Homework 4 & 5

Today, we have the great pleasure of using clustering and decision trees to solve the historic mystery the Federalist Papers. These papers include a series of 85 documents to encourage the citizens of New York to ratify the US Constitution. Although those documents were signed under the pin name “Publius”, the actual authors of Federalist Papers were Alexander Hamilton, James Madison, and John Jay. It has been documented that 52 of those essays were written by Hamilton, 15 by Madison, and 3 by Jay. Who wrote the remaining 11? It is claimed that either Hamilton or Madison wrote them. We will discover the true author in this project.

**Packages & Libraries**

install.packages("RColorBrewer")

install.packages("NLP")

install.packages("tm")

install.packages("proxy")

install.packages("dplyr")

install.packages("mclust")

install.packages("factoextra")

install.packages("tibble")

install.packages("fancyRpartPlot")

install.packages("rpart")

install.packages("rplot")

install.packages("rpart.plot")

install.packages("rattle")

library(RColorBrewer)

library(wordcloud)

library(slam)

library(quanteda)

library(SnowballC)

library(arules)

library(cluster)

library(stringi)

library(Matrix)

library(tidytext)

library(dplyr)

library(ggplot2)

library(mclust)

library(proxy)

library(tibble)

library(rpart)

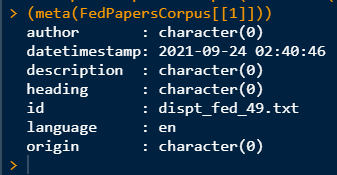
library(rpart.plot)

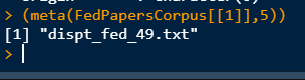
library(rattle)

