Toni Hanrahan

Federalist Papers Analysis – Decision Tree

**Introduction:**

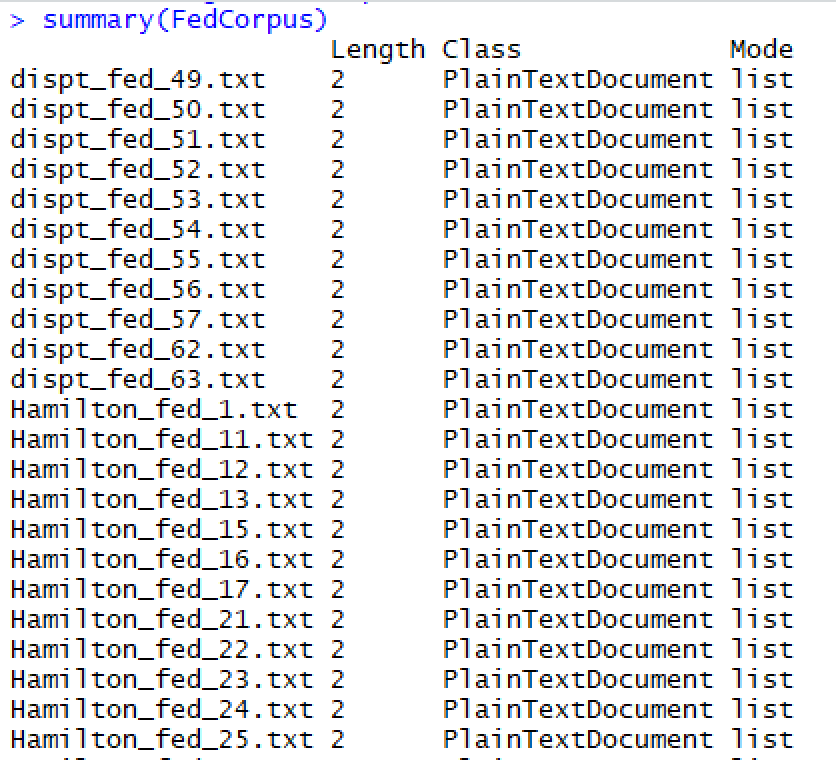
Authorship attribution is a modern way of determining the author of a document or written text when it is unclear who wrote it. This can be very useful when two or more people claim to have written something or when no one is willing (or able) to state that she or he wrote the piece. As the range of anonymous information increases, and plagiarism becomes more prevalent, the ability to determine authorship is a critical function.

From Shakespeare, to “The Night Before Christmas” to the Unibomber, data science has enabled analysts to identify the authors of many previously unknown or disputed works. A common technique to perform this analysis is to use unsupervised clustering. Data is grouped into clusters based on a measurement of similarity. This allows us to also be able to distinguish non-similar data points, effectively separating information into clear categories.

