Darrell Nelson II

Data Analytics (IST - 707)

04/23/2019

HW04: History Mystery

Introduction

Without them, there would have been no United States of America. The Founding Fathers, a group of predominantly wealthy plantation owners and businessmen, united 13 disparate colonies, fought for independence from Britain and penned a series of influential governing documents that steer the country to this day.

All the Founding Fathers, including the first four U.S. presidents, at one point considered themselves British subjects. But they revolted against the restrictive rule of King George III—outlining their grievances in the Declaration of Independence and won a stunning military victory over what was then the world's preeminent superpower.

The Founding Fathers is a term applied to those leaders who initiated the Revolutionary War and framed the Constitution. The Federalist Papers are a collection of 85 articles and essays written by three of the founders: Alexander Hamilton, James Madison, and John Jay under the pseudonym "Publius" to promote the ratification of the Constitution. However, with such prestige laid at the feet of the Founding Fathers human nature can arise. There are 11 articles that both Madison and Hamilton claim to have been the sole writers on. After Hamilton died in a duel, a list was found in his personals where he claimed authorship. Madison also publicly announced that the disputed papers were his after his presidential term, perhaps to avoid dishonoring the office with a personal feud. Who to believe: a slain Founding Father or a former President of the United States?

Analysis and Models

About the Data

There are 85 text documents in total:

- 11 papers which are disputed as being authored solely by Alexander Hamilton or James Madison
- 51 papers authored by Alexander Hamilton
- 3 papers coauthored by Alexander Hamilton and James Madison
- 5 papers authored by John Jay
- 15 papers authored by James Madison

There's a much larger sample size of Hamilton's work than Madison's. The larger size may skew the results. However, a reasonable case can be made that if Hamilton was the more advent writer than there is a higher probability that he either wrote or at least influenced the creation of the disputed

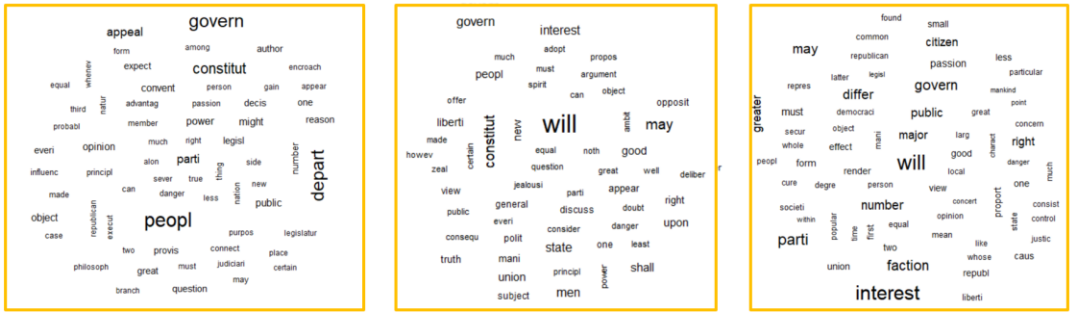
papers. Madison and Hamilton have worked together on these Federalist Papers coauthoring three of them. By working together it is quite possible the more adept writer "rubbed off" on the other. If that's true, another compelling method to solve this mystery would be to put the federalist papers in chronological order and see if the style/word choice of Hamilton gets closer to Madison's over time or vice versa.

All 85 of the Federalist Papers were renamed to correspond to their author(s), with the disputed articles being labeled as “disp”. All documents were brought into R Studio as a single corpus with 85 variables/documents. Extremely rare words and overly used words are ignored by excluding all words from the corpus that are used in less than 1% or more than 50% of the documents. Common English stop words were also removed by using the ISO English database. Upon inspection, the ISO’s database has 175 conventional words. An important note as the Federalist Papers were written before the 1800’s, these stop words will not disrupt the flavorful dialect of old English (King's English). In *Figure 1.* the most/least common root words in the corpus are displayed. Some of the least used root words were “abhor” and “absolve”; meaning to regard with disgust and declare free from blame respectively. The most common root words among all papers was “state” and “will”.

```
> (WordFreq[head(ord)]) # least used words in all documents
 abhorr  abject abraham   abreg  absenc  absolv
      1       1       1       1       1       1
> (WordFreq[tail(ord)]) # most used words in all documents
constitut       may     power    govern      will     state
      686       811       937      1040      1263      1662
```

Figure 1. Frequency List

The words with lowest counts hold the most weight as different authors may use distinct words to convey meaning. This rule applies strongly to this data-set as the papers have the same subject: influence voters to ratify the Constitution. The differences in analogies and turn of phrases should also help identify stylistic patterns. To help identify this, word clouds were generated (*Figure2.*) for the disputed authors and compared to one of the disputed texts. The root word ‘constitut’ is a high frequency word in the disputed and Hamilton text. The root word ‘govern’ is common across all three texts.



