Decision Tree

Jose Reyes

8.11.2020

#### Introduction

A famous man once said, “Philosophy is common sense with big words.” Was it James Madison who said it? Or was it Alexander Hamilton? Giving proper credit for literature can be confusing, and it’s especially confusing when two dead men are fighting for it.

In the curious case of the Federalist Papers, Alexander Hamilton and James Madison, two very famous figures in American History, have claimed authorship of 11 essays. To identify the rightful author(s) of these 11 essays, a data mining technique called clustering was previously conducted on all 85 Federalist Papers. While the results provided some clues to the authorship of these mysterious essays, there was no definitive conclusion as to who authored the papers. This time, another technique called decision tree will be used to predict authorship. Hopefully the results will provide more clues and solve one of the great mysteries of United States history.

#### Analysis and Models

#### About the Data

Before starting the clustering process, a data frame is created with information on each of the Federalist paper including each documents’ words and their frequencies. Of the 85 documents, 51 are known to be written by Hamilton, 15 by Madison, 5 by Jay, and 3 are by both Hamilton and Madison. The authorship of 11 of the documents are disputed.

The .txt files of all 85 Federalist Papers are loaded as a corpus, cleansed, and formed into a Document Term Matrix (DTM). The parameters for the DTM are defined below:

FedPapers <- DocumentTermMatrix(FedPapersCorpus, control = list(  
 stopwords = TRUE,  
 wordLengths=c(3,15),  
 removePunctuation = T,  
 removeNumbers = T,  
 tolower = T,  
 stemming = T,  
 remove\_separators = T,  
 stopwords = StopWords,  
 removeWords = STOPS,  
 removeWords = StopWords,  
 bounds = list(global = c(20,Inf))  
 ))

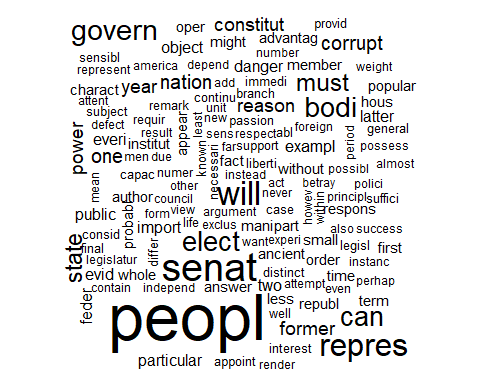
The parameters removed words less than 3 characters, more than 15 characters, and several “stop words” which generally don’t bring much value to text analysis. The lower bound of the word frequency was set to 20 and this additionally reduced the number of words used for analysis from 4,901 to 707.

### Explore Data

### 

### After the initial cleansing, the most common words in the Federalist Papers are state, will, and govern. The frequency drops exponentially from there.

Additionally, here’s a fun word cloud:



While it is interesting to see all the words visually and graphically, a decision tree analysis will be needed to adequately provide insight on the mysterious 11 Federalist Papers

### Normalization

The data is first normalized and labeled. Here, the disputed papers have been labeled “Who\_Am\_I?”

#Normalization  
N\_FedPapers <- apply(FedPapers\_Matrix, 1, function(i) round(i/sum(i),3))  
N\_FedPapers <- t(N\_FedPapers)

#Convert to DataFrame and label data  
DF\_FedPapers <- as.data.frame(N\_FedPapers)  
DF\_FedPapers <- DF\_FedPapers %>% rownames\_to\_column("Author")  
DF\_FedPapers[1:11,1] <- "Who\_Am\_I?"  
DF\_FedPapers[12:62,1] <- "Hamilton"  
DF\_FedPapers[63:65,1] <- "Hamilton\_Madison"  
DF\_FedPapers[66:70,1] <- "Jay"  
DF\_FedPapers[71:85,1] <- "Madison"

#### Train and Test Data

Because this a supervised learning technique, the model needs to first be trained on a subset of the data. 60% of the documents have been designated as train data while the other 40% will be used to test the model. All = 11 disputed papers were allocated to the test model.

#Make Train/Test Data  
numDisputed = 11  
numTotalPapers = nrow(DF\_FedPapers)  
trainRatio <- .6  
set.seed(11)  
sample <- sample.int(n=numTotalPapers-numDisputed, size = floor(trainRatio\*numTotalPapers), replace = FALSE)  
newSample = sample + numDisputed  
train <- DF\_FedPapers[newSample,]  
test <- DF\_FedPapers[-newSample,]

#### Decision Tree

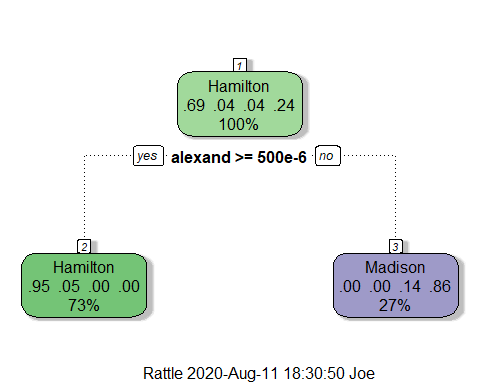
A decision tree splits the source set into subsets based on a certain attribute value test. This splitting of data occurs on a number of nodes, and the process is repeated until splitting no longer adds value to the prediction.

In this case, the decision tree will split the set of documents based on specific word frequencies. The model will first be trained using the training data set. It will then be used on the test data set and predict the authorship of the 11 disputed documents.