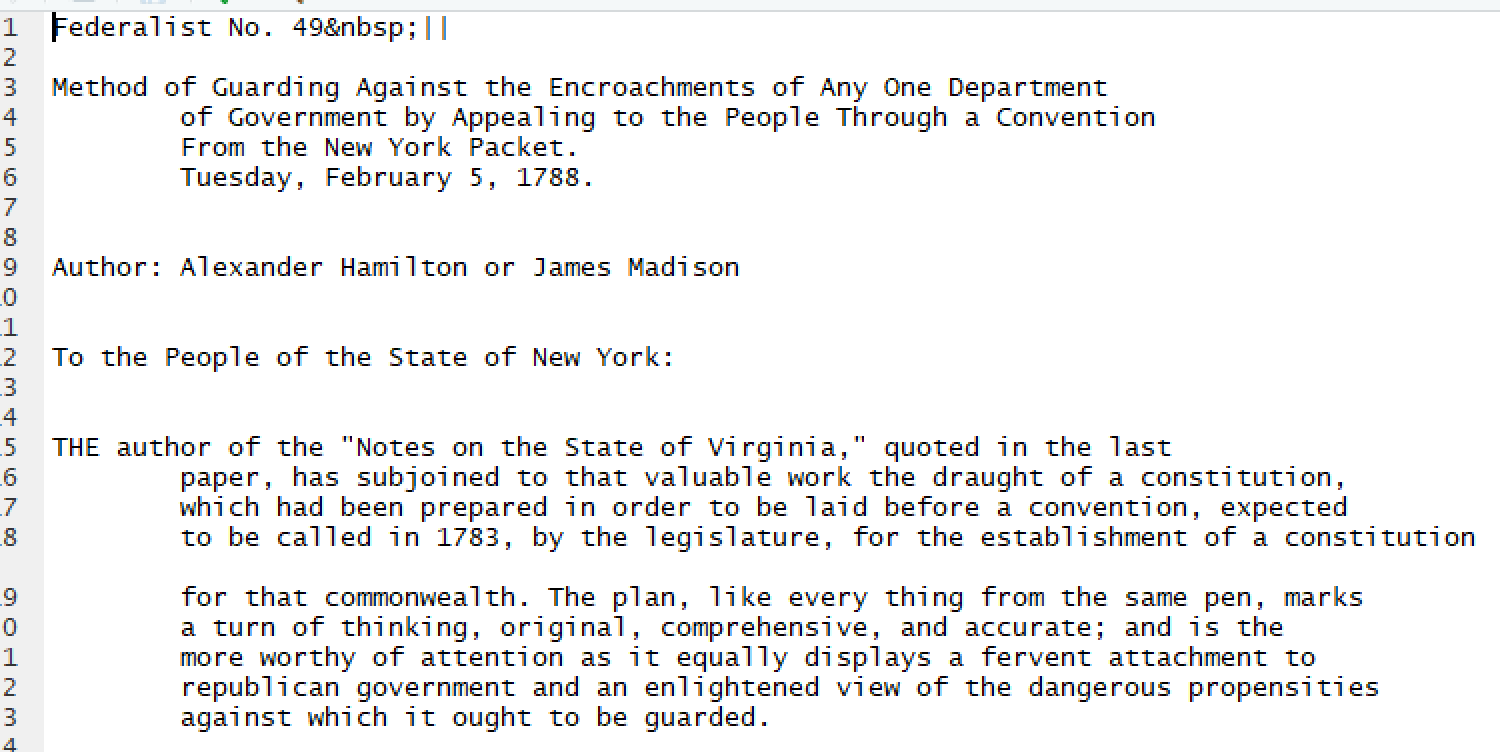
The *Federalist* Papers are a famous example of historical documents where authorship is under dispute. While the *Federalist* papers were actually written by Alexander Hamilton, John Jay, and James Madison, they were published under a pseudonym, presumably to protect these founding fathers from any repercussions from these controversial papers. And while many of the papers have since had their correct authors identified, eleven are still under dispute over two hundred years later.

