IST 707: Homework 5

**Introduction**

Although the United States (US) famously declared its independence in 1776, at that time it was merely a collection of 13 colonies. These colonies did not officially become part of the US until they chose to adopt the proposed US Constitution, which was written over a decade later in 1787 (National Constitution Center). The Constitution is one of the most famous written documents in history, however, it was highly controversial when it was first proposed.

Those in favor of ratifying the Constitution were called Federalist and those in opposition were the Anti-Federalist (National Constitution Center). Proponents from each party often plead their case with their respective colony by publishing essays in local newspapers (About the Federalist Papers). Perhaps the most famous collection of such essays is the Federalist Papers—a collection of 85 essays published in New York’s newspapers under the pen name “Publius” persuading New Yorkers to ratify the Constitution (About the Federalist Papers).

Today, “Publius” is believed to have been the literary persona of a collection of famed Federalist supporters (About the Federalist Papers). Historians attribute 74 of the essays’ authorship to a combination of Alexander Hamilton (51 sole-author, 3 co-author), John Jay (5 sole-author), and James Madison (15 sole-author, 3 co-author) (The Federalist Papers). However, the exact authorship of the remaining 11 essays is still widely disputed. The Library of Congress has elected to sponsor a series of studies in order to determine the most likely author of these disputed essays.

The first study in this series attributed the authorship of 3 (#49, #50, and #63) of the disputed essays to Alexander Hamilton. However, it was unable to definitively suggest the authors of the remaining disputed papers. The Library of Congress hopes that this second study will confirm/improve upon the previous findings.

**Analysis and Models**

**About the Data**

The original 85 Federalist Paper essays were used for this study. Each essay includes the body of text in addition to the author (if known). For the essays with a disputed author, “Disputed” was used as the author.

These documents were compiled as a corpus of text files with the primary indicator of authorship being the filenames. The nomenclature for the original filenames was as follows:

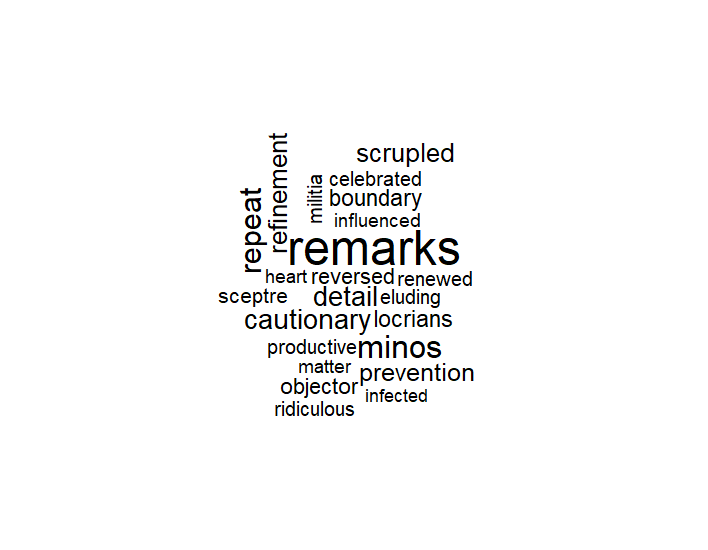
*[ Author ]\_[ fed] \_[ paper # ].txt :* (i.e. Hamilton\_fed\_1.txt)

In order to improve code readability and future visualizations, these filenames were modified to reflect only the author and paper # (i.e. Hamilton\_1).

The corpus of Federalist Papers was compiled into a document term matrix for further analysis. To eliminate “noise” within the data, stopwords were removed along with all punctuation, numbers, and separators. In order to further eliminate noise, all words included in the authors’ names were removed (i.e. “Alexander”,”Hamilton”, etc.). Additionally, all words were made into lower case, and limited to only those between 4 and 15 characters in length. The remaining data consisted of rows with filename labels and the frequencies of words occurring in the respective Federalist Paper.

Figure 1 (below) is a word cloud of the top 30 most frequent words among the 11 disputed Federalist Papers.

**Figure 1: Disputed Essays Word Cloud**



Initial observation of the data showed that the essays varied in length, so to support comparison of the document data within the document term matrix, each document’s word frequency was divided by the document’s total word count. This produced normalized frequencies of words across documents to eliminate document length bias in future analysis.

