**Introduction**

The Federalist Papers are a series of 85 essays written by Alexander Hamilton, John Jay, and James Madison between October 1787 and May 1788. The essays were published anonymously, under the pen name "Publius," in various New York state newspapers of the time. The Federalist Papers were written to urge New Yorkers to approve the proposed United States Constitution, which was drafted in Philadelphia in the summer of 1787. In petitioning for adoption of the Constitution over the Articles of Confederation, the essays explain provisions of the Constitution in detail. James Madison's Federalist no. 10 is said to be one of the most important and enduring statements of American political theory. Its reasoned statement explains what an expanding nation might do if it accepted the basic premise of majority rule, a balanced government of three separate branches, and a commitment to balance all the diverse interests through a system of checks and balances. The Federalist Papers are often used to help interpret the intentions of those drafting the Constitution.

The Federalist Papers were published primarily in two New York state newspapers: The New York Packet and The Independent Journal. They were reprinted in other newspapers in New York state and in several cities in other states. A bound edition, with revisions and corrections by Hamilton, was published in 1788 by printers J. and A. McLean. An edition published by printer Jacob Gideon in 1818, with revisions and corrections by Madison, was the first to identify each essay by its author's name. Because of its publishing history, the assignment of authorship, numbering, and exact wording may vary with different editions of The Federalist. 11 essays are authored by “Hamilton or Madison”. These are the famous essays with disputed authorship. Hamilton wrote to claim the authorship before he was killed in a duel. Later Madison also claimed authorship. Historians were trying to find out which one was the real author.

**About the Data**

The text files of the essays are brought in as a corpus. 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 are authored by “Hamilton or Madison”.

The corpus needs to be turned into a Document Term Matrix. Punctuation, numbers, stopwords, and separators need to be removed. The text needs to be lowercase.

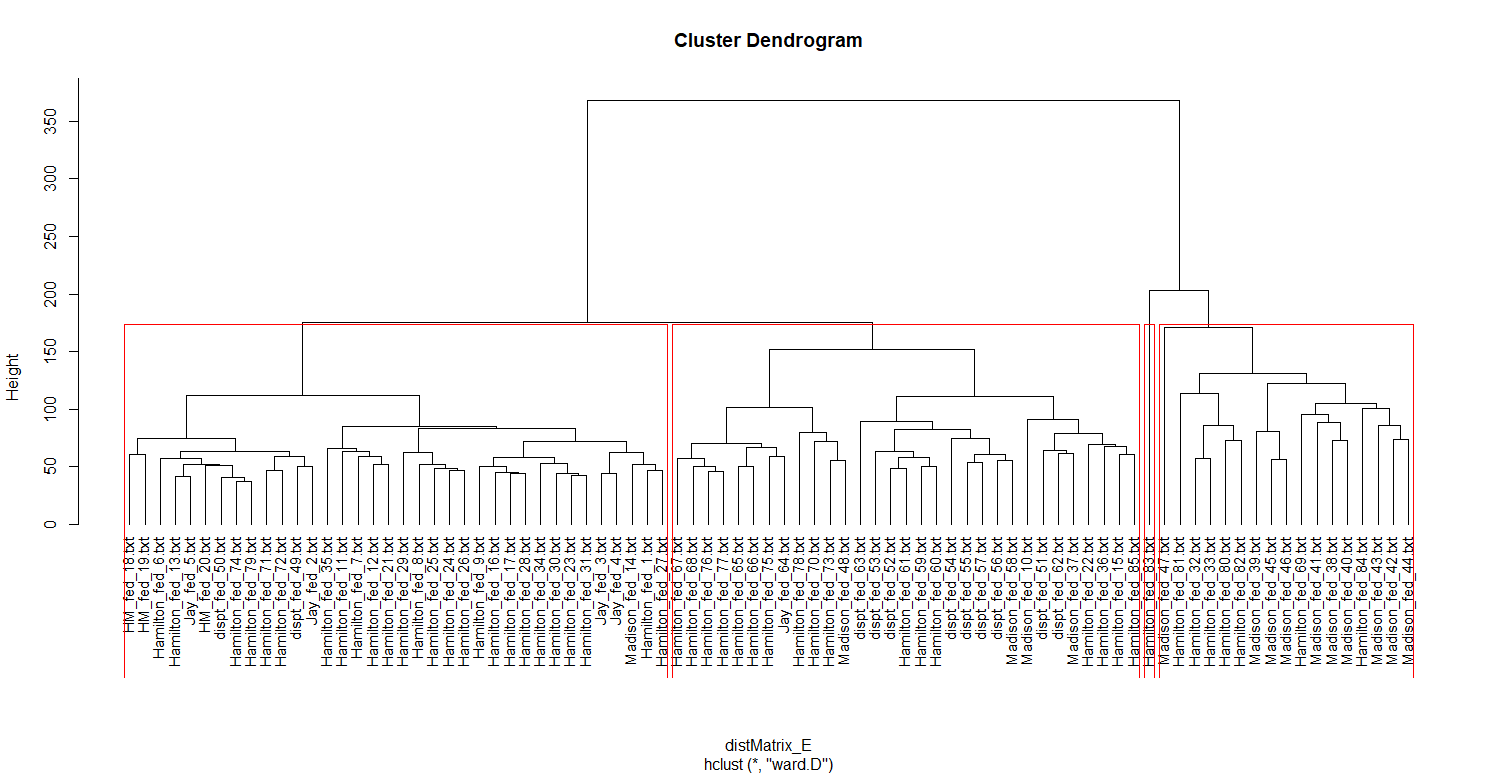
To perform distance measures the Document Term Matrix needs to be normalized and transposed.