**Analysis and Models**

*About the data*



*85 files in a corpus*

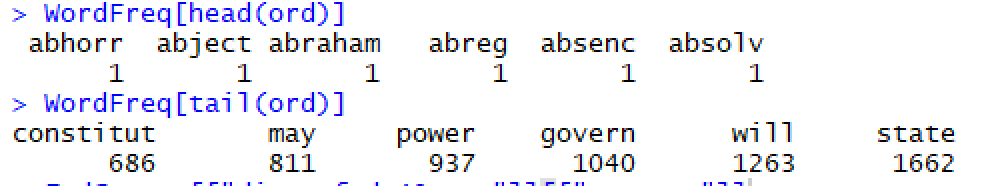


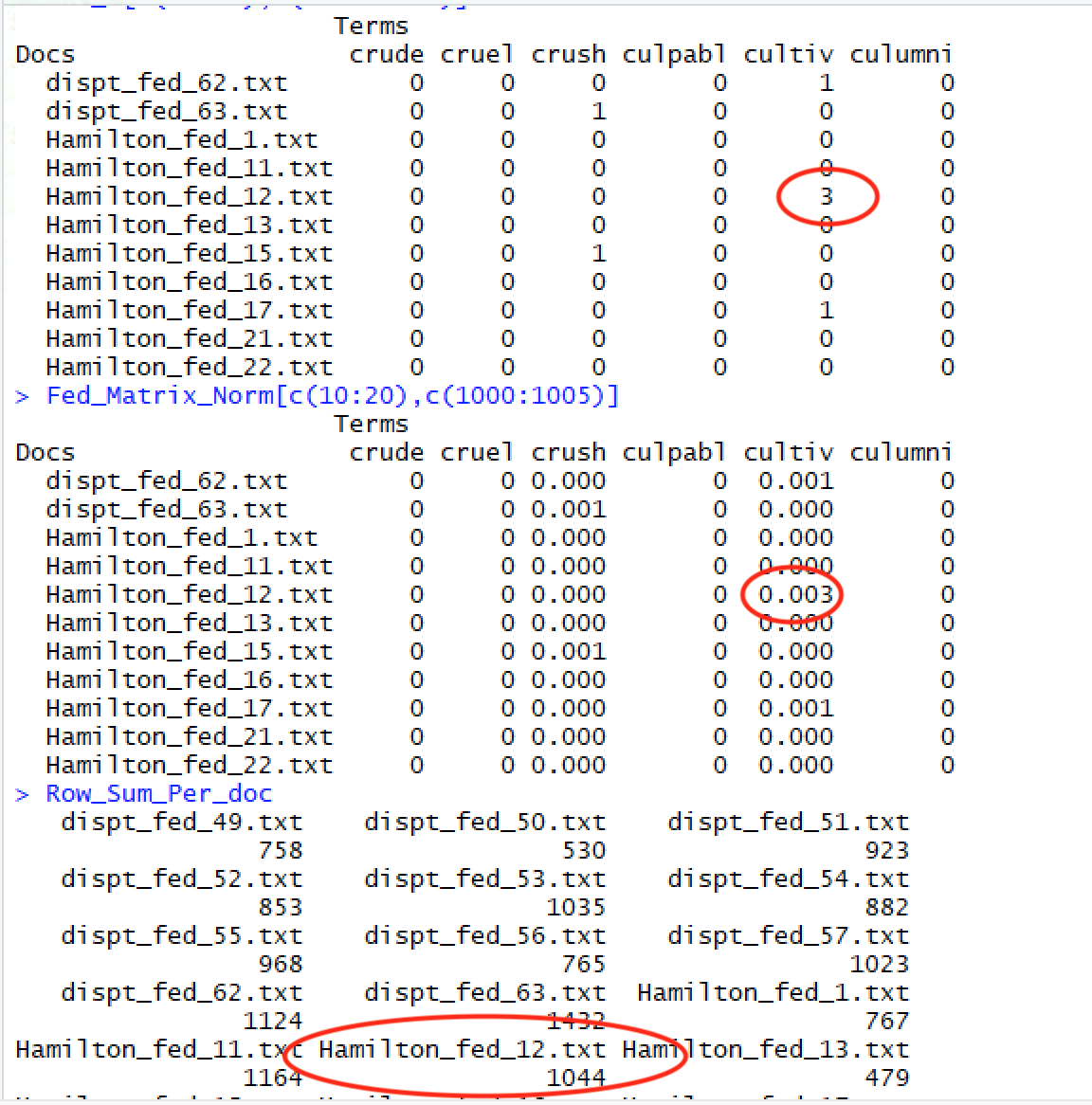
*Example of a text file*

The data used in this analysis originated from 85 individual text files. Each of these text files contains the text from each *Federalist* Papers. Together in the same folder these files comprise a corpus.

