Introduction

In the past five years, data science expert witnesses have become highly valuable for a wide range of legal cases. Data scientists have applied expertise to everything from DNA analysis to authorship attribution.

One of the first published papers on authorship attribution related to a set of documents created 232 years ago, on this very week on October 27th,1887. The Federalist Papers is a collection of 85 articles and essays written by Alexander Hamilton, James Madison, and John Jay under the pseudonym "Publius" to promote the ratification of the United States Constitution. John Jay wrote five papers, while Alexander Hamilton and James Madison wrote the remaining 80. Between the last two authors, there are conflicting accounts of which author wrote which paper. Most sources agree on the authorships of 65 papers (51 by Hamilton and 14 by Madison), but 15 papers are in dispute.

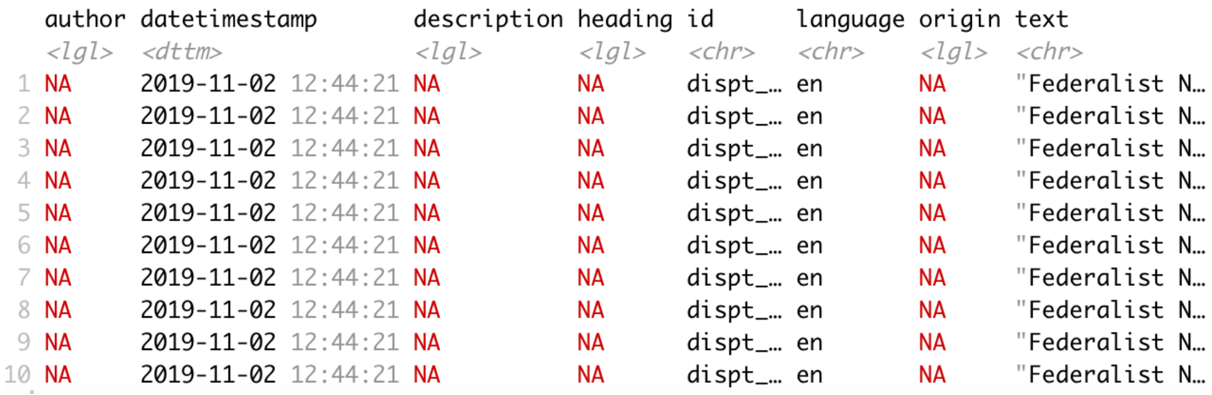
In one of the earliest examples of statistical text analysis, F. Mosteller and D. L. Wallace used a form of Naive Bayes classification to identify the authorship of the 15 disputed papers.

In honor of the anniversary, this paper will look to perform a similar analysis.

About the dataset

This dataset consists of 85 papers. For the purposes of this analysis, the papers written by John Jay and jointly by Madison and Hamilton were be removed.

When imported as a corpus, the dataset consisted of 2 primary elements, an id, which was the name of the document, and the text of the document. A corpus may also contain useful metadata such as author, language, date-time stamp, and heading. However, those elements were NULL for this Fed Paper dataset.



```
   author datetimestamp       description heading id      language origin text
   <lgl>  <dttm>              <lgl>       <lgl>   <chr>   <chr>    <lgl>  <chr>
 1 NA     2019-11-02 12:44:21 NA          NA      dispt_… en       NA     "Federalist N…
 2 NA     2019-11-02 12:44:21 NA          NA      dispt_… en       NA     "Federalist N…
 3 NA     2019-11-02 12:44:21 NA          NA      dispt_… en       NA     "Federalist N…
 4 NA     2019-11-02 12:44:21 NA          NA      dispt_… en       NA     "Federalist N…
 5 NA     2019-11-02 12:44:21 NA          NA      dispt_… en       NA     "Federalist N…
 6 NA     2019-11-02 12:44:21 NA          NA      dispt_… en       NA     "Federalist N…
 7 NA     2019-11-02 12:44:21 NA          NA      dispt_… en       NA     "Federalist N…
 8 NA     2019-11-02 12:44:21 NA          NA      dispt_… en       NA     "Federalist N…
 9 NA     2019-11-02 12:44:21 NA          NA      dispt_… en       NA     "Federalist N…
10 NA     2019-11-02 12:44:21 NA          NA      dispt_… en       NA     "Federalist N…
```

Figure 1: Tidy(corpus)

Data Pre-processing

Prep-processing the dataset focuses on rearranging the data so that it can be worked on it more easily. Tasks such as identifying the sparsity of the dataset and checking the amount of missing data were performed.

