Joyce Woznica

jlwoznic@syr.edu

November 8, 2019

Joyce Woznica  
Homework #5

Contents

[Introduction 3](#_Toc23874972)

[Data Preparation 4](#_Toc23874973)

[Data Load, Cleanse, Munge and Preparation 4](#_Toc23874974)

[Data Load 4](#_Toc23874975)

[Data Cleanse 5](#_Toc23874976)

[Build and Tune Decision Tree Models 8](#_Toc23874977)

[Decision Tree Model 1 8](#_Toc23874978)

[Decision Tree Model 2 10](#_Toc23874979)

[Decision Tree Model 3 13](#_Toc23874980)

[Prediction 15](#_Toc23874981)

List of Figures

[Figure 1: The Federalist Papers 3](#_Toc23874982)

[Figure 2: Example of Federalist Paper – Number 49 4](#_Toc23874983)

[Figure 3: Structure of Corpus created from Federalist Papers 5](#_Toc23874984)

[Figure 4: Setting up Term Frequencies and Removing Stop Words from Federalist Papers 5](#_Toc23874985)

[Figure 5: Inspecting the Document Term Matrix 6](#_Toc23874986)

[Figure 6: Normalizing the Word Matrix 6](#_Toc23874987)

[Figure 7: Note Responsible Author 7](#_Toc23874988)

[Figure 8: Remove Disputed Papers for Training and Test Sets 7](#_Toc23874989)

[Figure 9: Wordcloud Code for Each Author: Disputed, Hamilton, Madison, HM, Jay 7](#_Toc23874990)

[Figure 10: Wordcloud for Disputed, Hamilton, Joint (H/M), Jay and Madison Papers 8](#_Toc23874991)

[Figure 11: Initial Decision Tree 1 Model 8](#_Toc23874992)

[Figure 12: Decision Tree 1 rpart Details 9](#_Toc23874993)

[Figure 13: Plotting of Errors and Split – Tree 1 9](#_Toc23874994)

[Figure 14: Decision Tree and Confusion Matrix – Tree 1 10](#_Toc23874995)

[Figure 15: Initial Decision Tree 2 Model 10](#_Toc23874996)

[Figure 16: Decision Tree 2 rpart Details 11](#_Toc23874997)

[Figure 17: Plotting of Errors and Split – Tree 2 12](#_Toc23874998)

[Figure 18: Decision Tree and Confusion Matrix – Tree 2 12](#_Toc23874999)

[Figure 19: Initial Decision Tree 3 Model 13](#_Toc23875000)

[Figure 20: Decision Tree 3 rpart Details 13](#_Toc23875001)

[Figure 21: Plotting of Errors and Split – Tree 3 14](#_Toc23875002)

[Figure 22: Decision Tree and Confusion Matrix – Tree 3 14](#_Toc23875003)

[Figure 23: Prediction of the Disputed Articles 15](#_Toc23875004)

# 

# Introduction

There is a collection of 85 articles and essays written in the second half of 1780 by Alexander Hamilton, James Madison and John Jay under the pseudonym “Publius” called the “Federalist Papers” which were written to promote the ratification of the United States Constitution. The authorship was a closely guarded secret. After Hamilton’s death in 1804, the list of articles attributed to Hamilton became public.