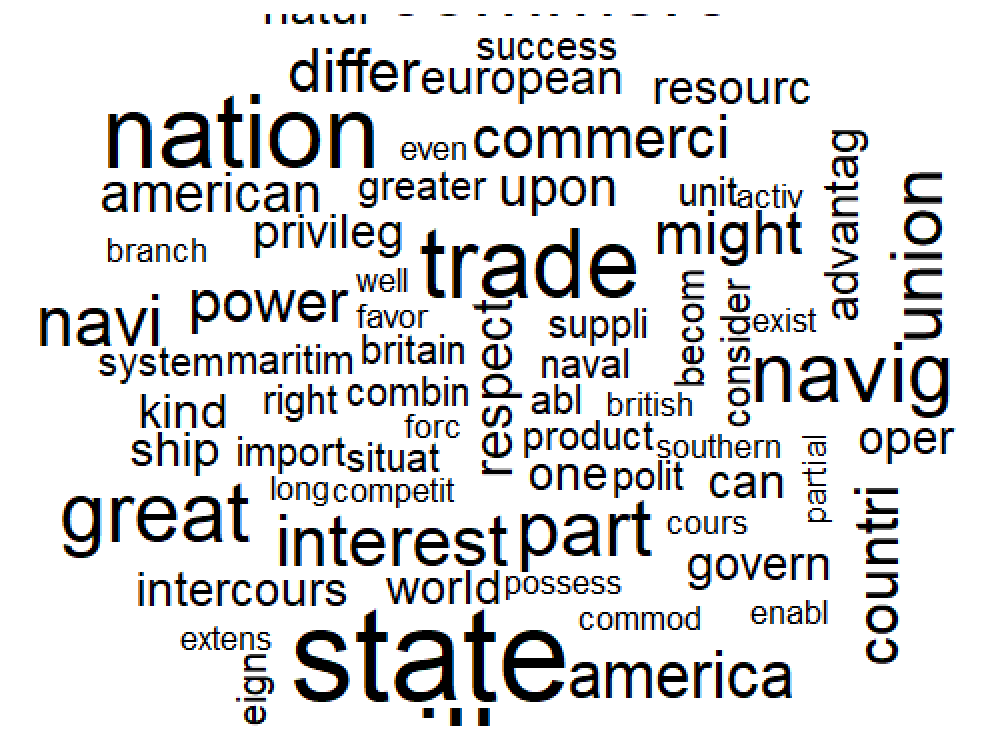
The data was next cleaned by reformatting it into a Document Term Matrix and removing punctuation, numbers and “stopwords”. Stop words are generally the most common words in a language, and add little value to being able to cluster and classify text.

The matrix was reviewed and the most frequent words were identified. This was done as a way to “sanity check” and ensure that the content of the matrix correctly included the original text files.



Normalizing the data is a critical next step. This ensures that longer documents are not given more preference or weight than shorter documents. It scales the values, allowing for an equal comparison. This was validated by comparing the original dataset to the normalized dataset and verifying the the new values were correctly calculated. For example, looking at the Row\_Sum\_Per\_doc the Hamilton\_Fed\_12 has a row sum of 1044. The value of the word "cultiv" in the original dataset is 3 and it is 0.003 in the normalized dataset. Since 3/1044 = 0.00287 (which rounds to 0.003), the calculation validates that the normalized data is correct.

Another way to visualize large bodies of text is via a word cloud. This provides directional information in a visual design about the most common words.



Lastly, a column for “author” was added (this will be the predictor variable) and test and training datasets were created. In order to test different scenarios, two different test/training sets were developed. The first set separated out the disputed texts into the test dataset, and the remaining texts were placed into the training dataset. The second set removed the disputed texts completely, and then split the remaining texts fairly equally between test and training (36 in test and 38 in training).

