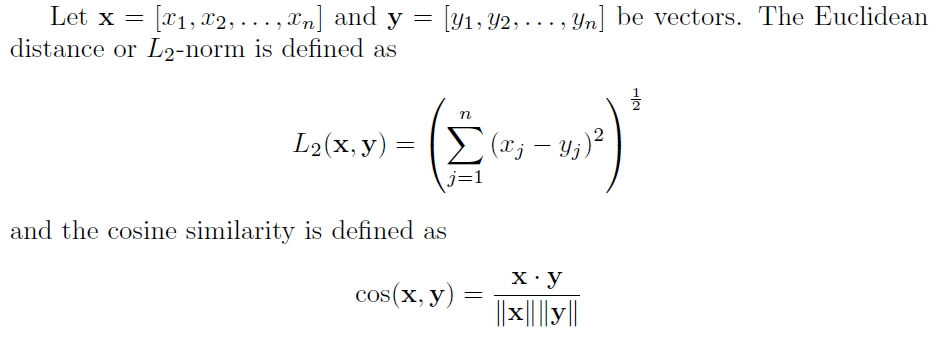
**Section 1.2: Analysis**

It was mentioned previously that the goal of the analysis would be to measure the “distance” between the documents to group or cluster them together with the hope that the Hamilton and Madison authored papers would be separated with the disputed papers spread between them. With the documents turned into data structures that can have math done them, clustering can begin.

Two types of clustering techniques were employed for the analysis. The first, *K*-means, requires specification of the number of clusters up front. In the case of the Federalist Papers, this is not an issue, since the goal is to determine which of two authors wrote the disputed papers, so two clusters is a logical decision. The second technique, hierarchical, begins by declaring each variable to be document to be in a cluster by itself and then connecting the closest two clusters, repeating this process until the entire collection is together. Each technique shall be addressed, in order.

***K*-means**

The basic idea of *K*-means clustering is that after the number of desired clusters is established, that number of centroids is chosen by the computer, and then each of the points in the dataset is assigned to the centroid it is closest to. The centroids of those groups are then redetermined, the points re-assigned, and this continues until none or very few of the centroids change. Several types of distance measure can be used, though Euclidean distance and Cosine similarity are the most common. Definitions of both follow.



Cosine similarity will be used later, but for the first part of the study Euclidean distance was employed[[4]](#footnote-4) and resulted in clusters that look roughly like this

A close up of a map

Description automatically generated

That’s paper 83 way out to the right. Cluster 1 above is the larger of the two, with 52 total papers including every disputed paper except for one (specifically paper 63) and exactly three Madison papers, leaving 75% of the papers in Cluster 1 being written by Hamilton. Cluster 2 had the remaining 25 papers, with an even split between Hamilton and Madison along with the single disputed paper. This leaves the authorship of paper 63 up for debate. Hence the turn to hierarchical clustering.

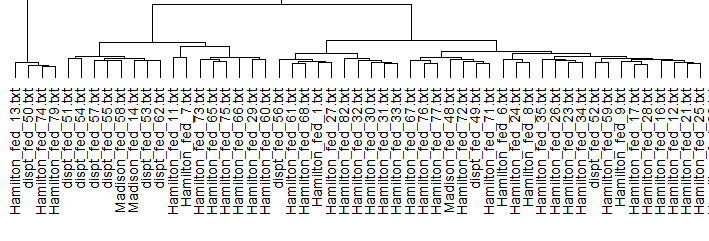
**Agglomerative Hierarchical clustering**

A bit more control over the distance measure is available (at least in R where this analysis was done) with hierarchical clustering. There isn’t much else really to say regarding the process in general (beyond that which was described at the beginning of this section), so here first is a diagram showing how the document cluster based on Euclidean distance, where the red rectangles indicate how the documents clustered.

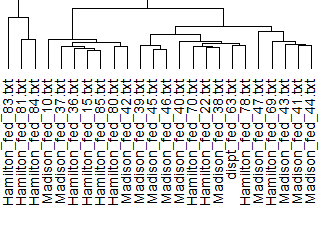
A screenshot of a social media post

Description automatically generated

To see how the method distributed the disputed papers, here is a closer view of the first cluster.



