**Introduction**

The Federalist Papers are a series of 85 essays arguing in support of the United StatesConstitution. [Alexander Hamilton](https://www.constitutionfacts.com/us-founding-fathers/about-the-founding-fathers/#ah), [James Madison](https://www.constitutionfacts.com/us-founding-fathers/about-the-founding-fathers/#jm), and [John Jay](https://www.constitutionfacts.com/us-supreme-court/supreme-court-justices/) were the authors behind the pieces, and the three men wrote collectively under the name of Publius to stay anonymous. At the time of publication, the authorship of the articles was a closely guarded secret. It wasn't until Hamilton's death in 1804 that a list crediting him as one of the authors became public. It claimed fully two-thirds of the essays for Hamilton. Many of these would be disputed by Madison later. The Federalist Papers are considered one of the most important sources for interpreting and understanding the original intent of the Constitution.

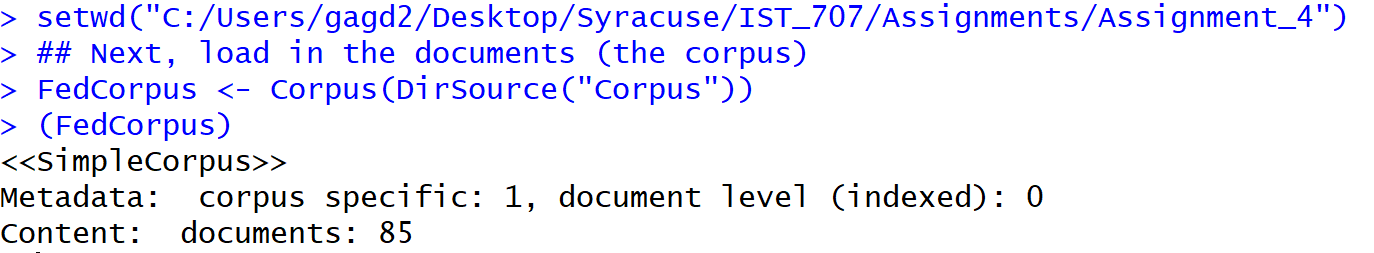
**Alexander Hamilton was the force behind the project and was responsible for recruiting James Madison and John Jay to write with him as Publius. John Jay was the author of five of the Federalist Papers. He would later serve as Chief Justice of the United States. Jay became ill after only contributed 4 essays and was only able to write one more before the end of the project. James Madison, Hamilton's major collaborator, later President of the United States and "Father of the Constitution", wrote 29 of the Federalist Papers, although Madison himself, and many others since then, asserted that he had written more. Various statisticians tried to find the authorship of the disputed papers and came out with their conclusions. One of the most famous analysis on the same was written by Mosteller and Wallace who analyzed the frequency distributions of common words in the papers to draw their conclusions. The analysis provided in this study will use k-means clustering technique to determine who wrote the disputed papers.**

**Analysis and Models**

**About the Data**

**The original essays are available for free in the web and can be downloaded easily. There is a total of 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, is 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.**

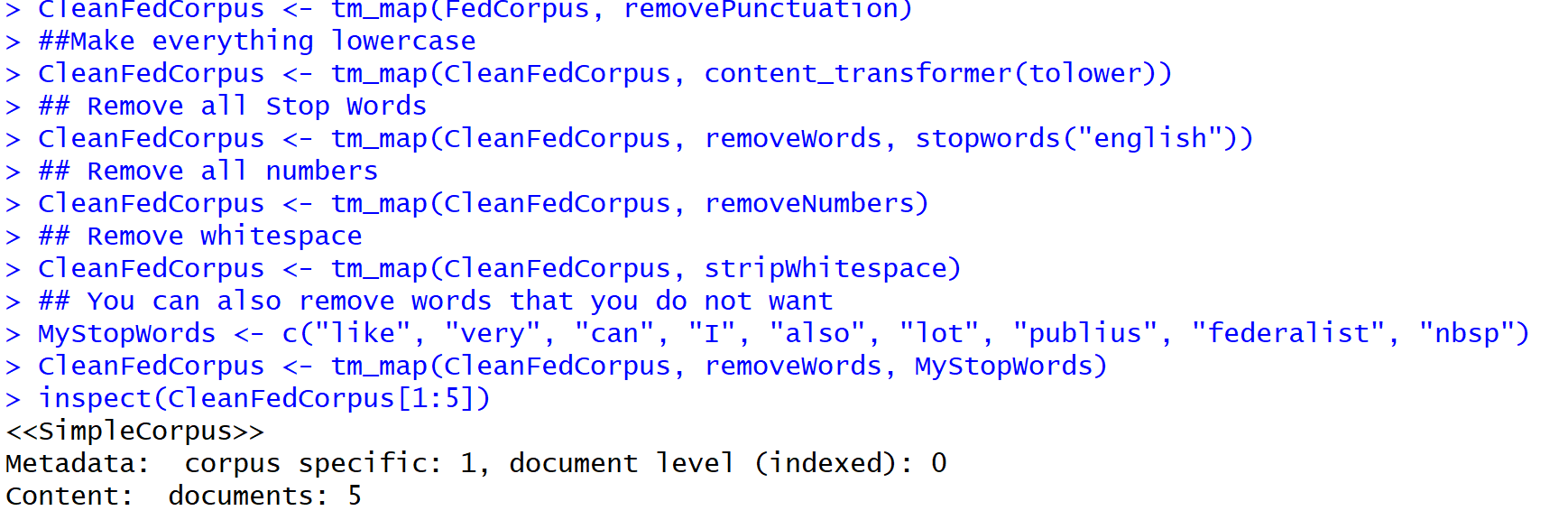
**To identify the authorship of the essays easily, the files were named with the author names who wrote it. For example, an essay written by Hamilton was named as ‘Hamilton\_fed\_1’. The essays that are under dispute for authorship are named as ‘dispt\_fed\_49’ with the last digits representing the file number. All the 85 essay files are stored in the same folder called ‘Corpus’. As always, data must be cleaned and preprocessed before analysis happens. First, documents are read into and converted into a format that can be used for the analysis. Necessary libraries are installed, working directory path is set and the files are read in and stored in a variable called *FedCorpus*.**