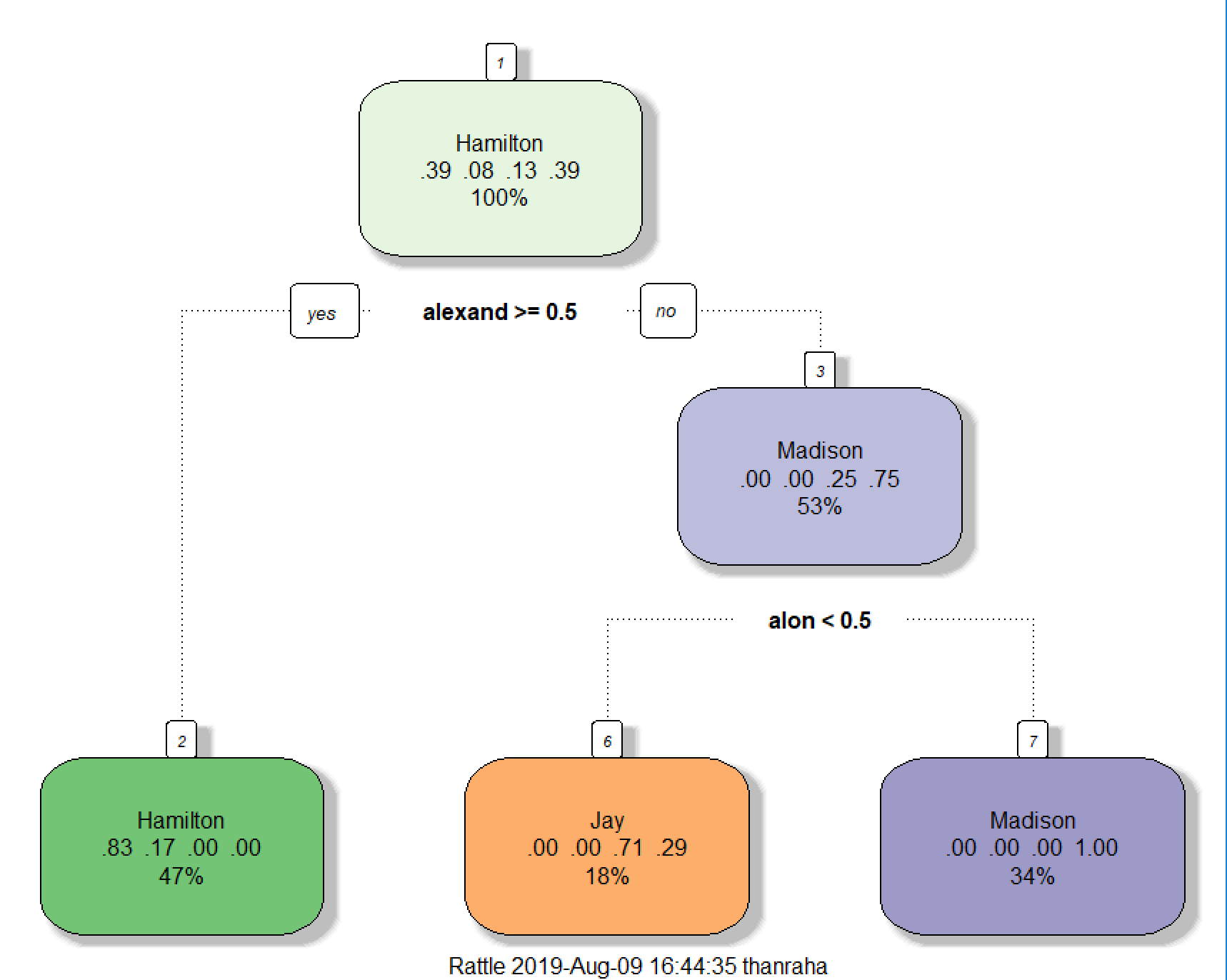
*About the models*

Decision trees were used to classify the data by author type: Hamilton, Madison, Jay or HM. The rpart program builds classification models that can be represented as binary trees. These models can then be used to predict the class label of unknown records. Once the training data correctly identifies texts from the four authors, it can then apply that to the disputed texts in order to determine authorship.

