Sammy Pardes  
IST707  
Homework 4  
2//2020

**Introduction**

In 1788, Alexander Hamilton, James Madison, and John Jay wrote what would come to be known as the Federalist Papers. The Papers were a collection of 85 essays written to encourage endorsement of the Constitution by the people of New York state. Their efforts ultimately led to the ratification of the Constitution, the cornerstone of our nation!

Although the Papers were all written about the same topic, each author had his own ideas for what the federal government should look. Even so, authorship has not been credited to twelve of the essays.

The National Archives Museum will be revamping their exhibit on the founding fathers in the spring of 2020. The museum would like to attribute authorship of the disputed papers at the grand opening to encourage visitors. To solve the mystery of the Federalist Papers, the National Archives Museum has enlisted the expertise of SLP Consulting.

**Analysis and Models**

**About the Data**

The National Archives provided a digitized collection of the Papers as a zip file. Each of the 85 papers was formatted as a separate txt file. The zip file was loaded into R as a corpus of 85 documents with the Corpus() function.

To remove some of the noise, a copy of the zip file without the Jay or co-authored papers was also created to examine the group of documents. Once the remaining 77 documents were analyzed successfully, the final 8 were added back into the corpus. The DocumentTermMatrix() function was used to convert the corpus into a document term matrix (DTM).

Many, many combinations of parameters were paired together to ultimately create the desired number of terms within the matrix. At first, it was theorized that using less frequent words to identify authorship by uniqueness of word choice. Since there were many thousands of words that were used infrequently within the papers, this created too much extraneous data. Alternatively, using only the most frequent words yielded better results. The bounds of the DTM were set to only include words that appeared in more than 60% and less than 98% of all terms. Stopwords, or commonly used words, punctuation, and numbers were ultimately left in the matrix. All words were converted to lowercase and “stemmed” to find their roots. This left 172 terms in the 85-paper corpus and 170 terms in the 77-paper corpus.

Using as.matrix(), the document term matrix was then converted into a matrix so that it could be used for distance measuring. Due to the varying lengths of each paper, the matrix had to be normalized. That is, each terms occurrence was converted from a count to a percentage to ensure that the frequency of the word was not impacted by the length of its document. The number of occurrences of every word was divided by the total number of words in the containing paper. This value was then multiplied by 100 to yield the percentage, from 1 to 100, of frequency. The word with the greatest frequency occurred in 10% of its parent document.

**Models**

Several distance measures were computed to determine the variance in the similarity of each document. Each distance measure, Euclidean, Cosine Similarity, and Manhattan, was then used in Hierarchical Clustering analysis. Using the corpus without Jay and co-authored papers, the number of clusters was set to 3 to accommodate the Hamilton, Madison, and disputed papers. When using the corpus with all 85 papers, the number of clusters was set to 5. This accounted for the Jay and H-M papers.

Re-using the three distance measures, heat maps of document similarity were created. Each of the distance measured produced very similar looking graphics.

A picture containing object

Description automatically generated

K Means Clustering was then performed with 3 and 5 clusters on the normalized matrix as well as on a version of the matrix where the rows and columns were transposed. This allowed the grouping of documents to be viewed in addition to the grouping of words.

**Results**

**Hierarchical Clustering**

This code was run multiple times and generated similar results each iteration. The dendrograms below illustrate the overarching trends. These show the results of the hierarchical clustering using the Manhattan distance measure for the 3 and 5 clustered matrices, with and without the Jay and H-M papers. It is quite clear that the Madison and disputed papers were grouped together in both diagrams. In the 5-cluster plot, all the Jay papers are grouped together along with a few papers from Hamilton and Madison. The co-authored papers are mostly grouped with Hamilton.

A close up of a logo

Description automatically generated

A screenshot of a cell phone

Description automatically generated

**K Means Clustering**

Excluding the Jay and co-authored papers displayed the following plots. Each time, it became clear that, while there is some overlap between the papers, the disputed papers are more closely related to Madison than they are to Hamilton. Two of the clusters are primarily filled with Hamilton papers and largely overlap one another. The third cluster is a greater distance from the other two and is almost exclusively populated by the disputed and Madison papers.

A close up of a map

Description automatically generated

Using the full 85-paper matrix, a very similar outcome occurs. Jay has his own cluster featuring a few papers from both Madison and Hamilton. Hamilton has two over-lapping clusters mostly to himself, while Madison has his cluster intertwined with the disputed cluster. This indicates that the Madison’s papers and disputed papers are most similar when compared to Jay and Hamilton.

A close up of a map

Description automatically generated

When the rows and columns were transposed, the cluster plot showed the words that were most like one another, rather than the documents. It can be inferred that the words in clusters 1 and 2 belong to the disputed and Madison papers because they are closest together.

A close up of a map

Description automatically generated

**Conclusion**

After a lengthy analysis, SLP Consulting has determined that the author of the twelve disputed papers was, decisively, James Madison. Time and time again, the models generated through Hierarchical and K Means clustering support this conclusion.

It is probable that Madison wished to remain anonymous due to his conflicting views from his peers about the Executive branch of his future federal government. Although Madison and Hamilton each wanted the Executive branch of the United States to be kept in check by the Legislative and Judicial branches, Hamilton wanted to give more power to the Executive branch. Further analysis of the papers may be done to confirm that the views of the disputed papers match those of Madison.

The National Archives Museum can now be confident that Madison wrote the disputed Federalist Papers. When opening their new exhibit, the National Archives Museum should attribute the disputed Federalist Papers to Madison.