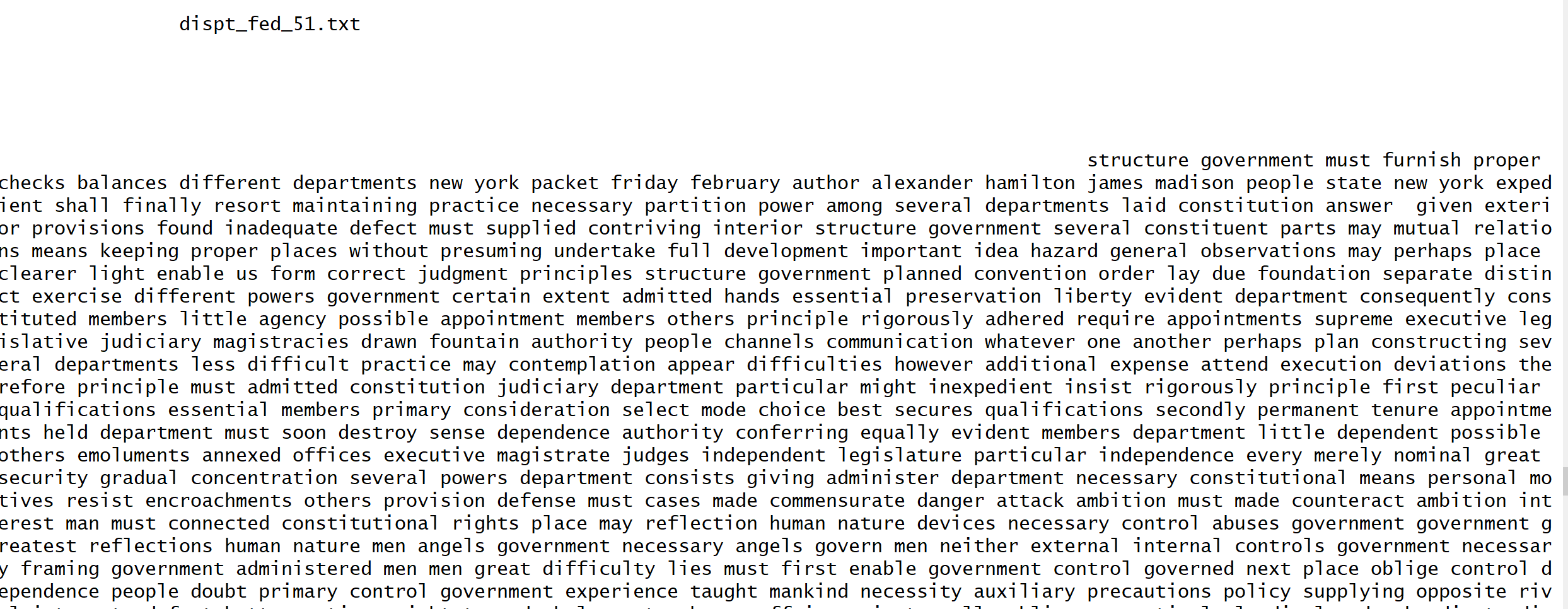
****

**When the variable is viewed, the above screenshot shows that 85 documents were read in successfully.**

**The names of the files are kept as a list so that they can be used later with the document term matrix. Text files usually have several unwanted words, punctuations, numbers and whitespaces that needs to be removed for effective analysis. Files will also have common and useless words like ‘and’, ‘the’, ‘or’ that need to be removed as well. In order to apply these transformations on all the text files at once, *tm\_map* function is used.**

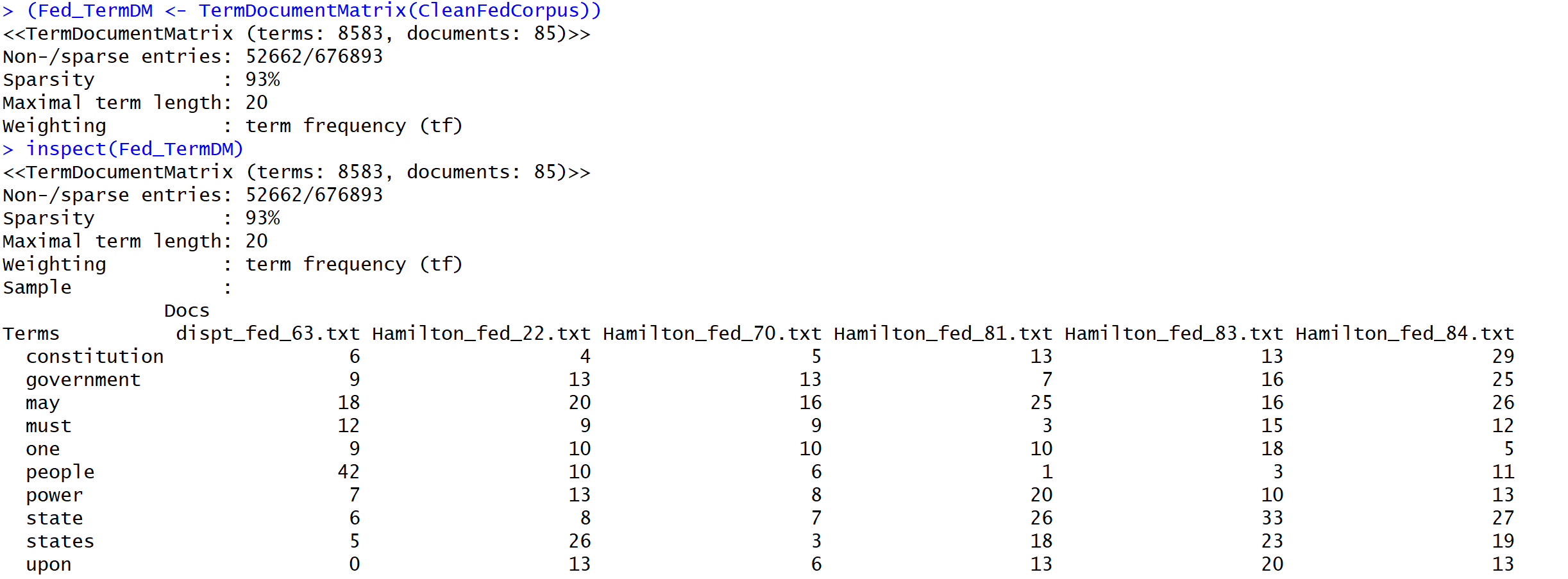
**The below commands remove punctuations, transforms all words into lower case, removes all stop words, numbers, whitespaces and also removes a list of unwanted words that were provided as a list.**

****

****

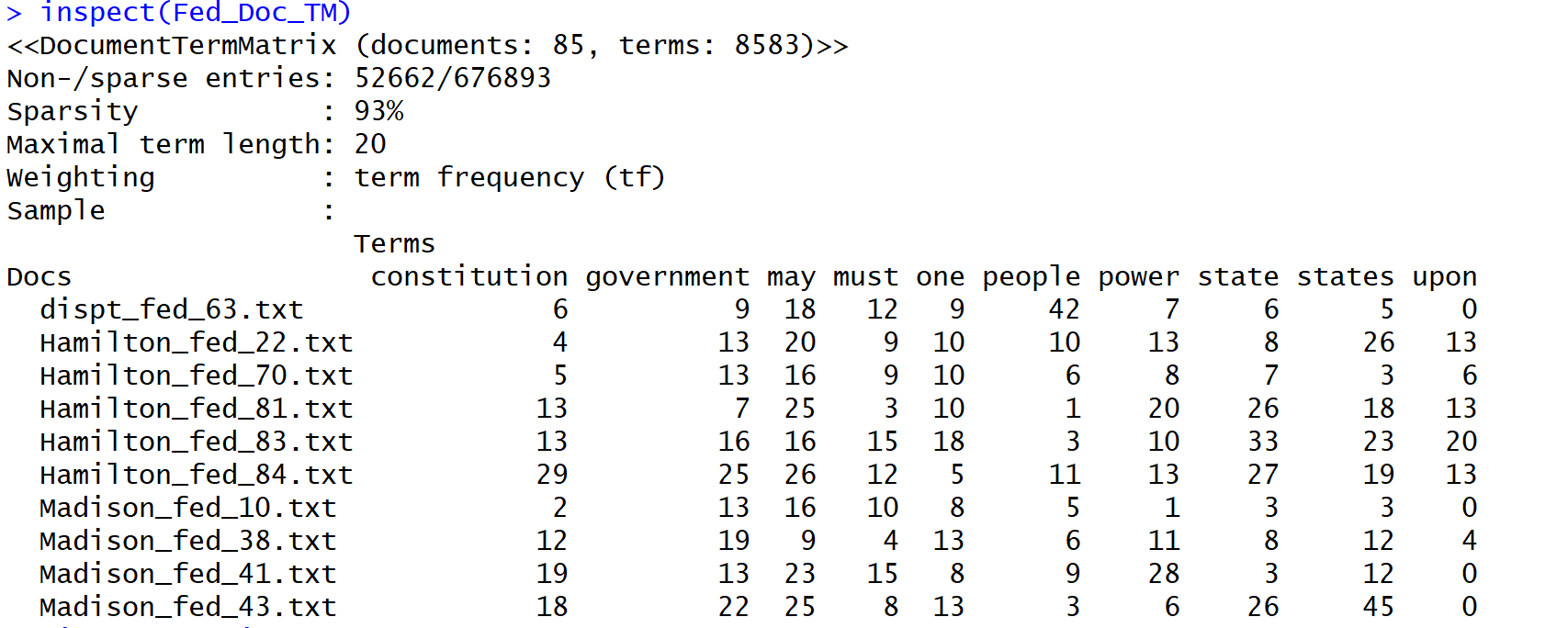
**A quick inspection of the cleaned file looks like the above screenshot where all the transformations have been applied.**

