Jeff Levesque

Professor Gates

IST 565

Homework #5

**Introduction**

Three founding fathers, Alexander Hamilton, James Madison, and John Jay greatly contributed to the ratification for the Constitution of the United States of America. In the papers they authored, The Federalist Papers, a consistent message was delivered in favor for signing the Constitution.

These founding fathers wrote the collection of 85 papers. However, 11 were disputed to have authorship of either Alexander Hamilton, or James Madison. This controversy motivated many to perform various analysis. In a previous result, using kmeans clustering, it was found that the disputed authorship belonged to Alexander Hamilton. However, it was also noted in the result, the analysis was likely flawed. Instead of discretizing the author, the removal would have produced better clustering results.

In this study, decision trees will be used to verify whether previous results were accurate. This will be done, by generating a train set using all papers except those of disputed authorship. Once a model is generated, a prediction will be made on the disputed authorship. Various error values, and tuning methods will be attempted to verify the integrity of the decision trees.

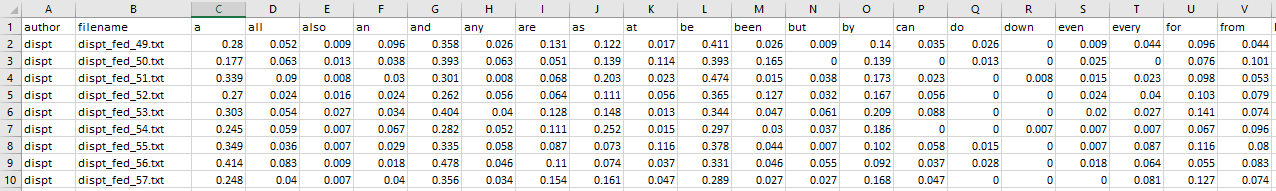
**Analysis**

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 (**dispt**)

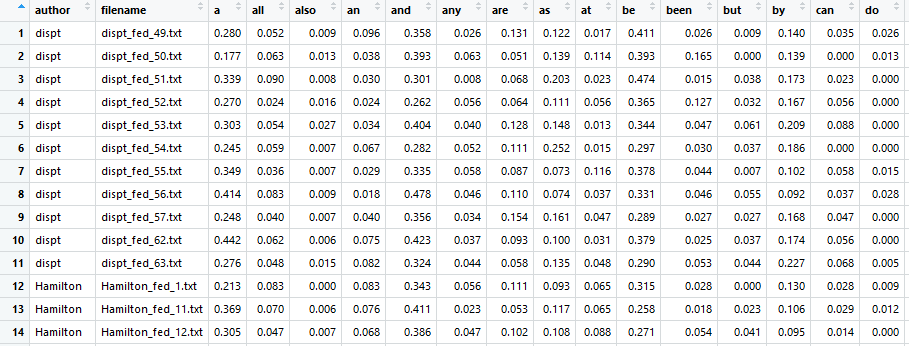
Each having a normalized frequency measure of 70 words:



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

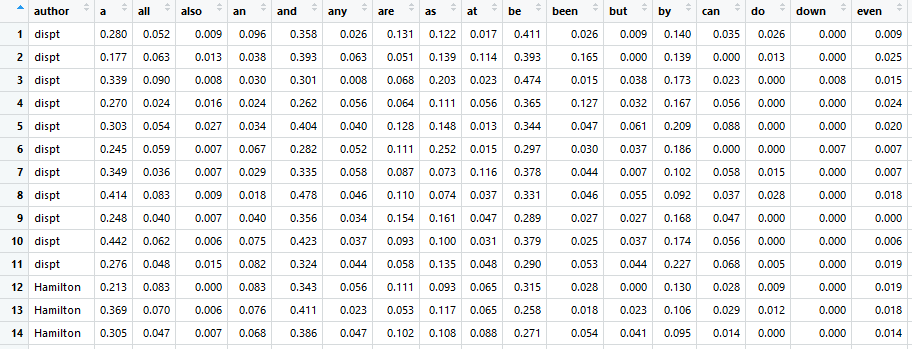
Processing:

The provided dataset was loaded into R as a dataframe, using standard read.csv():



Once loaded, some initial preprocessing was implemented. Specifically, the filename column was removed, since the supplied fedPapers85.csv, contained the same redundant file information within the normalized values:

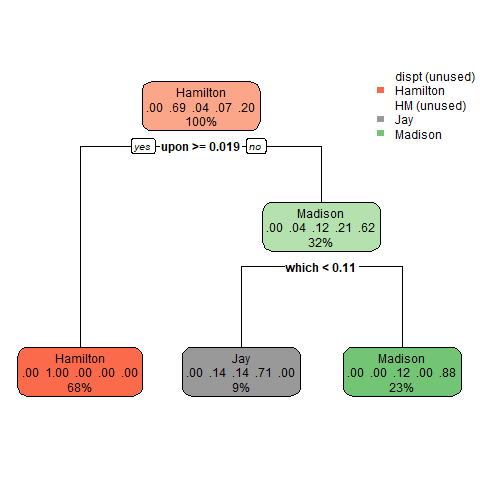
df = df[, -c(2)]



Next, the train set was assigned all row instances of the dataframe not containing dispt. Conversely, the test set was any row instance of the dataframe containing the dispt author. Prior to this segregation, an attempt was made to create a train set using 2/3 of all rows not disputed, while the test set was 1/3 of the remaining rows not disputed. However, the selection of the train set, was the first 2/3 articles not disputed. Since the supplied dataset grouped authors together, the test set contained only Madison articles. Therefore, this approach was largely overfit, and invalid. A better attempt would have to been to randomize the remaining articles not disputed for assignment into either the train or test set.

**Results**

Using the train set, a default decision tree was generated. The resulting tree using default values supplied by the rpart package, generated 3 levels:



It is also seen that two variables were used to construct the decision tree:

Classification tree:

rpart(formula = author ~ ., data = train, method = "class")

Variables actually used in tree construction:

[1] upon which

Since the train set did not include the dispt cases, the resulting decision tree does not contain any nodes with this value. Additionally, Hamilton and Madison coauthor cases were not significant, and were removed internally during the model generation. The “root node error”, used to measure the error rate at the node tree was 0.31, while the re-substitution error rate, used to measure the predictive performance was found to be 0.054.

Using the generated decision tree, the disputed cases were tested. The results indicate most of the disputed authorship to be owned by James Madison:

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

test prediction (probability)

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

dispt Hamilton HM Jay Madison

1 0 0.0000000 0.1176471 0.0000000 0.8823529

2 0 0.0000000 0.1176471 0.0000000 0.8823529

3 0 0.1428571 0.1428571 0.7142857 0.0000000

4 0 0.0000000 0.1176471 0.0000000 0.8823529

5 0 0.0000000 0.1176471 0.0000000 0.8823529

6 0 0.0000000 0.1176471 0.0000000 0.8823529

7 0 0.0000000 0.1176471 0.0000000 0.8823529

8 0 0.0000000 0.1176471 0.0000000 0.8823529

9 0 0.0000000 0.1176471 0.0000000 0.8823529

10 0 0.0000000 0.1176471 0.0000000 0.8823529

11 0 0.0000000 0.1176471 0.0000000 0.8823529

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

test prediction (class)

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

1 2 3 4 5 6 7 8 9 10 11

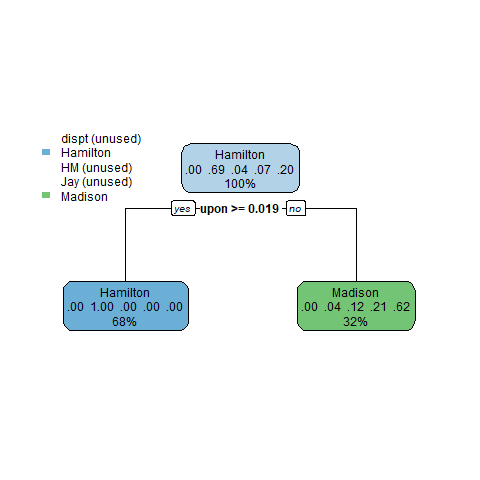
Madison Madison Jay Madison Madison Madison Madison Madison Madison Madison Madison

Levels: dispt Hamilton HM Jay Madison

Interestingly, the third disputed paper was found to be owned by John Jay. However, this is not possible, since the dispute belongs between Alexander Hamilton, and James Madison. Thus, it is immediately evident that the prediction contains errors.

