**Syracuse University**

**IST-707 Assignment 4**

ThulasiRam RuppaKrishnan

IST 707Section: 35

Professor Ami Gates

Contents

[Introduction 3](#_Toc7814370)

[Analysis and Models 5](#_Toc7814371)

[**About the data** 5](#_Toc7814372)

[**Models** 10](#_Toc7814373)

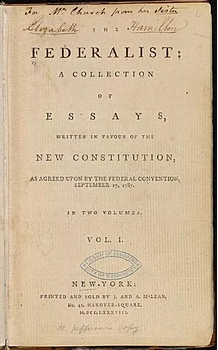
[Results 24](#_Toc7814374)

[Conclusion 29](#_Toc7814375)

## 

## **Introduction**

The Federalist (later known as The Federalist Papers) is a collection of 85 articles and essays written by Alexander Hamilton, James Madison, and John Jay under the pseudonym "Publius" to promote the ratification of the United States Constitution. The first 77 of these essays were published serially in the Independent Journal, the New York Packet, and The Daily Advertiser between October 1787 and April 1788. A two-volume compilation of these 77 essays and eight others was published as "The Federalist: A Collection of Essays", Written in Favor of the New Constitution, as Agreed upon by the Federal Convention, September 17, 1787, by publishing firm J. & A. McLean in March and May 1788.



The authors of The Federalist intended to influence the voters to ratify the Constitution. In "Federalist No. 1", they explicitly set that debate in broad political terms:

It has been frequently remarked, that it seems to have been reserved to the people of this country, by their conduct and example, to decide the important question, whether societies of men are capable or not, of establishing good government from reflection and choice, or whether they are forever destined to depend, for their political constitutions, on accident and force.

"Federalist No. 10" is generally regarded as the most important of the 85 articles from a philosophical perspective. In it, Madison discusses the means of preventing rule by majority faction and advocates a large, commercial republic. This is complemented by "Federalist No. 14", in which Madison takes the measure of the United States, declares it appropriate for an extended republic, and concludes with a great defense of the constitutional and political creativity of the Federal Convention. In "Federalist No. 84", Hamilton makes the case that there is no need to amend the Constitution by adding a Bill of Rights, insisting that the various provisions in the proposed Constitution protecting liberty amount to a "bill of rights." "Federalist No. 78", also written by Hamilton, lays the groundwork for the doctrine of judicial review by federal courts of federal legislation or executive acts. "Federalist No. 70" presents Hamilton's case for a one-person chief executive. In "Federalist No. 39", Madison presents the clearest exposition of what has come to be called "Federalism." In "Federalist No. 51", Madison distills arguments for checks and balances in an essay often quoted for its justification of government as "the greatest of all reflections on human nature."

According to historian Richard B. Morris, the essays that make up The Federalist Papers are an "incomparable exposition of the Constitution, a classic in political science unsurpassed in both breadth and depth by the product of any later American writer."

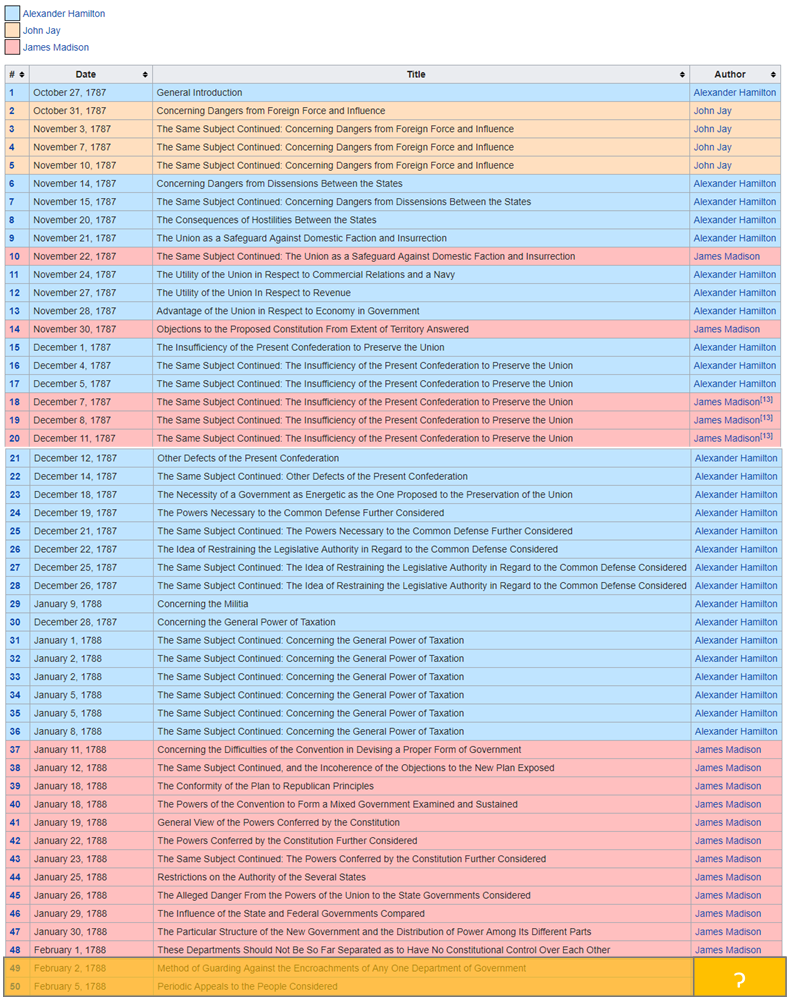
The authorship of certain of the Federalist essays was disputed from the beginning. Both Hamilton and Madison produced lists that claimed some of the same papers. There followed a series of lists, some claiming authorship for Madison and some for Hamilton.

The consensus of traditional scholarship, seconded by Mosteller and Wallace, allocates the papers: Hamilton 51 (1, 6-9, 11-13, 15-17, 21-36, 59-61, 65-85); Madison 29 (10, 14, 18-20, 37-58, 62, 63); Jay 5 (2-5, 64). Mosteller and Wallace set the boundary conditions for the subsequent non-traditional work – e.g., not using the Jay articles as a control. Most of these later practitioners do not select or prepare the input text as carefully as Mosteller and Wallace – and their selection and preparation were not as rigorous and complete as it should have been – as we will see.

## **Analysis and Models**

### **About the data**

The list of federalist papers is shown in **Table 1.1** and are highlighted with colors for each author.



****

**Table 1.1 Federalist papers by Author**

In the above list of papers, the ones which are highlighted is disputed between two authors Madison and Hamilton. Clustering techniques are applied here to solve the puzzle.

**Summary of the imported corpus data**

(summary(EssayCorpus))