Disputed Paper *Hamilton Paper* *Madison Paper*

Figure 2. Word Cloud Comparison

Descriptive statistics showed a large spread between word counts from paper to paper (*Figure 3.*). This may skew the results if we solely rely on word counts, therefore the wordcounts are normalized by dividing each word count in a paper by the total words in that paper.

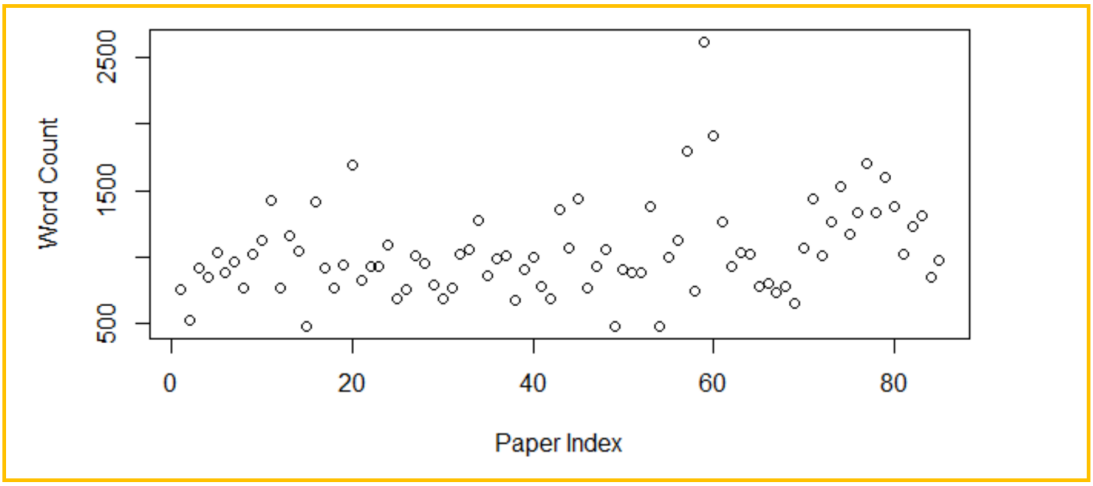


Figure 3. Word Count Chart

Clustering

Hierarchical Agglomerative Clustering (HAC)

HAC is an unsupervised machine learning technique that allows each individual data point to be its own cluster. Then, it groups the two closest clusters together and forms a new cluster. This process is repeated until all data points are grouped into one. This technique is useful when not knowing how many clusters a dataset should consist of. It also provides an excellent visual of the data being grouped from the micro to macro level allowing for a deeper investigation into the clustering methodology.

By transforming the data into a [85 x N] matrix (where each row is a paper and N is the amount of different root words in all of the papers) HAC can be implemented. Once data transformation is complete, the definition of distance must be decided upon. Is it from their respective medoids, centroids, farthest two points in two clusters, nearest two points in two clusters, etc.? There are many different techniques that can be implemented. For example, the Euclidean method measures absolute distance between two points, and the Cosine method measures the angle between two points.

Both methods were tested with the HAC algorithm with the objective function set to minimize Ward's method of minimum variance. "Ward's minimum variance criterion minimizes the total within-cluster variance. To implement this method, at each step find the pair of clusters that leads to minimum increase in total within-cluster variance after merging."[1]

The algorithm was also implemented with the objective function to minimize the maximum distance between clusters (farthest points), the minimum distance between clusters (closest points), the average

[1] (Tan, Steinbach, & Kumar, 2006)

distance of all pairs of members between two clusters, and the distance between centroids of two clusters.

The best methodology for most accurate groupings was the cosine method to measure maximum distance between two clusters of the normalized data.

Kmeans Clustering

K-means clustering was also used to group the data. K-means clustering is an unsupervised machine learning algorithm that randomly assigns *k* number of points to be the centroids for each cluster. Each point in the data-set is then assigned a cluster based off its nearest centroid. Once all data points are assigned, the centroid is recalculated based on the average value of all its members. All members are then re-assigned to their nearest centroid and the centroid is in turn recalculated. This process continues until the centroids reach equilibrium. The pitfalls of this technique are that the user has to be confident in how many clusters should be created and it needs to be run multiple times. Since the initial centroids are random guesses, there is a possibility that they will split one cluster into two or combine clusters when they should be separated out. A good practice to ensure quality results is to run the K-means algorithm multiple times.