**Analysis**

After removing the stopwords the most frequently occurring word are state, will, nation, trade, commerc, great, navig, part, union, America, interest, may, and navi.

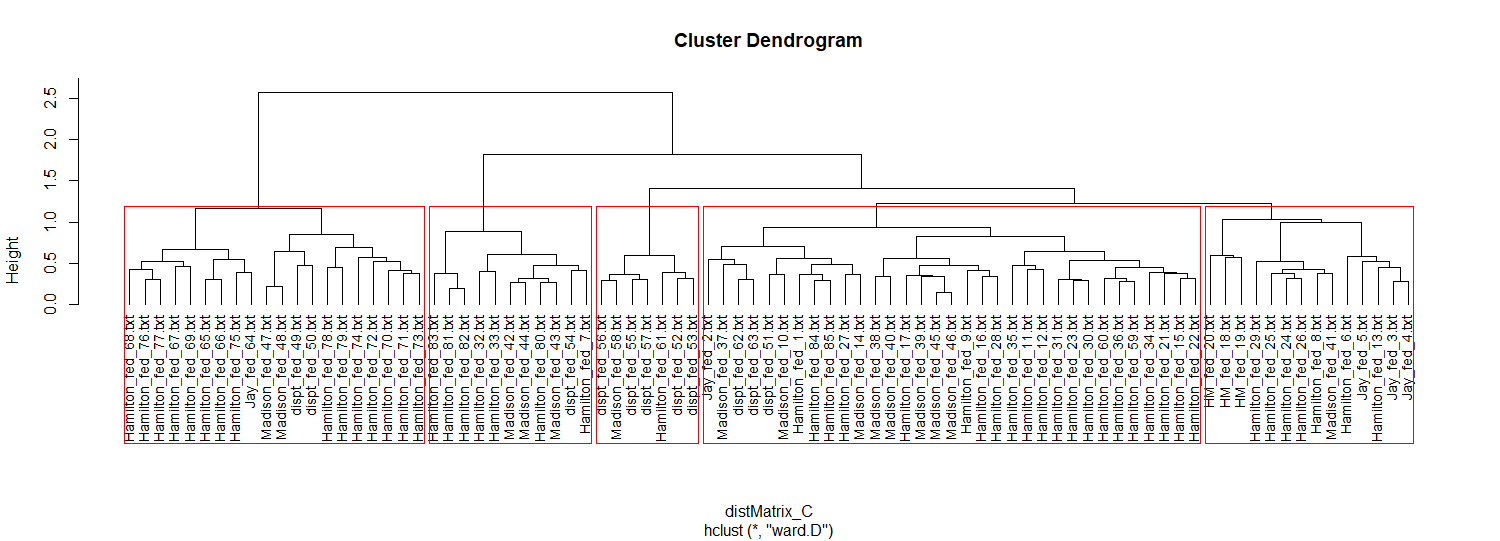


Hierarchal Euclidean Clustering



