Jeff Levesque

Professor Gates

IST 565

Homework #4

**Introduction**

The ratification of the Constitution of the United States of America was led by three founding fathers, Alexander Hamilton, James Madison, and John Jay. Collectively, they contributed to the 85 papers, known to this day as The Federalist Papers. These papers provided the groundwork, and the intent of the Constitution. However, each of the author would often sign their work as the latin ‘Publius’, rather than their own names.

It is known today, that of the 85 articles, 74 have been identified with an author(s). The remaining 11 controversial papers have been claimed to be written by either Hamilton or Madison. This controversy has led to various types of analysis. More recently, various types of text analysis approaches have been exercised to help estimate the likelihood of one author over another.

The analysis provided in this study will attempt to use some clustering techniques to determine distance measures, sum of square errors, and corresponding visualizations. An attempt will be made to justify whether the disputed papers are more likely to be written by Alexander Hamilton, or James Madison.

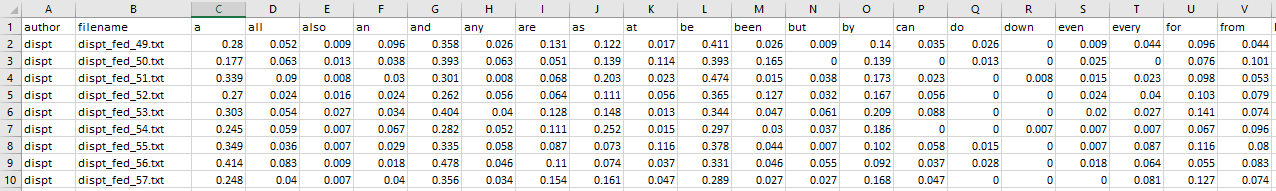
**Analysis**

Data Preparation:

A csv dataset representing instances authorship:

* Alexander Hamilton (**Hamilton**)
* James Madison (**Madison**)
* John Jay (**Jay**)
* Alexander Hamilton and James Madison (**HM**)
* Alexander Hamilton or James Madison (**disp**)

Each having a normalized frequency measure of 70 words:



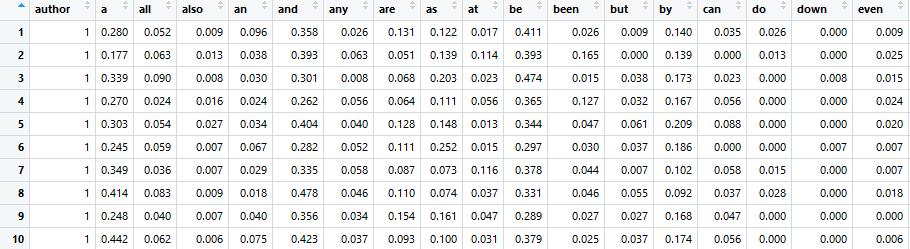
The second column titled filename, corresponds to a supplied textfile, containing the content for the respective paper, with respect to the given author.

Processing:

The provided dataset was loaded into R as a dataframe, using standard read.csv(). Once loaded, some initial preprocessing was implemented. Specifically, the author column was converted into a numeric type. Then, the second filename column was removed, since the supplied fedPapers85.csv, normalized the overall dataset:

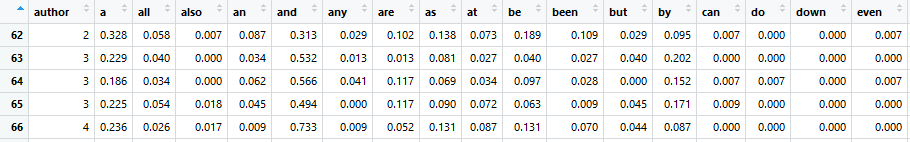
df[, 1] = as.numeric(df[, 1])

df = df[, -c(2)]



Some analysis was performed using the above adjusted dataset. However, in attempt to better distinguish between Hamilton and Madison for the disputed papers, all rows with the author set to both Hamilton and Madison (i.e. HM), indicated by the value 3, after the above preprocessing, was removed:

df2 = df[df$author != 3, ]