Notice that papers 50, 56, 49, and 52 are nearly exclusively grouped with Hamilton, while paper 51, 54, 57, 55, 53, and 62 are with Madison. Here is the closer view of the second cluster.

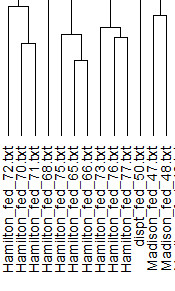


This cluster has paper 63, but it is in sub-clusters with both Hamilton and Madison. These two clusters are near exact matches to the *K*-means clusters, with the exception that *K*-means put paper 66 in cluster 2 while this technique placed it in cluster 1. Consider now using Cosine similarity:

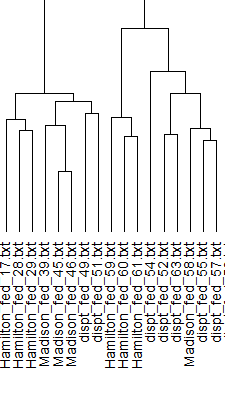
A screenshot of a social media post

Description automatically generated

Notice the much smaller cluster than with the Euclidean cluster.



In this case, paper 50 was paired most closely with papers 47 and 48, while in the Euclidean cluster it was with papers 74 and 79.[[5]](#footnote-5) Looking at the second cluster, this is just a portion of it, the interesting part.



The reason this part is interesting is because it has papers 49 and 51 connected to Madison whereas previously 49 was connected to Hamilton and 51 was only very loosely connected to Madison. What this clustering shares with its Euclidean counterpart is that papers 54, 52, 55, and 57 (all of which are disputed) are grouped together and papers 55 and 57 are paired with the same Madison paper, number 58.

**Section 2: Results**

As was shown in the previous section, various clustering techniques were employed to determine which of the Federalist Papers were most closely related. *K*-means clustering resulted in the following determination:

1. Disputed papers 49-57 and 62 clustered in a group with 42 author-determined papers, 93% of which were authored by Hamilton
2. Disputed paper 63 clustered in a group that had an even number of Hamilton and Madison papers

Because this is not sufficient to make a definitive conclusion, hierarchical clustering was then employed to paint a clearer picture. Hierarchical clustering using cosine similarity (because cosine similarity is more reliable for document data) revealed the following:

1. Papers 49-52, 54, 55, 57 and 63 are more connected to Madison than Hamilton
2. Papers 53 and 56 are connected but directly connected to either author

With the results of the two clustering techniques, the following is conjectured:

1. James Madison authored disputed papers 49-52, 54, 55, 57, and 63
2. Hamilton or Madison could have authored papers 53 and 56.

**Conclusion**

It was noted in the introduction that much more could be said about the Federalist Papers. For instance, paper 84 is notable in that it is an opposition to what would come to be the Bill of Rights, one of the foundational notions of the US Democracy. Regardless, having roughly established who wrote the disputed papers (with the exception of papers 53 and 56, which ironically both pertain to the House of Representatives) this opens further conversation about the authors and the papers themselves.

It would be interesting to see if other data mining techniques could reveal the author of papers 53 and 56. Perhaps Lin-Manuel Miranda knows.

1. Hamilton provided his list to his lawyer mere days before his infamous duel with Aaron Burr, and Madison suggested the discrepancies in the list were due to this [↑](#footnote-ref-1)
2. Corpus: a collection of written texts, especially the entire works of a particular author or a body of writing on a particular concept [↑](#footnote-ref-2)
3. After removing “stop words” such as ‘a’, ‘is’, ‘it’, etc., as well as excessively lengthy words. Also, all words are converted to lowercase, and numbers are removed [↑](#footnote-ref-3)
4. Partly due to limitations in R [↑](#footnote-ref-4)
5. In both cases, paper 50 is only associated to Hamilton papers, so this isn’t that big an issue [↑](#footnote-ref-5)