**Load Data**

**Graphical user interface, text

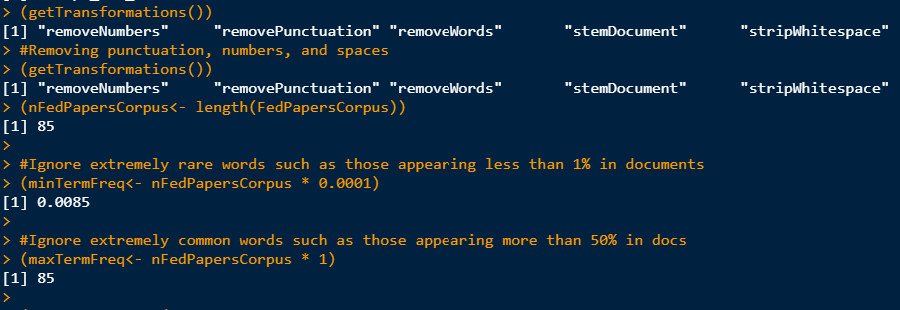
Description automatically generated**

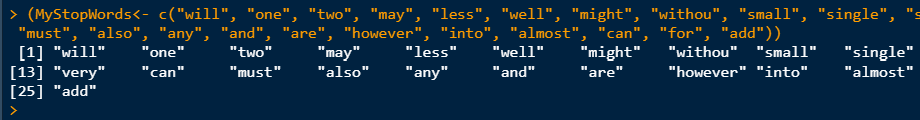


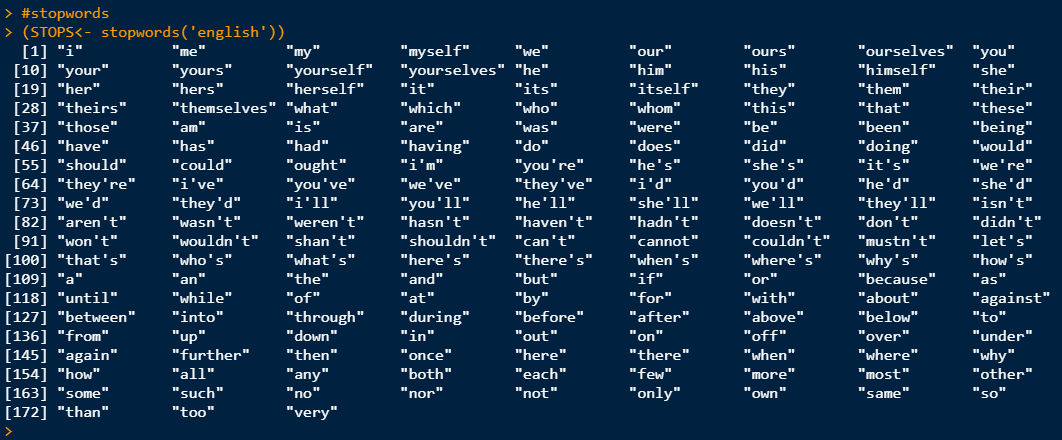


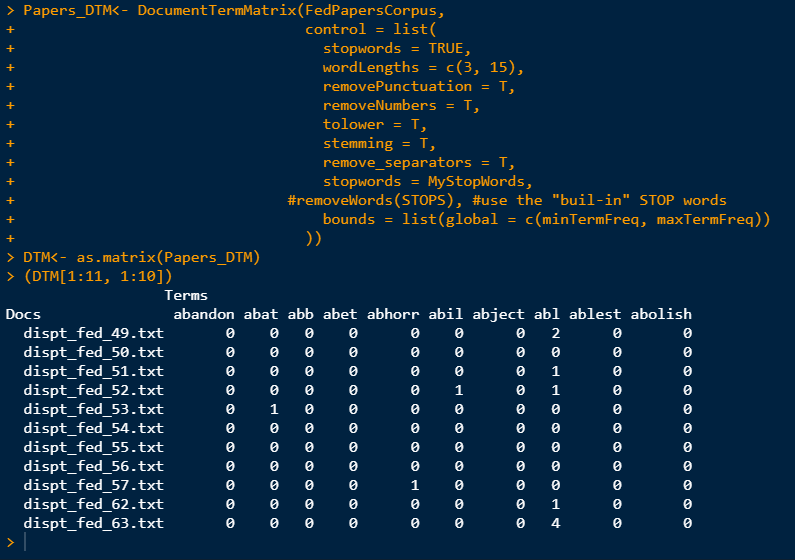
**Data Cleaning**

Now, we will investigate the data and vectorize it using DocumentTermMatrix. We will ignore words and very frequents. This process will perform much of the cleansing process.



