History Mystery - Classification

## Introduction

Determining authorship of the Federalist Papers is a common case used for teaching data mining and as such this report intends to evaluate who is the author of the disputed Federalist Papers where authorship is not clear using classification and decision tree modeling.

The assigment instructions from Syracuse University Data Analytics course included the following background:

Quote from the Library of Congress <http://www.loc.gov/rr/program/bib/ourdocs/federalist.html>

The Federalist Papers were a series of eighty-five essays urging the citizens of New York to ratify the new United States Constitution. Written by Alexander Hamilton, James Madison, and John Jay, the essays originally appeared anonymously in New York newspapers in 1787 and 1788 under the pen name “Publius.” A bound edition of the essays was first published in 1788, but it was not until the 1818 edition published by the printer Jacob Gideon that the authors of each essay were identified by name. The Federalist Papers are considered one of the most important sources for interpreting and understanding the original intent of the Constitution.

The following background was also provided with the assignment:

These are the famous essays with disputed authorship. Hamilton wrote to claim the authorship before he was killed in a duel. Later Madison also claimed authorship. Historians were trying to find out which one was the real author.

## Analysis and Models

## About the Data

85 essays were provided for this report.

The set of 85 essay files included the following:

* 51 by Hamilton
* 23 by Madison
* 11 essays, labeled as being authored by “Hamilton or Madison”, are the disputed essays that are the goal of this report.

## Getting the Data

The files were provided in a folder and in cvs (text) format. The files were uploaded to R in Corpus format to be conducive to text analysis.

## Data Cleaning

Next, the text was cleaned using the following steps:

* Remove Numbers
* Remove Punctuations
* Remove <1% Words
* Remove Very Common Words (Stop Words)

## Cleaning Results

After cleaning the data, we can do some high level data exploration.

First, we can confirm how many words are left to work with: 4,900

## [1] 4900

Next, we prepare the data for analysis thru vectorization and do some high-level data.

## Additional Data Preparation

##### Normalization

After cleaning the data, the data was further prepared with additional normalization.

##### Add Labels

Labels are added based on authorship: Madison, Hamilton, or disputed.

##### Additional Versions of Data

Matrix and DF versions of the data were created in preparation for different types of analysis:

## Exploratory Data Analysis

We can look at the top six words with the highest frequencies:

## constitut may power govern will state   
## 686 811 937 1040 1263 1662

##### Word Clouds

Word Clouds are an interesting word data exploration technique for text. Below are the wordclouds and top 50 word counts for the following essay segments:

Disputed Hamilton Madison

Word Cloud for the top fifty (50) words in the disputed essays:



The top 50 words by count in the disputed essays:

## peopl senat will may repres govern bodi   
## 42 24 19 18 18 16 15   
## can elect must measur state corrupt nation   
## 14 14 12 11 11 9 9   
## one constitut former power reason year assembl   
## 9 8 8 8 8 8 7   
## exampl two annual danger everi evid feder   
## 7 7 6 6 6 6 6   
## import latter object particular public advantag ancient   
## 6 6 6 6 6 5 5   
## answer appear author charact fact first hous   
## 5 5 5 5 5 5 5   
## institut less mani member might oper order   
## 5 5 5 5 5 5 5   
## part   
## 5

##### Hamilton vs. Madison Word Clouds

WordCloud comparison for the top fifty (50) words in the Hamilton and Madison essays:



*Madison*

*Hamilton*

## Models

Next, data is randomly selected to create train and test data sets.

### Decision Tree Model #1

The train to test ratio for data is set to 60%.