For the purpose of enabling future analysis, the aforementioned document term matrix was compiled as a data frame. Once in this format, the document names were separated into “Author” and “PaperNo” columns (example below).

Before: Hamilton\_1 : After: Author = Hamilton, PaperNo = 1

With the Author and PaperNo distinctions, the following subsets of data were created for later modeling:

|  |  |
| --- | --- |
| Subset Name | Subset Description |
| Disputed Papers | All data from Federalist Papers Data Frame where  Author = “Disputed” |
| Undisputed Papers | All data except for PaperNo\*\* from Federalist Papers Data Frame where Author <> “Disputed” |
| Training Data | Random sample without replacement of 2/3 of the Undisputed Papers subset |
| Test Data | Remaining 1/3 of the Undisputed Papers subset after Training Data sampling |
| *\*\*PaperNo excluded to prevent “noise” in word frequency-based predictions* | |

**Decision Tree Models**

**Unbound Decision Tree Model**

Initially, an unbound decision tree model was applied to the Federalist Papers data. This model used the Training Data subset to inform its decision tree creation and the Test Data to check its prediction accuracy. This model was given no parameters for restricting its decision tree creation. This model attempts to make splits based on word frequency to determine the most likely author of a paper. Once trained and tested, this model will be used to predict the authorship of the disputed papers.

**Pruned Decision Tree Model**

In addition to an unbound model, a pruned decision tree model was used to examine the essays. This model was also trained and tested on the Training Data and Test Data subsets. Unlike the unbound model, this model uses a complexity parameter to prune its decision trees. For the purpose of this study, .03 was used which causes the model to prune all splits that do not improve the overall R-squared value of the model by at least .03. This model will also be applied to the disputed papers once trained and tested.

**Parameterized-Pruned Decision Tree Model**

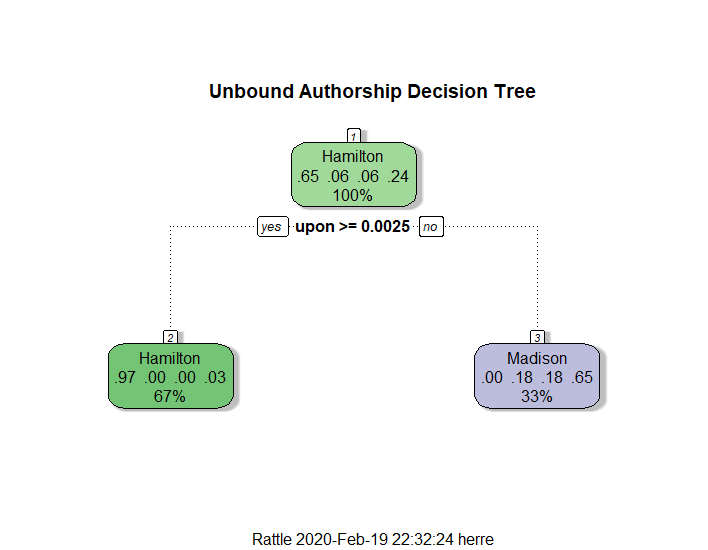
Furthermore, a parameterized-pruned decision tree model was used. This model likewise uses the Training Data and Test Data subsets to inform and test its decision tree creation. This model ingests a complexity parameter like the pruned model; however, it also has parameters for minimum number of splits and maximum depth. The minimum split determines the least amount of splits the decision tree must have, and the maximum depth limits the depth of each node in the model. For the purpose of this study: complexity parameter = .07, minimum split = 2, maximum depth = 5. Again, this model will be used to predict the disputed paper authorship once it has been trained and tested.

**Results**

**Unbound Decision Tree Model Results**

The resulting decision tree from the unbound decision tree model can be seen below (Figure 2). This decision tree classifies the authorship of essays as Hamilton or Madison with the majority (67%) of classifications being Alexander Hamilton.