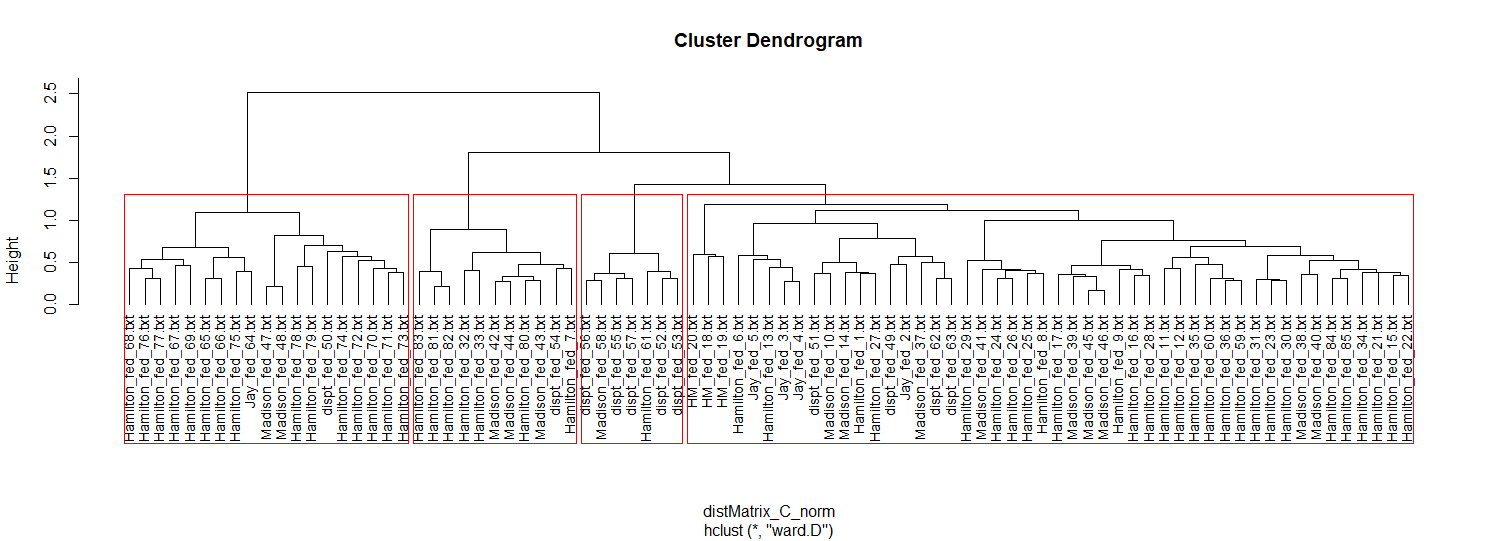
The 1st cluster is mostly Hamilton, the 2nd is mostly Hamilton, the 3rd just contains a single Hamilton essay, and the 4th cluster is pretty evenly split between Hamilton and Madison.