Length Class Mode

dispt\_fed\_49.txt 2 PlainTextDocument list

dispt\_fed\_50.txt 2 PlainTextDocument list

dispt\_fed\_51.txt 2 PlainTextDocument list

dispt\_fed\_52.txt 2 PlainTextDocument list

dispt\_fed\_53.txt 2 PlainTextDocument list

dispt\_fed\_54.txt 2 PlainTextDocument list

dispt\_fed\_55.txt 2 PlainTextDocument list

dispt\_fed\_56.txt 2 PlainTextDocument list

dispt\_fed\_57.txt 2 PlainTextDocument list

dispt\_fed\_62.txt 2 PlainTextDocument list

dispt\_fed\_63.txt 2 PlainTextDocument list

Hamilton\_fed\_1.txt 2 PlainTextDocument list

Hamilton\_fed\_11.txt 2 PlainTextDocument list

Hamilton\_fed\_12.txt 2 PlainTextDocument list

Hamilton\_fed\_13.txt 2 PlainTextDocument list

Hamilton\_fed\_15.txt 2 PlainTextDocument list

Hamilton\_fed\_16.txt 2 PlainTextDocument list

Hamilton\_fed\_17.txt 2 PlainTextDocument list

Hamilton\_fed\_21.txt 2 PlainTextDocument list

Hamilton\_fed\_22.txt 2 PlainTextDocument list

Hamilton\_fed\_23.txt 2 PlainTextDocument list

Hamilton\_fed\_24.txt 2 PlainTextDocument list

Hamilton\_fed\_25.txt 2 PlainTextDocument list

Hamilton\_fed\_26.txt 2 PlainTextDocument list

Hamilton\_fed\_27.txt 2 PlainTextDocument list

Hamilton\_fed\_28.txt 2 PlainTextDocument list

Hamilton\_fed\_29.txt 2 PlainTextDocument list

Hamilton\_fed\_30.txt 2 PlainTextDocument list

Hamilton\_fed\_31.txt 2 PlainTextDocument list

Hamilton\_fed\_32.txt 2 PlainTextDocument list

Hamilton\_fed\_33.txt 2 PlainTextDocument list

Hamilton\_fed\_34.txt 2 PlainTextDocument list

Hamilton\_fed\_35.txt 2 PlainTextDocument list

Hamilton\_fed\_36.txt 2 PlainTextDocument list

Hamilton\_fed\_59.txt 2 PlainTextDocument list

Hamilton\_fed\_6.txt 2 PlainTextDocument list

Hamilton\_fed\_60.txt 2 PlainTextDocument list

Hamilton\_fed\_61.txt 2 PlainTextDocument list

Hamilton\_fed\_65.txt 2 PlainTextDocument list

Hamilton\_fed\_66.txt 2 PlainTextDocument list

Hamilton\_fed\_67.txt 2 PlainTextDocument list

Hamilton\_fed\_68.txt 2 PlainTextDocument list

Hamilton\_fed\_69.txt 2 PlainTextDocument list

Hamilton\_fed\_7.txt 2 PlainTextDocument list

Hamilton\_fed\_70.txt 2 PlainTextDocument list

Hamilton\_fed\_71.txt 2 PlainTextDocument list

Hamilton\_fed\_72.txt 2 PlainTextDocument list

Hamilton\_fed\_73.txt 2 PlainTextDocument list

Hamilton\_fed\_74.txt 2 PlainTextDocument list

Hamilton\_fed\_75.txt 2 PlainTextDocument list

Hamilton\_fed\_76.txt 2 PlainTextDocument list

Hamilton\_fed\_77.txt 2 PlainTextDocument list

Hamilton\_fed\_78.txt 2 PlainTextDocument list

Hamilton\_fed\_79.txt 2 PlainTextDocument list

Hamilton\_fed\_8.txt 2 PlainTextDocument list

Hamilton\_fed\_80.txt 2 PlainTextDocument list

Hamilton\_fed\_81.txt 2 PlainTextDocument list

Hamilton\_fed\_82.txt 2 PlainTextDocument list

Hamilton\_fed\_83.txt 2 PlainTextDocument list

Hamilton\_fed\_84.txt 2 PlainTextDocument list

Hamilton\_fed\_85.txt 2 PlainTextDocument list

Hamilton\_fed\_9.txt 2 PlainTextDocument list

HM\_fed\_18.txt 2 PlainTextDocument list

HM\_fed\_19.txt 2 PlainTextDocument list

HM\_fed\_20.txt 2 PlainTextDocument list

Jay\_fed\_2.txt 2 PlainTextDocument list

Jay\_fed\_3.txt 2 PlainTextDocument list

Jay\_fed\_4.txt 2 PlainTextDocument list

Jay\_fed\_5.txt 2 PlainTextDocument list

Jay\_fed\_64.txt 2 PlainTextDocument list

Madison\_fed\_10.txt 2 PlainTextDocument list

Madison\_fed\_14.txt 2 PlainTextDocument list

Madison\_fed\_37.txt 2 PlainTextDocument list

Madison\_fed\_38.txt 2 PlainTextDocument list

Madison\_fed\_39.txt 2 PlainTextDocument list

Madison\_fed\_40.txt 2 PlainTextDocument list

Madison\_fed\_41.txt 2 PlainTextDocument list

Madison\_fed\_42.txt 2 PlainTextDocument list

Madison\_fed\_43.txt 2 PlainTextDocument list

Madison\_fed\_44.txt 2 PlainTextDocument list

Madison\_fed\_45.txt 2 PlainTextDocument list

Madison\_fed\_46.txt 2 PlainTextDocument list

Madison\_fed\_47.txt 2 PlainTextDocument list

Madison\_fed\_48.txt 2 PlainTextDocument list

Madison\_fed\_58.txt 2 PlainTextDocument list

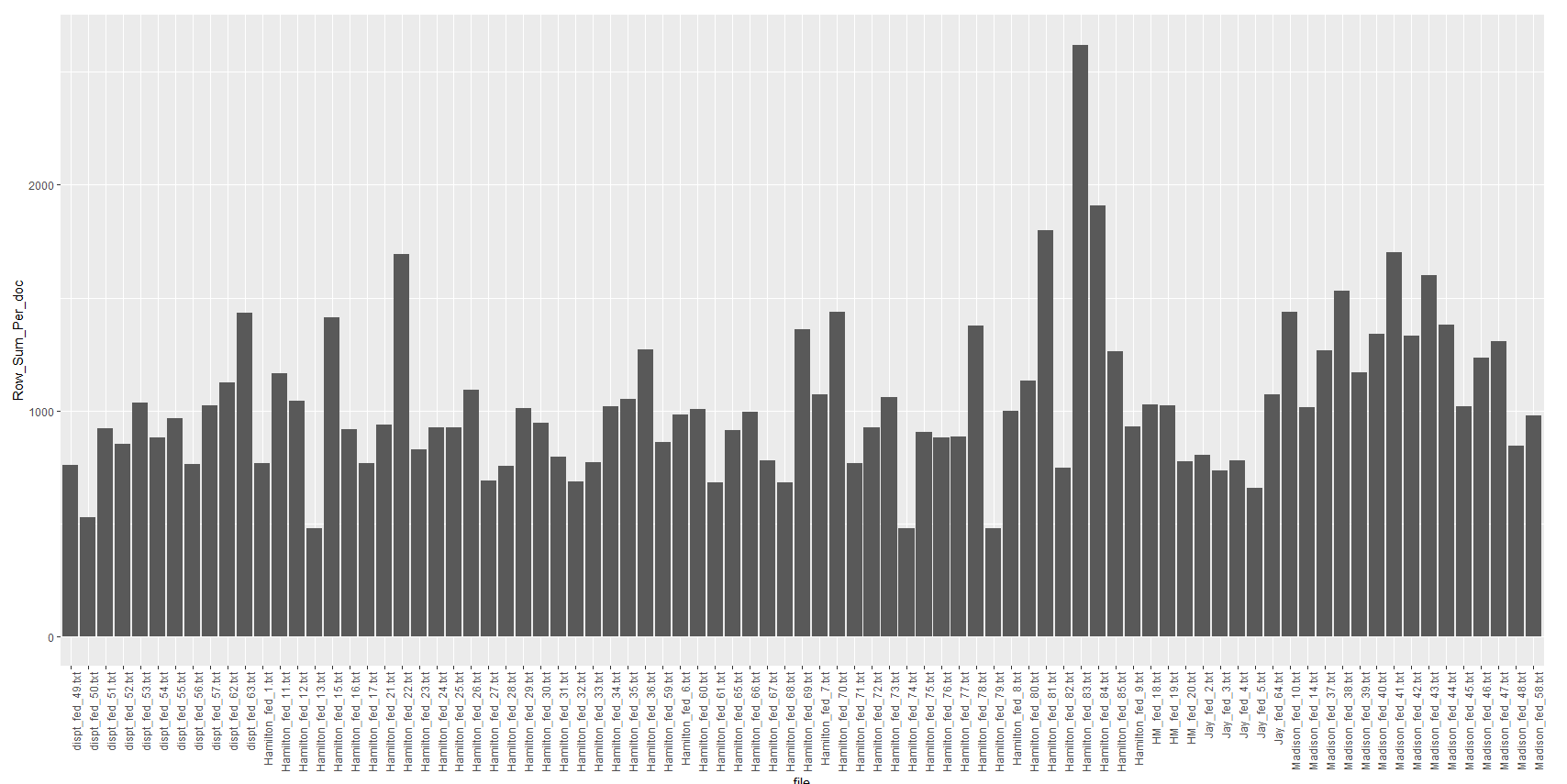
**Table 1.2 Papers**

**Inspect the Document Term Matrix**



**Table 1.3 Document Term Matrix**

**Word count by papers**



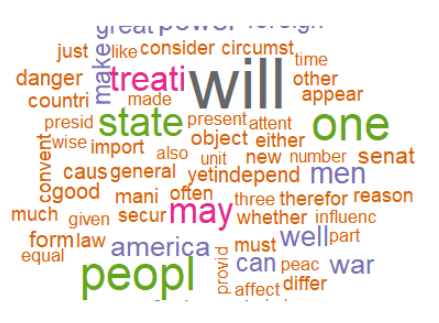
**Fig 1.1 Word count by papers**

**Word Clouds**

Fig 1.2,1.3,1.4 and 1.5 shows the word cloud of federalist papers by authors



**Fig 1.2 Word Cloud on Disputed Papers Fig 1.3 Word Cloud on Hamilton Papers**



**Fig 1.4 Word Cloud on Jay Papers Fig 1.5 Word Cloud of Madison Papers**

### **Models**

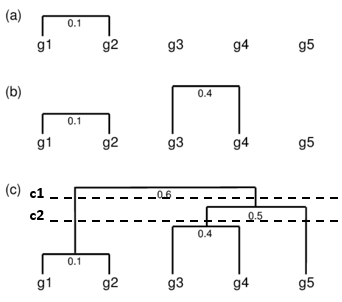
Clustering is the classification of data objects into similarity groups (clusters) according to a defined distance measure. It is used in many fields, such as machine learning, data mining, pattern recognition, image analysis, genomics, systems biology, etc. Machine learning typically regards data clustering as a form of unsupervised learning.

#### **Clustering Technique**

There are two clustering techniques used for cluster analysis

**Hierarchical clustering**

It is an iterative process where the records or points with the closest distance are clustered in the beginning. New clusters are generated by identifying and joining nearest groups together again and again until it reached to one cluster with all data points as illustrated in **Fig 2.1**



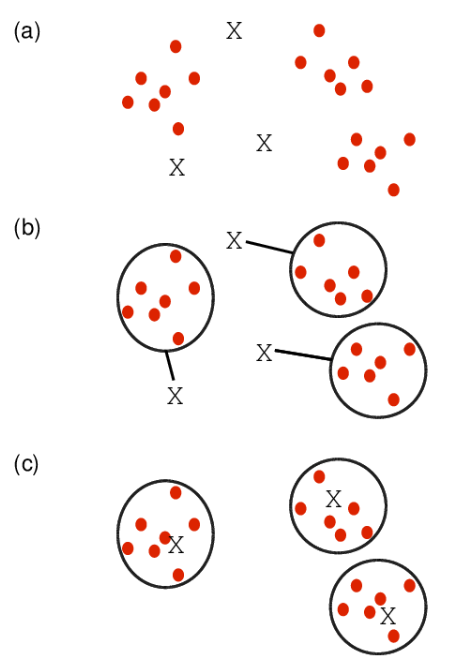
**Fig 2.1 Hierarchical Clusters**

The optimum number of clusters can be obtained by cutting the tree horizontally under each node. Cutting the final tree at c1 will result in two significant clusters below the horizontal line, and c2 will result in three clusters below the horizontal line as shown in **Fig 2.1**.

*Source* [*http://girke.bioinformatics.ucr.edu/GEN242/pages/mydoc/Rclustering.html*](http://girke.bioinformatics.ucr.edu/GEN242/pages/mydoc/Rclustering.html)

**K means clustering**

Unlike hierarchical clustering, K means require a number of clusters (k) as an input. This algorithm then randomly assigns items to the k clusters. Calculate new centroid for each of the k clusters and the distance of all items to the k centroids. Then assign items to the closest centroid. Repeat this process until clusters assignments are stable as illustrated in **Fig 2.2**

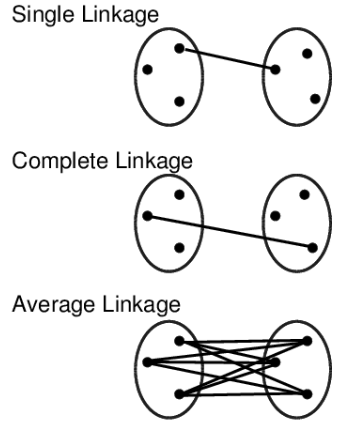


**Fig 2.2 K Mean Clustering**

*Source* [*http://girke.bioinformatics.ucr.edu/GEN242/pages/mydoc/Rclustering.html*](http://girke.bioinformatics.ucr.edu/GEN242/pages/mydoc/Rclustering.html)

#### **Cluster Linkage**

Distance between two clusters can be derived by using single, complete or average linkage methods as shown in **Fig 2.3**. Single linkage uses the minimum distance between two points whereas the complete linkage used the maximum distance between two points. Average linkage calculates an average of the distance between all points between the clusters



**Fig 2.3 Cluster Linkage**

*Source* [*http://girke.bioinformatics.ucr.edu/GEN242/pages/mydoc/Rclustering.html*](http://girke.bioinformatics.ucr.edu/GEN242/pages/mydoc/Rclustering.html)

#### **Data standardization**

Center and standardize

* 1. Center: subtract from each value the mean of the corresponding vector
  2. Standardize: divide by standard deviation

Center and scale with the **scale ()** function

* 1. Center: subtract from each value the mean of the corresponding vector
  2. Scale: divide centered vector by their *root mean square* (*rms*):

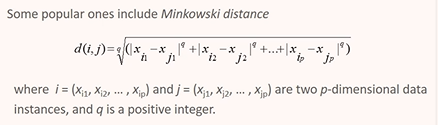
xrms=1n−1∑i=1nxi2−−−−−−−−−−−√xrms=1n−1∑i=1nxi2

#### **Distance Methods**

Following distance methods can be used to cluster group of records or data points into multiple buckets so that a cluster of points are closer to each other and the distance between the clusters are far apart to make a clear distinction of various subgroups in the dataset. For text documents, every word is treated as an attribute or column, and every row is a text document that we want to group them.

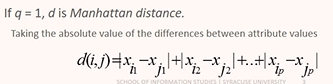
**Minkowski distance**

Minkowski distance between two points or rows can be explained using the following mathematical equation



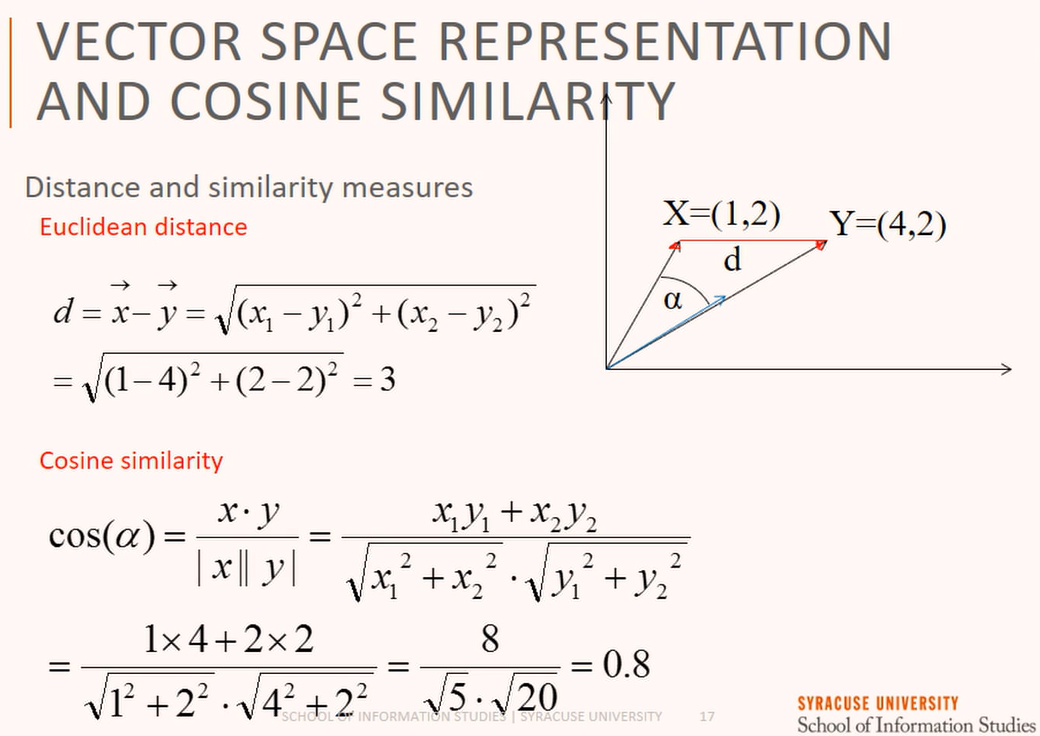
**Manhattan distance**

When q=1 in the Minkowski distance equation, then the equation derives Manhattan distance



**Euclidean distance**

When q=2 in the Minkowski distance equation, then the equation derives Euclidean distance. This is one of the popular distance measure technique used in a wide application.



**Cosine similarity**

Cosine similarity is the angle of measure between two points from the origin which is a different way of accessing the distance between rows or points and the equation to calculate the angle is as follows

