Nicholas Louis Brown

Professor Wallace

IST 707

September 25, 2018

Author Attribution Using K-means and HAC Clustering Techniques

**Executive Summary:**

The federalist papers are a series of papers with a disputed history: no one is sure whether Alexander Hamilton, James Madison, or John Jay authored the documents. Using clustering techniques, we are able to gain insight into the possible answer to this historical mystery. The process used to attribute authors to the documents was twofold: first, the k-means clustering algorithm and HAC algorithm was used to create four and five distinct clusters each which contained the known works of Hamilton, Madison, Jay, and Hamilton & Madison as well as the disputed documents. Second, we were able to successfully make a cluster plot, allowing us to make educated guess as to who authored the disputed papers. Using these methods, we concluded that the author of the disputed papers is most likely James Madison. Our first and strongest piece of evidence, the K-means cluster plot (Figure 1), showed the disputed papers assigned to clusters three and four similar to the known Madison papers. The numbers serve to further this theory: Of the 11 works with disputed authorship, seven of them were assigned to cluster 3 and four were assigned to cluster 4, (a ratio of 7:4). This ratio of works is very similar to that of James Madison, where out of 15 known Madison documents, nine were assigned to cluster 3, and six were assigned to cluster 4 (a ratio of 3:2). Looking at the table below it is even more apparent that Madison is the best fit as all other options such as Hamilton or Jay exist only in clusters 1 and 2. Although we can guess that Madison wrote the papers, more data would be useful so that we could more confidently assume these trends to be accurate.

**K-Means Cluster Plot:**

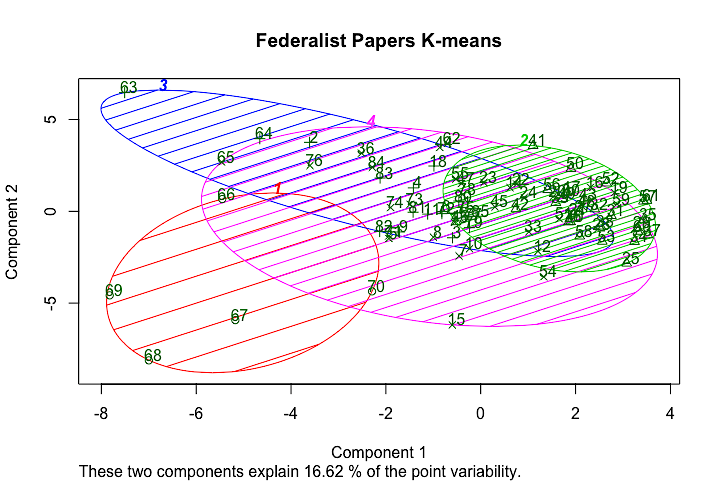
****

Figure - Kmeans Cluster Plot

**Methodology:**

**Preprocessing:**

|  |  |
| --- | --- |
| Cluster Number(s) | Author |
| 1 | John Jay |
| 2, 4 | Hamilton |
| 3, 4 | Madison |
| 4, 3 | Hamilton & Madison |

To preprocess the data for clustering analysis I first imported the csv of tokenized word vectors into R studio as a data frame and removed the first two columns of the data as they would not be useful in clustering classification. These two columns contained the authors name and the document name respectively.

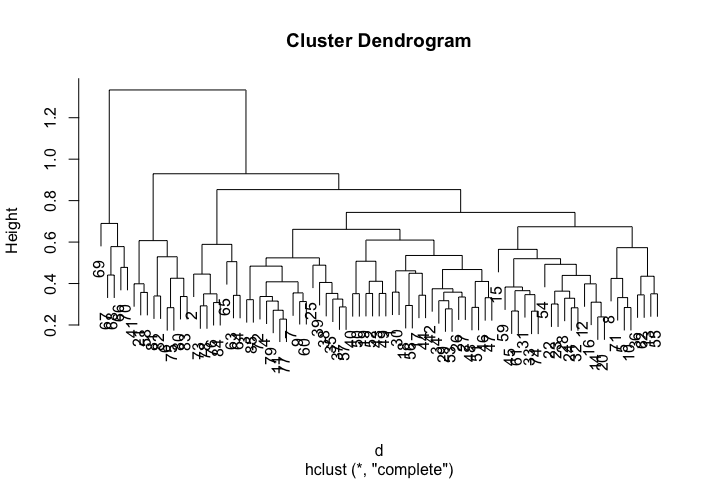
**K Means Cluster:**

The k-means model was easily implemented in R in a total of three lines of code. Using our stripped-down data frame (myData\_unlabel) containing only the clustering data, a k-means model was made. The decision to use 4 clusters (n-1) was made because there was only a possibility of the papers being authored by one of four people.

model\_r <- kmeans(myData\_unlabel,4)

cluster\_assignment <- data.frame(myData,model\_r$cluster)

View(cluster\_assignment)

**HAC Algorithm:**

Our HAC algorithm used to compute the Euclidian distance between each point of data generated the dendrogram below. Again, this was a very simple process in R requiring a total of three lines of code. Because the dendrogram is significantly less intuitive and was only very helpful in confirming our k-means clustering results we chose to abstain from including it in the executive summary. The dendrogram shows our four categories with the final category (Madison & Hamilton) being split into two smaller categories, presumably the works of Madison and Hamilton as individuals. This dendrogram really helped to reinforce how we really can’t be certain as to who wrote the articles without more data.

d = dist(as.matrix(myData\_unlabel))

hc=hclust(d)

plot(hc)

**Conclusion:**

Though we concluded that James Madison wrote the disputed papers it is important to reiterate that this is only one conclusion using one method of data analysis techniques. Ideally, more data could be gathered using other known works to help improve the clustering accuracy. Additionally, the co-authored works should be excluded next time to make for more clear boundaries as to authorship. Because we included the co-authored works in our cluster we cannot be absolutely certain that James Madison alone wrote the disputed papers.