Hierarchal Cosine Clustering



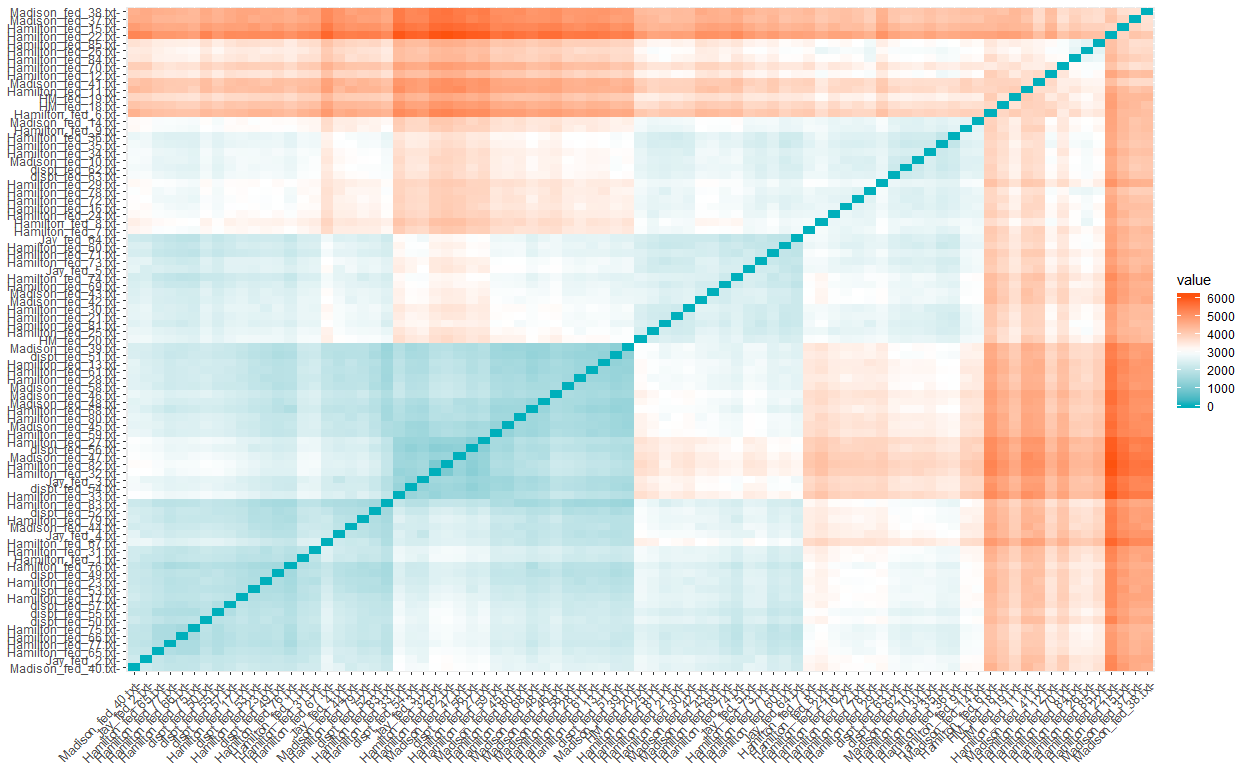
The 3rd cluster contains most of the essays in question.