**Results**

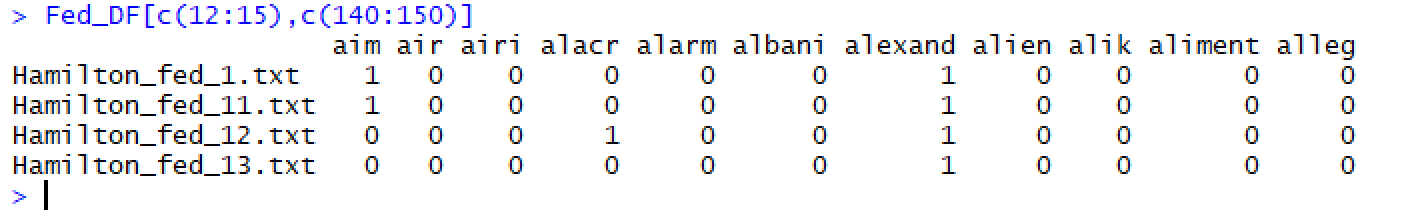
*Initial Decision Tree Results*

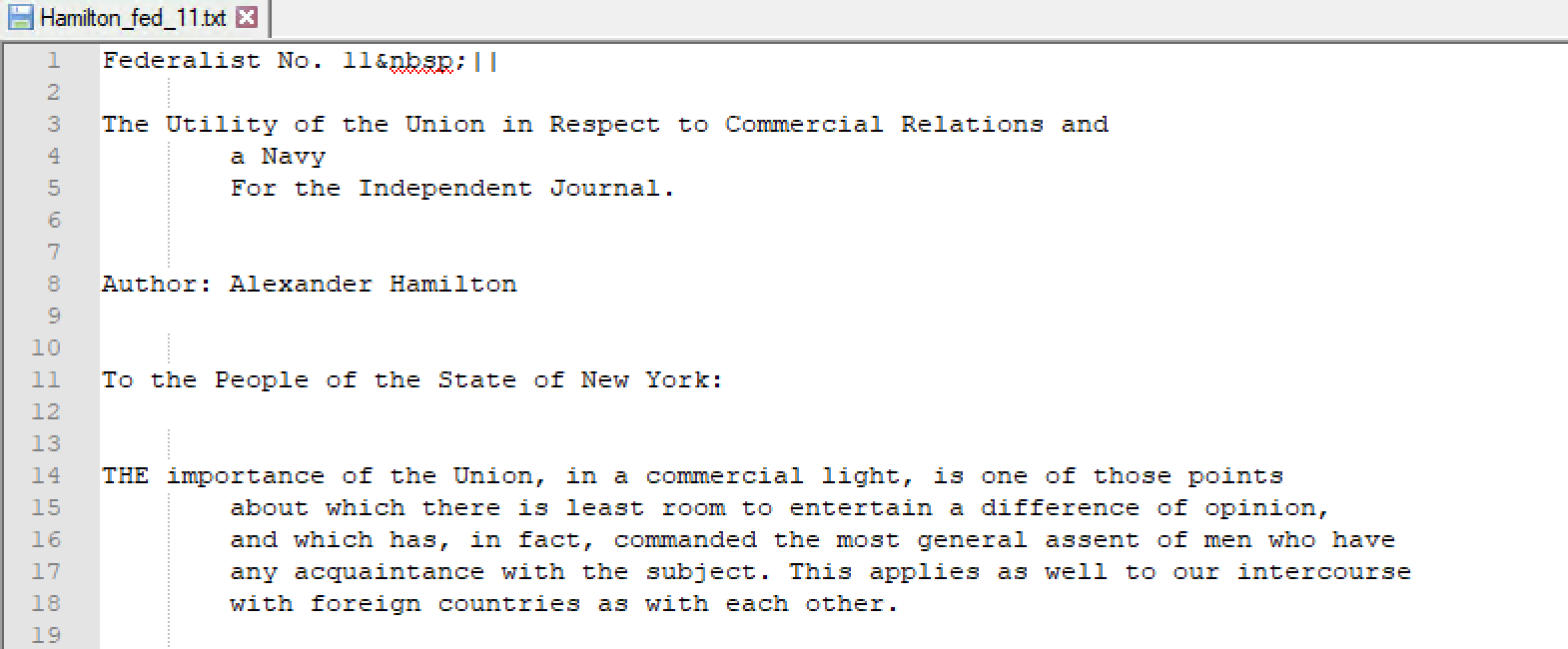
The initial results from the decision tree highlighted some unexpected results. The variables that were chosen as the primary splits appeared to contain the author names (or partial names). This was concerning since there was already a column established for the author name.