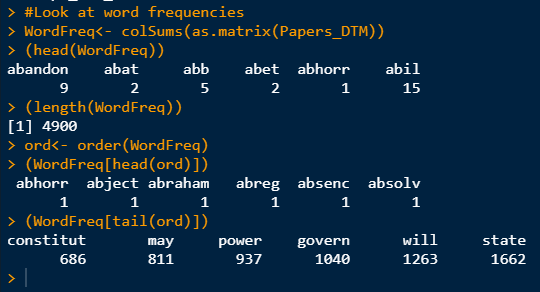


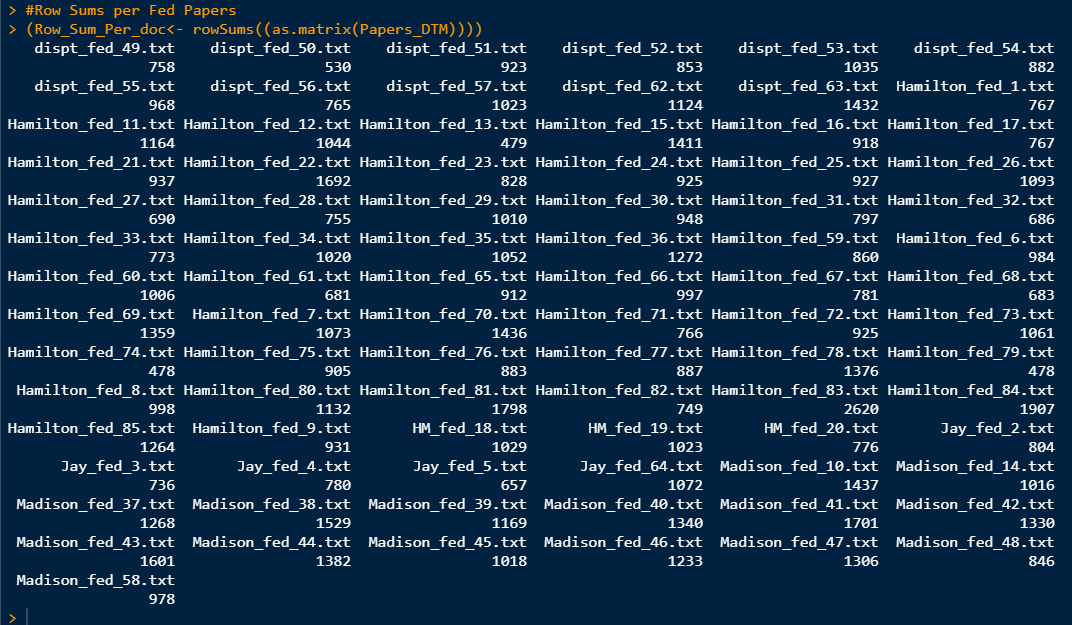




**Inspect Initial Cleaning Results**

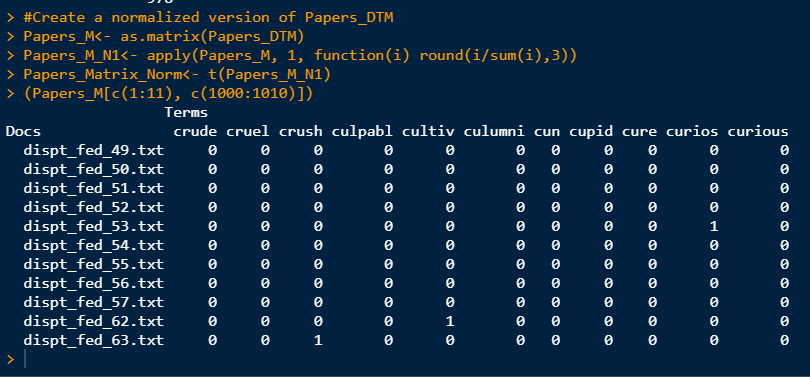
Here, we will investigate the initial results of data cleaning. Depending on the results, we may decide to do further cleaning (perhaps add more stop words).

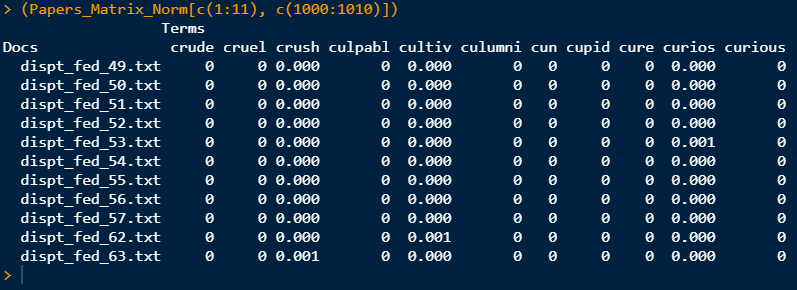




**Normalization**

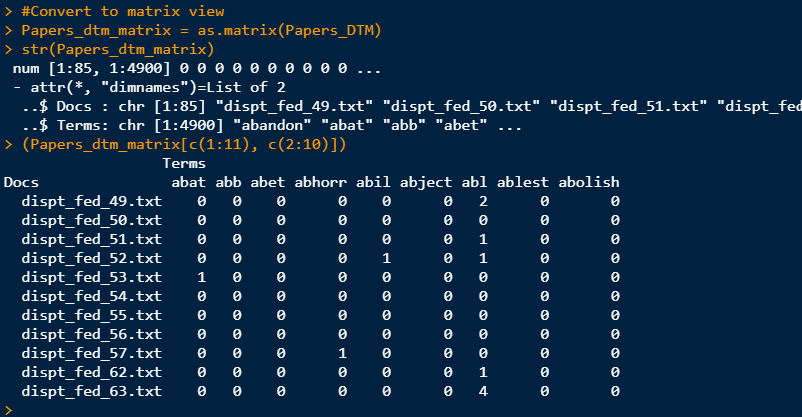
When processing text data, it is often beneficial to normalize word vectors prior to applying standard analysis techniques.