Once the needed dataframes were defined, the kmeans cluster was created using the stats package. Additionally, the expectation maximization (i.e. EM), and hierarchical clustering (i.e. hclust) were implemented through the mclust package. It is important to note, to reduce syntax, a custom R package was created to allow successive packages to be installed, and loaded with one line of code:

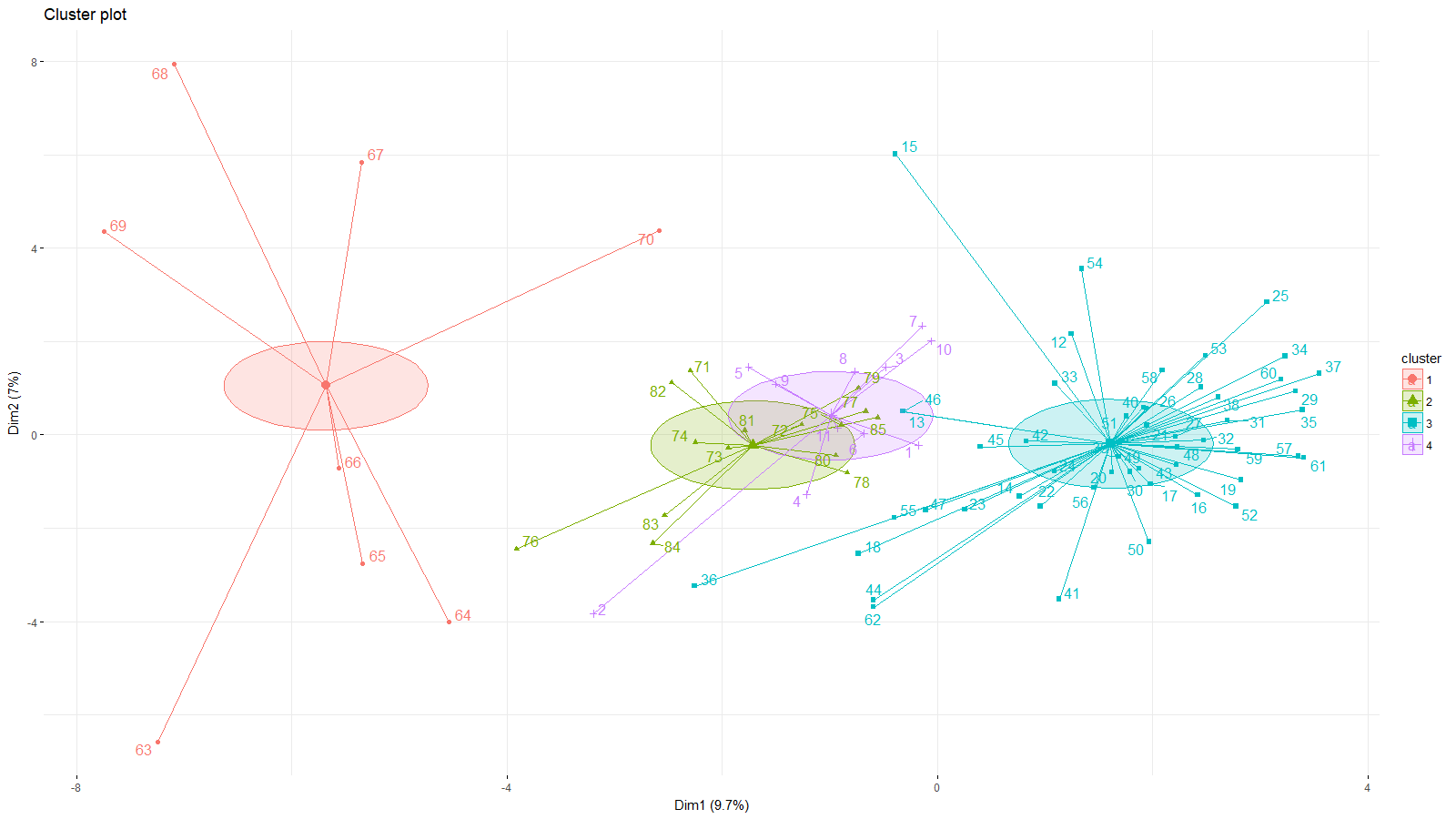
load\_package(c('stats', 'flexclust', 'mclust', 'ggplot2'))

Additionally, the overall R script was duplicated. Specifically, the first case involved generating the kmeans, em, hclust using four clusters, including the Hamilton and Madison coauthoring case. The second case eliminated the latter, which reduced the needed cluster to three.