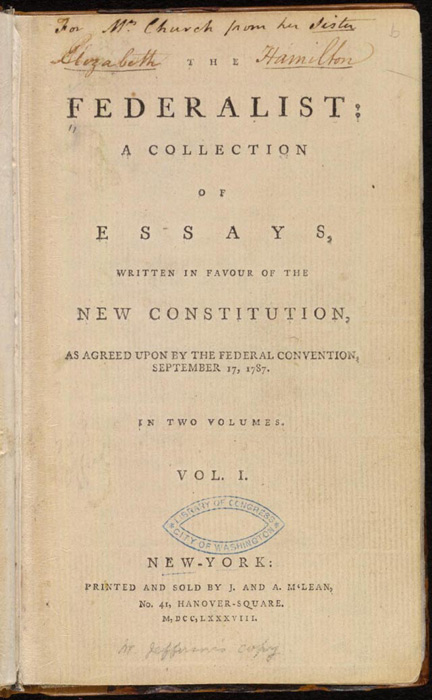


Figure 1: The Federalist Papers

However, who wrote each paper is still mysterious and unfortunately, the truth is has been lost. It has been noted that each author had some unique characteristics to his writing. This could be the use of particular words, phrasing or writing style. Using statistical methods, the goal is to determine which papers can be attributed to which author and how different information in the text draws that conclusion.

This document outlines the data collected, the loading process, any modifications required to the data and why, initial visualizations and information about the data, and the results of decision tree modeling to arrive at the true author of the disputed papers.

# Data Preparation

Each “paper” is a text file (.txt) that is numbered and the content look similar to that represented in the following figure.

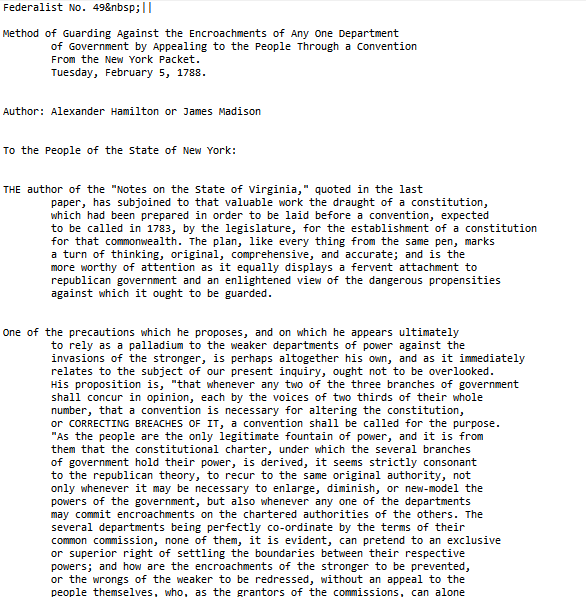


Figure 2: Example of Federalist Paper – Number 49

Each paper is labelled by its contributing author or suspected author. There are 11 disputed papers where authorship is unknown. There are 15 Madison papers, 51 Hamilton papers, 3 Hamilton/Madison (jointly written by Hamilton and Madison) and 5 Jay papers in the data set.

Before inspecting the articles and attempting to make some determination of the author of the papers, there is some cleaning and work that needs to be done on the data set to prepare it for the statistical functions that will be used.

## Data Load, Cleanse, Munge and Preparation

### Data Load

To load the federalist papers, we utilize the *corpus* function. This function creates a corpus from the documents which can then be manipulated for statistical analysis. The resulting 85 documents create a structure like that in the following figure. The text of the paper can be found in the “content” section. The metadata contains the paper name, or ID.