**Now that the data has been cleaned, some preprocessing is done to convert the data into a format that is required for the analysis. In order to find the frequency of each word in the files, all files must be viewed as a document matrix. There are two ways in which this can be viewed. Term Document Matrix (TDM) function transforms the terms(words) into rows and documents(filenames) into columns whereas the Document Term Matrix (DTM) function transforms the terms(words) into columns and documents(filenames) into rows. Both ways are explored here to look at different perspectives. Below is a screenshot of the TDM that was created from the cleaned corpus file.**

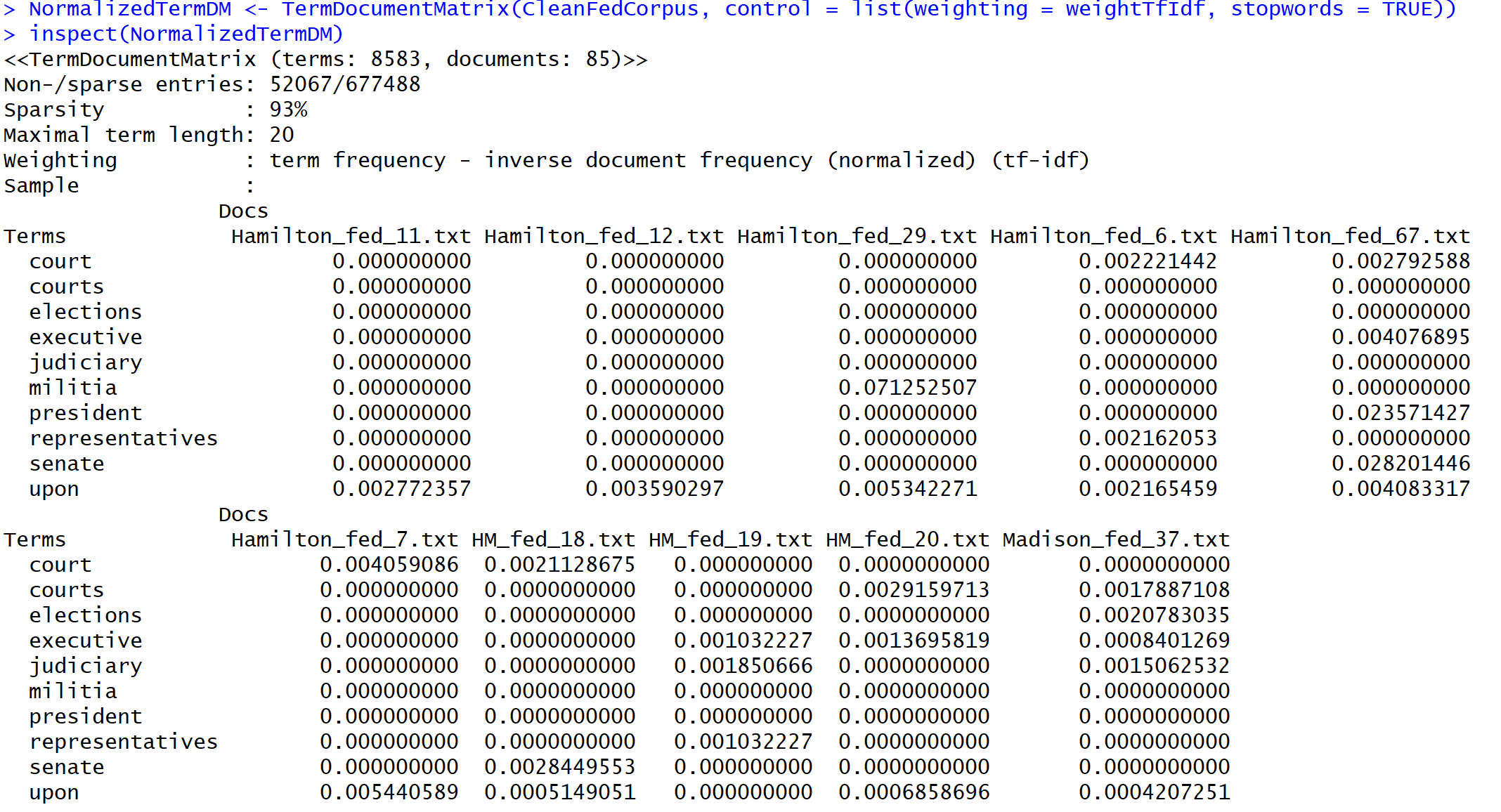
****

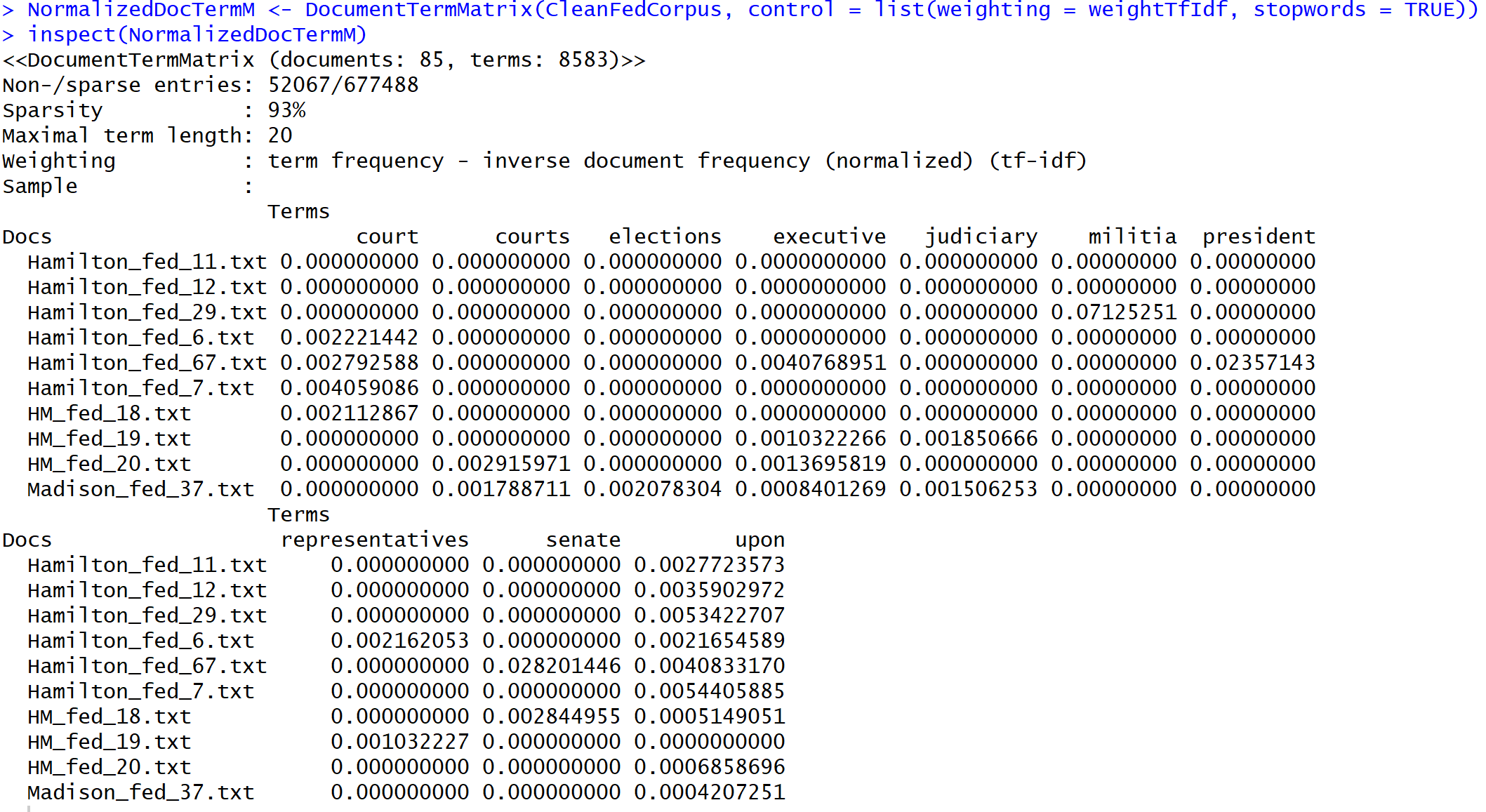
The column names show the file names and the row names show the words in the documents. The values are the frequencies of each word in a specific document.