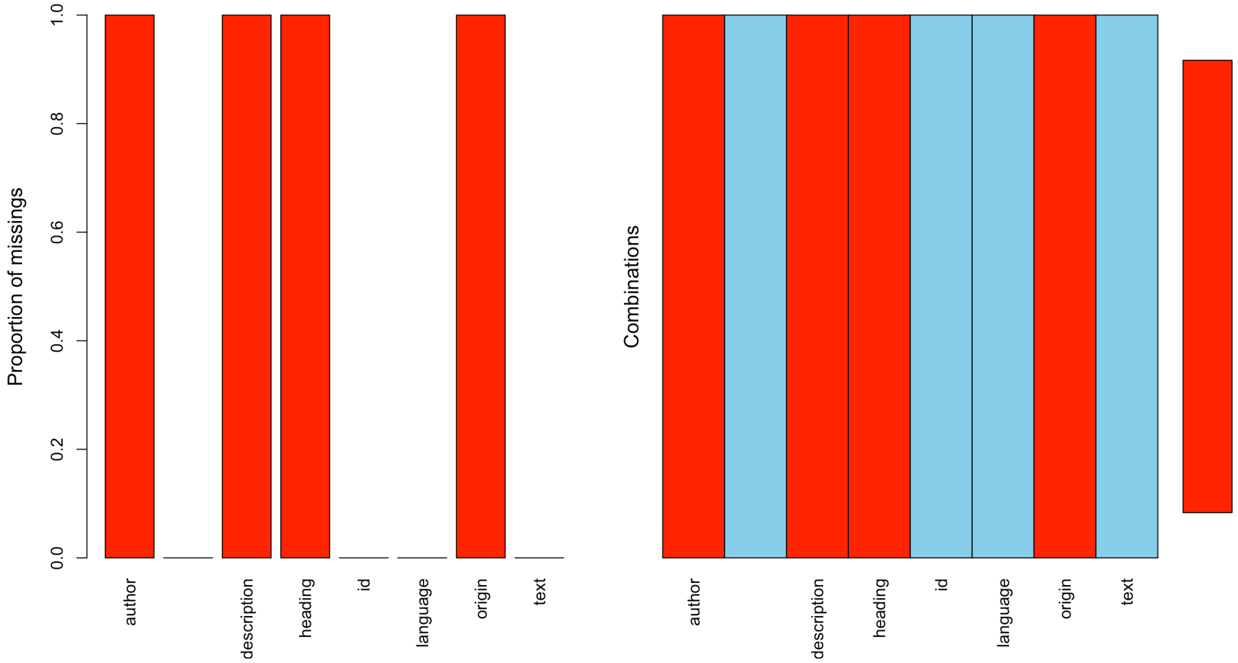


Fig 2: check for missing data, Red represents missing data

Using the text mining package, the corpus was cleaned to remove special characters, punctuation, and removing numbers.

Remove Stop Words

Common words (like "a" "and" "the," for example) were removed, as they serve to muddy the dataset and will not help to identify the author of the disputed text.

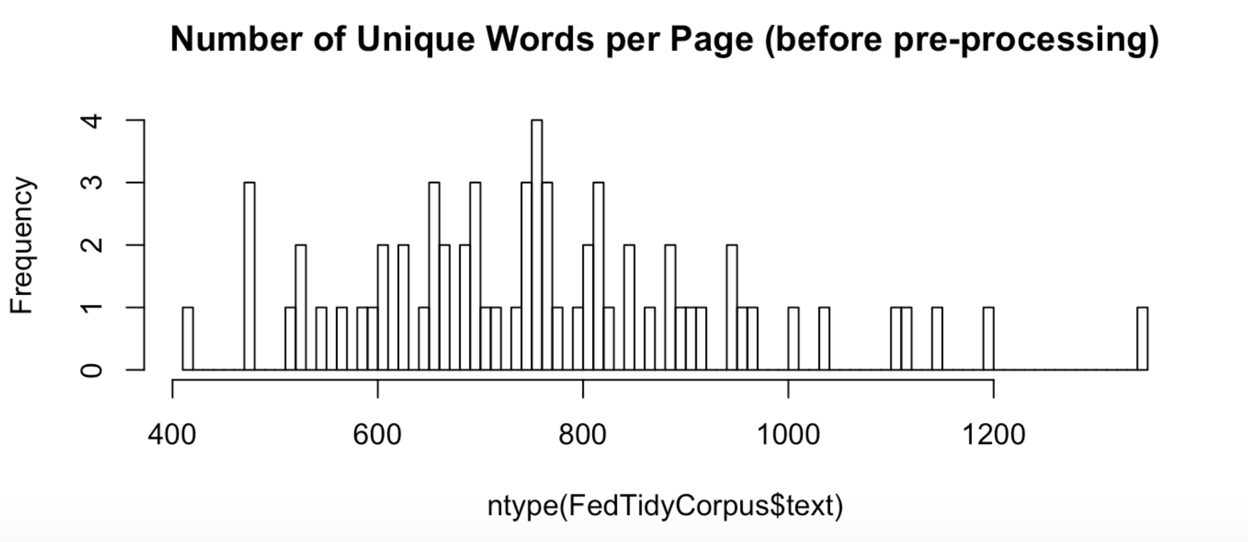


Figure 3: Unique words per page

Stem
Stemming is the process of converting the words of a sentence to its non-changing portions. For example, amusing, amusement, and amused above, the stem would be amus.