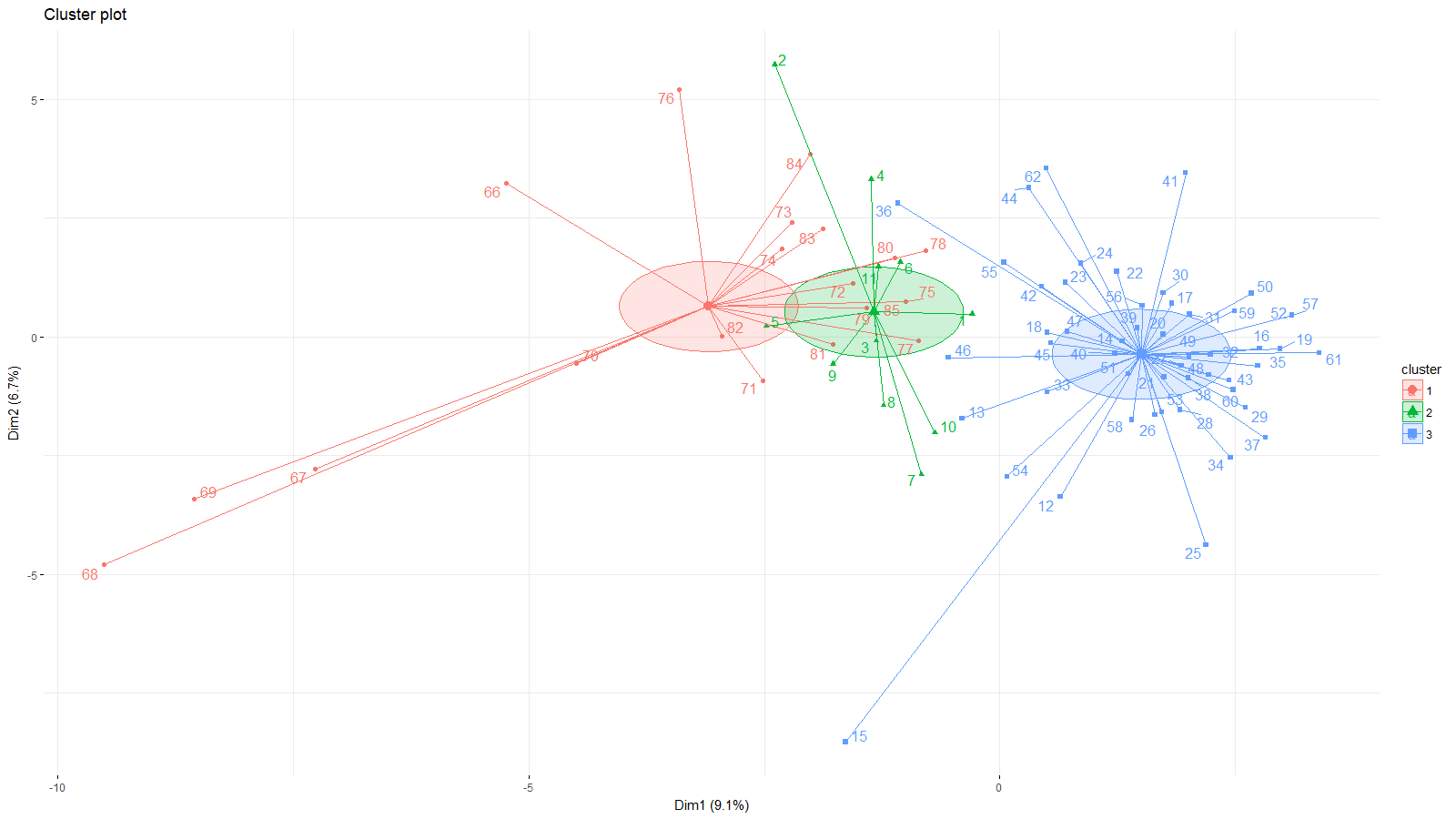
**Results**

First Case: four clusters, all authors including Hamilton and Madison

The generated kmeans clustering resulted with two clusters heavily overlapping one another, while another cluster partially overlapping a former cluster:



However, a second case could consider removing the Hamilton and Madison authored papers, and only consider single authored papers. This would mean only three clusters would exist:



When reviewing both cases, it is noticeable that the disputed papers are more closely related to Alexander Hamilton. Specifically, these articles are closer to the centroid associated with the Hamilton’s cluster. However, when comparing between the two above cases, it appears that the first case (with joint author cluster), produced better results for the ARI randIndex, while the second case had a much lower sum of squares by cluster:

Case 1 (kmeans result subset):

Within cluster sum of squares by cluster:

[1] 3.4275893 1.4444369 5.4981349 0.9230551

(between\_SS / total\_SS = 92.7 %)

Available components:

[1] "cluster" "centers" "totss" "withinss" "tot.withinss" "betweenss" "size"

[8] "iter" "ifault"

===========================================================

randIndex: measure between author, and cluster partitions.

Note: values vary between -1 to 1

===========================================================

ARI

0.9913004

Case 2 (kmeans result subset):

Within cluster sum of squares by cluster:

[1] 7.5691860 0.9230551 5.4981349

(between\_SS / total\_SS = 90.9 %)

Available components:

[1] "cluster" "centers" "totss" "withinss" "tot.withinss" "betweenss" "size"

[8] "iter" "ifault"

===========================================================

randIndex: measure between author, and cluster partitions.

Note: values vary between -1 to 1

===========================================================

ARI

0.9543314

Additionally, upon reviewing the results, standardizing the author column as a numeric attribute may have skewed the results. Specifically, when reviewing a subset of the kmeans cluster means, for the first case:

Cluster means:

author a all also an and any are as at

1 3.625 0.1798750 0.03850000 0.014625000 0.03337500 0.6460000 0.03025000 0.08412500 0.1280000 0.03912500

2 5.000 0.2698000 0.05533333 0.011066667 0.05946667 0.4196667 0.02980000 0.07486667 0.1340667 0.02953333

3 2.000 0.3156078 0.05376471 0.004784314 0.08080392 0.3394902 0.04674510 0.07254902 0.1177255 0.04882353

4 1.000 0.3039091 0.05554545 0.011272727 0.04845455 0.3560000 0.04218182 0.09590909 0.1380000 0.04700000

Values in the author column are significantly higher than other columns. This is the same case for the full result set, as well as for case 2. Removing the author attribute will likely allow other attributes stabilize its weighted influence on the overall cluster. Additionally, the computed expectation maximization resulted in a single partition for both cases. This could likely be impacted by the same author column.

|  |  |
| --- | --- |
| Case 1: expectation maximization | Case 2: expectation maximization: |
|  |  |

Lastly, the following hierarchical cluster visualizes how papers can be aggregated:

|  |  |
| --- | --- |
| Case1 | Case2 |
|  |  |

**Conclusions**

From this investigation, it has been found that the disputed authorship appears more closely related to Alexander Hamilton, then James Madison. When reviewing historical generalizations, it appears that Hamilton was largely active between his time in the army, as well as congress. His experience, coupled with the perceived notion that states needed to be governed more closely, likely gave way to his stance on The Federalist Papers, and it’s needing to be ratified.

James Madison on the other hand was known to be very detail oriented, and at times reclusive. It would be interesting to understand why Hamilton wrote so many more papers than Madison and understand the relationship between the two founding fathers. Specifically, why did Madison write more papers with Hamilton, than himself writing his own papers independently? Similarly, why was there a dispute between authorship, rather than agreement on co-authorship?

Overall, this study has provided some good first steps for initial exploratory analysis. However, obtaining additional volumes, or papers authored by Madison, could help further understand and differentiate from the Federalist Papers written by Hamilton. Similarly, since the study was to differentiate between the two authors, John Jay could potentially be removed during the study, and analysis.