Hierarchal Cosine Clustering with a Normalized Matrix



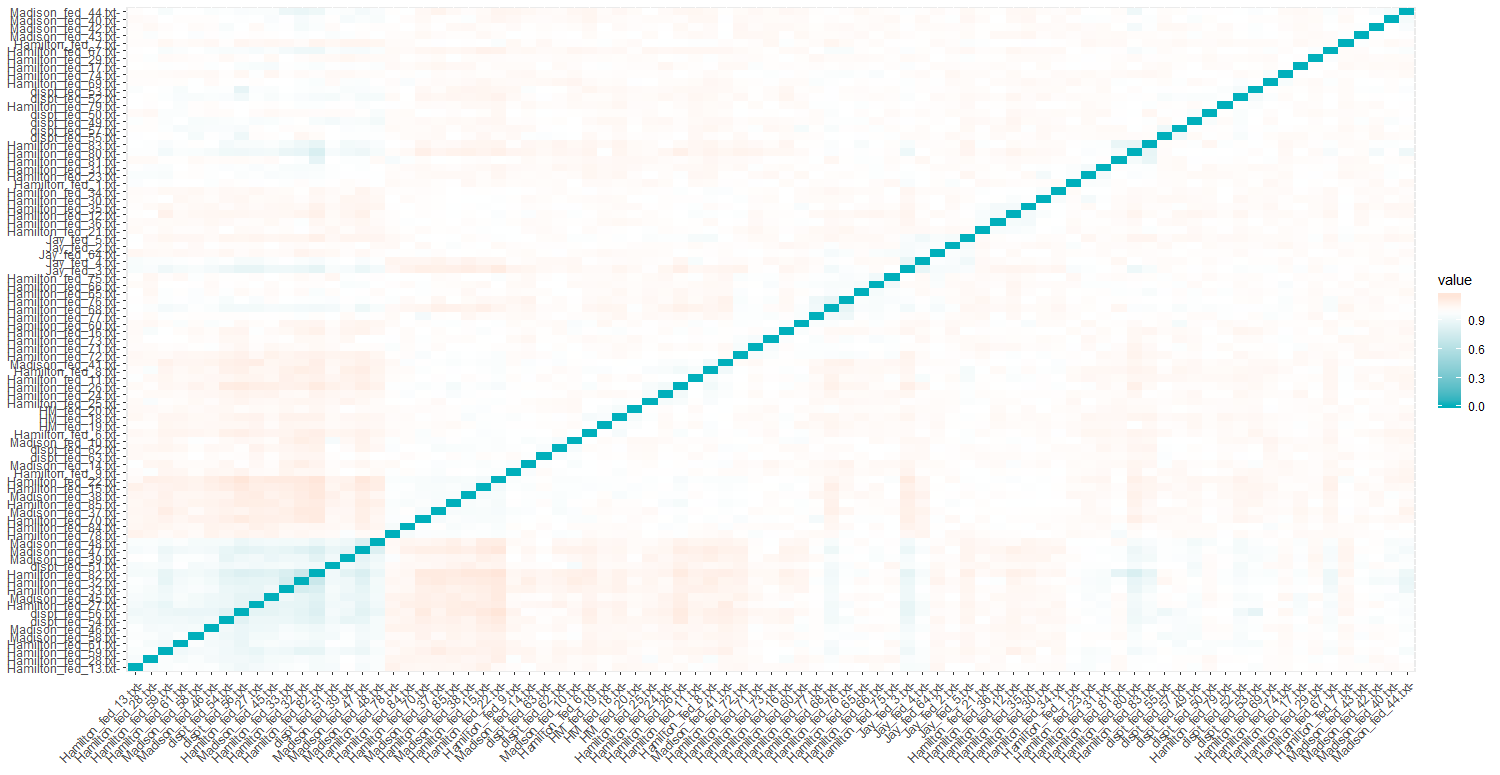
This clustering model does not seem to helpful in identifying the unknown authors either.