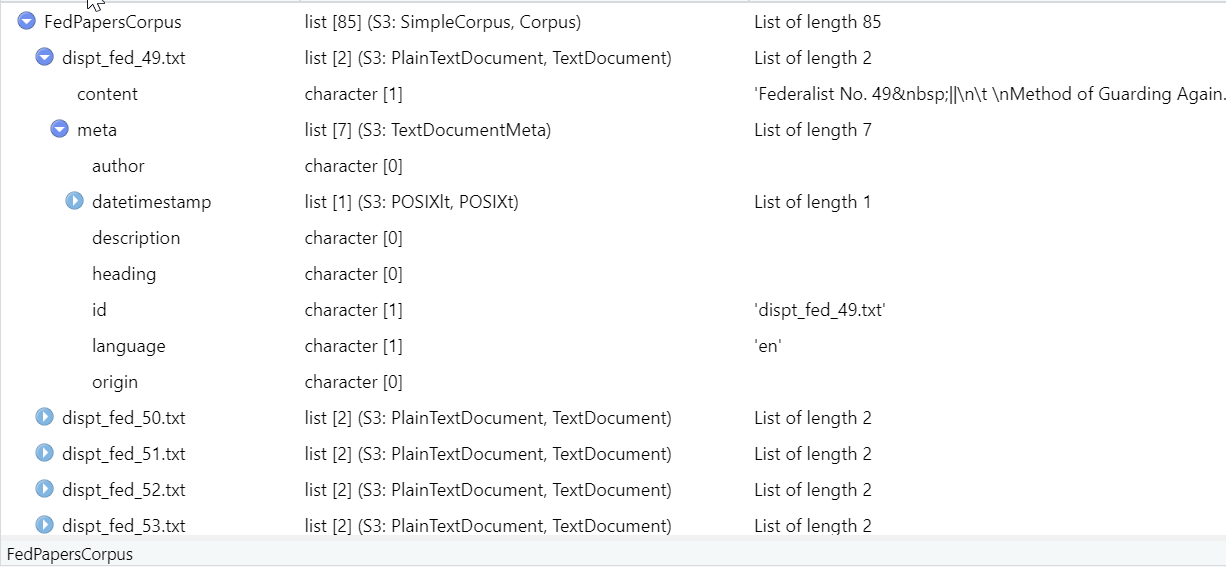


Figure 3: Structure of Corpus created from Federalist Papers

### Data Cleanse

Some initial cleanup needs to be done on the papers to remove very rare and very common words as well as removing stop words such as common conjunctions like “and”. In addition, all punctuation, numbers and separators are removed as shown in the figure below. Stemming was also used to help combine words with similar roots. To properly remove the unnecessary words, convert things to lower case, and more, the *tm* functions were used as shown in the following figure.

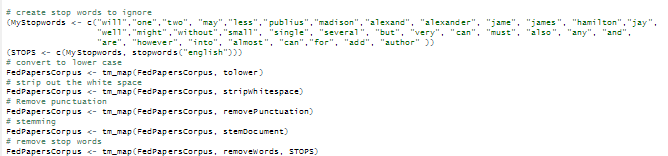
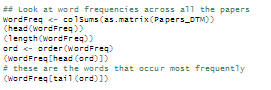


Figure 4: Setting up Term Frequencies and Removing Stop Words from Federalist Papers

Now that a Document Term Matrix has been created, the data can be explored and various statistical methods can be applied to the words to begin to review patterns in the data set. First, the document term matrix can be explored as well as looking at word frequencies and other details.



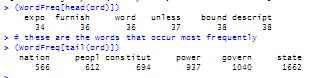
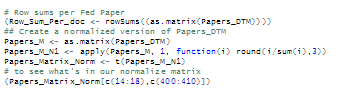


Figure 5: Inspecting the Document Term Matrix

These initial inspections show that there are 414 words (using the *length* function) that appear most frequently in the papers and the most frequent words are “state” and “govern” along with “constitut” (which is likely words like “constitution”), “power”, “peopl” (for people) and “nation”.

Next, a normalized version of the word matrix is created as show in the following figure.



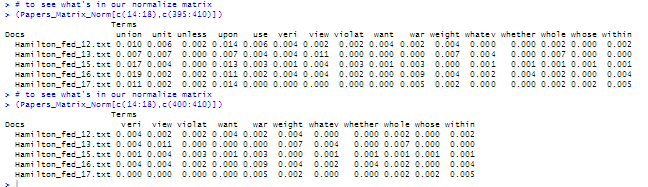
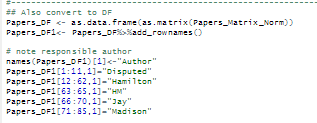


Figure 6: Normalizing the Word Matrix

The words associated with each paper are then noted by the responsible author, not the paper itself. This will help the training of the decision tree. The results of these steps can be found in the following figure.