Below is a screenshot of the DTM that was created from the cleaned corpus file. Here, the rows are the file names and the columns are the words found in these files.



The next step is to normalize these documents. Unless the words in all these files are of equal size, any type of analysis done on these files will be a waste. Hence, in order to normalize, a word’s frequency is divided by the total number of words in that document. In addition to that, the weights of the terms also need to be normalized where the frequently appearing words will have to be weighed down while scaling up the rare words. Both these approaches are applied to the matrices by specifying the weighting parameter. Below is a screenshot of the commands and the outputs for the same.





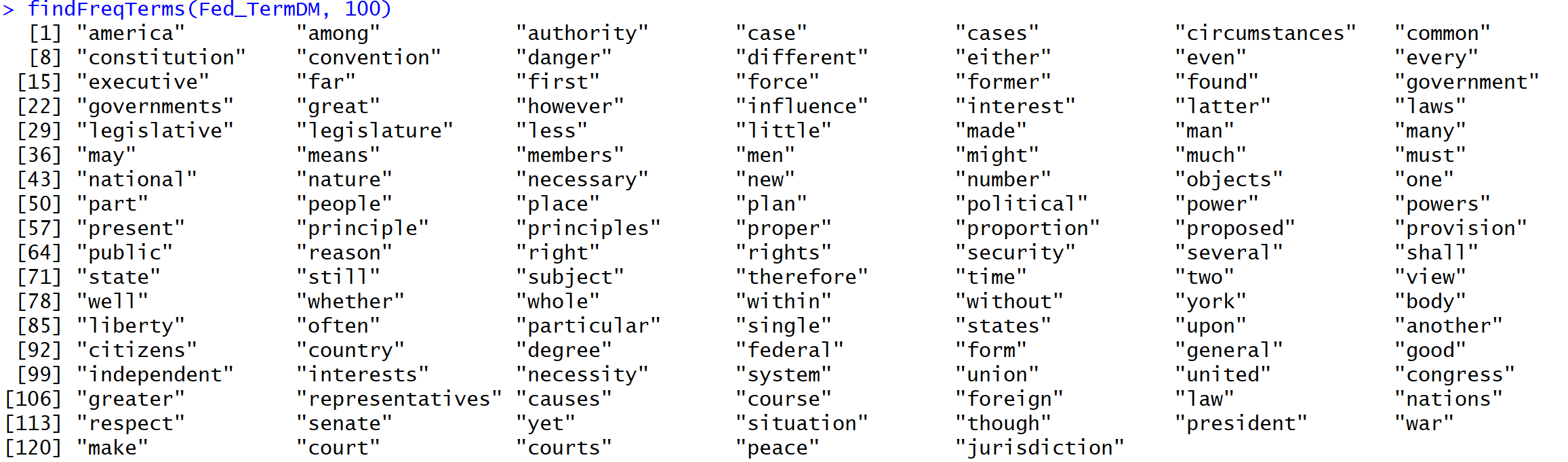
The above outputs show the weights applied to the terms in the matrices.

Now that the data has been cleaned and transformed into the necessary format, further analysis can be done to answer our questions. A couple of popular clustering techniques will be used for the analysis of the federalist papers.

**Clustering Techniques**

As per the definition provided in the book ‘Introduction to Data Mining’, cluster analysis groups data objects based only on information found in the data that describes the objects and their relationships. The goal of cluster analysis is to group similar objects in a cluster and different from the objects in other groups. There are different types of clustering techniques and different types of clusters. Two such techniques are used in this paper to understand the data and cluster similar files in groups.

To get a descriptive idea of the data, an overall frequency of the words can be calculated and viewed. For example, the below snapshot shows the words that have appeared more than 100 times in the entire corpus.

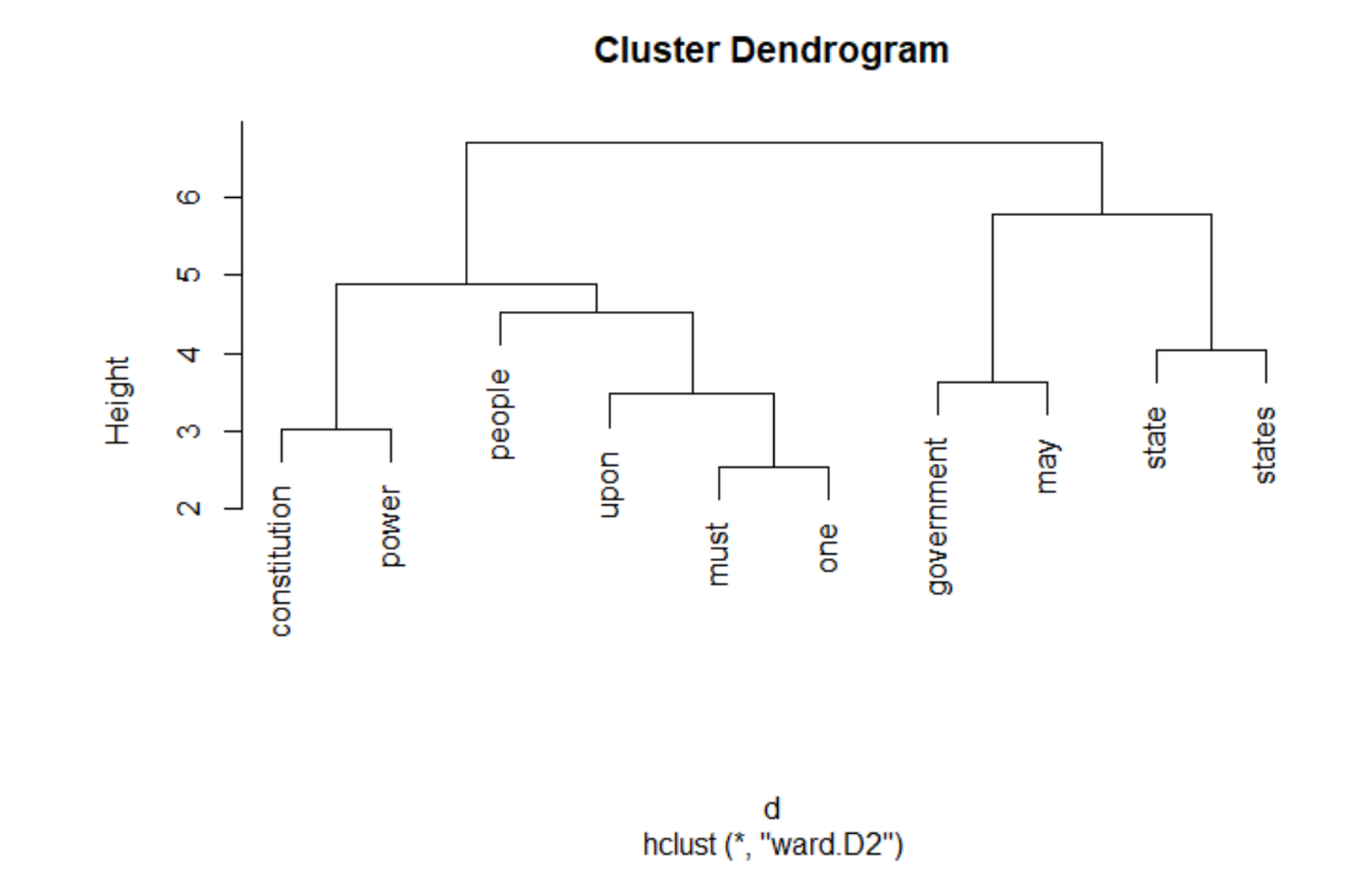


Given that these documents are related to the U.S constitution, the frequently used words in these papers are all related to the same.