On a similar note, a second decision tree was generated using a single tuning parameter. Specifically, control = list(maxdepth = 1) was implemented to reduce the earlier three levels to one. This allowed the John Jay node, on the third level to be removed, since this study aims at building a model to distinguish a binary case between Alexander Hamilton, and James Madison.

The resulting decision tree depicts a trivial two-level binary tree:



It is also seen that one variable was used to construct the decision tree:

Classification tree:

rpart(formula = author ~ ., data = train, method = "class", control = list(maxdepth = 1))

Variables actually used in tree construction:

[1] upon

The “root node error”, used to measure the error rate at the node tree was 0.31 (unchanged), while the re-substitution error rate, used to measure the predictive performance was found to be 0.121 (increased).

Using the generated decision tree, the disputed cases were tested. The results indicate most of the disputed authorship to be owned by James Madison:

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

test prediction (probability)

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

dispt Hamilton HM Jay Madison

1 0 0.04166667 0.125 0.2083333 0.625

2 0 0.04166667 0.125 0.2083333 0.625

3 0 0.04166667 0.125 0.2083333 0.625

4 0 0.04166667 0.125 0.2083333 0.625

5 0 0.04166667 0.125 0.2083333 0.625

6 0 0.04166667 0.125 0.2083333 0.625

7 0 0.04166667 0.125 0.2083333 0.625

8 0 0.04166667 0.125 0.2083333 0.625

9 0 0.04166667 0.125 0.2083333 0.625

10 0 0.04166667 0.125 0.2083333 0.625

11 0 0.04166667 0.125 0.2083333 0.625

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

test prediction (class)

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

1 2 3 4 5 6 7 8 9 10 11

Madison Madison Madison Madison Madison Madison Madison Madison Madison Madison Madison

Levels: dispt Hamilton HM Jay Madison

Although John Jay was removed from the train set, which eliminated potential errors of him being predicted, the re-substitution error rate doubled. Due to the nature of the supplied dataset, the decision tree was not largely branching. Therefore, there was not many opportunities for tuning. For example, limiting the number of branch levels in the tree only provided a small option. Additionally, since Hamilton dominated in quantity of written papers, limiting the number of nodes per bin, would not produced nearly as high of an impact as desired.

**Conclusions**

From this investigation, it has been found that the disputed authorship appears more closely related to James Madison, then Alexander Hamilton. These results are quite contrary to the previous finding. However, as indicated in this study, implementing proper train, and test sets would have provided additional error measures, for the corresponding decision tree models. But, due to the relatively small dataset size, further splitting the provided dataset into a train and test set, would have reduced the already small dataset for training a model. This could likely increase the chances of overfitting.

As mentioned in a previous study, obtaining additional works produced by James Madison, to complement the initial dataset, would have produced a richer context to work with. The same case for obtaining more writing samples for Alexander Hamilton would be valid. The best steps forward for both studies could likely be an integration of more data, as well as an adjustment to each preprocessing steps.

Overall, it seems the results produced by this study, are more likely to better predict, than the clustering study. This is due to the fact of the incorrect assumption to retain the author attribute during clustering. Nevertheless, both approaches have concluded the same notion. Techniques implemented for prediction, are often limited by the size of a dataset. As shown in the result of this study, the decision trees were constructed using at most two words. Perhaps, only measuring word frequency, followed by a normalization technique, is not enough by itself. Measuring different arrangement of words, with corresponding distance measures relative to one another, could potentially be an improvement to these studies. Though, it has been said many times, the simplest solution is often found best.