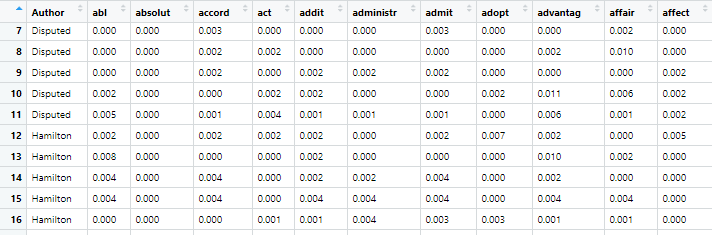


Figure 7: Note Responsible Author

Finally, it is important to removed the disputed papers from the training and test sets so that verification can be performed and the final determination of authorship of the disputed papers can occur. This is done by creating a new data frame that does not include any papers by author “Disputed” as shown in the following figure. This figure also shows the creation of the training and test sets to be used for the decision trees.

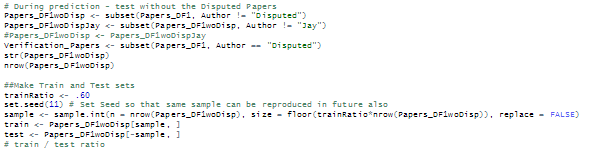


Figure 8: Remove Disputed Papers for Training and Test Sets

To visualize some of the information obtained about these words by the specific authors. This was done by running “word clouds” for each known author as shown in the following figures.

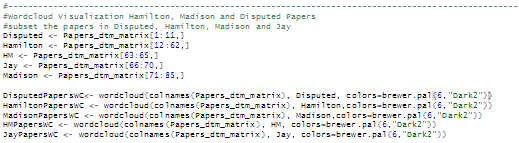


Figure 9: Wordcloud Code for Each Author: Disputed, Hamilton, Madison, HM, Jay

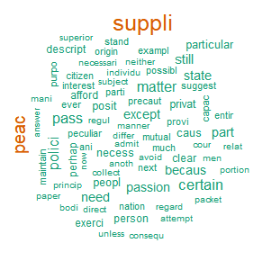


Figure 10: Wordcloud for Disputed, Hamilton, Joint (H/M), Jay and Madison Papers

The larger words represented in these wordclouds are the ones that appear more often. This type of visualization helps to see what words each author uses most often which will help narrow down the authorship.

# Build and Tune Decision Tree Models

## Decision Tree Model 1

In order to study the papers and assist in determining the actual authorship, decision trees were used and modified to arrive on a model with a good chance to predict the disputed papers with some accuracy.

The code used can be found in the following figure.

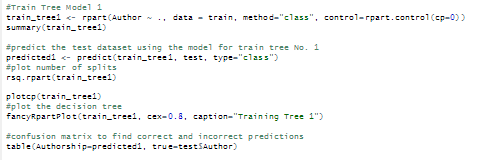


Figure 11: Initial Decision Tree 1 Model

Several plots and details about this tree are output from the code shown in the previous figure. For example, as shown in Figure 12, the variables of importance (words) are represented in the summary of the *rpart* tree model. In addition, the primary and surrogate splits are displayed along with a very high probability (96.7) for papers were the occurrence of “upon” is at least .35% of the words used in the paper.

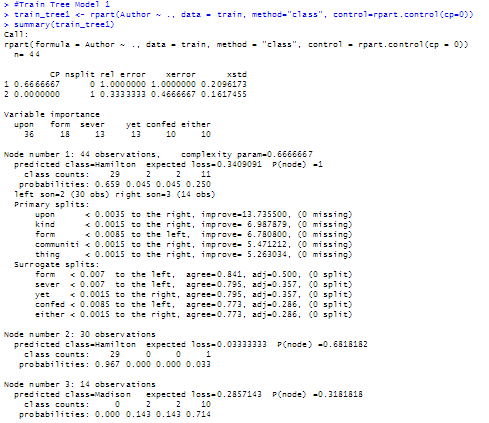


Figure 12: Decision Tree 1 rpart Details

In addition to the summary information, details on the errors and CP are displayed using the *plotcp* functionality with details in the output from *rsq.rpart* against the tree model.