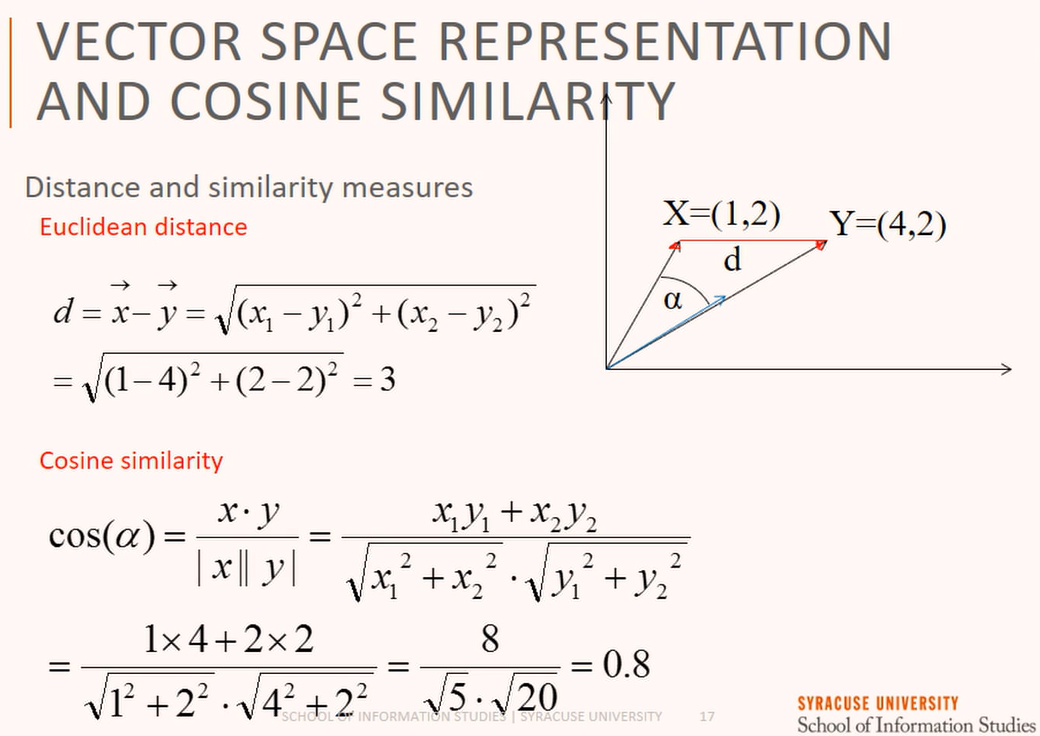
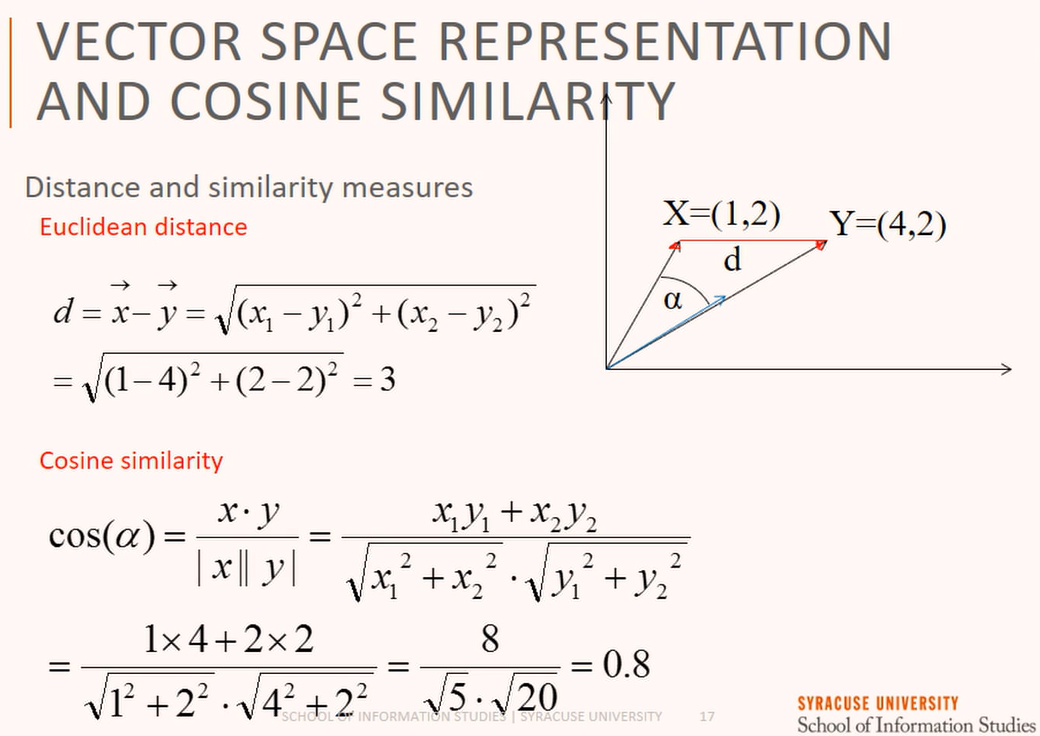


Fig 2.1 illustrates distance and angular measure between two points

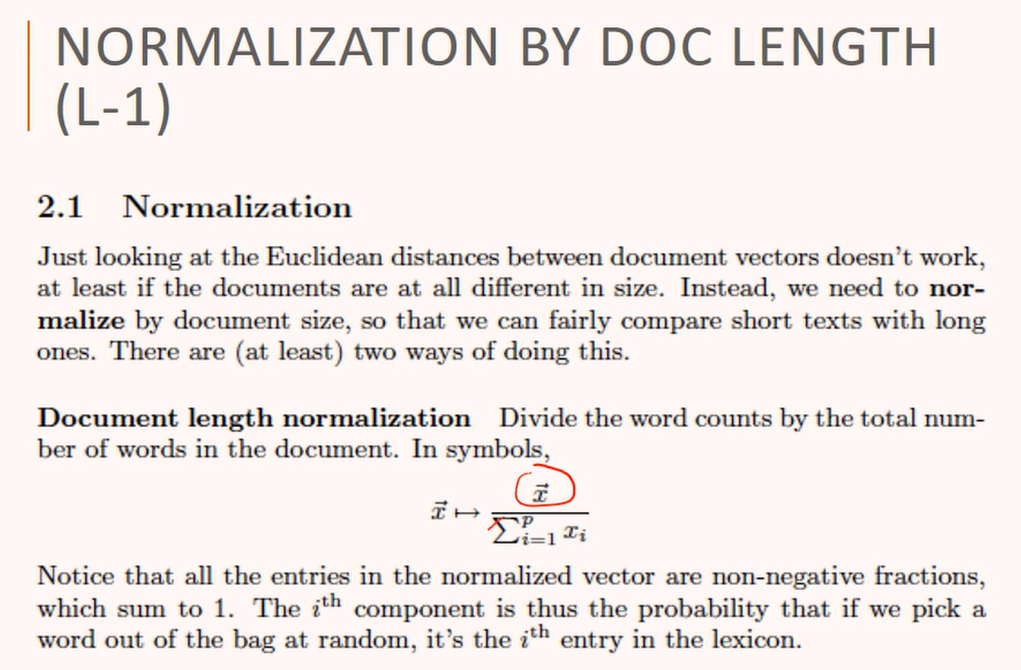


**Fig 2.1**

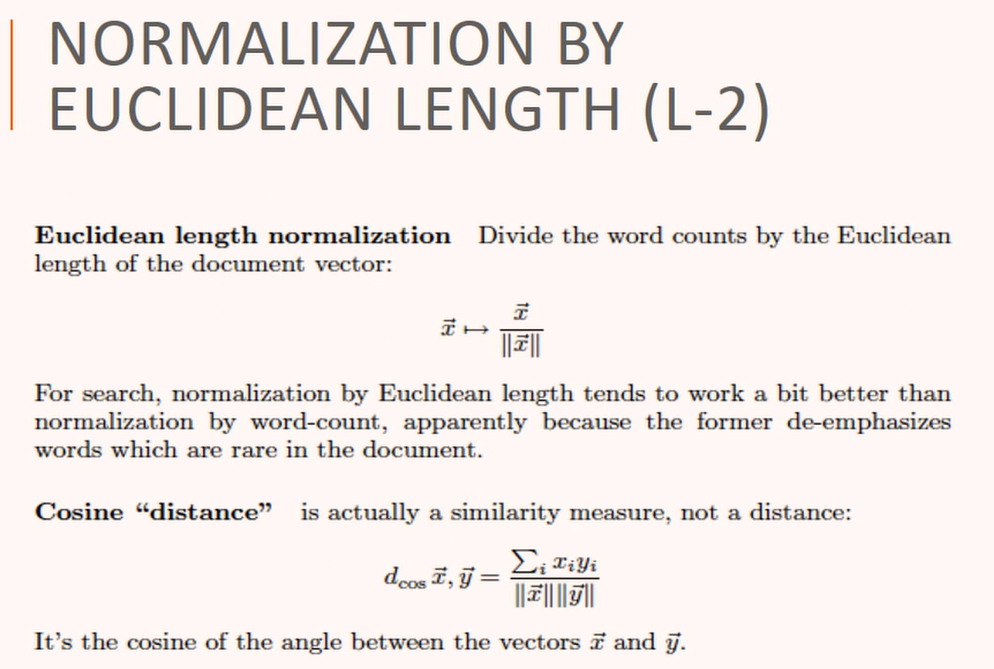
#### **Normalization for Text document**

Euclidean distances between document is not enough because each document comes as variety of sizes. Normalization by document size is more important and one of the below techniques can be used to achieve normalization by document length.

**1: Normalization by length**

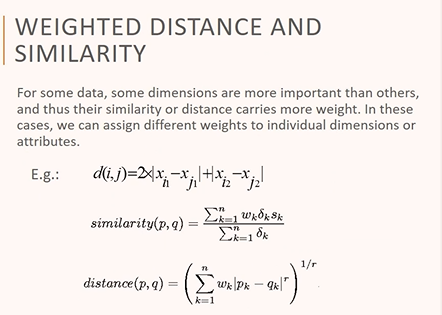


**2: Normalization by Euclidean length**



#### **Weighted Distance and Similarity**

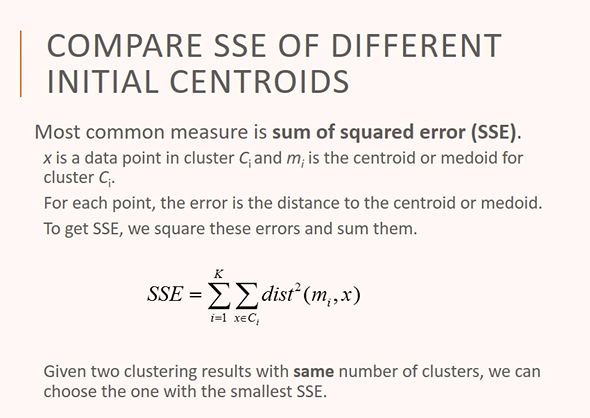
Some attributes or dimensions in a dataset are more important than others, and hence similarity or distance carries more weight. In this case we can assign different weights to individual attributes by using the following equation



#### **How to access accuracy of different distance and similarity measures**

Sum of squared error (SSE) is a measure which helps in accessing results of different clustering results, x is a data point in cluster Ci and mi is the centroid or medoid for cluster Ci

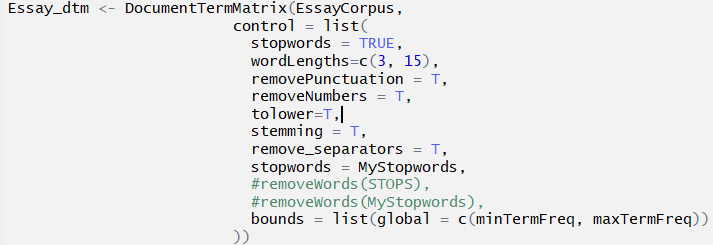
For each point, the error is the distance to the centroid or medoid. SSE can be derived using the following equation.



#### **Model 1: Hierarchical Clustering using Euclidean Distance and ward.D linkage method**

**Data preparation: Hamilton and Madison papers**

Hamilton, Madison and disputed papers are read as corpus elements and a matrix with only Hamilton and Madison papers are built on top of it to make it suitable for processing



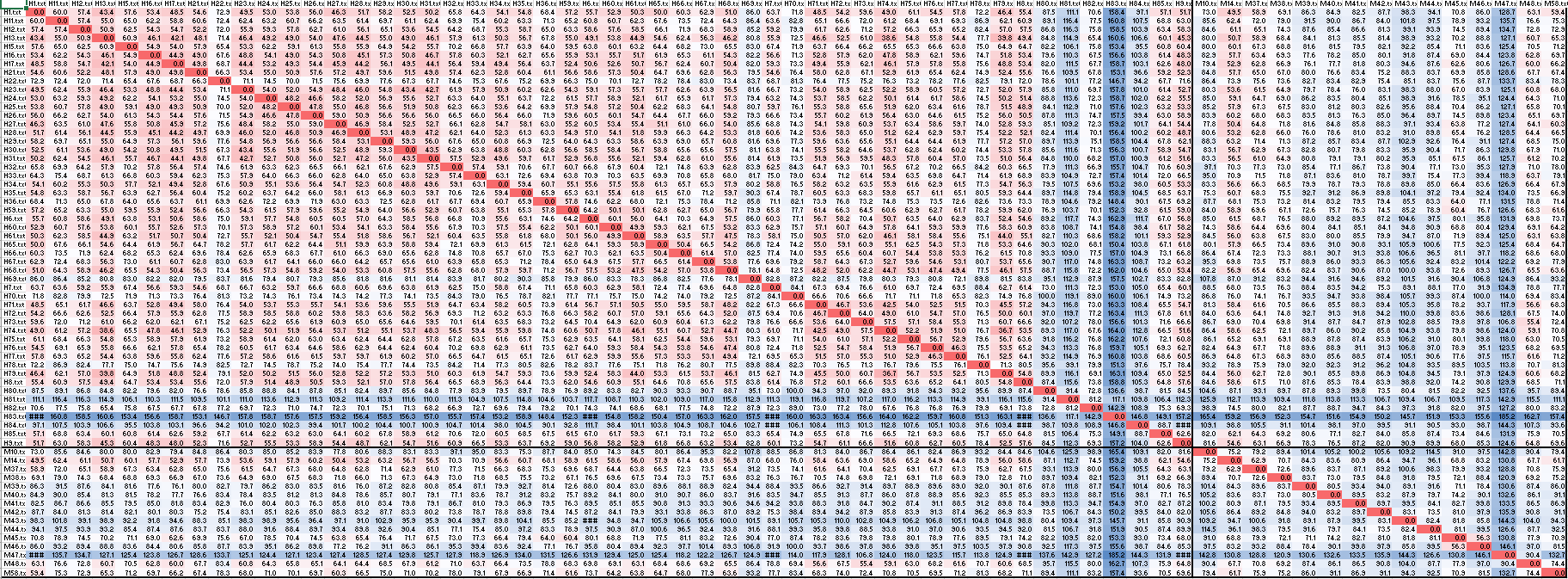




**Distance calculation: Hamilton and Madison papers**

Distance between Hamilton and Madison papers are calculated using Euclidean distance equation and the distance matrix is shown in **Fig. 2.1**