Manhattan



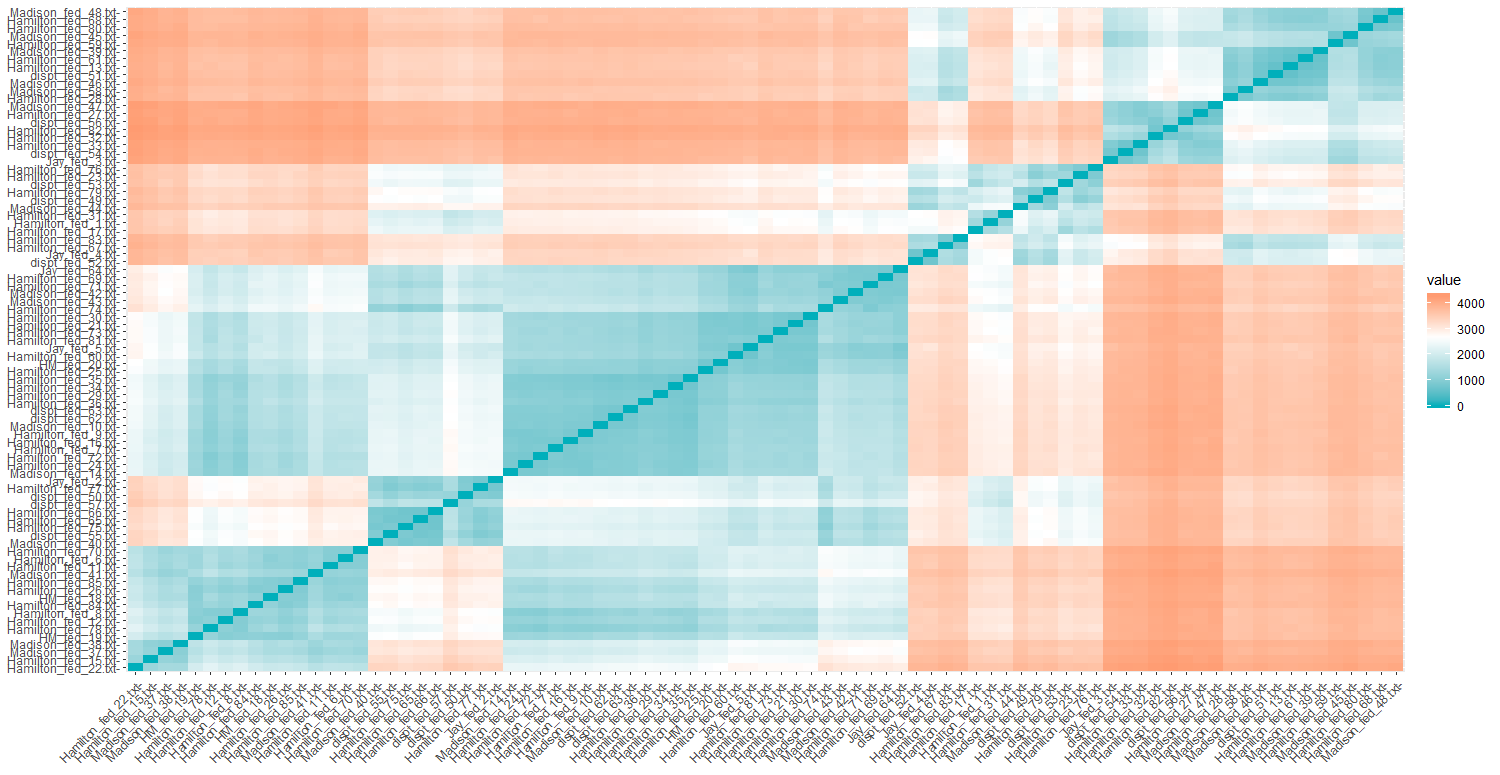
This model shows the grouping of the 3 different clusters.