Remove Additional Stop Words
Word frequencies were reviewed again, and additional stop words were removed.

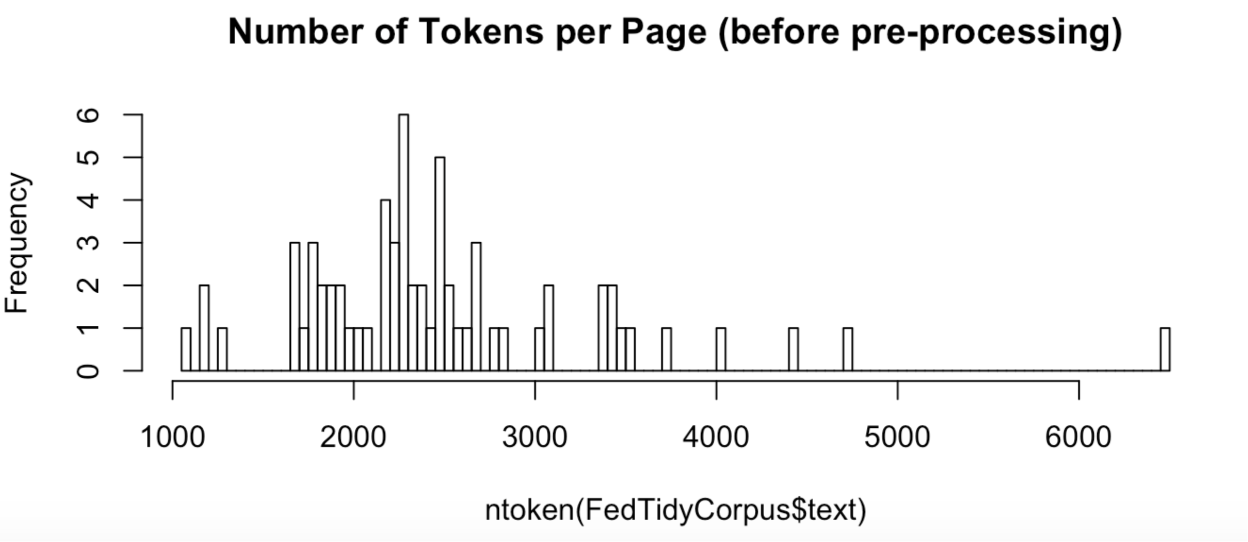
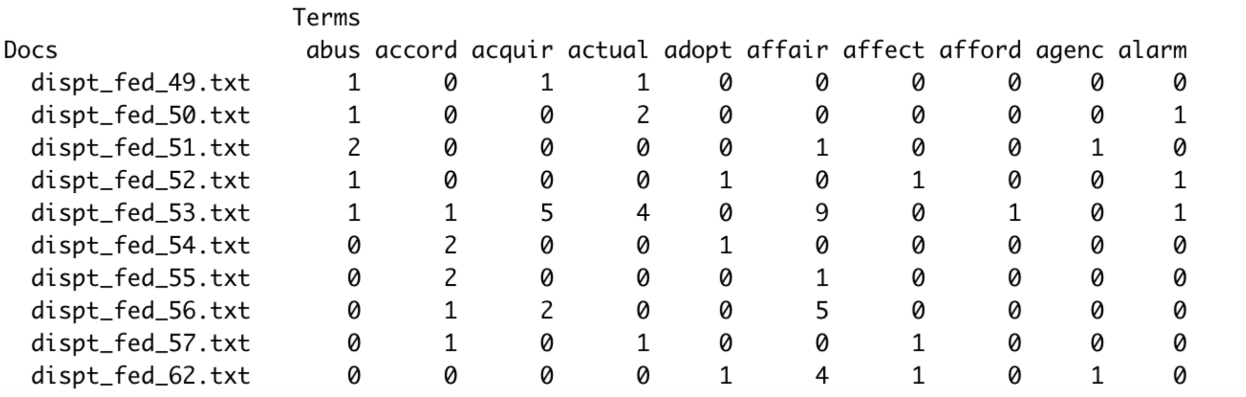