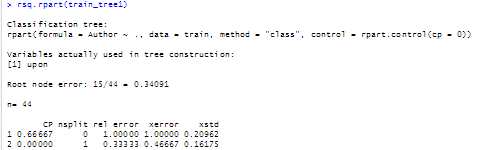
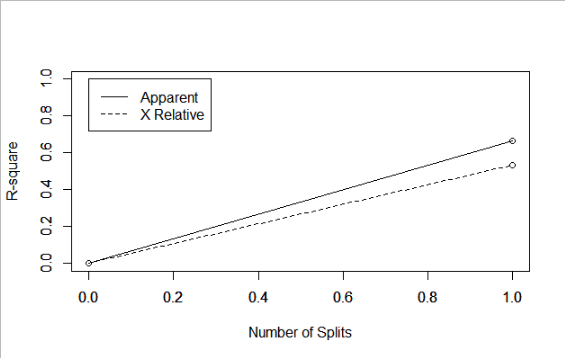
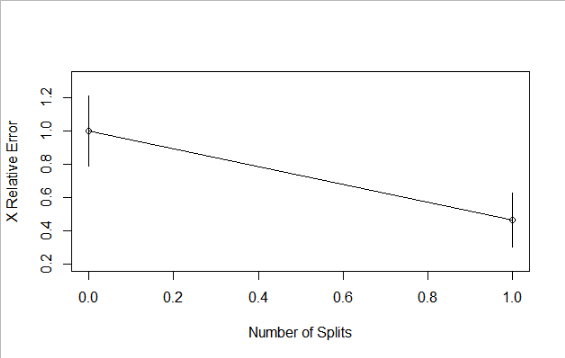
   
 

Figure 13: Plotting of Errors and Split – Tree 1

The resulting tree is shown in the following figure along with the confusion matrix. The model predicts quite well with all Hamilton papers correct, but the Madison, Jay and joint papers are not predicted with the same accuracy.

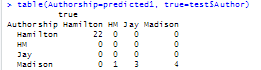
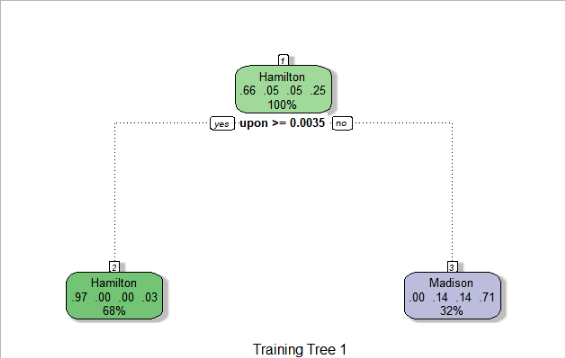


Figure 14: Decision Tree and Confusion Matrix – Tree 1

These same steps were repeated using different parameters for the *rpart* function to determine the best model with the best prediction capability. In this case, the exercise was repeated four times with varying parameters. Outputs for these additional decision trees are in the following figures.

## Decision Tree Model 2

For the next model, the minimum splits and maximum depth were modified to provide a different tree with more nodes.

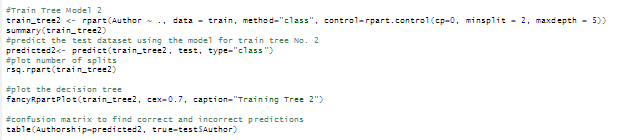


Figure 15: Initial Decision Tree 2 Model

Several plots and details about this tree are output from the code shown in the previous figure. For example, as shown in Figure 16, the variables of importance (words) are represented in the summary of the *rpart* tree model. Here it is shown that the word “upon” continues to be of significance.