## [1] 0.6

##### Decision Tree Model #1

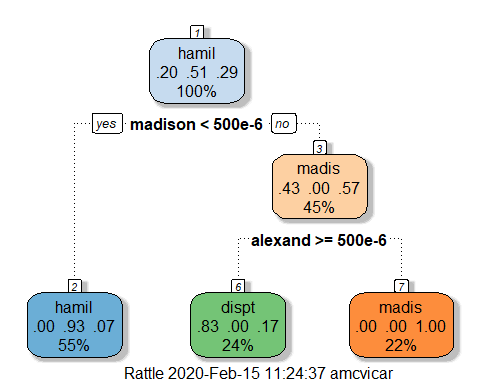
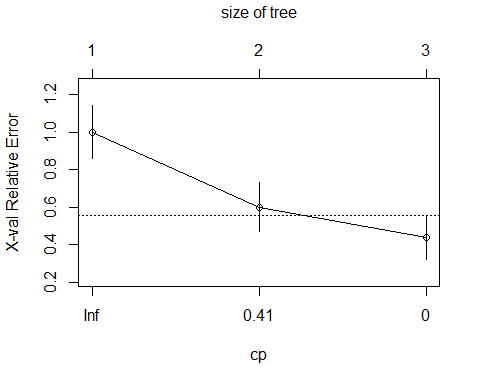
## Call:  
## rpart(formula = Author ~ ., data = train, method = "class", control = rpart.control(cp = 0))  
## n= 51   
##   
## CP nsplit rel error xerror xstd  
## 1 0.52 0 1.00 1.00 0.1428011  
## 2 0.32 1 0.48 0.60 0.1301583  
## 3 0.00 2 0.16 0.44 0.1174901  
##   
## Variable importance  
## madison jame upon kind second among alexand   
## 15 15 13 9 9 8 8   
## hamilton accomplish appli definit determin   
## 8 4 4 4 4   
##   
## Node number 1: 51 observations, complexity param=0.52  
## predicted class=hamil expected loss=0.4901961 P(node) =1  
## class counts: 10 26 15  
## probabilities: 0.196 0.510 0.294   
## left son=2 (28 obs) right son=3 (23 obs)  
## Primary splits:  
## madison < 5e-04 to the left, improve=16.35392, (0 missing)  
## upon < 0.0025 to the right, improve=16.32947, (0 missing)  
## jame < 5e-04 to the left, improve=14.91885, (0 missing)  
## alexand < 5e-04 to the right, improve=13.89886, (0 missing)  
## hamilton < 5e-04 to the right, improve=13.89886, (0 missing)  
## Surrogate splits:  
## jame < 5e-04 to the left, agree=0.980, adj=0.957, (0 split)  
## upon < 0.0025 to the right, agree=0.922, adj=0.826, (0 split)  
## kind < 5e-04 to the right, agree=0.804, adj=0.565, (0 split)  
## second < 5e-04 to the left, agree=0.804, adj=0.565, (0 split)  
## among < 0.0015 to the left, agree=0.784, adj=0.522, (0 split)  
##   
## Node number 2: 28 observations  
## predicted class=hamil expected loss=0.07142857 P(node) =0.5490196  
## class counts: 0 26 2  
## probabilities: 0.000 0.929 0.071   
##   
## Node number 3: 23 observations, complexity param=0.32  
## predicted class=madis expected loss=0.4347826 P(node) =0.4509804  
## class counts: 10 0 13  
## probabilities: 0.435 0.000 0.565   
## left son=6 (12 obs) right son=7 (11 obs)  
## Primary splits:  
## alexand < 5e-04 to the right, improve=7.971014, (0 missing)  
## hamilton < 5e-04 to the right, improve=7.971014, (0 missing)  
## regular < 5e-04 to the left, improve=6.688963, (0 missing)  
## possess < 0.0015 to the right, improve=6.097999, (0 missing)  
## union < 0.0025 to the left, improve=5.590062, (0 missing)  
## Surrogate splits:  
## hamilton < 5e-04 to the right, agree=1.000, adj=1.000, (0 split)  
## accomplish < 5e-04 to the left, agree=0.783, adj=0.545, (0 split)  
## appli < 5e-04 to the right, agree=0.783, adj=0.545, (0 split)  
## definit < 5e-04 to the left, agree=0.783, adj=0.545, (0 split)  
## determin < 5e-04 to the left, agree=0.783, adj=0.545, (0 split)  
##   
## Node number 6: 12 observations  
## predicted class=dispt expected loss=0.1666667 P(node) =0.2352941  
## class counts: 10 0 2  
## probabilities: 0.833 0.000 0.167   
##   
## Node number 7: 11 observations  
## predicted class=madis expected loss=0 P(node) =0.2156863  
## class counts: 0 0 11  
## probabilities: 0.000 0.000 1.000