**Agglomerative Hierarchical Clustering Technique**

This clustering technique produces 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. This algorithm computes the proximity between two clusters and forms a matrix of all the proximities found. Next, it merges the closest two clusters and updates the proximity matrix to reflect the proximity between the new cluster and the original clusters. This process is repeated until only one cluster remains. Different techniques can be used to calculate the proximity measure. In this paper, *Euclidean* distance method and Ward’s method is used to calculate the proximity between two clusters. Ward’s method calculates the proximity as the increase in the squared error that results when two clusters are merged.

A hierarchical clustering can be displayed graphically using a dendrogram which displays both the cluster and sub cluster relationships and the order in which they were merged. The below figure shows the dendrogram of the words clustered from the federalist papers before normalizing the same.

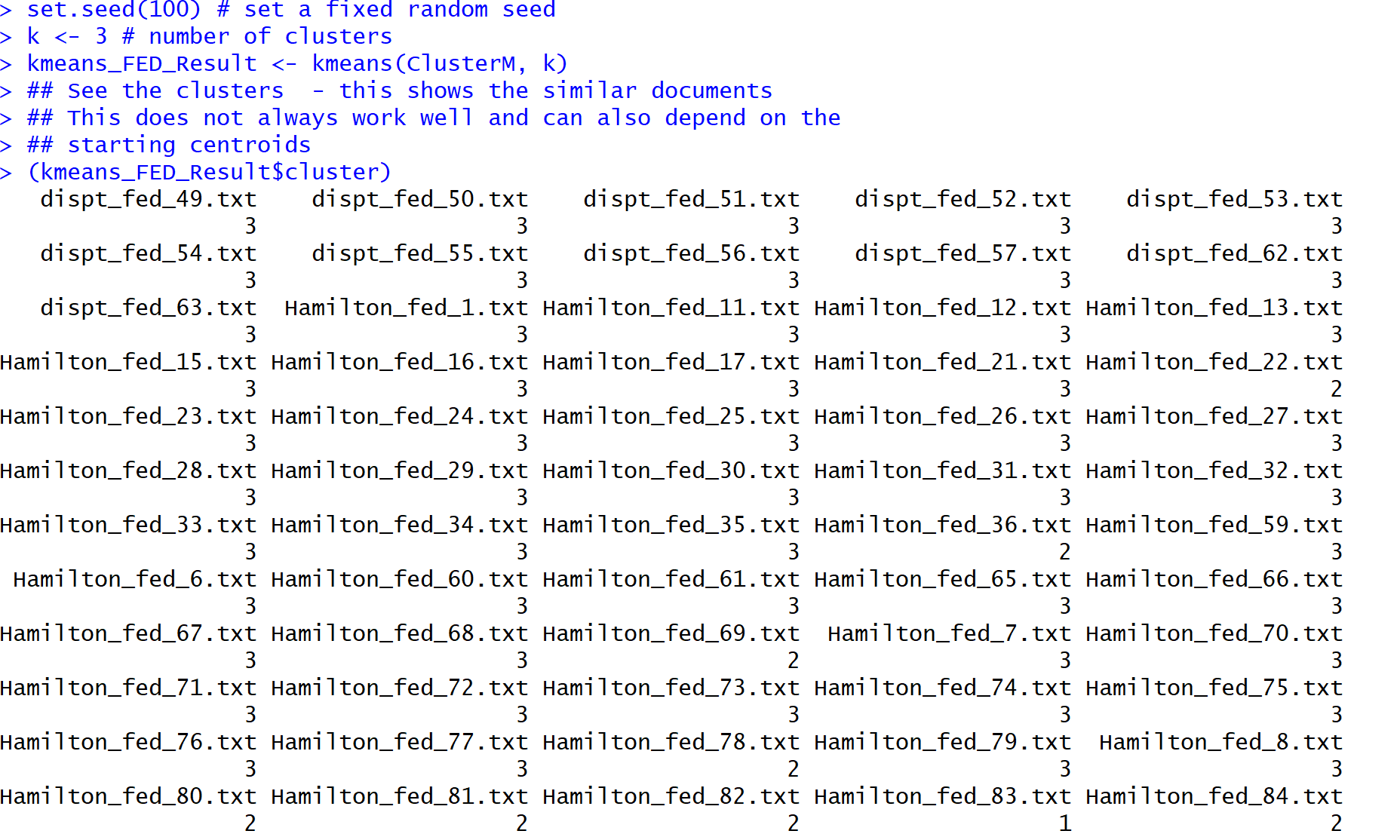


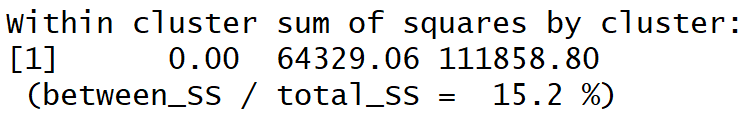
The above dendrogram shows that the words ’government’ and ‘may’ belong to one cluster, ‘constitution’ and ‘power’ belong to another, ‘must’ and ‘one’ form another. Later, the cluster with words ‘must’ and ‘one’ are combined with another word ‘upon’ to form another cluster and so on.

**K-means clustering Technique**

K-means technique is a prototype-based, partitional clustering technique that attempts to find a user-specified number of clusters(K), which are represented by their centroids. This technique can be used to confirm business assumptions about what types of groups exist or to identify unknown groups in complex data sets. Once the algorithm has been run and the groups are defined, any new data can be easily assigned to the correct group. The algorithm iterates between two steps. First, K points are selected as initial centroids where K is a user-specified parameter. Second, K clusters are formed by assigning each point to its closest centroid. The centroid of each cluster is recalculated until the centroids don’t change. This technique will be useful to cluster the federalist papers into groups that could help identify who wrote the disputed papers.

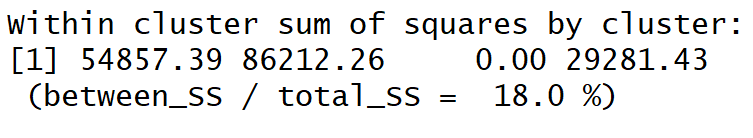
In order to cluster the documents, the TDM is transposed and an initial number of 3 clusters is provided as a parameter to the k-means function. Viewing the results below, all the disputed document papers are in cluster 3 along with some documents written by Hamilton and some written by Madison.