Figure 4: Word frequencies histogram

Create Document Term Matrix

Finally, a document-term matrix (DTM) was generated. The matrix lists all occurrences of words appearing in the corpus. In a DTM, documents are represented by rows and the terms (or words) by columns. If a word occurs in a particular document n times, then the matrix entry for corresponding to that row and column is n, if it doesn't occur at all, the entry is 0.



```
                     Terms
Docs                  abus accord acquir actual adopt affair affect afford agenc alarm
  dispt_fed_49.txt       1      0      1      1     0      0      0      0     0     0
  dispt_fed_50.txt       1      0      0      2     0      0      0      0     0     1
  dispt_fed_51.txt       2      0      0      0     0      1      0      0     1     0
  dispt_fed_52.txt       1      0      0      0     1      0      1      0     0     1
  dispt_fed_53.txt       1      1      5      4     0      9      0      1     0     1
  dispt_fed_54.txt       0      2      0      0     1      0      0      0     0     0
  dispt_fed_55.txt       0      2      0      0     0      1      0      0     0     0
  dispt_fed_56.txt       0      1      2      0     0      5      0      0     0     0
  dispt_fed_57.txt       0      1      0      1     0      0      1      0     0     0
  dispt_fed_62.txt       0      0      0      0     1      4      1      0     1     0
```

Figure 5: Document-term Matrix

Modeling

Supervised Learning

Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs. It infers a function from labeled training data consisting of a set of training examples.

The goal of this analysis will be to employ the supervise learning technique Decision Trees to see if the results can be more definitive.

Decision Trees

Decision tree learning is a method commonly used in data mining. The goal is to create a model that predicts the value of a target variable based on several input variables. What makes decision trees useful is its clarity of information representation. The "knowledge" learned by a decision tree through training is directly formulated into a hierarchical structure. This structure holds and displays the knowledge in such a way that it can easily be understood, even by non-experts.

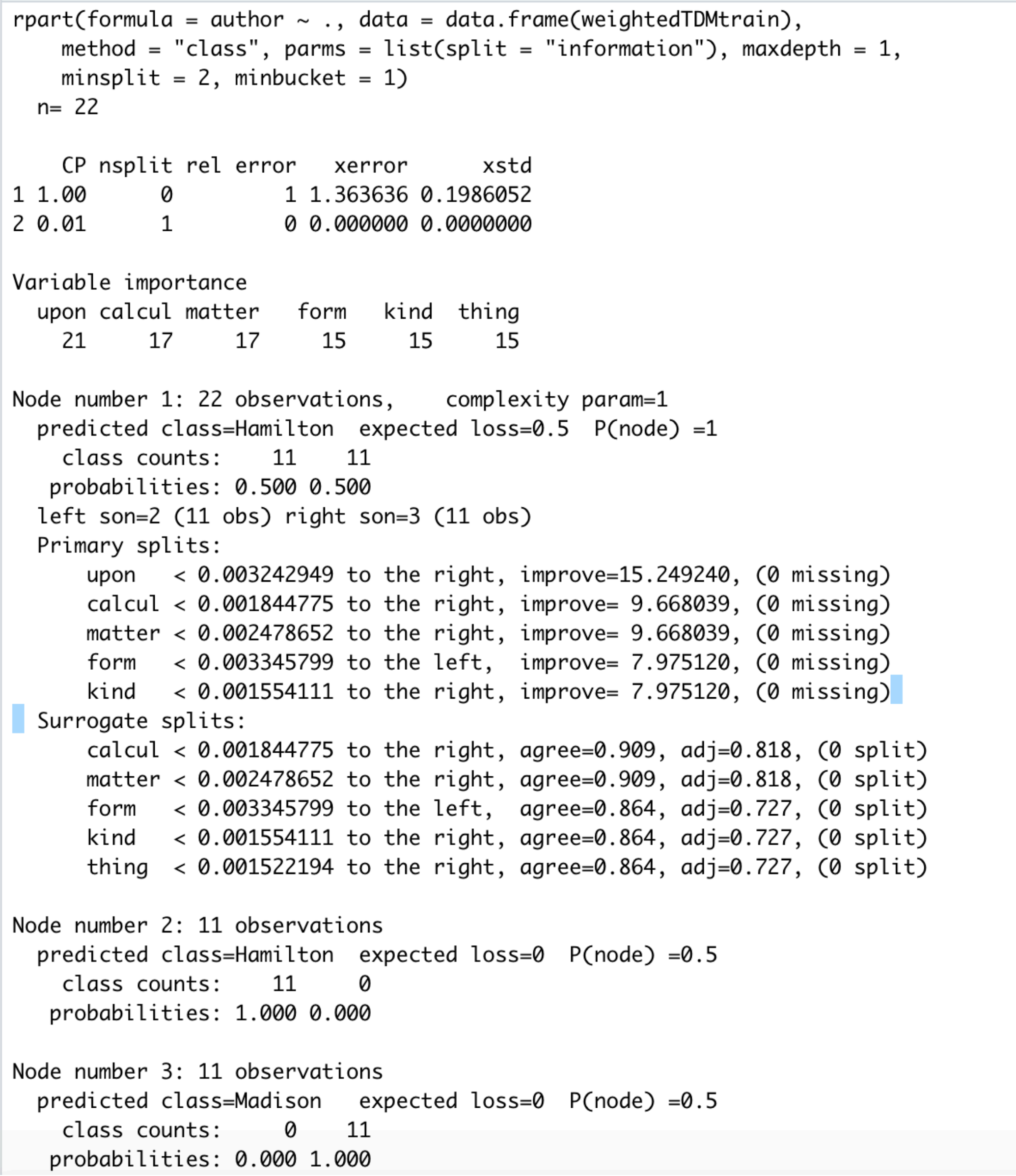
Key Steps:

1. Identify the class label. The class label is the value the decision tree is looking to predict. For this analysis, the class label will be the author of the disputed papers. Either Hamilton or Madison.

2. Split the data into two sets, one for training the data and another for testing the model. The data was split based on the paper's author. The training set contained papers written by Hamilton and Madison, and the test set contained only the disputed papers.

3. Train the model using the training dataset
4. Predict the class label using the test dataset and model generated in step 3.
5. Train/Test the corpus using both the generated term-document matrix and a weighted term-document matrix (WtIDF - Weight a term-document matrix by term frequency-inversese document frequency).
6. Validate both models using 10-k fold cross validation.

Training model: Summary(fit) – using weighted term-document matrix



```
rpart(formula = author ~ ., data = data.frame(weightedTDMtrain),
    method = "class", parms = list(split = "information"), maxdepth = 1,
    minsplit = 2, minbucket = 1)
  n= 22

    CP nsplit rel error   xerror      xstd
1 1.00      0         1 1.363636 0.1986052
2 0.01      1         0 0.000000 0.0000000

Variable importance
  upon calcul matter   form   kind  thing
    21     17     17     15     15     15

Node number 1: 22 observations,    complexity param=1
  predicted class=Hamilton  expected loss=0.5  P(node) =1
    class counts:    11    11
   probabilities: 0.500 0.500
  left son=2 (11 obs) right son=3 (11 obs)
  Primary splits:
      upon   < 0.003242949 to the right, improve=15.249240, (0 missing)
      calcul < 0.001844775 to the right, improve= 9.668039, (0 missing)
      matter < 0.002478652 to the right, improve= 9.668039, (0 missing)
      form   < 0.003345799 to the left,  improve= 7.975120, (0 missing)
      kind   < 0.001554111 to the right, improve= 7.975120, (0 missing)
  Surrogate splits:
      calcul < 0.001844775 to the right, agree=0.909, adj=0.818, (0 split)
      matter < 0.002478652 to the right, agree=0.909, adj=0.818, (0 split)
      form   < 0.003345799 to the left,  agree=0.864, adj=0.727, (0 split)
      kind   < 0.001554111 to the right, agree=0.864, adj=0.727, (0 split)
      thing  < 0.001522194 to the right, agree=0.864, adj=0.727, (0 split)

Node number 2: 11 observations
  predicted class=Hamilton  expected loss=0  P(node) =0.5
    class counts:    11     0
   probabilities: 1.000 0.000

Node number 3: 11 observations
  predicted class=Madison   expected loss=0  P(node) =0.5
    class counts:     0    11
   probabilities: 0.000 1.000
```