**Figure 2: Unbound Decision Tree**



This model used 3 nodes to decide the most important words in determining authorship were: upon, composing, several, although, sovereigns, and wish. The unbound model had a train error rate of 35.29% and test error rate of 8.70%. The confusion matrix from using the model to predict the Test Data is:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Actual Author(s)** | | | |
| **Predicted Author(s)** | Hamilton | Hamilton & Madison | Jay | Madison |
| Hamilton | 18 | 0 | 0 | 0 |
| Hamilton & Madison | 0 | 0 | 0 | 0 |
| Jay | 0 | 0 | 0 | 2 |
| Madison | 0 | 0 | 0 | 3 |

This model appears to be quite accurate in predicting authorship with only two misclassifications of Jay-authored papers as Madison-authored papers. The results of predicted the disputed authorship using this model were:

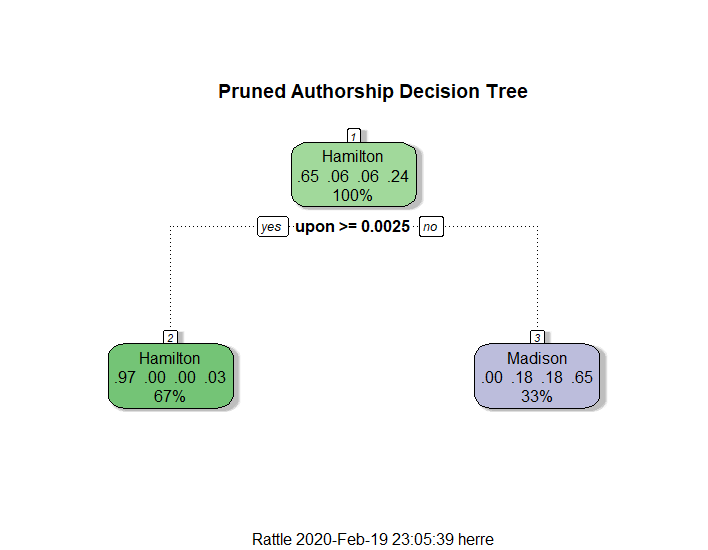
|  |  |
| --- | --- |
| **Disputed Paper** | **Predicted Author** |
| Fed Paper 49 | Madison |
| Fed Paper 50 | Madison |
| Fed Paper 51 | Madison |
| Fed Paper 52 | Madison |
| Fed Paper 53 | Madison |
| Fed Paper 54 | Hamilton |
| Fed Paper 55 | Madison |
| Fed Paper 56 | Madison |
| Fed Paper 57 | Madison |
| Fed Paper 62 | Madison |
| Fed Paper 63 | Madison |

These results contradict previous study outcomes attributing definitive authorship of Federalist Papers #49, #50, and #63 to Alexander Hamilton. Aside from #54 (Hamilton) all other disputed papers are predicted to have been authored by James Madison according to this model.

**Pruned Decision Tree Model Results**

The resulting decision tree from the pruned decision tree model can be seen below (Figure 3). This decision tree also classifies the authorship of essays as Hamilton or Madison with the majority (67%—same as unbound model) of classifications being Alexander Hamilton.

**Figure 3: Pruned Decision Tree**



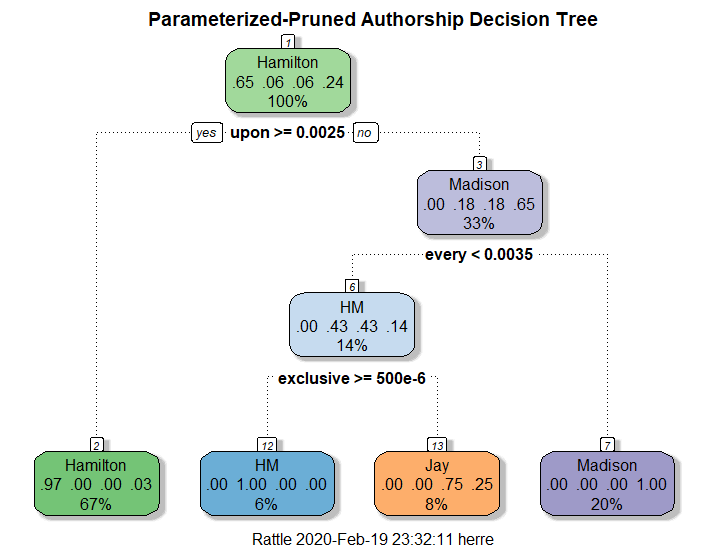
This model resulted in the same important words in determining authorship, test and error rates, confusion matrix, and disputed paper author predictions as the unbound model—indicating that the pruning may not have impacted the model.