**Fig 2.1 Euclidean Distance measure for Hamilton and Madison papers**

**Outliers: Hamilton and Madison papers**

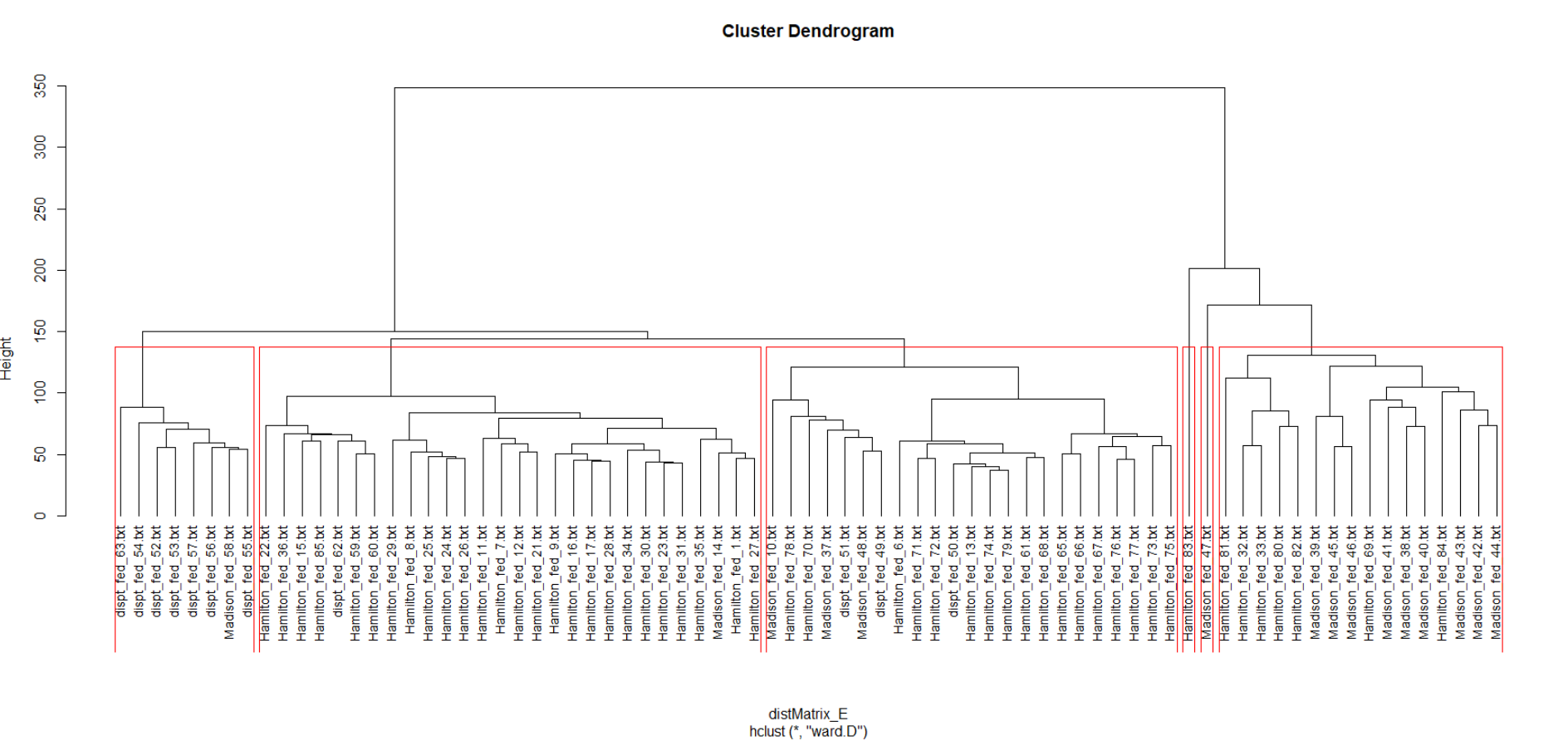
Hamilton Papers 🡪 69,80,81,82,83,84

Madison Papers 🡪 14,37,48,58

**Model specification: Hamilton and Madison papers**

Hierarchical cluster is generated by using the above distance matrix and ward.D linkage method. Six clusters are observed that are grouped closely to each author group as shown in **Fig 2.2.**

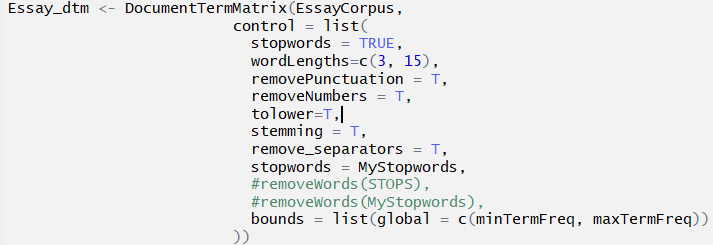




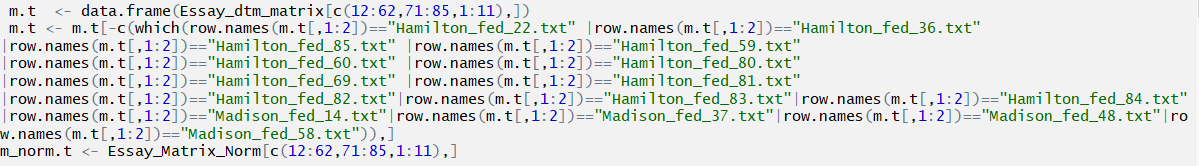
**Fig 2.2 Hierarchical cluster for federalist papers**

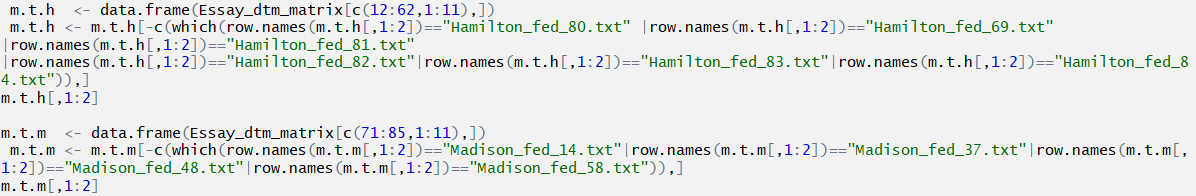
**Data preparation: Hamilton, Madison and disputed papers after removing outliers**

All the papers are read as corpus elements and a matrix with Hamilton, Madison and disputed papers are built on top of it to make it suitable for processing





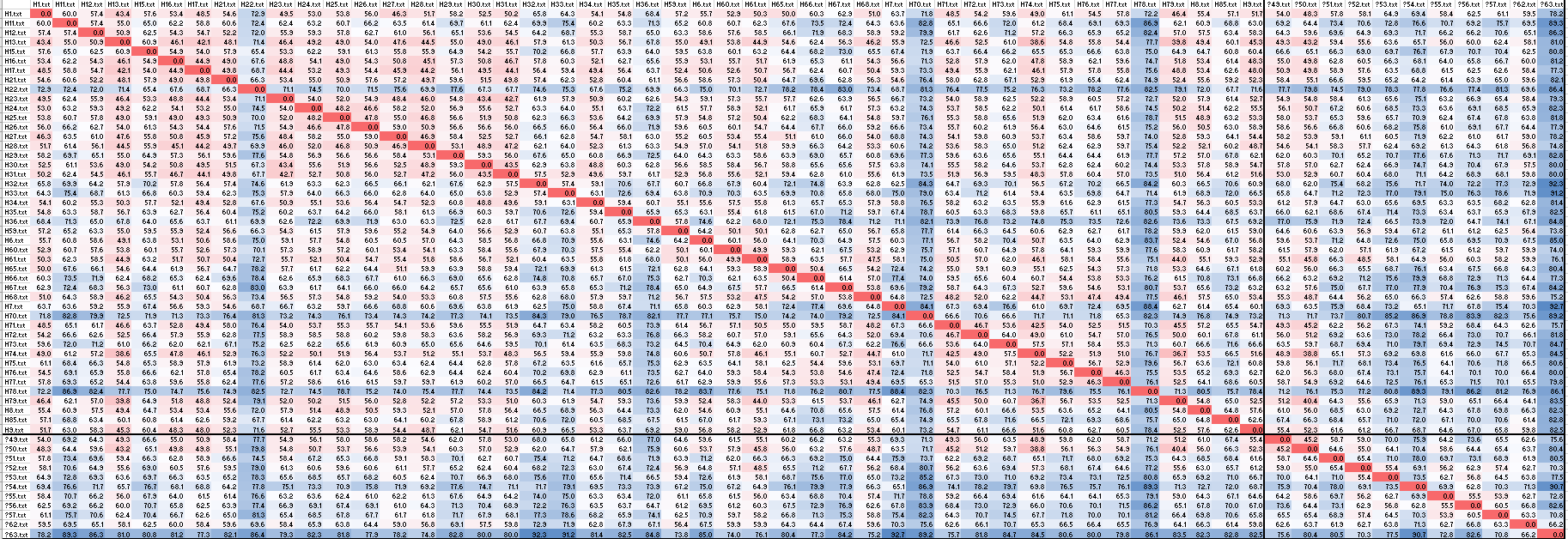




**Distance calculation: Hamilton and disputed papers**

Distance between Hamilton and disputed papers are calculated using Euclidean distance equation and the distance matrix is shown in **Fig. 2.3**



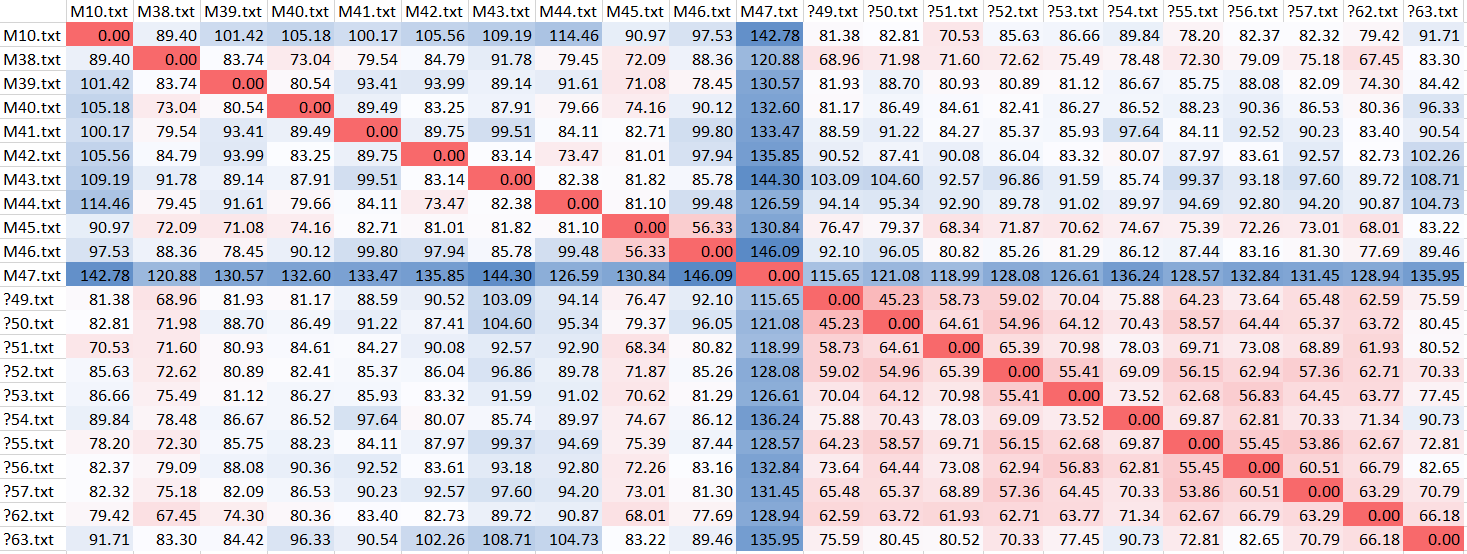


**Fig 2.3 Euclidean Distance measure for Hamilton and dispute papers**

**Distance calculation: Madison and disputed papers**

Distance between Madison and disputed papers are calculated using Euclidean distance equation and the distance matrix is shown in **Fig. 2.4**



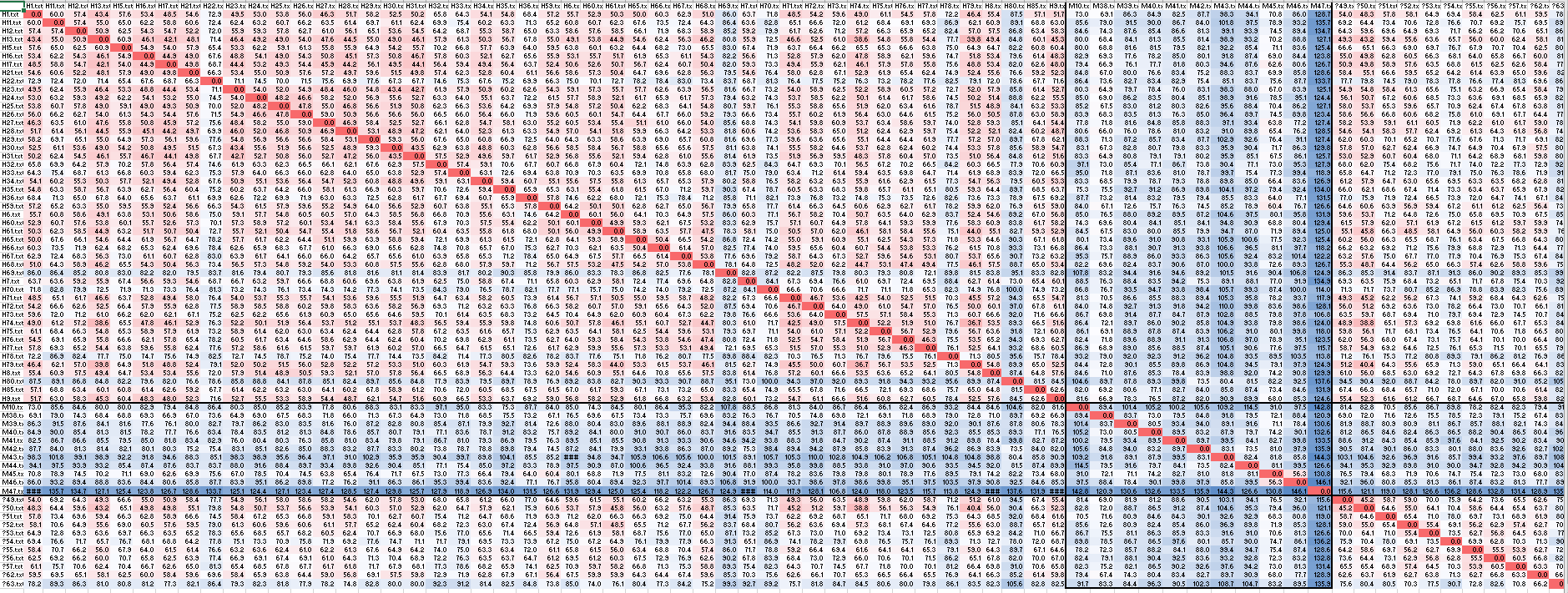


**Fig 2.4 Euclidean Distance measure for Madison and dispute papers**

**Distance calculation: Hamilton, Madison and disputed papers**

Distance between Madison and disputed papers are calculated using Euclidean distance equation and the distance matrix is shown in **Fig. 2.5**