##Decision Tree Model #1 Prediction

##   
## Classification tree:  
## rpart(formula = Author ~ ., data = train, method = "class", control = rpart.control(cp = 0))  
##   
## Variables actually used in tree construction:  
## [1] alexand madison  
##   
## Root node error: 25/51 = 0.4902  
##   
## n= 51   
##   
## CP nsplit rel error xerror xstd  
## 1 0.52 0 1.00 1.00 0.14280  
## 2 0.32 1 0.48 0.60 0.13016  
## 3 0.00 2 0.16 0.44 0.11749

## Warning in rsq.rpart(train\_tree1): may not be applicable for this method

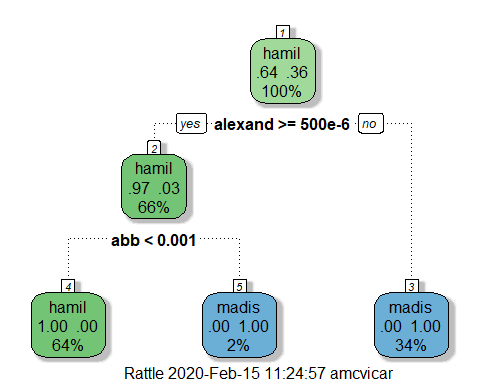
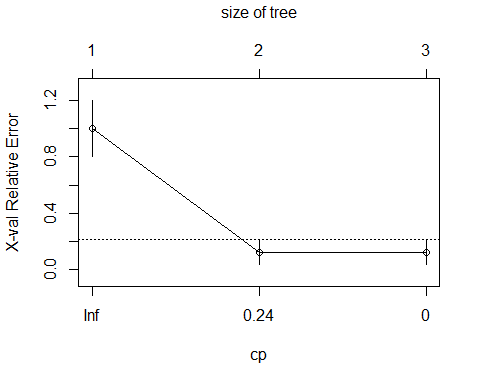
##### 

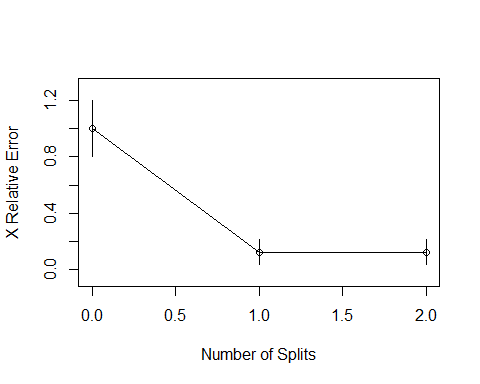
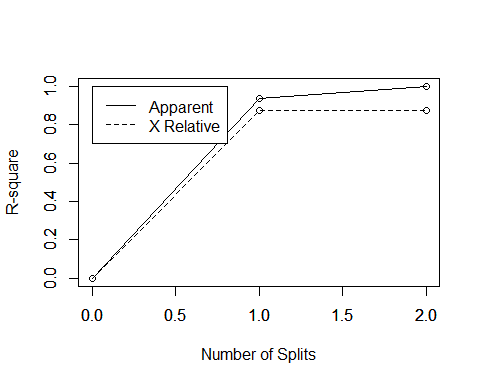


##### Decision Tree Model #1 Confusion Matrix

## true  
## Authorship dispt hamil madis  
## dispt 1 0 1  
## hamil 0 25 3  
## madis 0 0 4

## Decision Tree Model #2

In model 2, the data is divided by author and a dataset that only contains papers with determined authors of Madison or Hamilton is used.



## Decision Tree Model #2 Confusion Matrix

Model #2, the model that used only positively identified authorship essays, had the following accuracy results when applied to test data.

## true  
## Authorship hamil madis  
## hamil 21 0  
## madis 2 7

## Decision Tree Model #2 Disputed Prediction Matrix

Model #2 was then applied to the data set containing only disputed papers and predicted all 11 cases were written by Hamilton.

## true  
## Authorship dispt  
## hamil 11  
## madis 0

## Conclusion

Based on the earlier clustering analysis that indicated that some of the eleven disputed essays very well may have been written by Madison, but high level number analysis indicated a statistical favoring of guessing Hamilton for every paper, I have to conclude that the current decision tree model being applied is picking up the same statistical pattern and it is not granular enough.