Fig 6: Summary(fit) – weighted term-document matrix

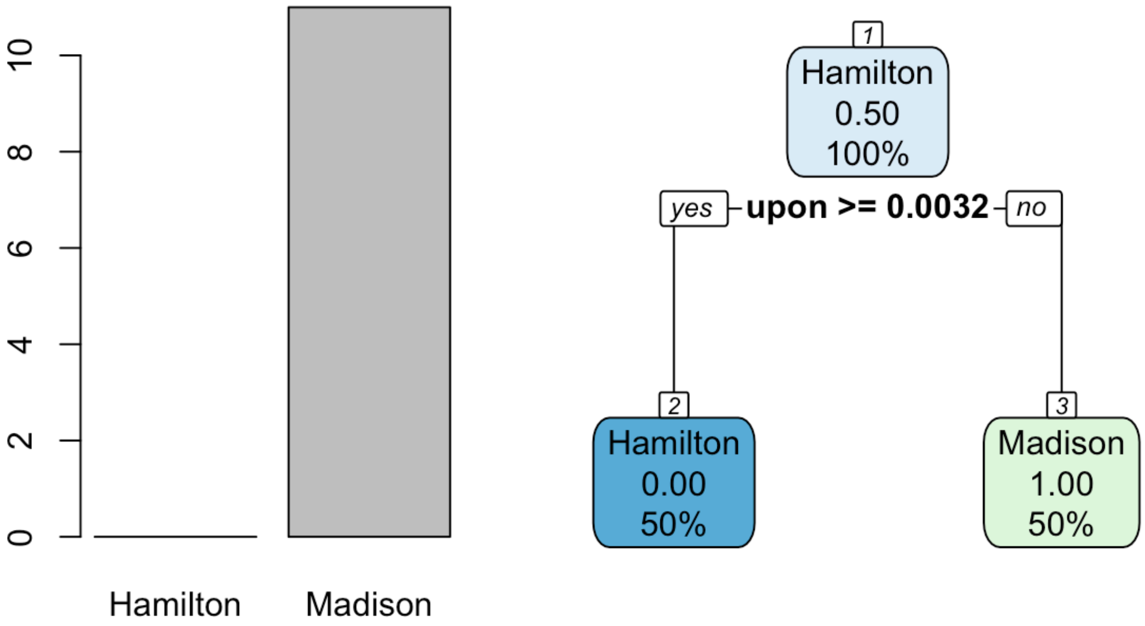


Figure 7: Plot of the weighted model

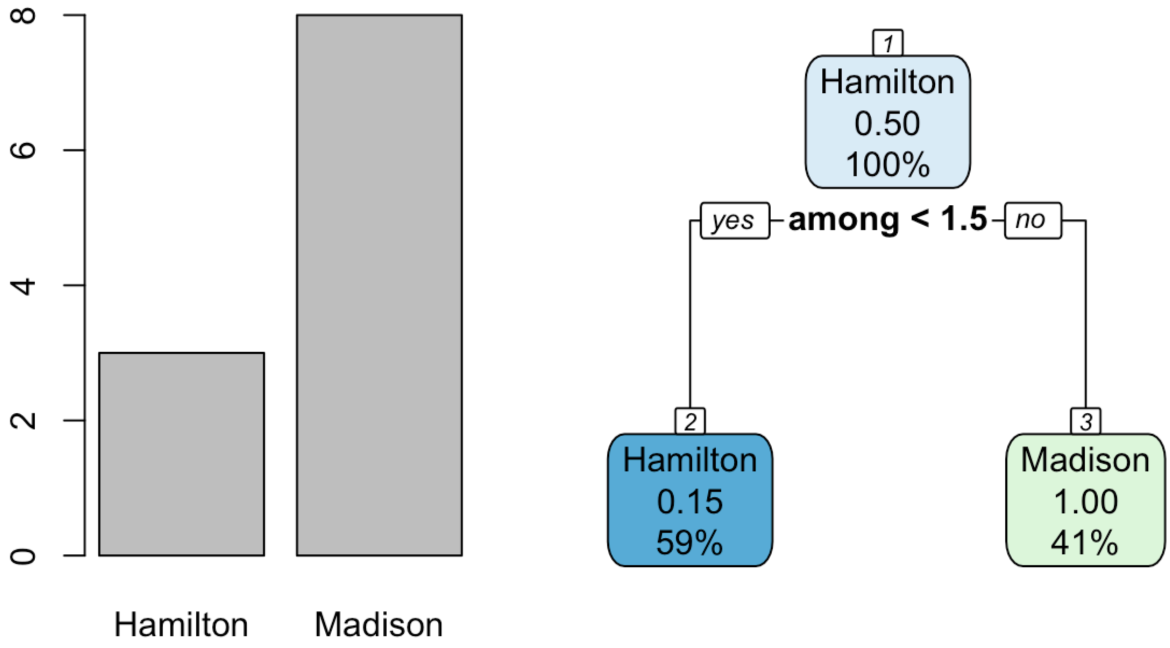
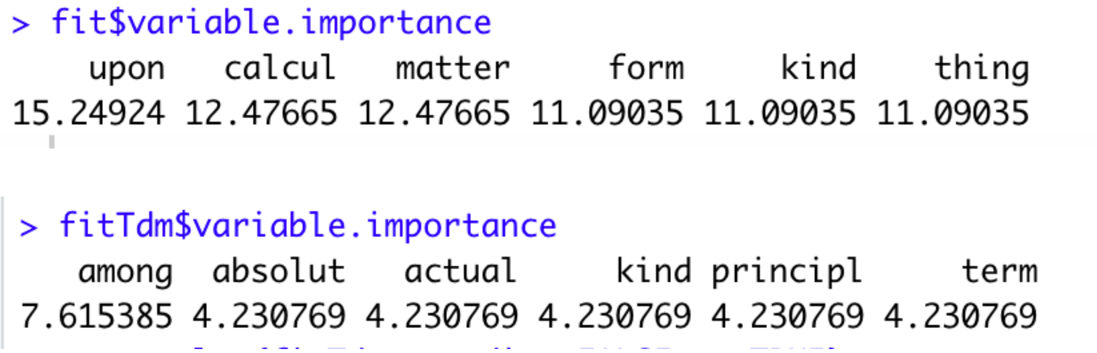