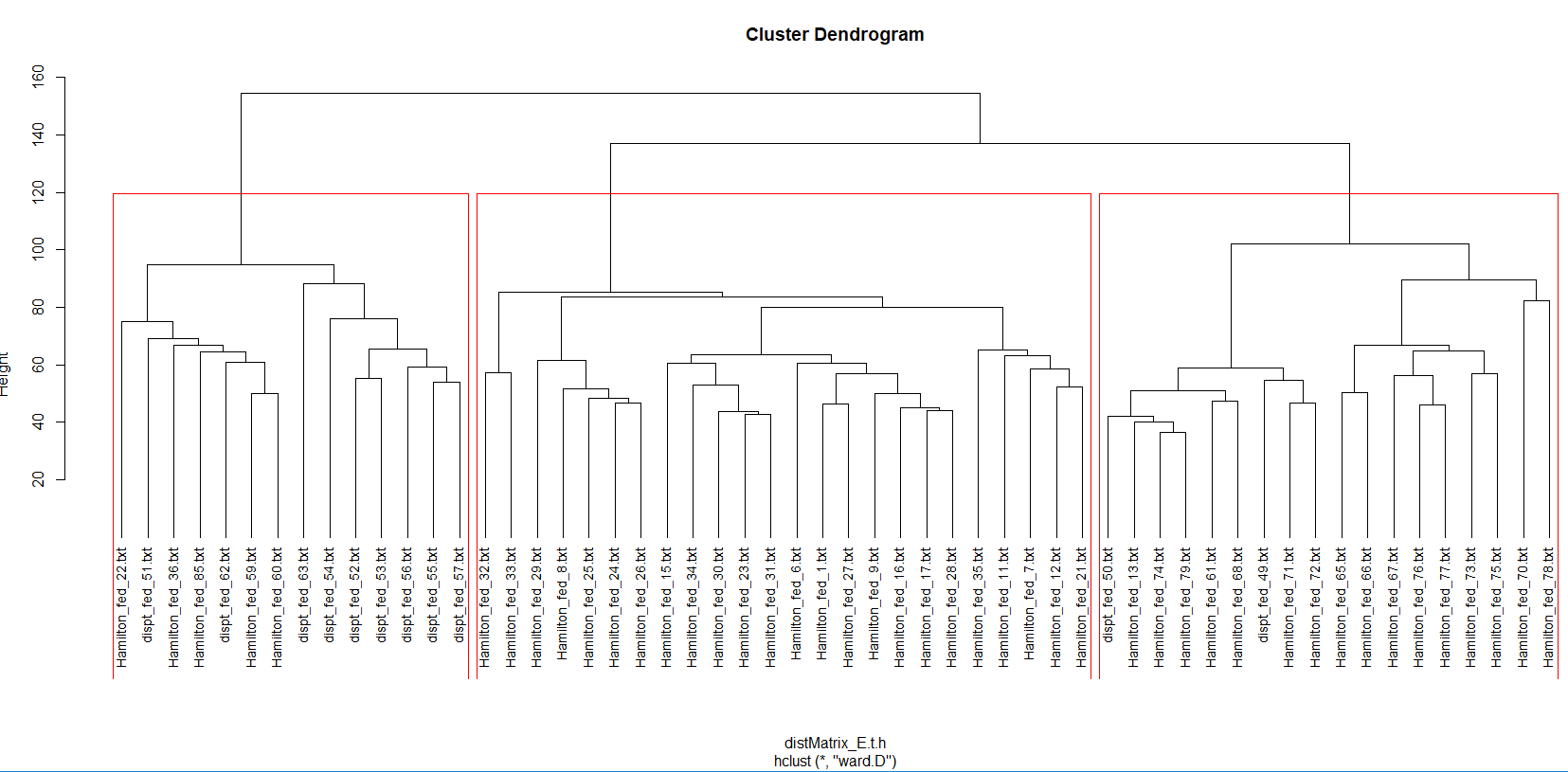




**Fig 2.5 Euclidean Distance measure for Hamilton, Madison and dispute papers**

**Model specification: Hamilton and disputed papers**

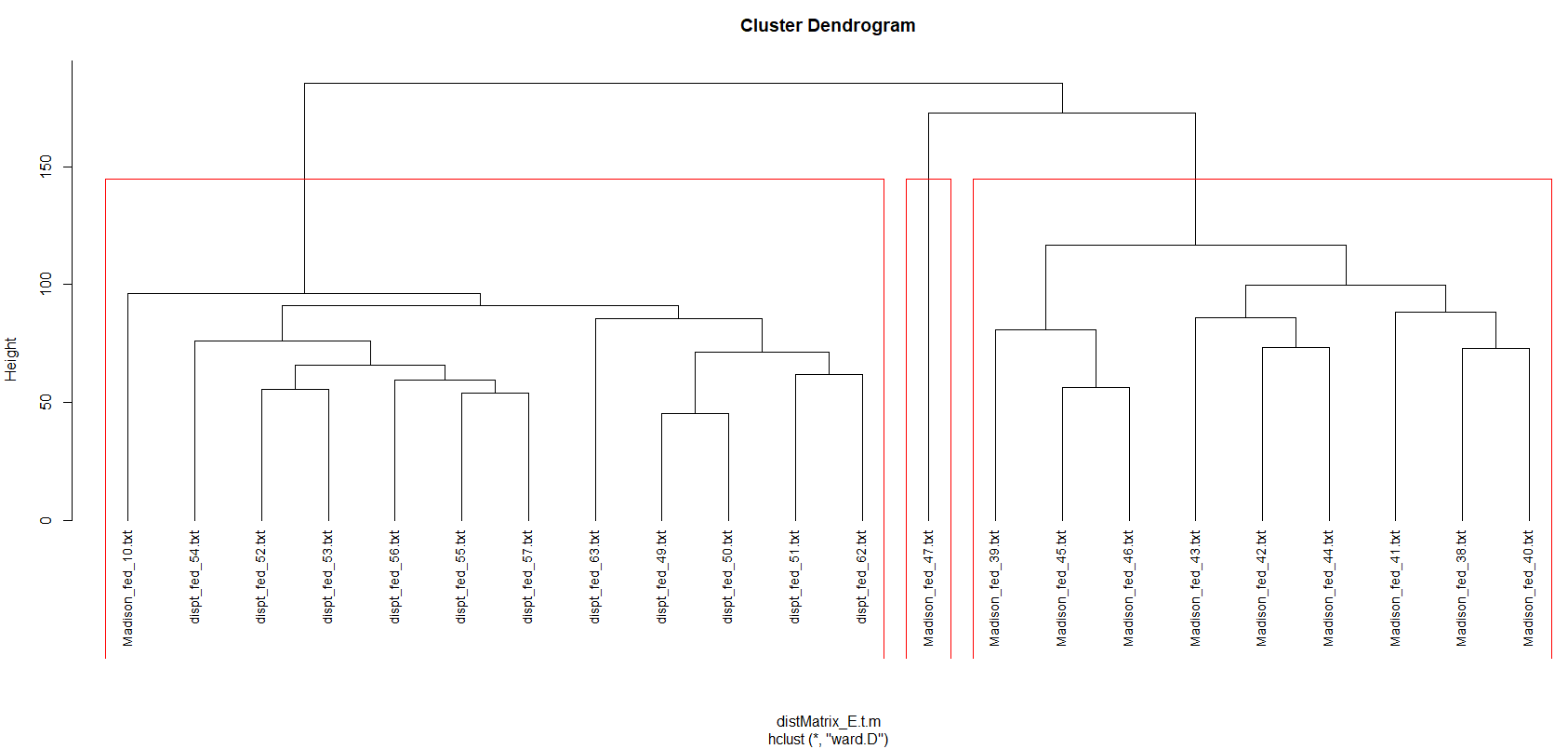
Hierarchical cluster is generated by using the above distance matrix and ward.D linkage method. Four major clusters are observed that are grouped closely to each author group as shown in **Fig 2.6**



**Fig 2.6 Hierarchical cluster of Hamilton and disputes**

**Model specification: Madison and disputed papers**

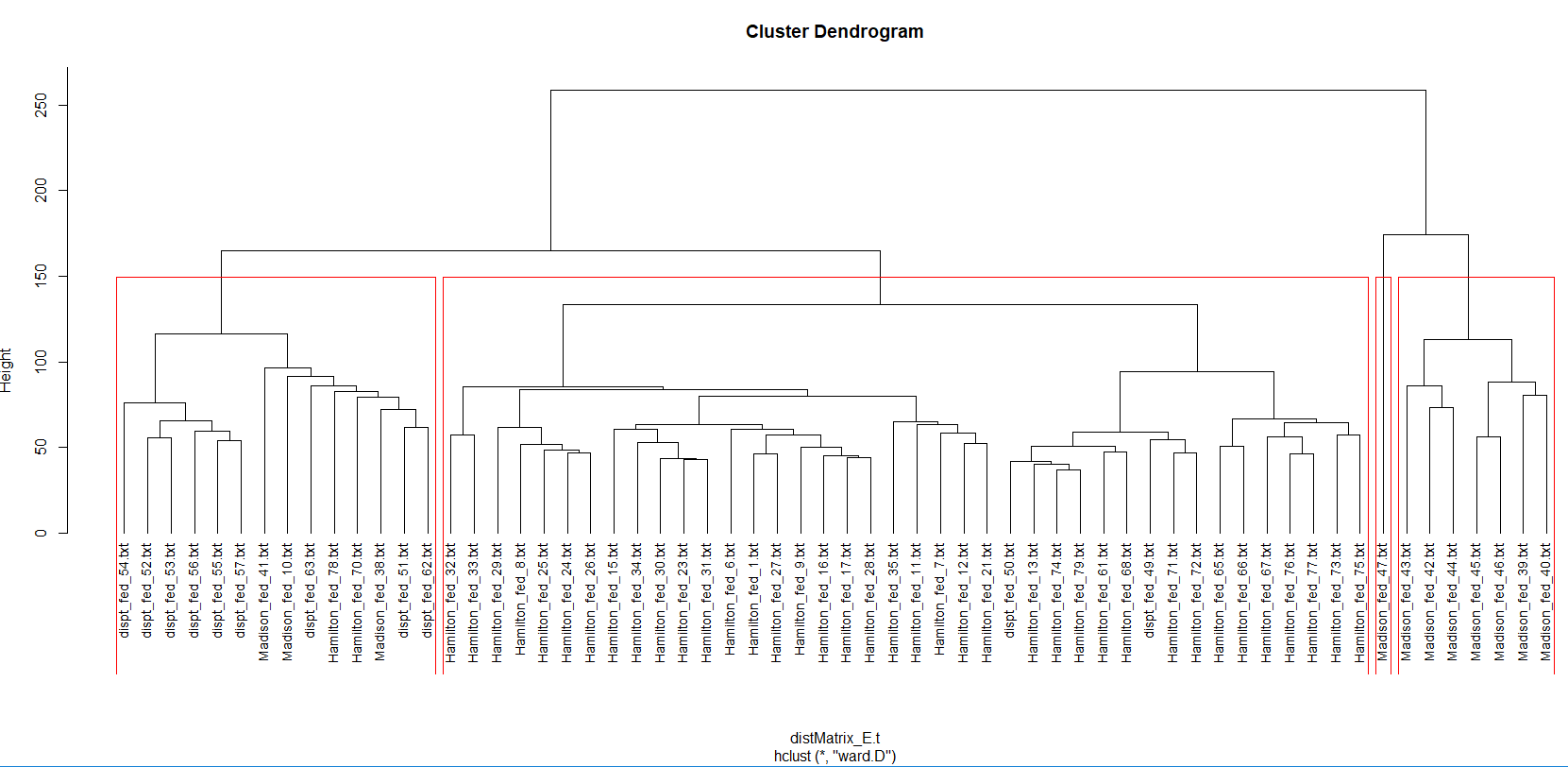
Hierarchical cluster is generated by using the above distance matrix and ward.D linkage method. Four major clusters are observed that are grouped closely to each author group as shown in **Fig 2.7**