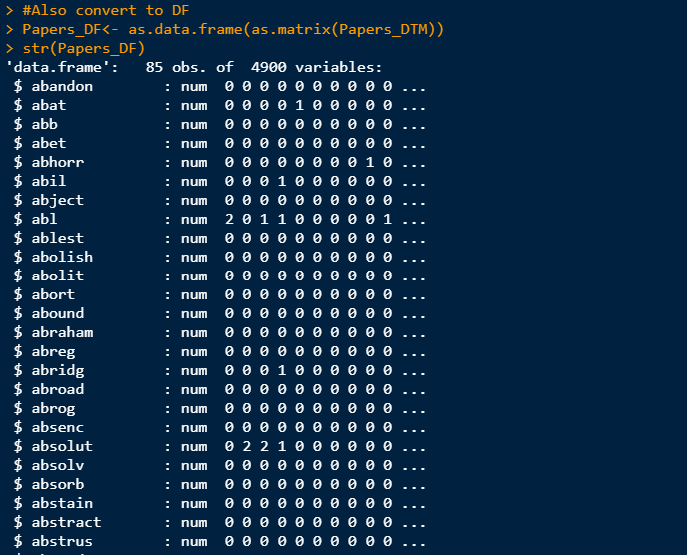


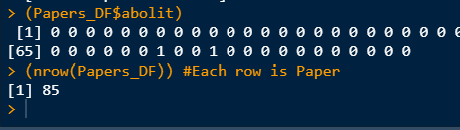


From the line of code '(Row\_Sum\_Per\_doc<- rowSums((as.matrix(FedPapersDTM)))) above, we can see that dispt\_fed\_text\_53/txt has a row sum of 1035. So, we can confirm correctness. For word "curious" we should have 1/1034 = 0.001 rounded, which is what we have.

**Data Structures**





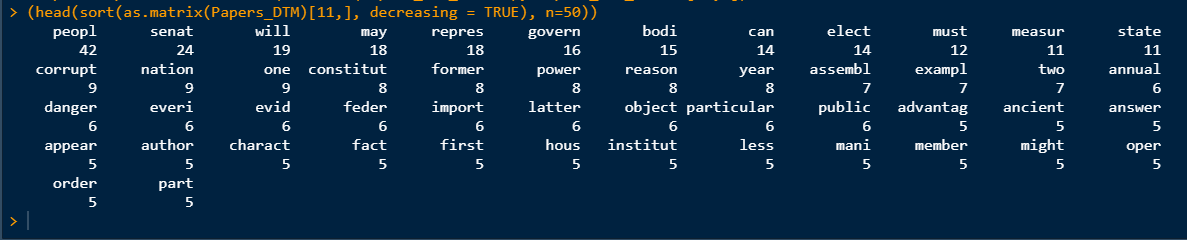


**Word Cloud Example**

****

Text

Description automatically generated

****

****

Text, letter

Description automatically generated

****

Text

Description automatically generated

**Analysis**

**.**

Now that the data has been vectorized, analysis can take place. We will apply some distance metrics and see how the data clusters. Note: Cosine distance usually works well with high dimensional data

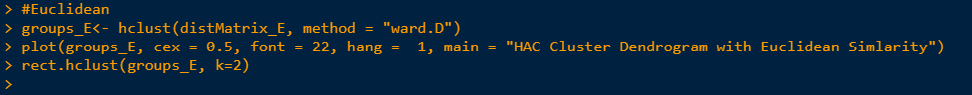
Distance Metrics

**Text

Description automatically generated**

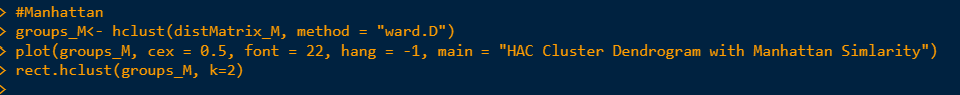
Clustering

Below are some Hierarchical Algorithm Clustering results.