As with the HAC algorithm above, which method to calculate distance in highly dimensional data is entirely up to the user. Four different distance measures were used to calculate clusters and each clustering algorithm was ran 5 times for a total of 20 different models. Each model was evaluated on how well it generated clusters around known data (e.g. was it able to group most of the Hamilton articles together).

The best K-means model for this data-set implemented the Pearson's correlation coefficient (r) which measures the strength of association between data points.

Results

After screening various methods the best HAC model was the cosine method to measure the maximum distance between two clusters of the normalized data (*Figure 4.*).

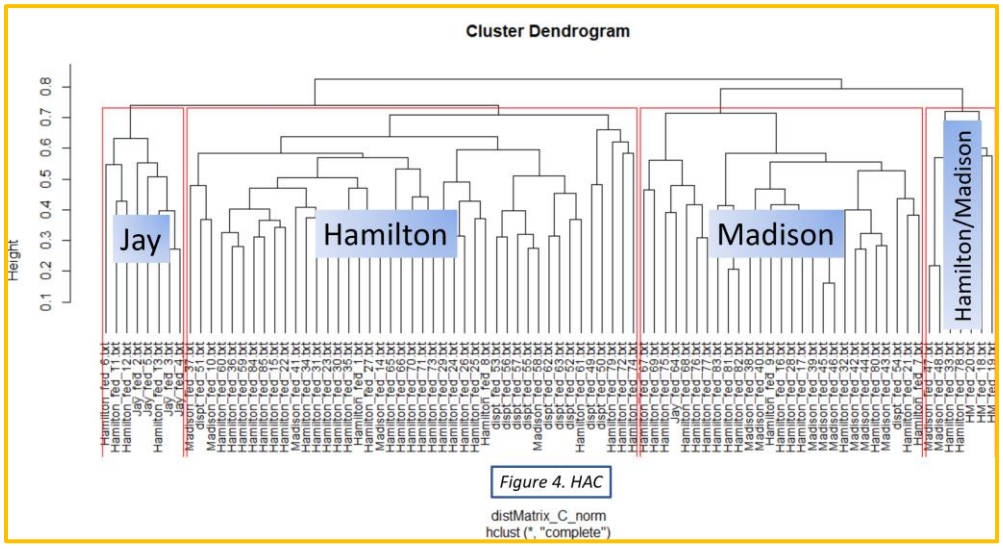


Figure 4. HAC

The above dendrogram gives a visual representation of the hierarchy system built from the selected HAC mode. There are 4 distinct clusters:

1. "Jay"
    a. Comprised of 8 papers in total; contains 4 out of 5 of John Jay's papers (80%) and 4 out of 51 of Alexander Hamilton's papers.
2. "Hamilton"
    a. Comprised of 43 papers in total; 28 of 51 of Hamilton's, 10 are disputed, and 5 out of 15 of James Madison's.
3. "Madison"
    a. Comprised of 27 papers in total; 8 out of 15 of Madison's, 1 out of 5 of Jay's, 1 disputed paper, and 17 of 51 of Hamilton's papers.
4. "Hamilton/Madison"
    a. Comprised of 7 papers in total; 3 of 3 of the Hamilton/Madison coauthored papers, 2 of 15 of Madison's, and 2 of 51 of Hamilton's.

The table below looks at each cluster and the percentage of total papers created by a specific author in that cluster:

	Clusters			
Authors	Jay	Hamilton	Madison	Hamilton/Madison
Jay	80%	0%	20%	0%
Hamilton	8%	55%	33%	4%
Madison	0%	33%	53%	13%
H/M	0%	0%	0%	100%
Disputed	0%	91%	9%	0%

Table 1. % of Authors Work in Cluster

For example, the first cluster "Jay" in blue contained 80% of all the papers written by author John Jay in grey and 8% of all the papers written by Alexander Hamilton. Each cluster has a significant majority of one author's papers which is why the names of each cluster reflects one author.

The highest performance from the K-means algorithm grouped papers based off of their Pearson's correlation coefficient (r) which measures the strength of association between data points (*Figure 5.*).