**Fig 2.7 Hierarchical cluster of Madison and disputes**

**Model specification: Hamilton, Madison and disputed papers**

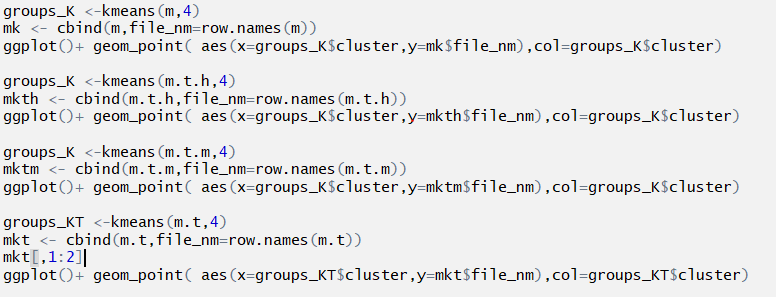
Hierarchical cluster is generated by using the above distance matrix and ward.D linkage method. Four major clusters are observed that are grouped closely to each author group as shown in **Fig 2.8**

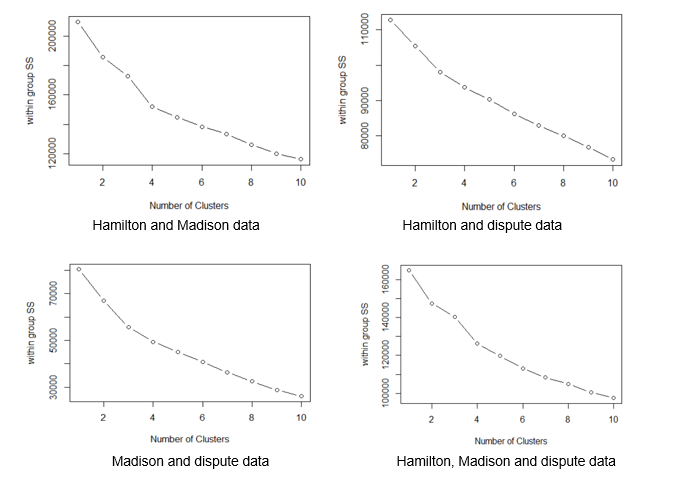


**Fig 2.8 Hierarchical cluster of Hamilton, Madison including disputes**

#### **Model 2: K Means Clustering using Euclidean Distance and ward.D linkage method**

Optimal number of clusters for K means is derived and observed that using 4 clusters are more efficient and are shown in