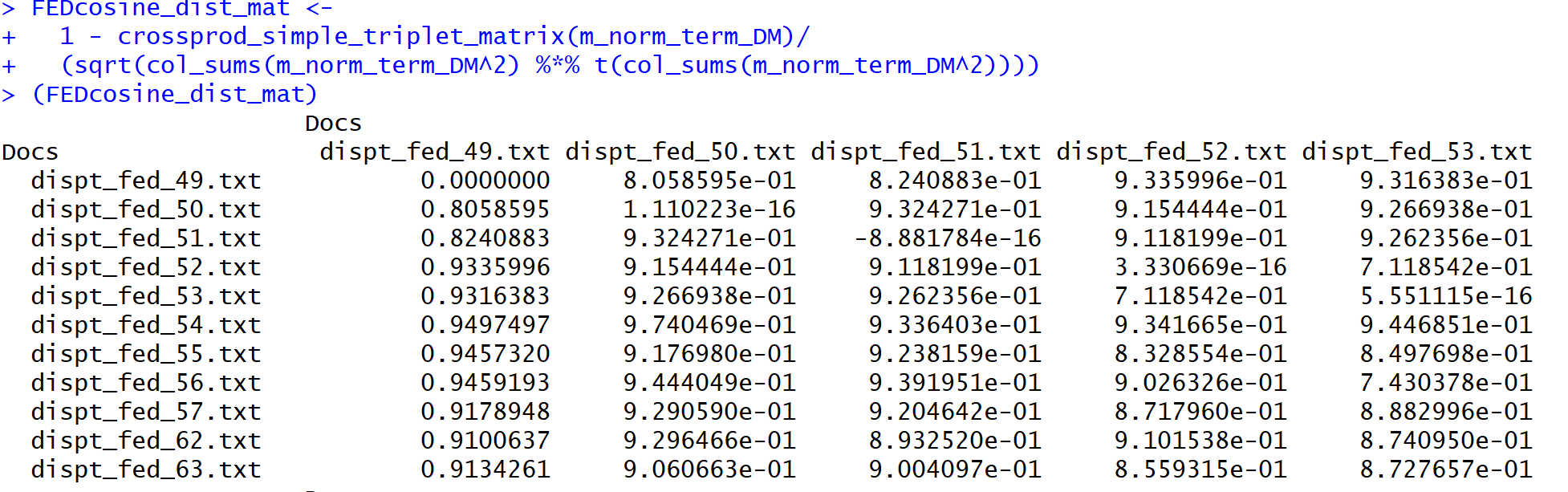
When random initialization of centroids is used, different runs of K-means produce different total Sum of Squared Errors (SSE). One common technique used to address this problem is to perform multiple runs, each with a different set of random initial centroids, and then select the set of clusters with the minimum SSE.

Keeping this in mind, if K is set to 4, the SSE increases to 18% which is not good.

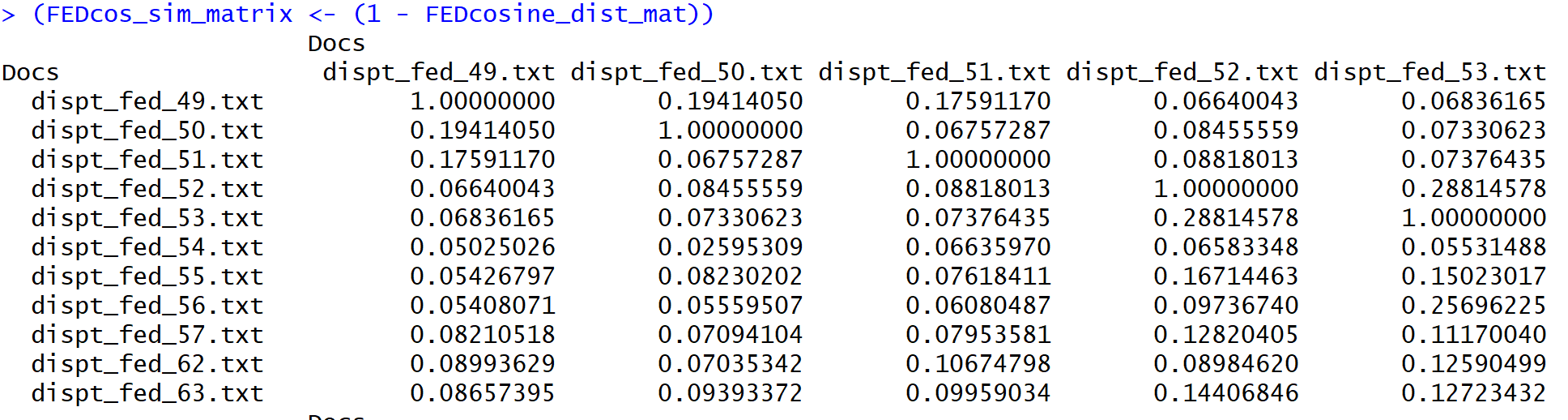
****

Even with repeated runs, there are problems of ineffective clustering with randomly selected initial centroids. In order to get a better fit, cosine proximity function is used instead of the Euclidean function that was used before. The objective function of the cosine function is the maximize the sum of the cosine similarity of an object to its cluster centroid.

By calculating the cosine distance function of the normalized TDM, below results are obtained.

****

The above results show the dissimilarity of the clusters. Rather, in order to find out the means of those documents that are closer to or like each other, the below command is used.