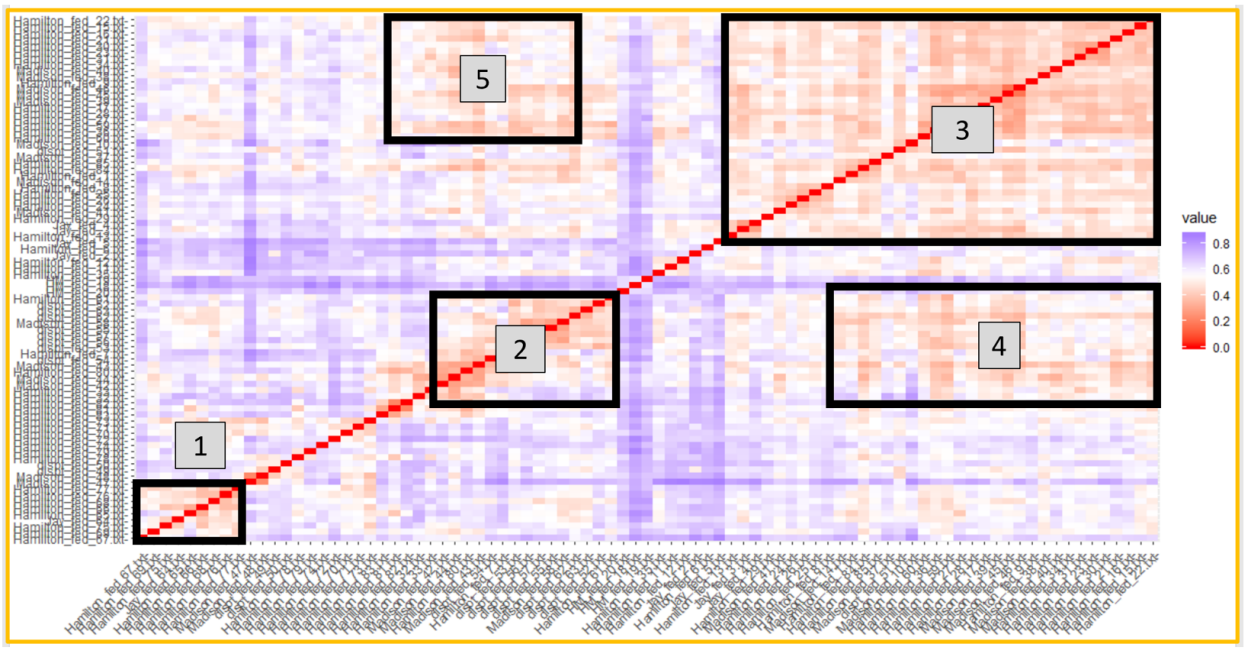


Figure 5. K-means

Based off the Figure above there are 5 distinct clusters. Three of which fall along the line of symmetry (red line) which tells us that the papers nearest each other on the x-axis are highly associated with themselves on the y-axis. Cluster 1 is primarily composed of Hamilton's work, cluster 2 is primarily composed of disputed text with only a small percentage of either Hamilton or Madison texts. This cluster does not provide enough distinction between Hamilton and Madison to make a clear distinction between the two authors. Cluster 3 is mainly Hamilton papers associating with each other. The more interesting clusters are the ones that are off the red line. Cluster 4 groups a lot of Hamilton's papers on the x-axis with 8 of the disputed papers on the y-axis. Cluster 5 groups Hamilton and Madison texts in equal amounts with each other. Cluster 4 contains the most important information to the overarching question of: Which author wrote the disputed papers? This analysis suggests that at least 8 of the 11 disputed papers were authored by Hamilton which agrees with the HAC model as well.

Conclusion

Using clustering techniques to predict the appropriate author of the Federalist Papers came back with surprisingly clear results. The HAC algorithm clustered the 85 papers into 4 groups based on word frequency and each cluster contained a majority of work from a single author. Meaning that their writing styles and choice of vocabulary were varied enough to make appropriate distinctions. The first cluster "Jay" contained 80% of all of John Jay's papers, the "Hamilton" cluster contained 55% of all of Alexander Hamilton's work, the "Madison" cluster contained 53% of all of James Madison's work, and the last cluster which was chosen to distinguish papers that were coauthored by Hamilton and Madison grouped all of those papers correctly. It is key to note that in the "Hamilton/Madison" cluster, not a single disputed paper was found. This means that it is less likely that these disputed papers were papers that Hamilton and Madison worked own together and someone wanted to take all the credit. It is more likely a majority of these papers were written by one distinct author. This author was: Alexander Hamilton!

The K-means clustering analysis also supports this hypothesis. In Cluster 4 there is a strong association between 8 of the disputed papers and Hamilton's writing. While the other 3 are lost in the noise, a closer inspection also provides strong evidence that Hamilton might have authored 2 more of the papers as well. Two algorithms with different methodologies have strongly suggested that Hamilton is the true author.

Although, it is quite possible that Madison wrote a few of the distinct papers the vast majority should confidently be assigned to Hamilton. This also makes sense historically: Hamilton was the one that first wrote to claim the papers before he was killed in a duel. After Hamilton's death, and experiencing the fame and fortune being a Founding Father brought to you, Madison sought to claim the papers to boost his own prestige. As traitorous as that sounds, remember that the Founding Fathers were the most traitorous lot there ever was in the Kingdom of Great Britain. God Bless America!