Pearson



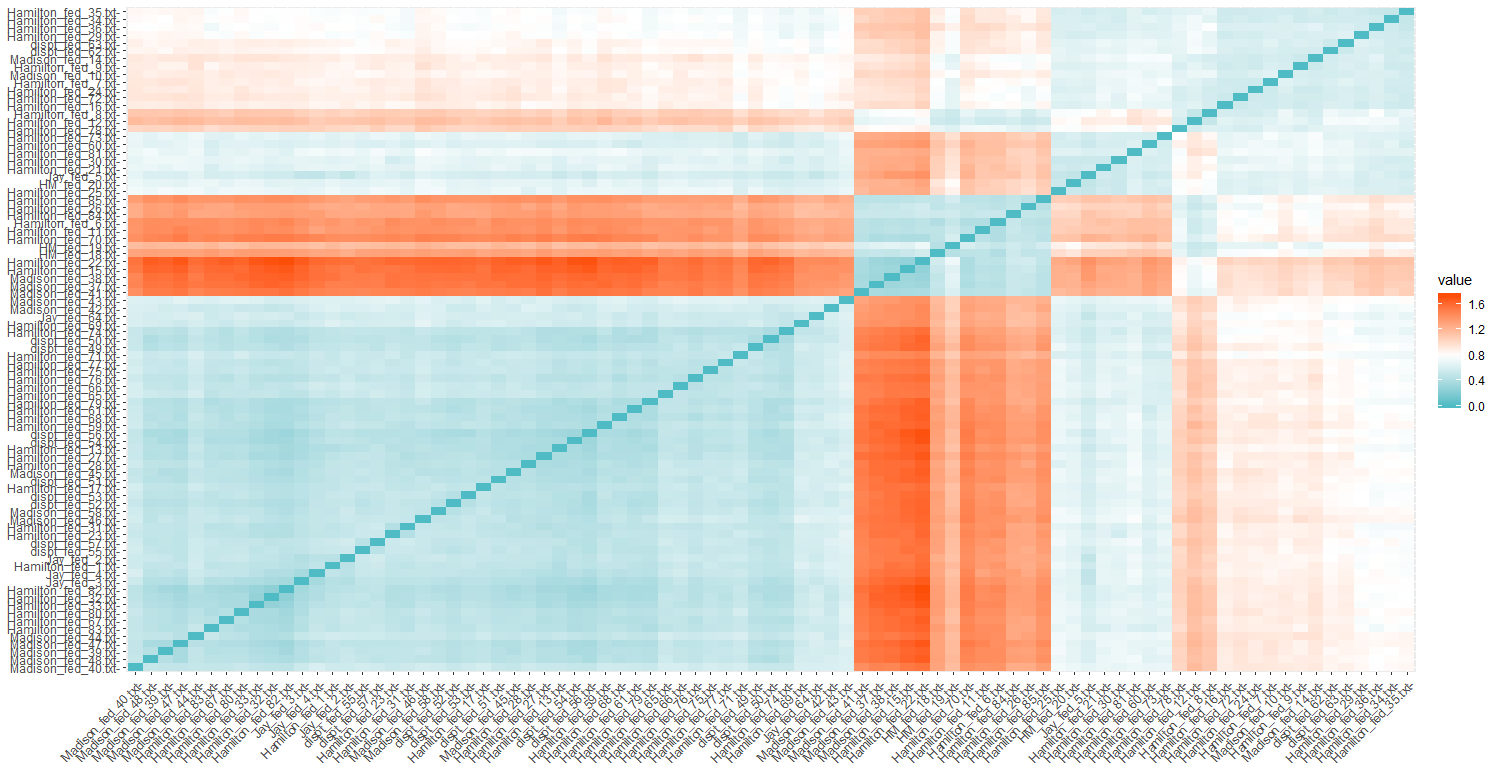
This model it is hard to differentiate between clusters.

Canberra



The model clearly shows 6 clusters.

Spearman



This model can be broken in to 3 or 4 clusters.

K-means Clustering



This model clearly has distinct separate clusters, but the authors are mixed in the clusters.

Expectation Maximization

The EM algorithm is used to find [maximum likelihood](https://en.wikipedia.org/wiki/Maximum_likelihood) parameters of a [statistical model](https://en.wikipedia.org/wiki/Statistical_model) in cases where the equations cannot be solved directly.