Increasing the complexity parameter to prune more from a decision tree of this size may not provide meaningful insights (parameterized-pruned model will use slightly higher complexity parameter), so the results of this model are mere redundancies (not confirmations) of the unbound model results.

**Parameterized-Pruned Decision Tree Model Results**

The resulting decision tree from the unbound decision tree model can be seen below (Figure 4). This decision tree classifies the authorship of essays with splits accounting for all four authorship potentials, the majority of which are still attributed to Hamilton.

**Figure 4: Parameterized-Pruned Decision Tree**



This model determined the most important words in signaling authorship as: upon, composing, several, although, sovereigns, wish, every, often, best, objection, real, state, exclusive, kept, objects, plan, powers, and therefore. This model had the same train error rate as the unbound model (35.29%) However, it had an improved train error rate of only 4.35% (see confusion matrix below).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Actual Author(s)** | | | |
| **Predicted Author(s)** | Hamilton | Hamilton & Madison | Jay | Madison |
| Hamilton | 18 | 0 | 0 | 0 |
| Hamilton & Madison | 0 | 0 | 0 | 0 |
| Jay | 0 | 0 | 2 | 0 |
| Madison | 0 | 0 | 1 | 2 |

With seemingly improved accuracy compared to the unbound model (only one misclassification), the results of predicted the disputed authorship using this model were:

|  |  |
| --- | --- |
| **Disputed Paper** | **Predicted Author** |
| Fed Paper 49 | Madison |
| Fed Paper 50 | Madison |
| Fed Paper 51 | Madison |
| Fed Paper 52 | Madison |
| Fed Paper 53 | Madison |
| Fed Paper 54 | Hamilton |
| Fed Paper 55 | Madison |
| Fed Paper 56 | Madison |
| Fed Paper 57 | Madison |
| Fed Paper 62 | Madison |
| Fed Paper 63 | Madison |

These results reinforce the unbound model’s contradiction of the previous study. Using the results of both the unbound and the parameterized-prune model, it can be concluded that Federalist Papers #49, #50, and #63 remain disputed; however, the remaining essays can be attributed to Hamilton (#54) and Madison (#51-53,#55-62).

**Conclusions**

In summation, before the United States officially came to be, the original states (then colonies) had to elect to ratify the proposed Constitution. In order to persuade or dissuade residents to ratify the Constitution, politicians often wrote anonymous essays in local newspapers. Perhaps the most famous collection of these essays is the Federalist Papers, published under the pen name “Publius” in New York in support of ratification (About the Federalist Papers). Publius is believed to have been used by Alexander Hamilton, James Madison, and Jon Jay who later claimed some of the essays. However, of the 85 essays comprising the Federalist Papers, 11 have disputed authorship. To help clarify, the Library of Congress initiated a series of studies of the documents (About the Federalist Papers).

The initial study concluded that Alexander Hamilton authored Federalist Papers #49, #50, and #63. However, the study was unable to make any definitive conclusions about the authorship of the remaining disputed essays. The Library Congress wanted the second study to confirm and/or improve upon the first results.

After analyzing the data, it appears that all the disputed essays, apart from #54 (authored by Alexander Hamilton), were authored by James Madison. This contradicts the previous study results but provides more insight on previously undetermined authorship.

If the Library of Congress wishes to act on the results of this study, it would likely retain the disputed authorship status of Federalist Papers #49, #50, and #63. However, it could assign #54 to Alexander Hamilton, and the remaining disputed essays to James Madison.

**Citations**

About the Federalist Papers. (n.d.). Retrieved February 8, 2020, from www.congress.gov/resources/display/content/About+the+Federalist+Papers

The Federalist Papers. (n.d.). Retrieved February 8, 2020, from https://www.congress.gov/resources/display/content/The+Federalist+Papers

National Constitution Center. (n.d.). *The Constitution of the United States: Fast facts*. Retrieved February 8, 2020, from constitutioncenter.org/learn/educational-resources/historical-documents/constitution-fast-facts