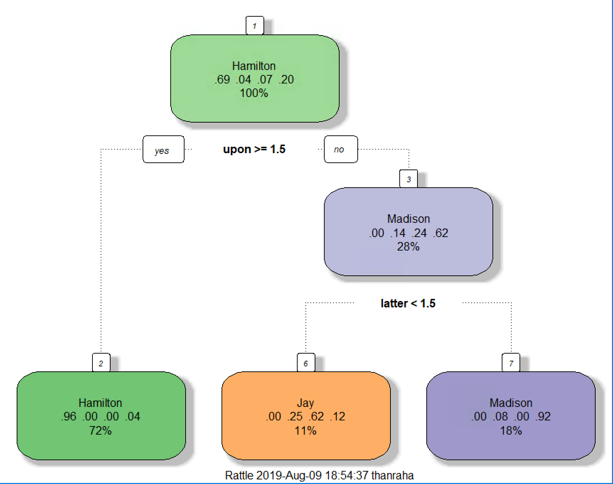
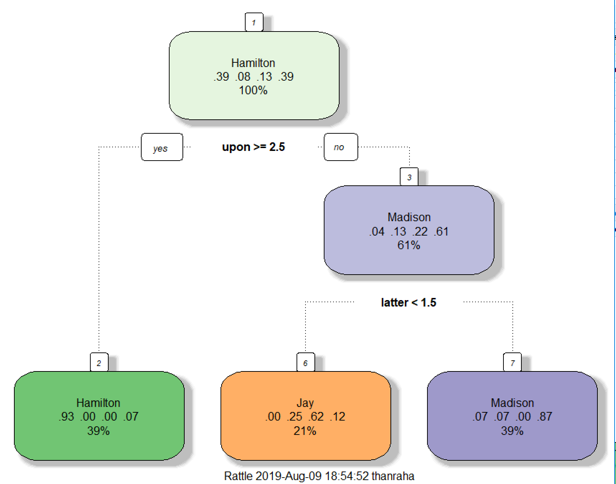
|  |  |  |
| --- | --- | --- |
| **## decision tree 1** |  | **## decision tree 2** |
| Primary splits: |  | Primary splits: |
| alexand < 0.5 to the right, improve=22.18468, (0 missing) |  | alexand < 0.5 to the right, improve=12.76316, (0 missing) |
| hamilton < 0.5 to the right, improve=22.18468, (0 missing) |  | hamilton < 0.5 to the right, improve=12.76316, (0 missing) |
| madison < 0.5 to the left, improve=21.24421, (0 missing) |  | jame < 0.5 to the left, improve=12.76316, (0 missing) |
| upon < 1.5 to the right, improve=20.16896, (0 missing) |  | madison < 0.5 to the left, improve=12.76316, (0 missing) |
| jame < 0.5 to the left, improve=19.62886, (0 missing) |  | upon < 2.5 to the right, improve=10.43997, (0 missing) |
| Surrogate splits: |  | Surrogate splits: |
| hamilton < 0.5 to the right, agree=1.000, adj=1.00, (0 split) |  | hamilton < 0.5 to the right, agree=1.000, adj=1.000, (0 split) |
| upon < 0.5 to the right, agree=0.932, adj=0.75, (0 split) |  | upon < 0.5 to the right, agree=0.868, adj=0.722, (0 split) |
| madison < 0.5 to the left, agree=0.892, adj=0.60, (0 split) |  | account < 0.5 to the right, agree=0.789, adj=0.556, (0 split) |
| jame < 0.5 to the left, agree=0.878, adj=0.55, (0 split) |  | jame < 0.5 to the left, agree=0.789, adj=0.556, (0 split) |
| although < 0.5 to the left, agree=0.865, adj=0.50, (0 split) |  | madison < 0.5 to the left, agree=0.789, adj=0.556, (0 split) |

Further inspection revealed that these author names were also included in the body of the texts. So, the data needed to be further cleaned before the decision trees could be run again.