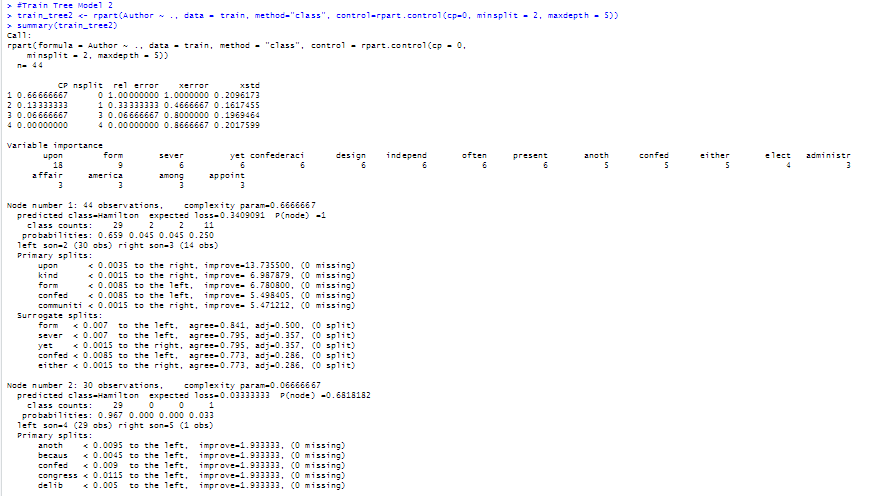
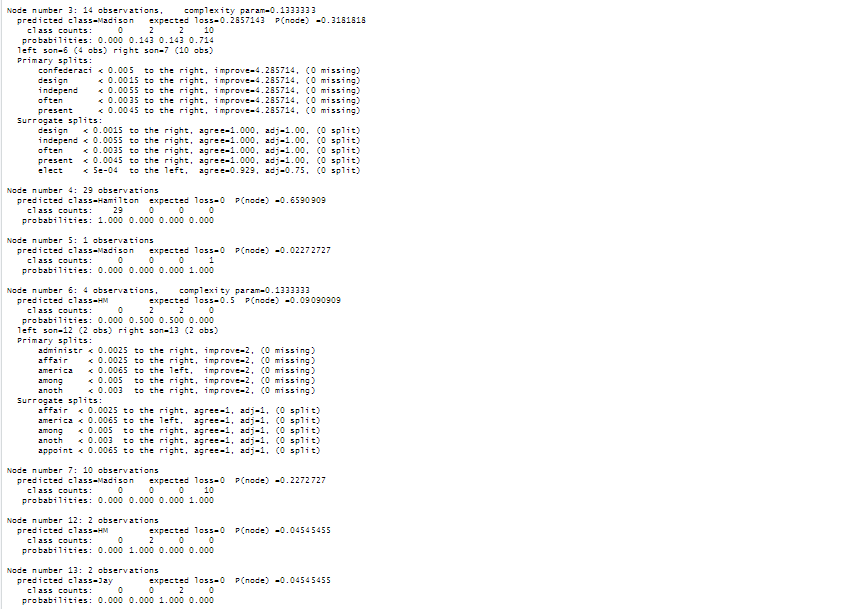
  


Figure 16: Decision Tree 2 rpart Details

In addition to the summary information, details on the errors and CP are displayed using the *plotcp* functionality with details in the output from *rsq.rpart* against the tree model.

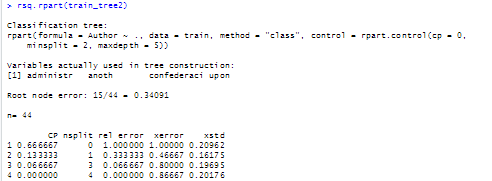
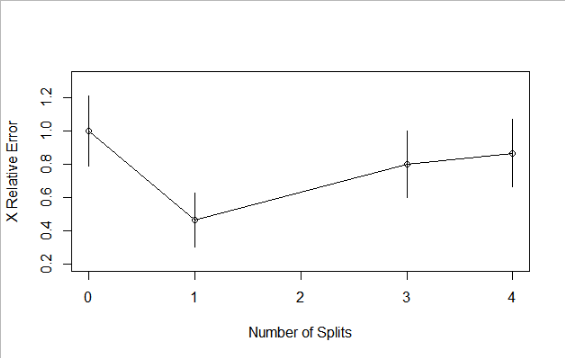
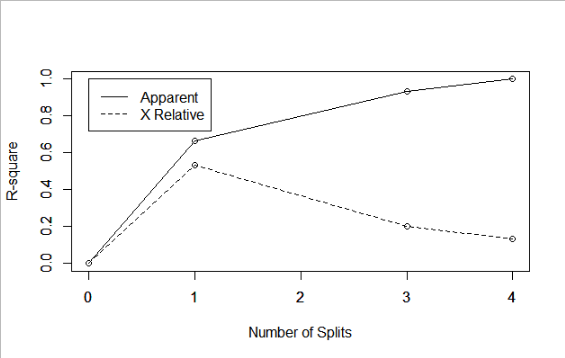
  