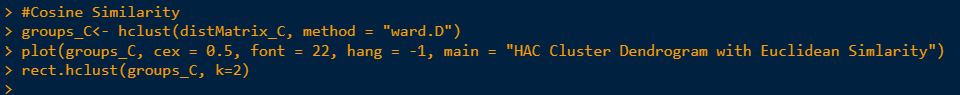
A picture containing diagram

Description automatically generated



Diagram

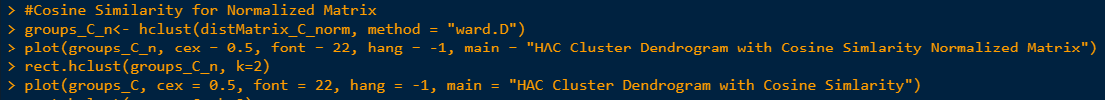
Description automatically generated



Diagram

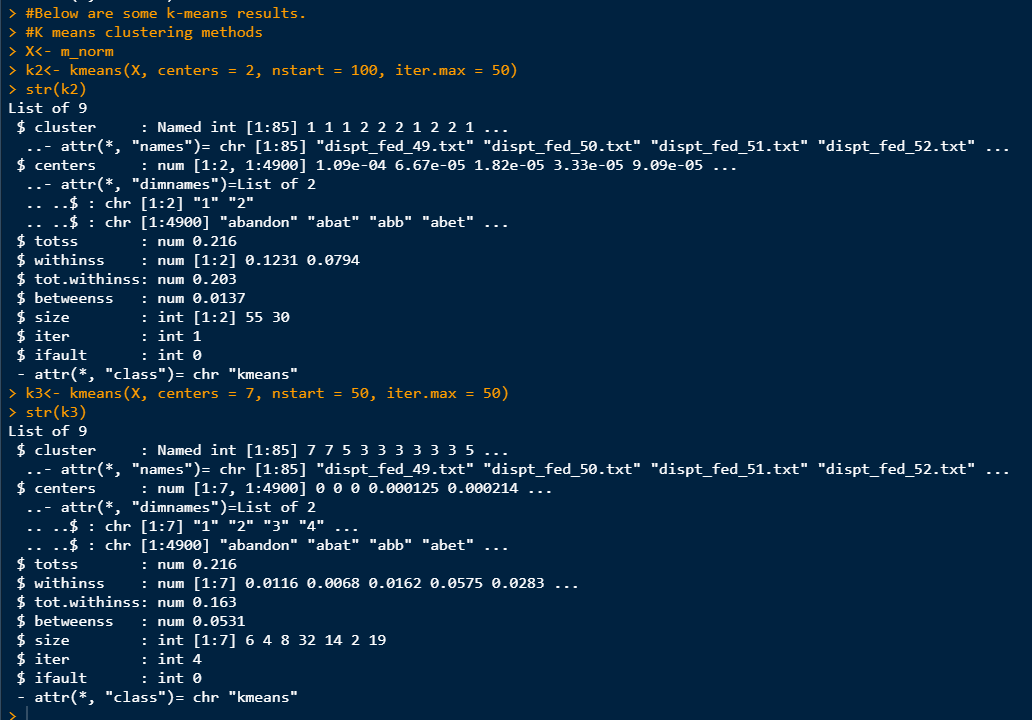
Description automatically generated with low confidence

The HAC Cosine Similarity Algorithm works the best. It clusters most of the disputed articles with Madison.