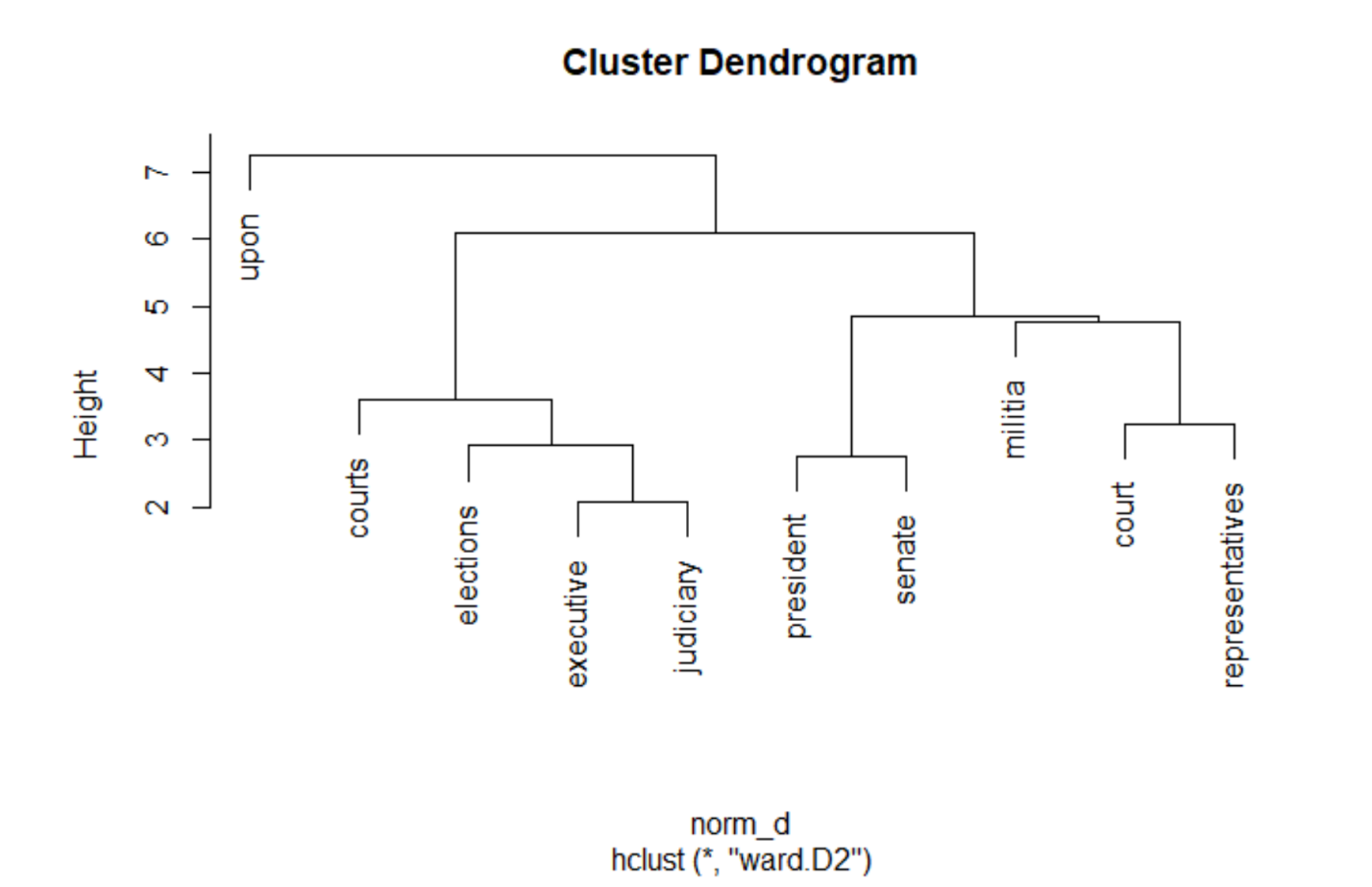


The above output shows that values closer to 1 show the similarity of the documents and those that are closer to 0 show the dissimilarity.

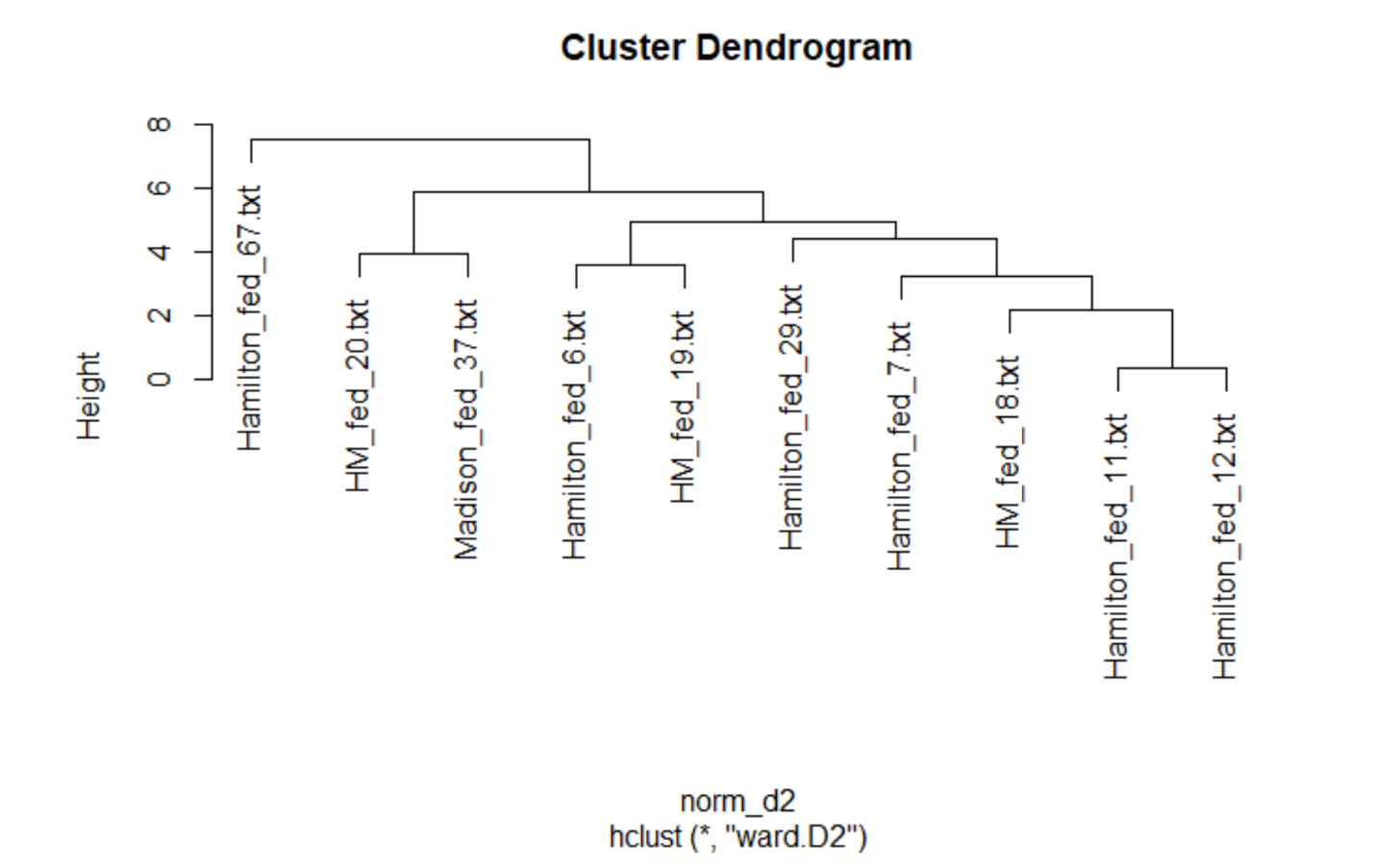
**Results**

**Agglomerative Hierarchical Clustering Results**

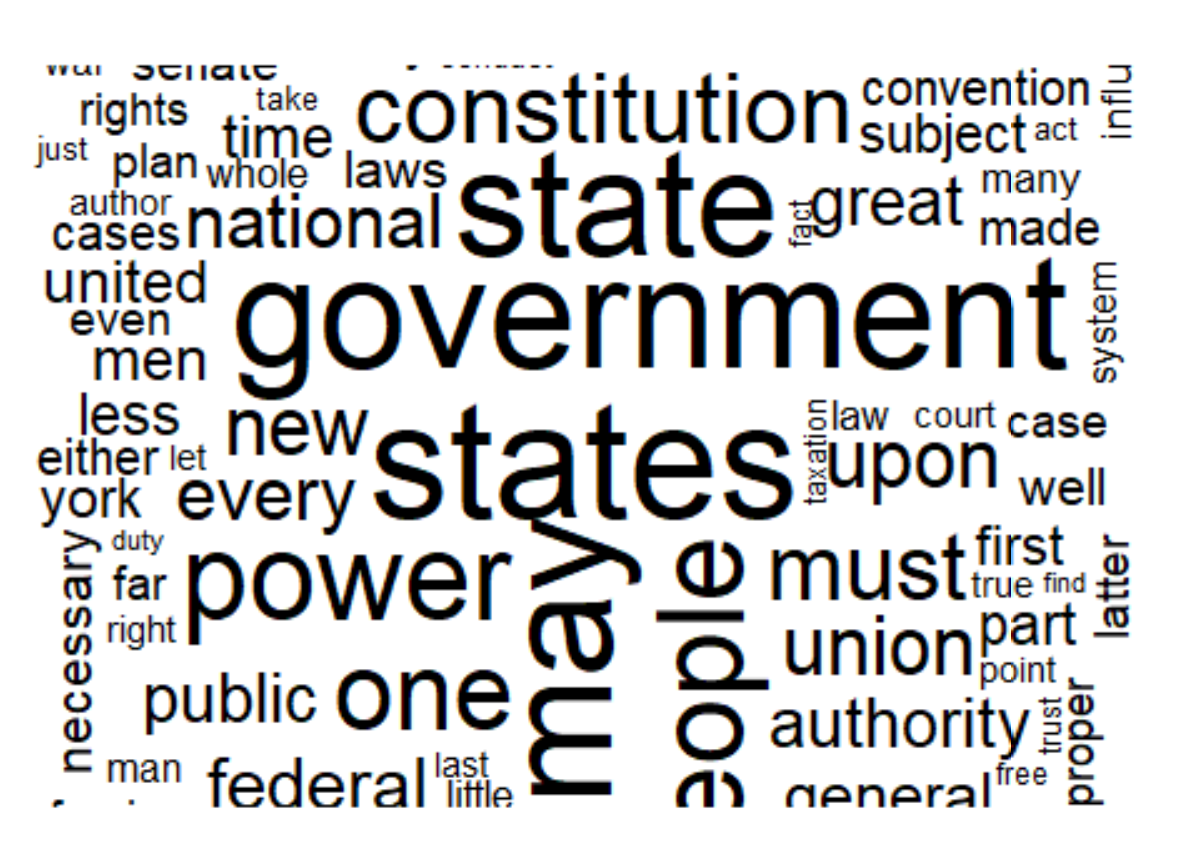
When the dendrogram is visualized using the normalized TDM, the results are different. Terms like ‘executive’ and ‘judiciary’ are grouped into the same cluster which is in turn clustered with the term ‘elections’ and so on. This shows how important normalization is in order to get the right insights.