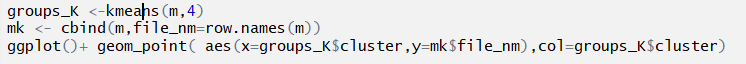


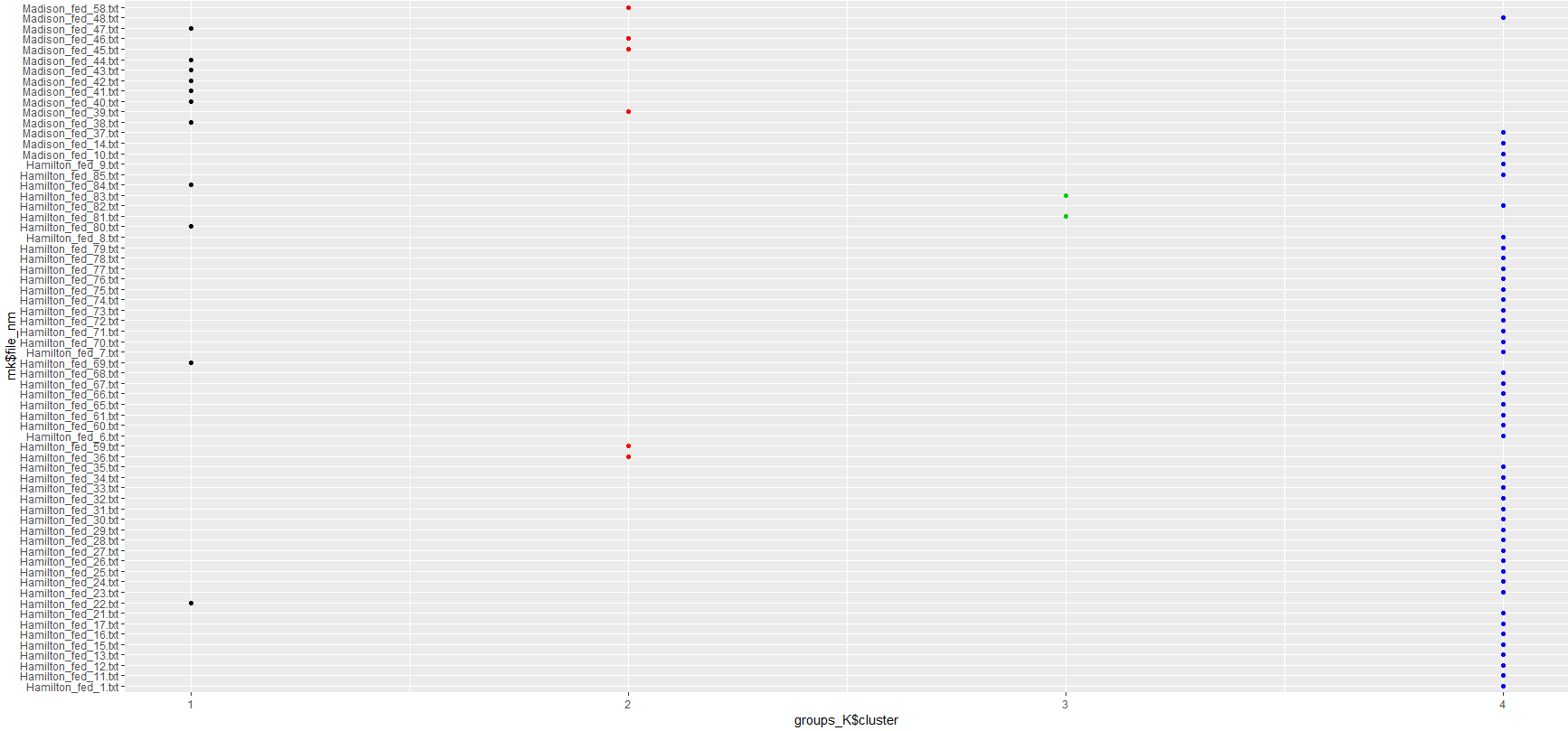


**Fig 2.9 WSS plot**

**Model specification: Hamilton and Madison papers**

K means models with 4 clusters for Hamilton and Madison papers are as follows

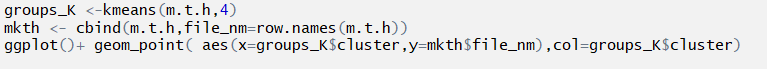


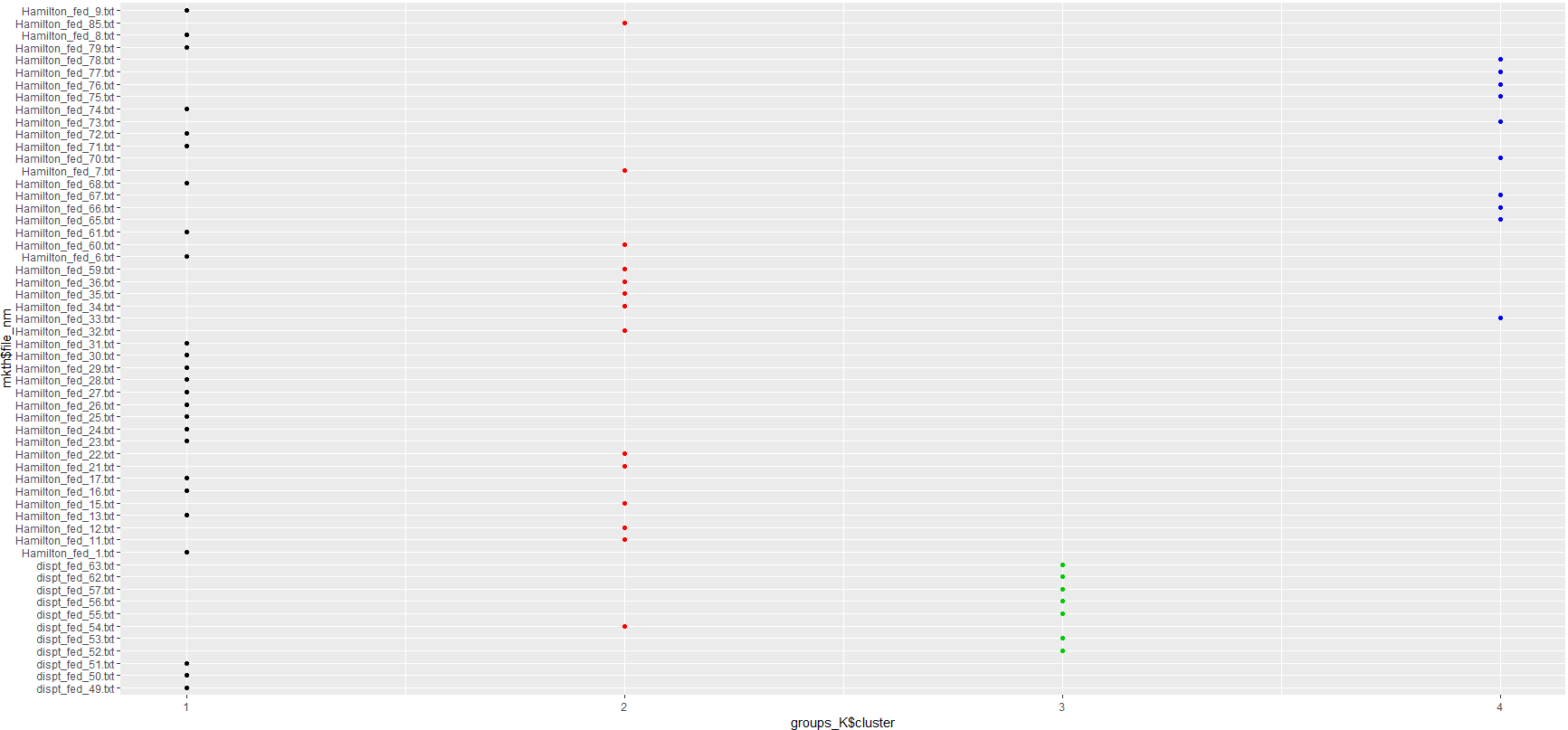


**Fig 2.10 K means clusters for Hamilton and Madison papers**

**Model specification: Hamilton and dispute papers**

K means models with 4 clusters for Hamilton and dispute papers are as follows



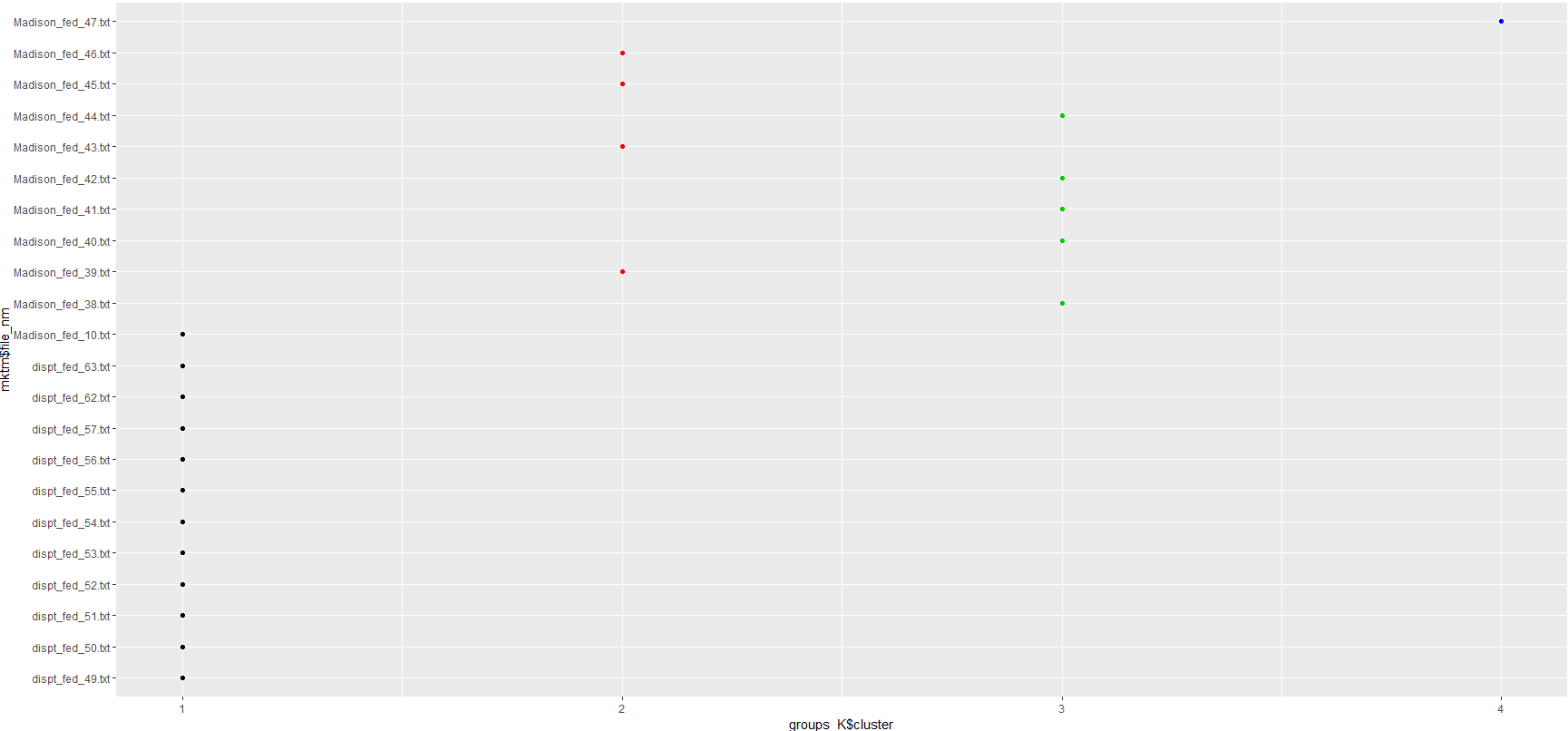


**Fig 2.11 K Means cluster for Hamilton and dispute papers**

**Model specification: Madison and dispute papers**

K means models with 4 clusters for Madison and dispute papers are as follows

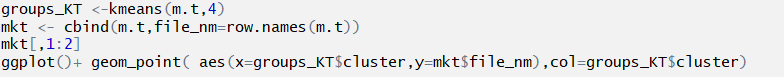


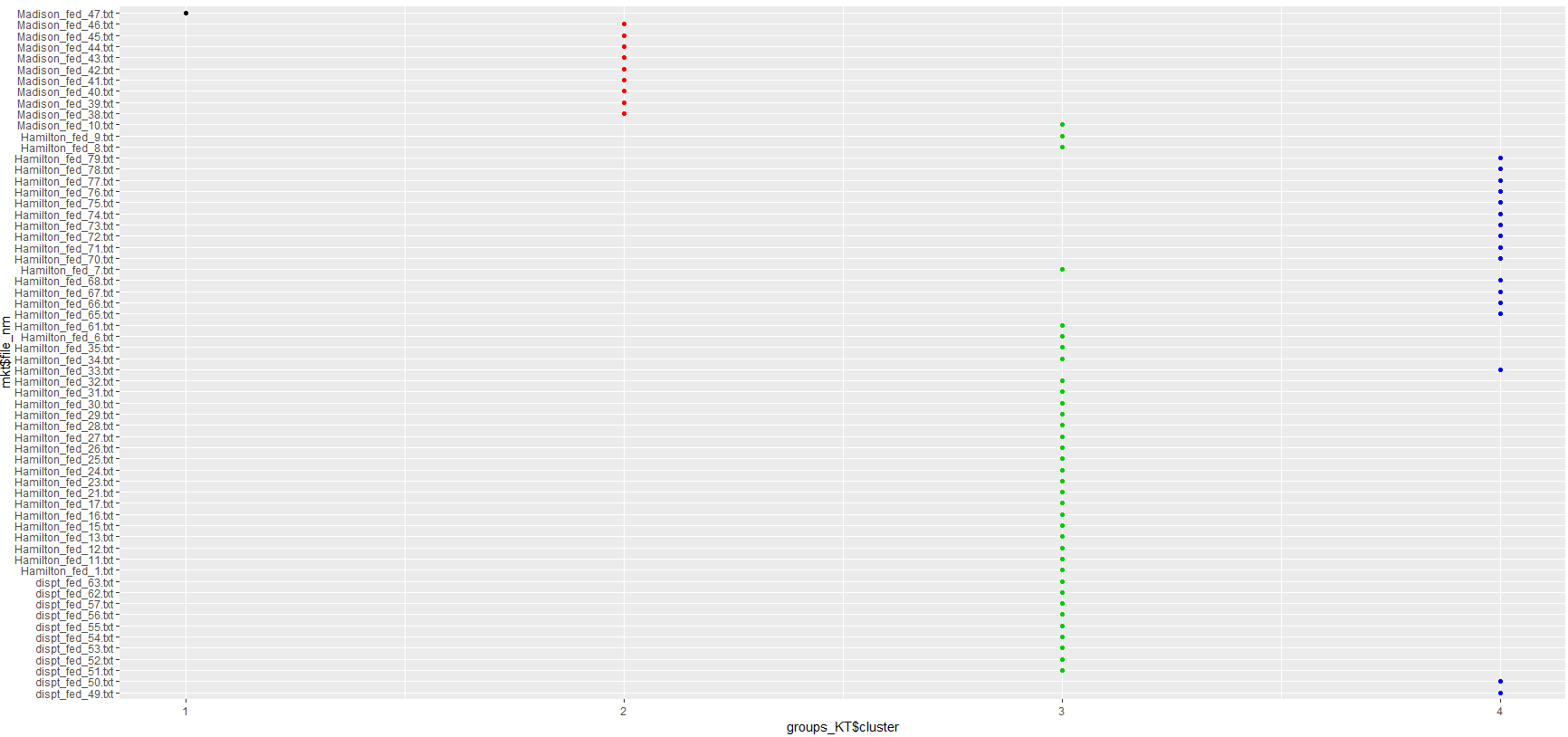


**Fig 2.12 K Means cluster for Madison and dispute papers**

**Model specification: Hamilton, Madison and dispute papers**

K means models with 4 clusters for Hamilton, Madison and dispute papers are as follows





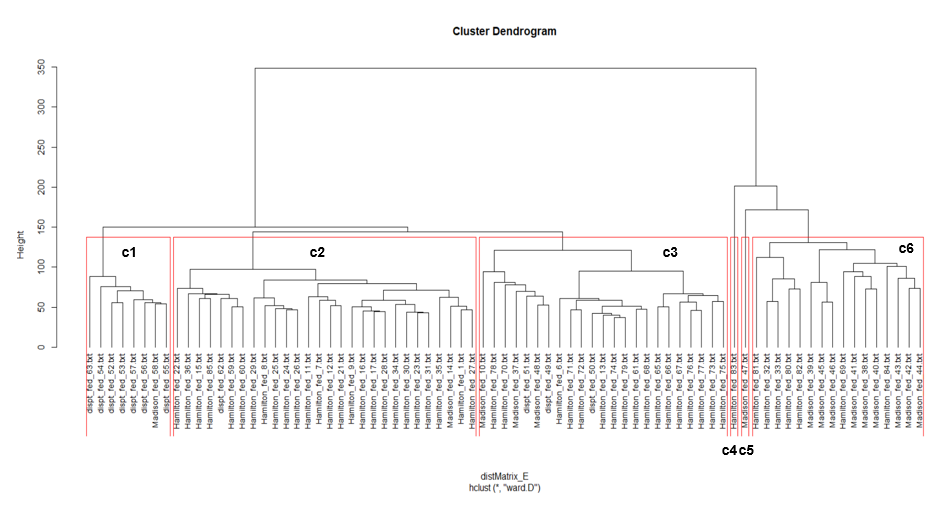
**Fig 2.13 K Means cluster for Hamilton, Madison and dispute papers**

## **Results**

#### **Model 1: Results of Hierarchical Clustering using Euclidean Distance and ward.D linkage method**

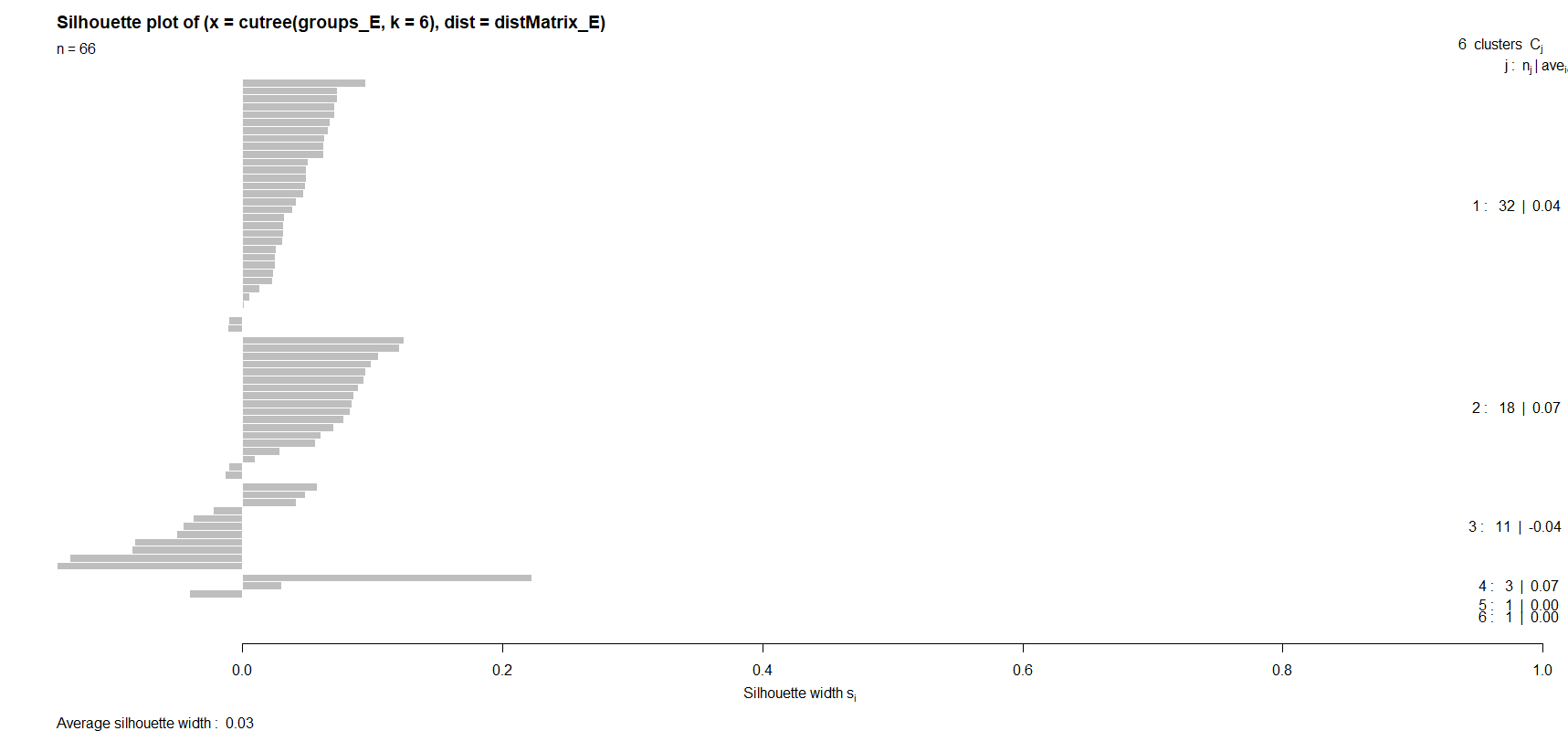
**Model Results: Hamilton and Madison papers**

In the six clusters observed in **Fig 3.1**, majority of Hamilton papers are clustered in c2 and c3, c6 has got more of Madison Papers. c4 and c5 are outliers where it stands out as a single cluster. Cluster c1 has got more of dispute papers but with one Madison paper. Even though c1t is grouped with one of the Madison paper, it is closely linked with Hamilton papers up in the hierarchy

.

**Fig 3.1 Hierarchical cluster Hamilton, Madison and dispute papers**

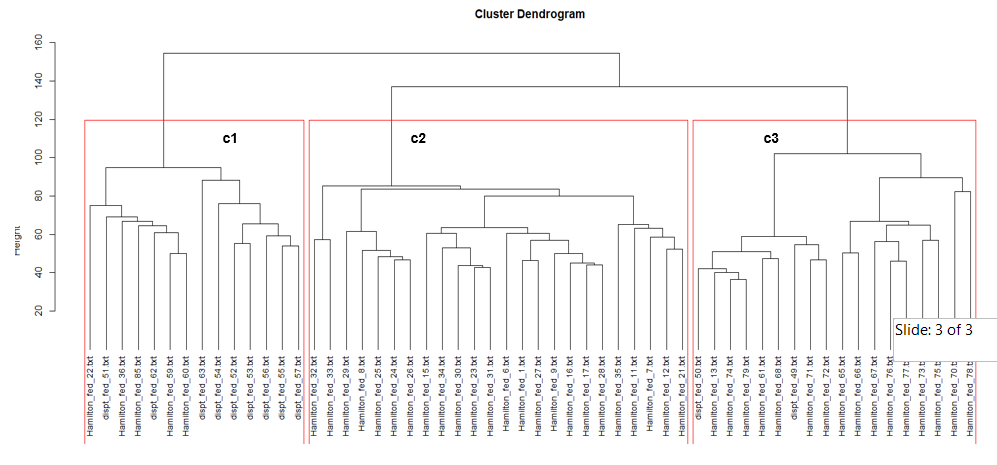
In **Fig 3.2,** Silhouette plot on all the above 6 clusters are plotted to see the variations within the cluster and outliers