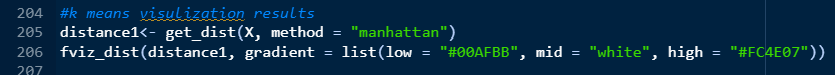


A picture containing diagram

Description automatically generated



**K-Means Visualization Results**



Chart

Description automatically generated



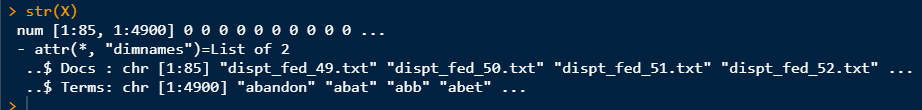
Chart, line chart

Description automatically generated

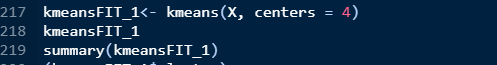


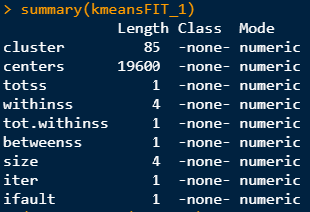
Chart

Description automatically generated

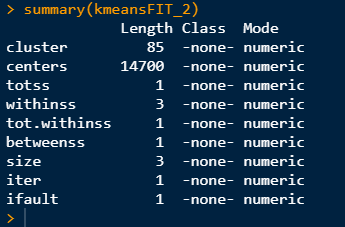


K-Means



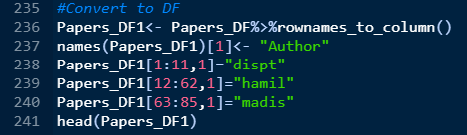




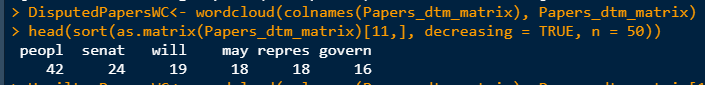


**Label the Data**

Below we label the data, prepare for modeling, and create some wordclouds.



Wordcloud Visualization Hamilton, Madison, and Disputed Papers



Text, letter

Description automatically generated



Text, letter

Description automatically generated



Text

Description automatically generated

**Experimental Design**

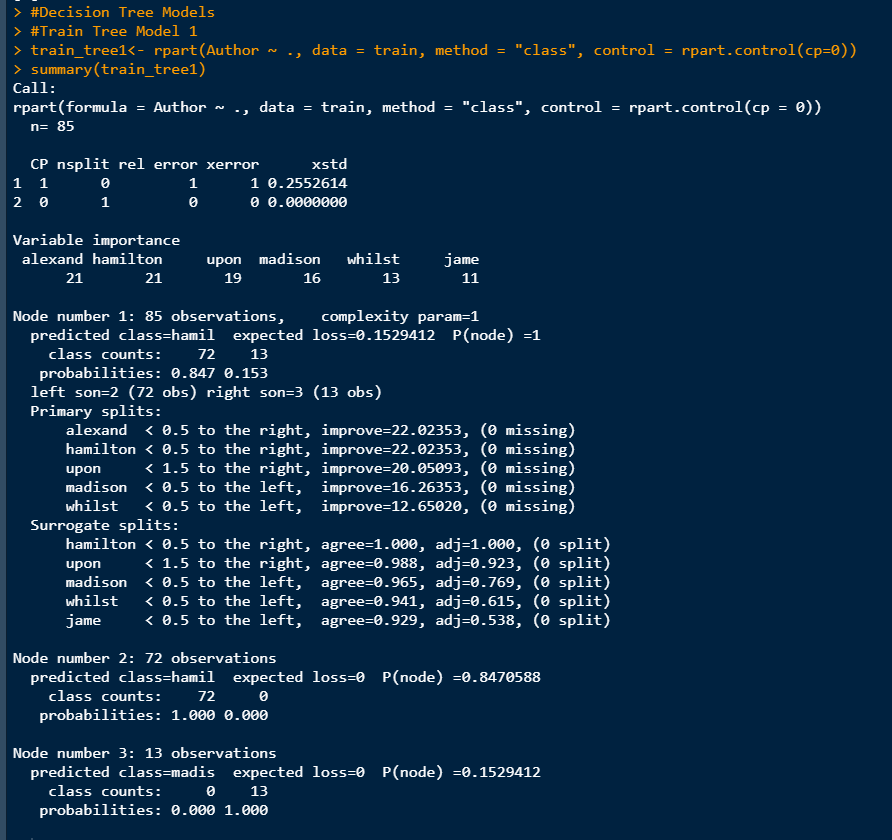
In this section, we will design an experiment. Initially, we randomly select a train and test set for validation.