Figure 17: Plotting of Errors and Split – Tree 2

The resulting tree is shown in the following figure along with the confusion matrix. The model predicts quite well with all Hamilton papers correct again. In addition, the prediction of the Jay papers improved.

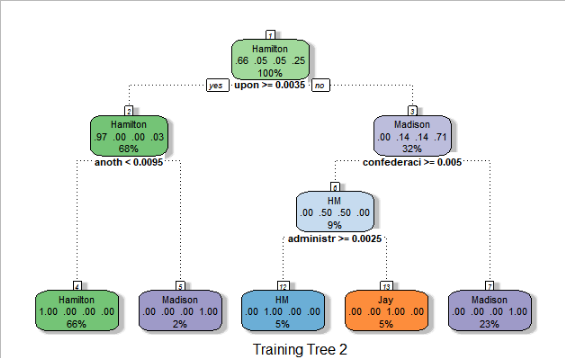
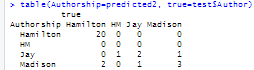
  


Figure 18: Decision Tree and Confusion Matrix – Tree 2

## Decision Tree Model 3

For the next model, the minimum splits and maximum depth were modified again to provide a different tree with more nodes.

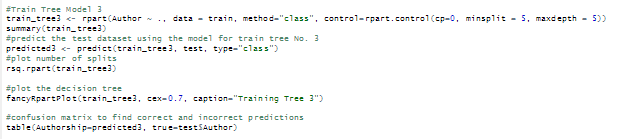


Figure 19: Initial Decision Tree 3 Model

Several plots and details about this tree are output from the code shown in the previous figure. For example, as shown in Figure 20, the variables of importance (words) are represented in the summary of the *rpart* tree model. Here it is shown that the word “upon” continues to be of significance.

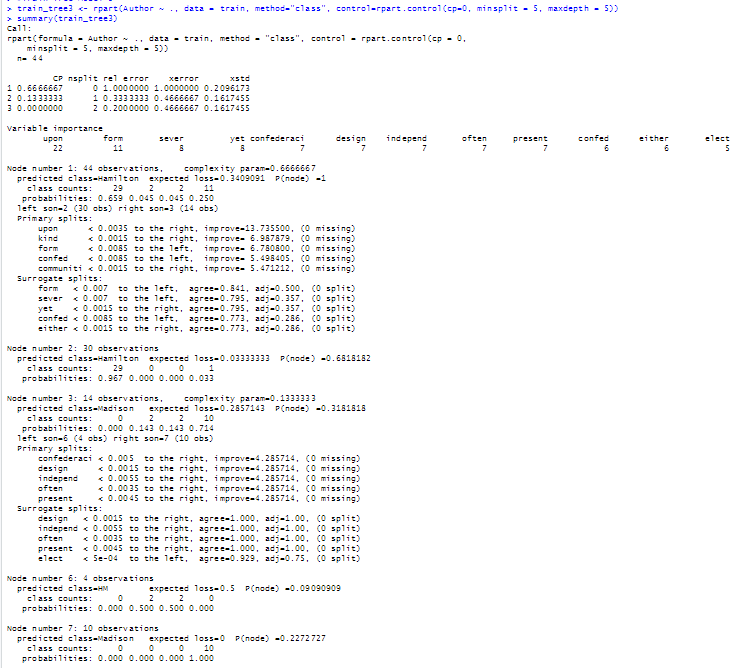


Figure 20: Decision Tree 3 rpart Details

In addition to the summary information, details on the errors and CP are displayed using the *plotcp* functionality with details in the output from *rsq.rpart* against the tree model.