Figure 8: Plot of normal term-document matrix model

Variable Importance:

Variable Importance (VI) represents the statistical significance of each variable in the data concerning its effect on the generated model. VI is each predictor ranking based on the contribution predictors make to the model. Based upon the training and testing model used the term significance varied



```
> fit$variable.importance
    upon    calcul   matter     form     kind    thing 
15.24924 12.47665 12.47665 11.09035 11.09035 11.09035 
```

```
> fitTdm$variable.importance
   among  absolut   actual     kind principl     term 
7.615385 4.230769 4.230769 4.230769 4.230769 4.230769 
```

Figure 9: Variable Importance

Prediction

As illustrated in figure 10, the results very much depended on the term frequency.

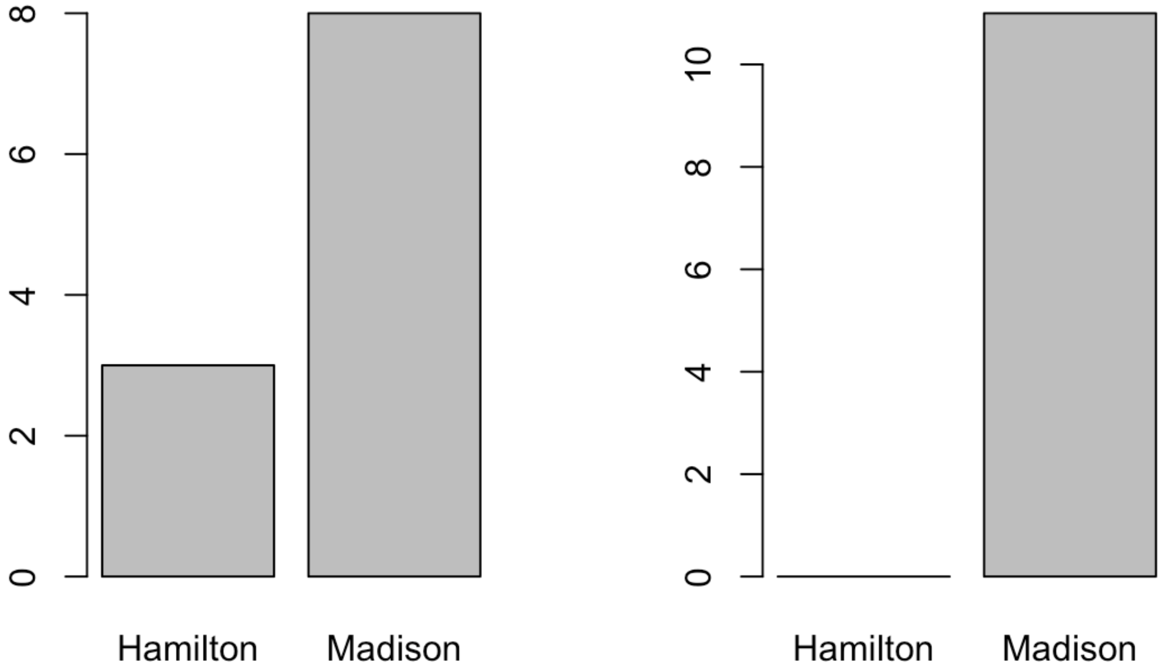


Figure 9: Plot of disputed papers prediction models

Results

Prior Analysis – Unsupervised Learning

Previously, this dataset was analyzed using the following unsupervised methods:

- Kmeans
- Hierarchical Clustering
- Pairwise - Matching using Burrow's Delta

Of the three approaches, Burrow's Delta provided the most definitive results as to who the document other authors may be. As noted in Figure 5 below, the clusters for Madison and the disputed papers are closer than Hamilton's clusters.