After all the author names were removed from the data, the decision trees showed more reasonable results.

|  |  |  |
| --- | --- | --- |
| **## decision tree 1** |  | **## decision tree 2** |
| Primary splits: |  | Primary splits: |
| upon < 1.5 to the right, improve=20.168960, (0 missing) |  | upon < 2.5 to the right, improve=10.439970, (0 missing) |
| form < 6.5 to the left, improve= 8.634933, (0 missing) |  | edit < 0.5 to the right, improve= 5.529825, (0 missing) |
| whilst < 0.5 to the left, improve= 8.497094, (0 missing) |  | mclean < 0.5 to the right, improve= 5.529825, (0 missing) |
| although < 0.5 to the left, improve= 8.367480, (0 missing) |  | readili < 0.5 to the right, improve= 5.529825, (0 missing) |
| latter < 2.5 to the left, improve= 8.180023, (0 missing) |  | intend < 0.5 to the right, improve= 5.216020, (0 missing) |
| Surrogate splits: |  | Surrogate splits: |
| form < 4.5 to the left, agree=0.838, adj=0.429, (0 split) |  | intend < 0.5 to the right, agree=0.789, adj=0.467, (0 split) |
| wish < 0.5 to the left, agree=0.824, adj=0.381, (0 split) |  | man < 0.5 to the right, agree=0.789, adj=0.467, (0 split) |
| side < 2.5 to the left, agree=0.811, adj=0.333, (0 split) |  | matter < 0.5 to the right, agree=0.789, adj=0.467, (0 split) |
| whilst < 0.5 to the left, agree=0.811, adj=0.333, (0 split) |  | thing < 1.5 to the right, agree=0.789, adj=0.467, (0 split) |
| although < 0.5 to the left, agree=0.797, adj=0.286, (0 split) |  | agre < 0.5 to the right, agree=0.763, adj=0.400, (0 split) |

Interestingly, the word clouds changed considerably after the author names were removed. This may indicate that author names were frequent enough to impact the distribution of other words.

*Word cloud* ***before*** *removing author names from text Word cloud* ***after*** *removing author names from text*

After ensuring that the decision tree predicted the correct author in the training dataset with reasonable accuracy, the decision tree was run again on the test data set. The tree model predicted that Hamilton wrote one text, Jay wrote 6 and Madison wrote 4.

|  |  |
| --- | --- |
| **Predicted Authors of Disputed Texts** | |
| **Author** | **Text** |
| Hamilton | 1 |
| HM | 0 |
| Jay | 6 |
| Madison | 4 |

**Conclusions**

Decision trees can be a powerful classification tool, but can produce unexpected results if the data is not properly cleaned. And since no “error” is produced, it can be very easy to deliver incorrect results inadvertently. When the author names were inadvertently left in the body of the text during the first part of this analysis, the tree that was produced looked reasonable on the surface. It was only through detailed review that it became clear more data cleansing was necessary.

After the successful second run, the authors were able to be identified with reasonable accuracy. The test dataset, with only the disputed authors, was then run through the same decision tree in order to predict who wrote these papers. Of the eleven disputed papers, it was predicted that Hamilton wrote one of them, Jay wrote six and Madison wrote four.

However, these results should continue to be tested. Of the undisputed papers where authorship was known, Jay wrote far fewer than either Hamilton or Madison, so it seems unusual that he would be the author of a larger quantity of the disputed papers. But with such a small sample size, that could be plausible.

To ensure confidence in the results, and as a general best practice, it is important to perform multiple runs and test a variety of scenarios. In addition, combining several different techniques can also result in better outcomes. Understanding the data and analyzing the structure and contents are a key part of determining the best clustering approach.