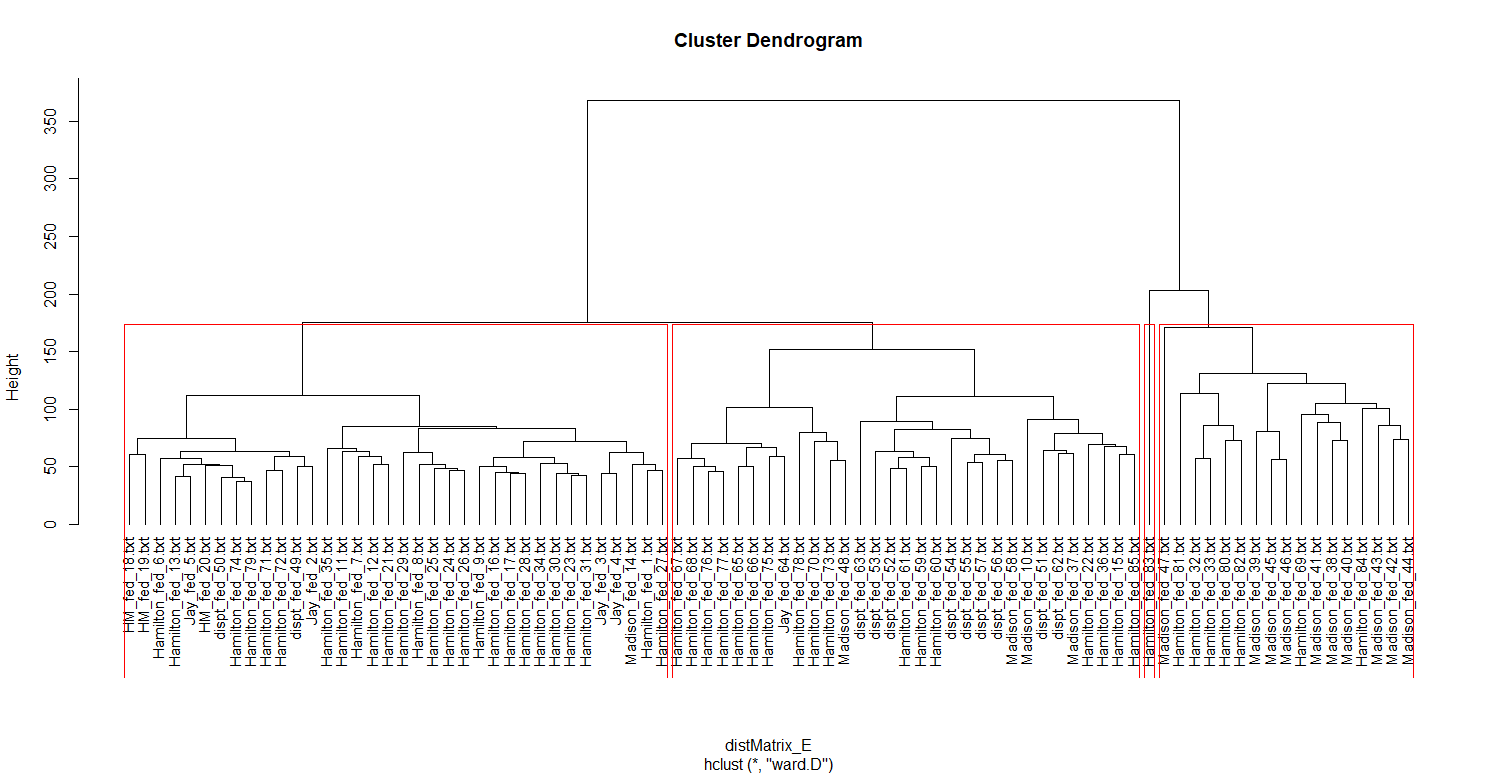
The clustering table produced results of which makes sense based on the fact 51 essays were written by Hamilton, 15 by Madison, 3 by Hamilton and Madison, 5 by Jay.

|  |  |
| --- | --- |
| 1 | 66 |
| 2 | 16 |
| 3 | 1 |
| 4 | 2 |

**Results**

To determine the authors of the essays that were unknown, the hierarchal model using Euclidean distance produces the easiest to read results.

* 50: Hamilton
* 49: Hamilton
* 63: Hamilton
* 53: Hamilton
* 52: Hamilton
* 54: Hamilton
* 55: Hamilton
* 57: Hamilton
* 58: Madison
* 51: Madison
* 62: Madison



**Conclusion**

The analysis is difficult, and the answer may never be determined. As always, additional data would be useful. Information on where the article was first printed would be helpful. There are different versions of the articles and essays available. Hamilton made corrections in the bound version which could be misleading the data analysis in to grouping articles to Hamilton’s writing. Articles were likely hand-written and sent to a printer, some similarities could be because they came from the same publisher or printer.

The hierarchal Euclidean clustering model provided the clearest results. It is important to experiment with several different models every time. Different data sets will have clearer results with different models.

Next steps are to compare the results with other students/mathematician results. Another thought would be to test the models that produced less than favorable results by removing the text files written by Jay to see if the models produce clearer results.