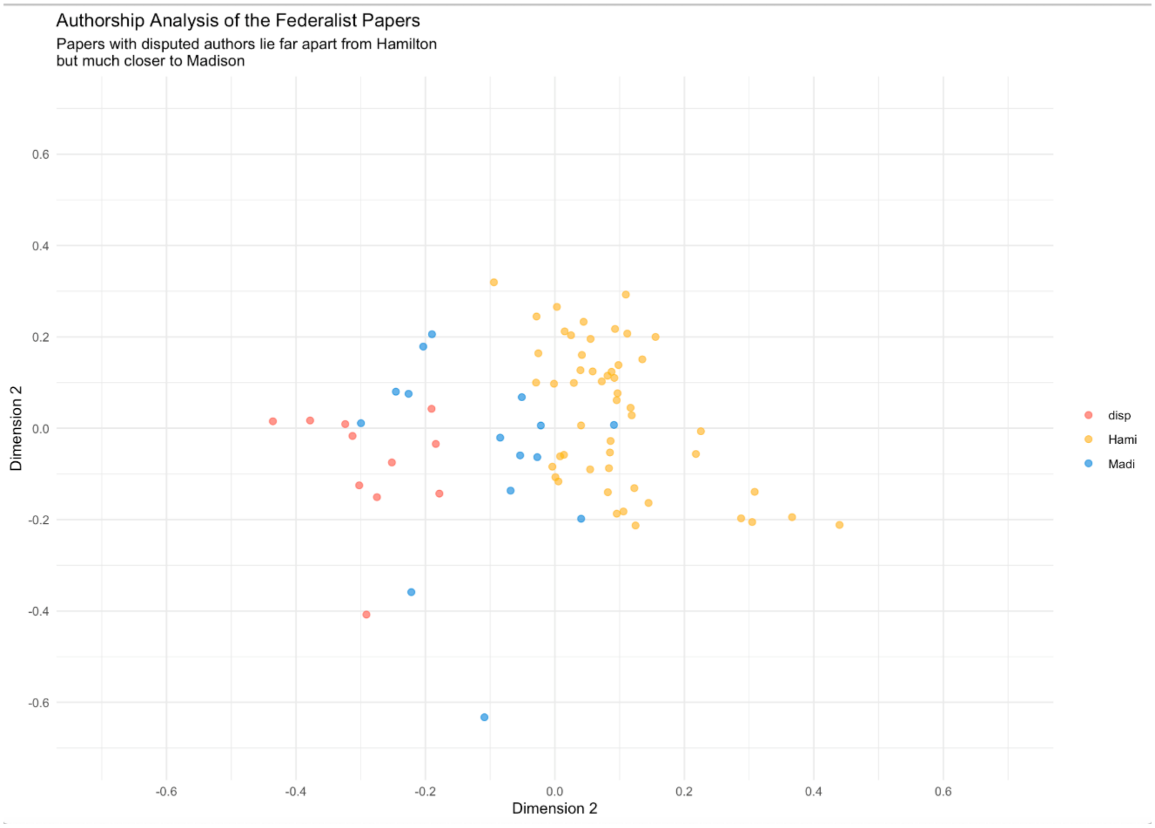


Fig 5: Burrow’s Delta Clusters

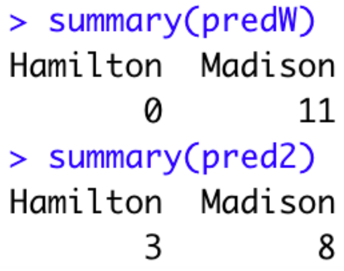
Validating the Results using K-Fold Validation

Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample.

The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation. When a specific value for k is chosen, it may be used in place of k in reference to the model, such as k=10 becoming 10-fold cross-validation.

Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. That is, to use a limited sample in order to estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model.

The results of the k-fold validation mirrored the analysis.



```
> summary(predW)
Hamilton  Madison 
       0       11 
> summary(pred2)
Hamilton  Madison 
       3        8 
```

Figure 11: Output of the 10k fold prediction

Stopwords

While the results of all four modeling techniques point to Madison as being the primary author. The results can easily be manipulated by included or excluded certain stop words when the document matrix is generated.

Goodness of Split

The decision tree uses entropy and information gain to select a feature that gives the best split. Generally, the two impurity measures (Gini index and entropy) usually yield similar results. This proved to be true for this analysis as well. Using the CoreLEARN package, confirmed the word "Upon" was the best word to split on as the Gini Index and information gain measured to 1.

Conclusion

The goal of this paper was to analyze the Federalist Papers, to determine if it is possible to identify who wrote the disputed papers. Previous methods using clustering did not provide a definitive answer. Based upon the initial analysis, it appears it may have been co-authored or mostly written by Madison as suggested by F. Mosteller and D. L. Wallace.. However, using decision trees, the primary author of the disputed papers appears to be Madison.