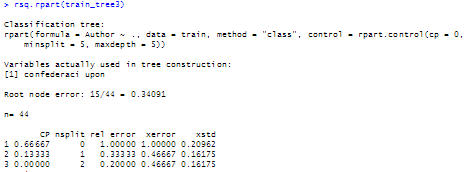
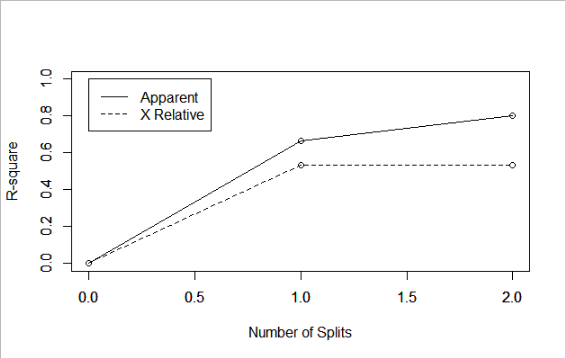
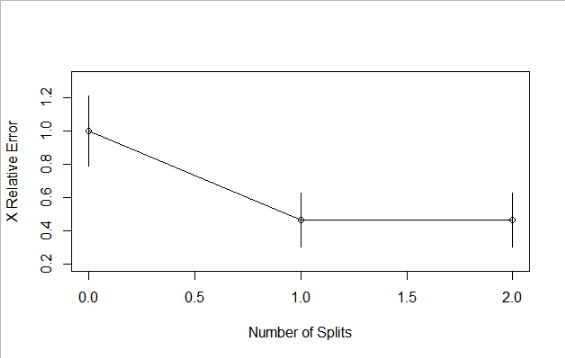
  
 

Figure 21: Plotting of Errors and Split – Tree 3

The resulting tree is shown in the following figure along with the confusion matrix. The model predicts quite well with all Hamilton papers correct again. In this case, the prediction of the Madison papers improved, but not the joint authored or Jay papers.

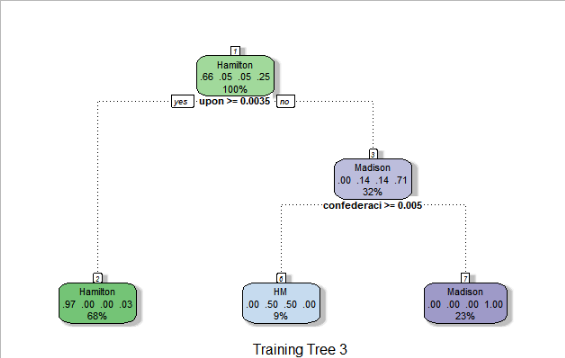
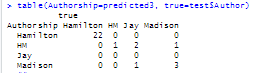
  


Figure 22: Decision Tree and Confusion Matrix – Tree 3

# Prediction

Finally, selecting the strongest model, which was determined to be decision tree 2, prediction was done for the eleven disputed papers.

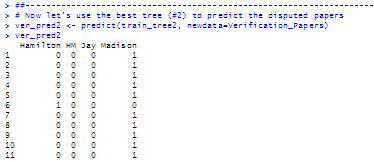


Figure 23: Prediction of the Disputed Articles

Using the 2nd decision tree created, the disputed papers are shown to be authored by Madison with one possibly authored by Hamilton.

It is clear that decision trees and working with common words that appear in papers written by the same author is a solid method for predicting the author of the disputed papers. Decision trees were both easier to understand and provided a better model for predicting the disputed papers.