The same visualization is applied to the DTM to visualize similar documents that form a cluster. It’s quite interesting to see how Madison’s paper 37 is clustered along with paper 20 that was jointly written by Madison and Hamilton. Likewise, Hamilton’s papers 11 and 12 are clustered together with paper 18 that was jointly written by Madison and Hamilton.

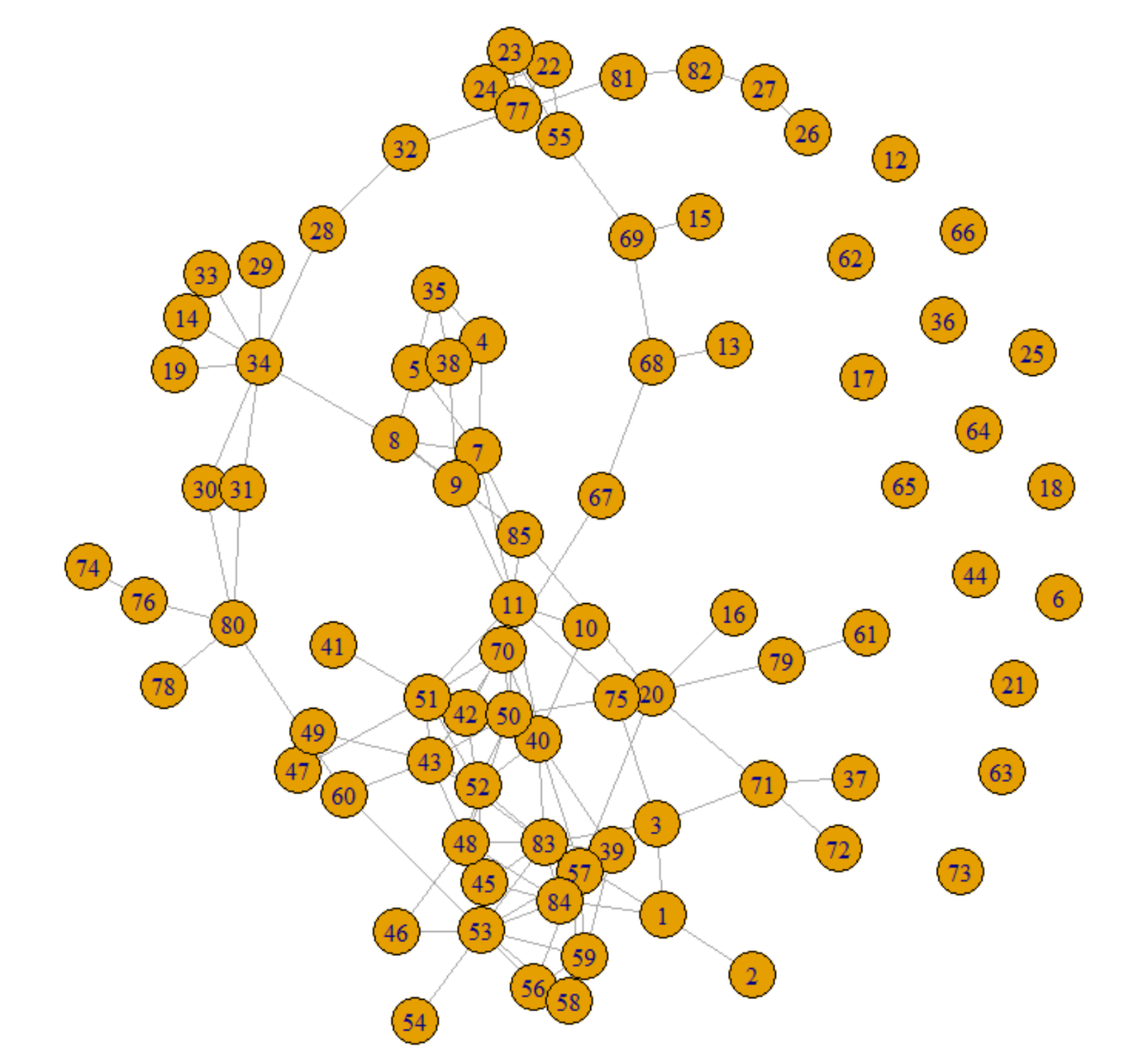


Packages like ‘*word cloud’* can also be used to visualize the frequency of words in these papers. It is easy to infer from the below picture that the terms *government, state, may, people, constitution* are some of the most frequently used words in these papers.



**K-means Clustering Results**

To answer the most important question of which document the disputed papers are the most similar, the values in the above matrix are updated to 0 if it is less than 0.15 and updated to 1 if it is more than or equal to 0.15. To visualize the same as a network, network matrix function is used and the same is plotted using a graph. To add the column names as vertices and remove loops, functions ‘graph from adjacency matrix’ and ‘simplify’ are used. When plotted, the below graph is outputted.



In the matrix above, vertices 1 through 11 show the disputed papers, 12-62 show papers authored by Hamilton, 63-65 show papers authored by both Hamilton and Madison, 66-70 were authored by Jay and lastly, 71-85 were authored by Madison. The graph shows that out of the 11 disputed papers, papers 1,2,3 and 7 were written by Madison since they are closer to vertices of the papers written by Madison. Papers 4,5 and 10 are closer to those vertices of the papers written by Hamilton. Papers 6, 8, 9 and 11 are closer to both the types of vertices of the papers written by Hamilton and Madison. Thus, the K- means clustering technique is not able to identify all the papers accurately.

The divided results could also throw some light as to why there was a dispute for authorship in the first place. It doesn’t look like either of the authors are trying to steal each other’s work. Due to the collaboration spanning over several years, it could be possible that their work got mixed up and the amount of work each author put on these different disputed papers varied.

**Conclusion**

**Even though the above techniques provide some conclusion as to who wrote which paper, one cannot speak to the accuracy of these results. Hence, different approaches and techniques need to be applied to this data to compare the various results. Even if other techniques could identify the authorship, one will never know the accuracy of it unless the authors themselves tell the truth. However, the above techniques were useful and helped gather some insights that could be more valuable if clubbed with more data or techniques. Another suggestion to try would be to remove the papers written by Jay to see if that helps to form better clusters. It would also be interesting to mine other books or papers written by the two authors to ascertain their style of writing and usage of words which in turn would help to analyze these federalist papers better.**

**The idea of using cluster analysis for text analysis developed from the pioneering work done by Mosteller and Wallace in the 1960’s. They analyzed the frequency distributions of high frequency words in the Federalist papers to draw their conclusions. An important observation to note from their work is the fact that they not only used statistics but involved domain experts in their studies to get different insights. Hence, further refinement of our analysis using Mosteller’s work will be imperative in finding more conclusive and accurate results.**