### Train Model 1

The decision tree can be customized using several arguments, including the complexity parameter (CP), the minimum number of observations that must exist for a split to be attempted (minsplit), and the maximum number of splits (maxdepth). For the first model, the default arguments have been set.

#Train Model 1  
Model\_1 <- rpart(Author~.,data = train, method = "class", control = rpart.control(cp=0))  
summary(Model\_1)



Interestingly, Model 1 only contains 1 internal node and separates the data based on the prevalence of the word ‘alexand.’ Because of this, the data is only split into two categories: Hamilton and Madison. It can be already be assumed that it will incorrectly predict documents authored by Jay and Hamilton/Madison.

The model is then used to predict the authorship of documents in the test data set.

#Use Model 1 on test data set  
Model\_1\_Pred <- predict(Model\_1, test, type="class")  
table(test$Author,Model\_1\_Pred)

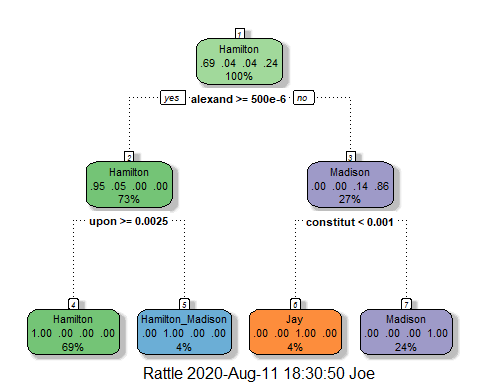
## Model\_1\_Pred  
## Hamilton Hamilton\_Madison Jay Madison  
## Hamilton 16 0 0 0  
## Hamilton\_Madison 1 0 0 0  
## Jay 0 0 0 3  
## Madison 0 0 0 3  
## Who\_Am\_I? 11 0 0 0

Model 1 correctly predicted 16 Hamilton documents and 3 Madison documents. It incorrectly predicted 1 Hamilton/Madison document and 3 Jay documents as expected. Lastly, it predicted that all 11 disputed documents were authored by Hamilton.

### Train Model 2

An additional model was trained with a cp value of 0, minimum split of 2, and maximum depth of 5.

#Train Tree Model 2  
Model\_2 <- rpart(Author~.,data = train, method = "class", control = rpart.control(cp=0, minsplit = 2, maxdepth = 5))  
summary(Model\_2)



This time, Model 2 has 3 internal nodes and can recognize documents authored by Jay and Hamilton/Madison in addition to the terminal nodes of Model 1.

## Model\_2\_Pred  
## Hamilton Hamilton\_Madison Jay Madison  
## Hamilton 16 0 0 0  
## Hamilton\_Madison 0 1 0 0  
## Jay 0 0 2 1  
## Madison 0 0 0 3  
## Who\_Am\_I? 2 9 0 0

Like Model 1, Model 2 correctly identified 16 Hamilton documents and 3 Madison documents. It also correctly identified 1 Hamilton/Madison document and 1 Jay document. Overall, it only incorrectly identified one Jay document and the overall accuracy is an indicator of a pretty strong model.

This time, the model predicted that 9 of the disputed documents are Hamilton/Madison while the other 2 were authored by Hamilton.

### Results

The following table provides a summary of results for each disputed Federalist Paper:

|  |  |  |
| --- | --- | --- |
| **Federalist Paper Number** | **Model 1 Prediction** | **Model 2 Prediction** |
| 49 | Hamilton | Hamilton/Madison |
| 50 | Hamilton | Hamilton |
| 51 | Hamilton | Hamilton/Madison |
| 52 | Hamilton | Hamilton/Madison |
| 53 | Hamilton | Hamilton/Madison |
| 54 | Hamilton | Hamilton |
| 55 | Hamilton | Hamilton/Madison |
| 56 | Hamilton | Hamilton/Madison |
| 57 | Hamilton | Hamilton/Madison |
| 62 | Hamilton | Hamilton/Madison |
| 63 | Hamilton | Hamilton/Madison |

Firstly, it should be noted that Model 2 produced more accurate results overall and contained terminal nodes for Jay and Hamilton/Madison while Model 1 only has terminal nodes for Hamilton and Madison.

For Papers 50 and 54, both models agree that Hamilton is the rightful author. However, Model 2 predicted that the other 9 papers are authored by Hamilton/Madison. This leads to two potential and interesting conclusions:

1. Model 2 is just as confused as everyone; unsure if the papers are written by Hamilton or Madison.
2. The papers were truly written by both Hamilton and Madison

Conclusion 2 is interesting to think about – perhaps Hamilton and Jay were both claiming authorship because they both took part in writing it! And this is perhaps why it has been so difficult to determine the authorship because the papers contain elements of both their writing styles. Overall, it is difficult to tell definitively from these results and will likely need more analysis from other techniques such as neural networks.

#### Conclusion

To determine the author of 11 Federalist Papers, a data mining technique called decision tree was performed. The model was trained to create a number of nodes which separated the documents based on their word frequencies. The model was then used on the train data to predict the authorship of the 11 disputed documents.

Prior to modelling, the .txt files of all 85 documents were loaded, cleansed, and loaded into a data frame to allow for clustering analysis. Words that are common or rare were removed to allow for proper analysis. Two decision tree models with different parameters were created and performed on the dataset.

Model 2 produced more accurate results and concluded that Hamilton authored two documents and Hamilton/Madison authored the other 9. Model 1 predicted that Hamilton authored all 11. Overall, it is hard to make strong conclusions solely from the modelling results and will likely require more analysis.