**Fig 3.2 Silhouette plot Hamilton, Madison and dispute papers**

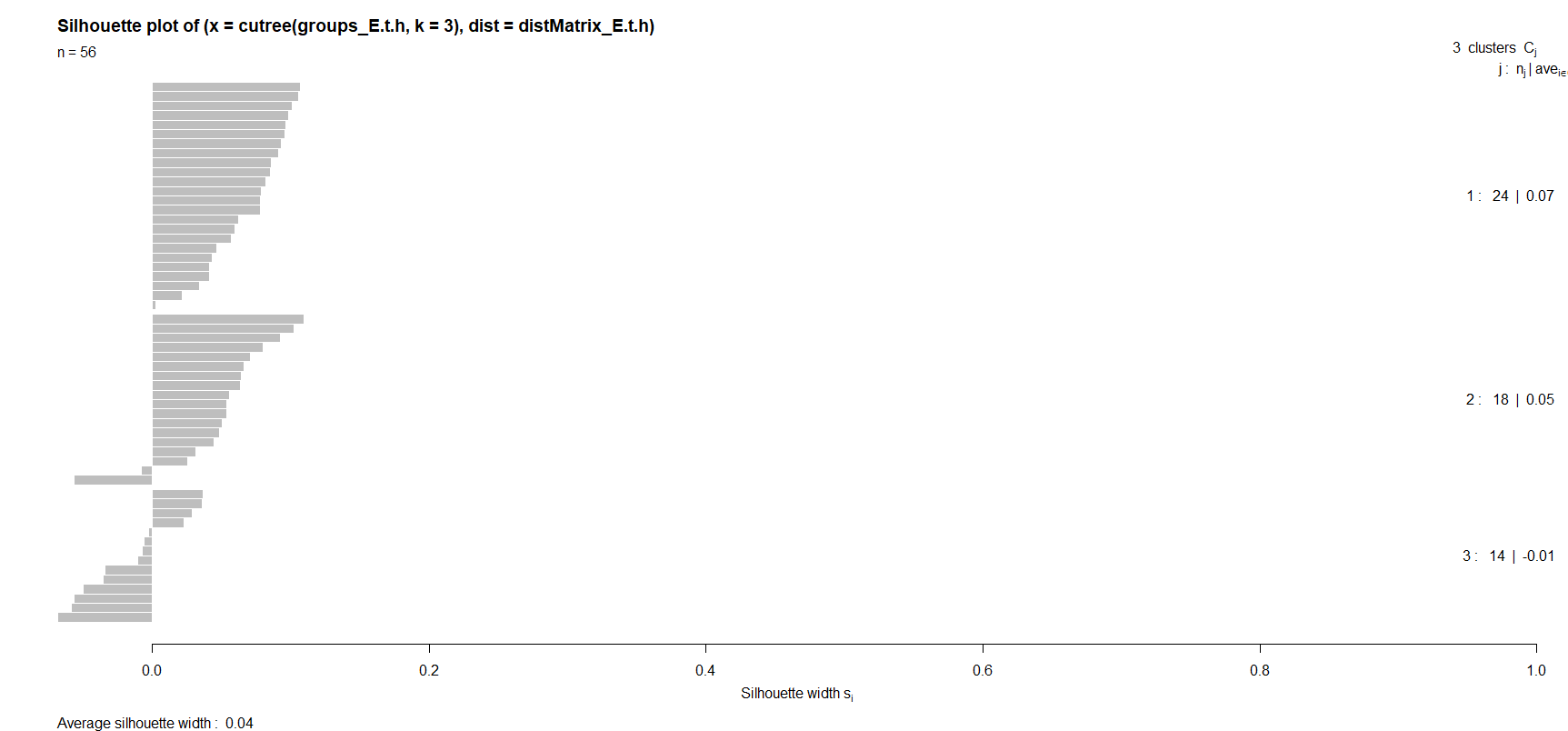
**Model Results: Hamilton and disputed papers**

In the three clusters observed in **Fig 3.3**, Hamilton papers are clustered in c2 and c3, c1 has got more of dispute papers with some Hamilton papers but are standing out as separate cluster.



**Fig 3.3 Hierarchical cluster Hamilton and dispute papers**

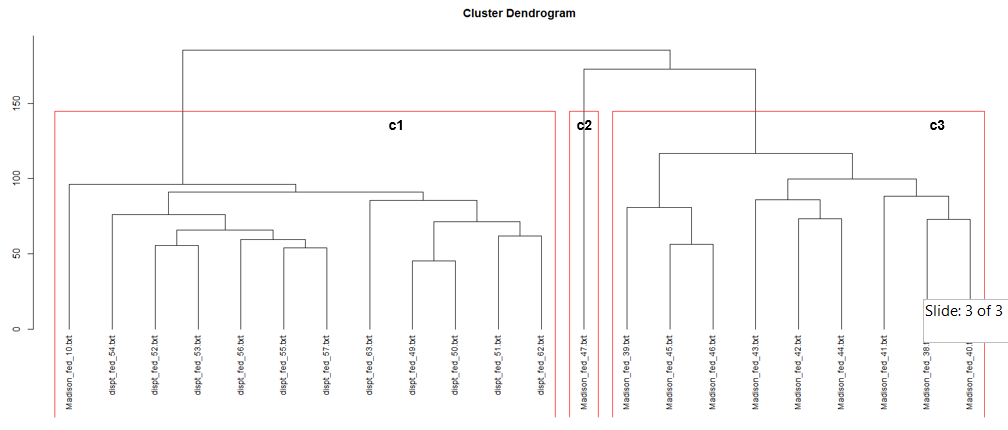
In **Fig 3.4,** Silhouette plot on all the above 3 clusters are plotted to see the variations within the cluster and outliers



**Fig 3.4 Silhouette plot Hamilton and dispute papers**

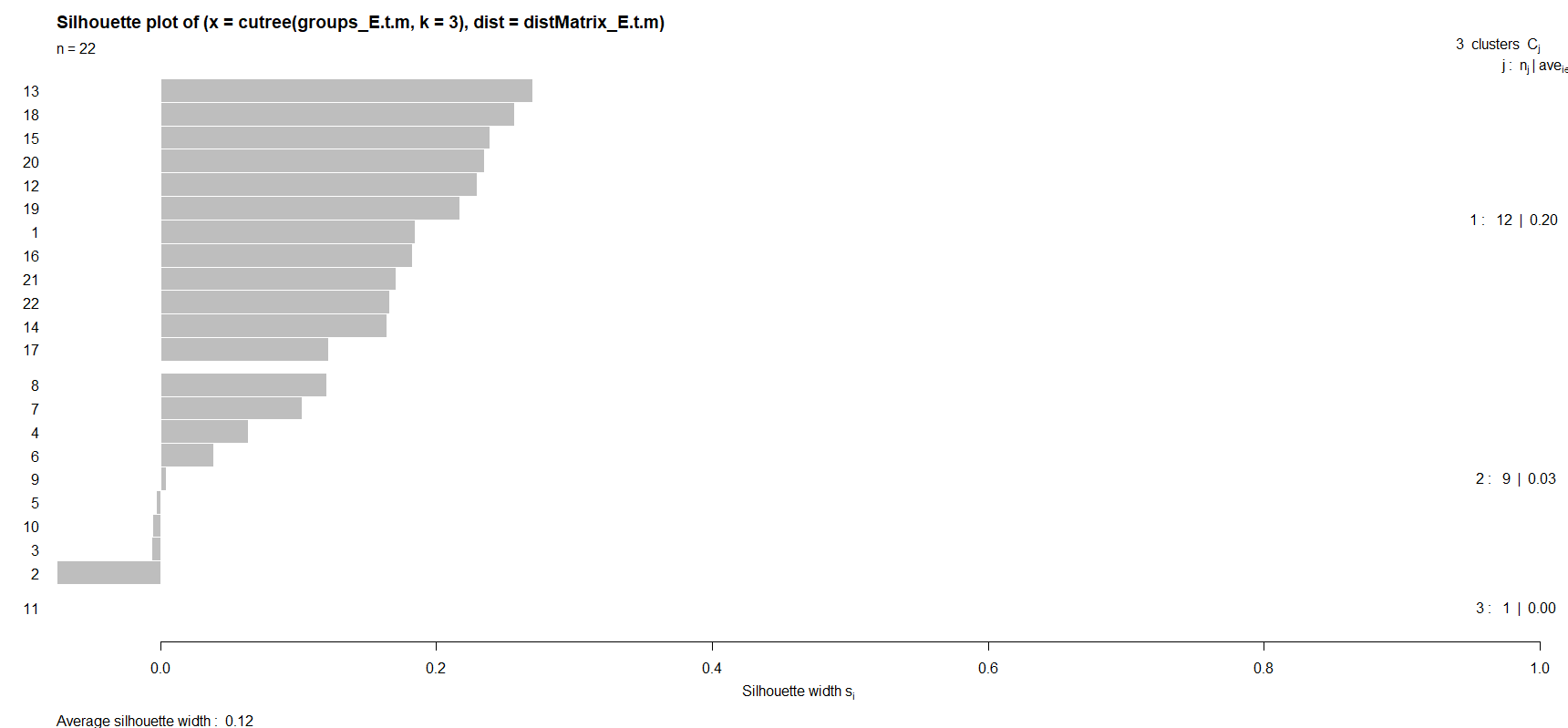
**Model Results: Madison and disputed papers**

In the three clusters observed in **Fig 3.5**, cluster c1 and c3 are two major clusters and clearly separating out Madison papers from dispute papers. Cluster c2 is an outlier standing out as separate cluster



**Fig 3.5 Hierarchical cluster Madison and dispute papers**

In **Fig 3.6,** Silhouette plot on all the above 3 clusters are plotted to see the variations within the cluster and outliers

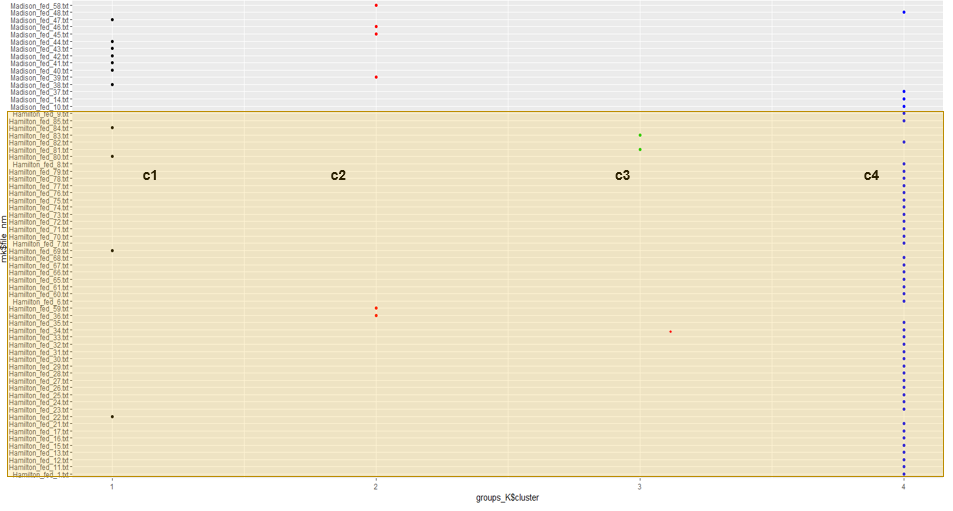


**Fig 3.6 Silhouette plot Madison and dispute papers**

#### **Model 2: Results of K Means Clustering using Euclidean Distance and ward.D linkage method**

**Results of Hamilton and Madison papers using K means**

Most of the Hamilton papers are falling in c4 cluster and few of them distributed in c1, c2 and c3. Similarly, most of Madison papers are in c1 cluster and few found in c2 and c4 cluster as shown in **Fig 3.7**



**Fig 3.7 K means clusters for Hamilton and Madison papers**

**Results of Hamilton and dispute papers using K means**

Most of the dispute papers are grouped and standing out in a separate c3 cluster and few observed in c1 along with Hamilton as shown in **Fig 3.8**

**Fig 3.8 K Means cluster for Hamilton and dispute papers**

**Results of Madison and dispute papers using K means**

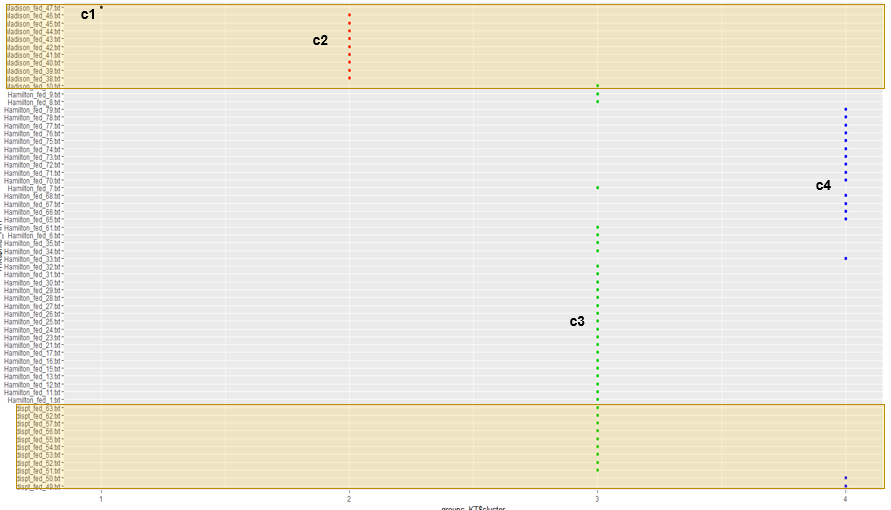
As shown in **Fig 3.9** all the dispute papers are standing out from Madison clusters in a separate cluster c1. Majority of the Madison papers are in c2 and c3, one exception is found in c1 and c4.



**Fig 3.9 K Means cluster for Madison and dispute papers**

**Results of Hamilton, Madison and dispute papers using K means**

As shown in Fig.3.10, Madison papers are grouped in c1 and c2 clusters, Hamilton papers are grouped in c3 and c4 clusters and the dispute papers are again grouped in c3 and c4 clusters.



**Fig 3.10 K Means cluster for Hamilton, Madison and dispute papers**

## **Conclusion**

**Hamilton papers**

Hamilton papers are significant in numbers when compared with other authors. Most of the Hamilton papers exhibit the same pattern except for few that are distinct from the general pattern. This shows that Hamilton has got a similar writing pattern and the variation on his essays are not very large.

**Madison papers**

Madison papers are not very large in number like Hamilton. With fewer papers compared to Hamilton, Madison papers are falling into two to three patterns. This indicates that Madison had a very different writing pattern in each of his essay showing more considerable variations when compared with all of his essays.

**Dispute papers**

Dispute paper shows a similar pattern like Hamilton papers. All the dispute papers when compared with each other shows fewer variations and can be grouped with the writing style of Hamilton. It didn’t show more significant differences like Madison papers within its group. Also, the writing style of dispute papers are very different from Madison style and couldn’t group them under the same bucket of Madison papers. This suggests more inclination towards the authorship of Hamilton and not Madison