Sammy Pardes  
IST707  
Homework 5  
2/10/2020

**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. To encourage visitors, the museum would like to attribute authorship of the disputed papers at the grand opening. With the exhibit opening in a few short weeks, the National Archives Museum has asked SLP Consulting to re-examine its original determination of authorship to be even more sure the writer is correctly credited.

**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. Another corpus of 66 documents was created containing only the Hamilton and Madison papers, for preliminary analysis.

The corpus of papers was then converted into a Document Term Matrix (DTM) with the DocumentTermMatrix() function. The parameters set on the DTM removed separators, converted the words to lowercase, stemmed the words, and removed select stop words. Only words between 60% and 98% frequency were included in the DTM. This left a DTM containing 173 terms, and 25% sparsity for the Hamilton/Madison papers, and 189 terms with 41% sparsity for all papers. The terms were then normalized to account for variations in document length using the weightTfIdf() function.

Once the DTM was created, it was converted into a matrix and then to a data frame. The row names of the documents were determined and converted into author names. At first, this only included Hamilton and Madison. Using the second corpus the Jay, disputed, and co-authored papers were added.

For the Hamilton/Madison only data frame, the number of rows was determined and then the indices were randomized. 2/3 of the randomized rows were selected to create a training data set. The remaining 1/3 was used as the test data set. For the data frame with all 85 papers, the data frame was split up into three sections. The first contained every third row beginning at row 1. The second held every third row starting at row 2, and the third data frame consisted of every third row from row 3 onward. The second and third data frames were combined to create the test data. The first data frame was used as the training data set.

**Models**

Both the Hamilton/Madison only and complete set of Federalist Papers were used to generate decision trees with rpart(). The added “author” column was set as the variable to be predicted and all variables, or words, were used as potential predictors. To refine the results, the minimum number of observations for each split was set to 2 and the complexity parameter was set to 0.

Once the models were generated with rpart(), the predict() function was used on the test data set to test the accuracy of the model. For the Hamilton/Madison only papers, the model was spot on. Each of the 22 papers were attributed to the correct author.

A close up of text on a black background

Description automatically generated

The snippet below show that with all the documents together, while the model misidentifies a few of the authors, the disputed papers were mostly attributed to Madison or vice versa.

A close up of text on a black background

Description automatically generated

fancyRpartPlot() was used to generate the graphical representation of the decision trees.

**Results**

Looking at only the Hamilton and Madison papers, when a word with the root “upon” was used at least 0.22% of the time in a paper, the paper was likely to be written by Hamilton. If “upon” occurred less frequently, there author was probably Madison. If the root “consid” appeared at less than 0.22% frequency, then this all but confirmed that Madison wrote the paper. In only 2% of cases, Hamilton wrote the paper if “upon” was < 0.22% frequent and “consid” > 0.22% frequent.

A close up of a device

Description automatically generated

Using this model on the entirety of the Federalist Papers produced the below decision tree. With a similar initial indicator, in 56% of cases, if “upon” occurred in at least 0.49% of a paper, it was likely written by Hamilton. All the disputed cases fell under the Madison branch of the tree.

After the first node, if the root “doe” showed up with at least at 0.59% frequency and “given” with less than 0.03%, the paper was very likely written by Madison. Otherwise, it’s probable that the paper was either written by Jay or co-authored by Madison and Hamilton. Here, the determining factor is the frequency of the root word “act”. Jay used this term more often than Hamilton/Madison. Interestingly, the Hamilton-Madison co-authored papers fall right in the middle of the resulting tree nodes, with Hamilton and Madison on the bookends of the tree.

It is known for certain that Hamilton wrote at least 60% of the papers and Madison wrote at least 17%. This model accurately displays these proportions with Hamilton at 59% and Madison at 17%.

A picture containing indoor

Description automatically generated

**Conclusion**

SLP Consulting has determined that the author of the twelve disputed papers was, decisively, James Madison. Although the decision tree analysis on its own may not be enough evidence, together with k-means and hierarchical clustering, this conclusion is strongly supported.

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.