Text

Description automatically generated

**Classification**

We are now ready to train and test using classifiers. Below we use a few different decision tree models.

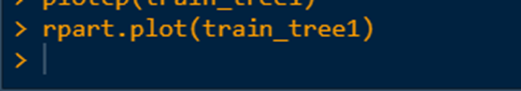


Text

Description automatically generated

Chart, scatter chart

Description automatically generated



Chart, line chart

Description automatically generated



Chart

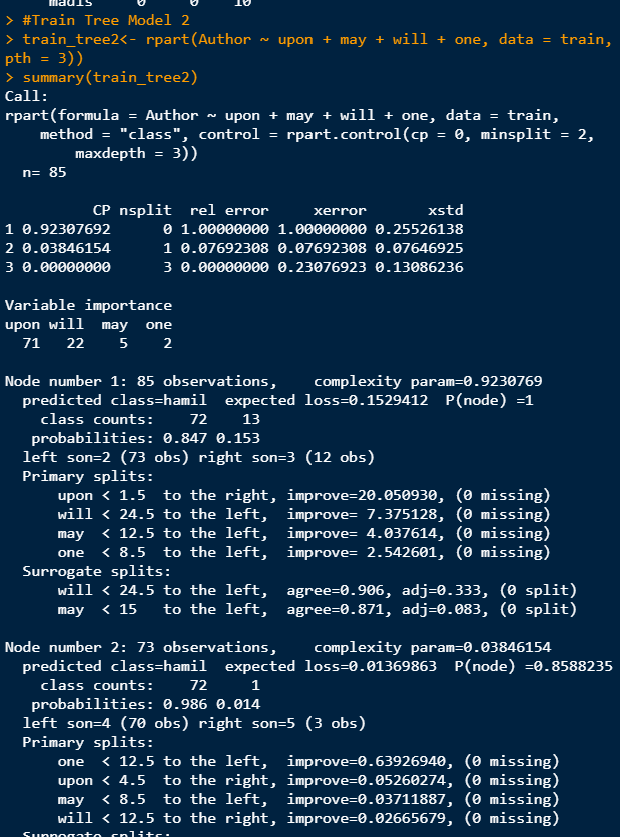
Description automatically generated with medium confidence

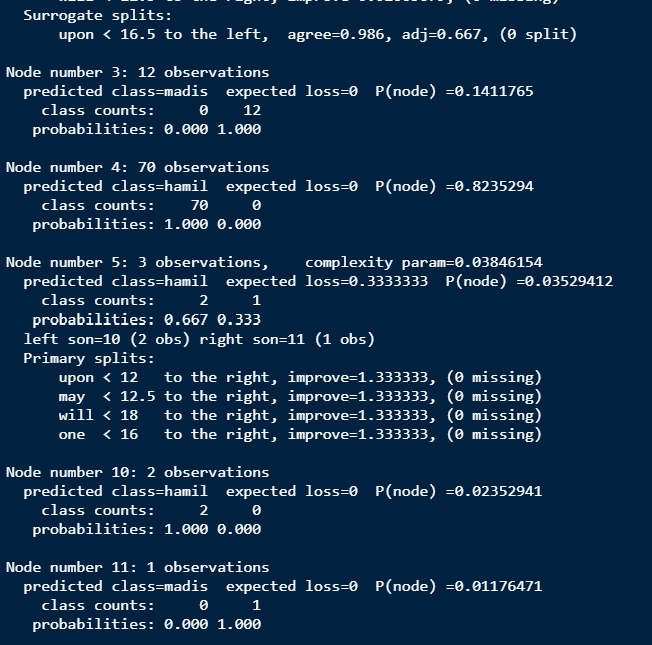
A picture containing text, clock

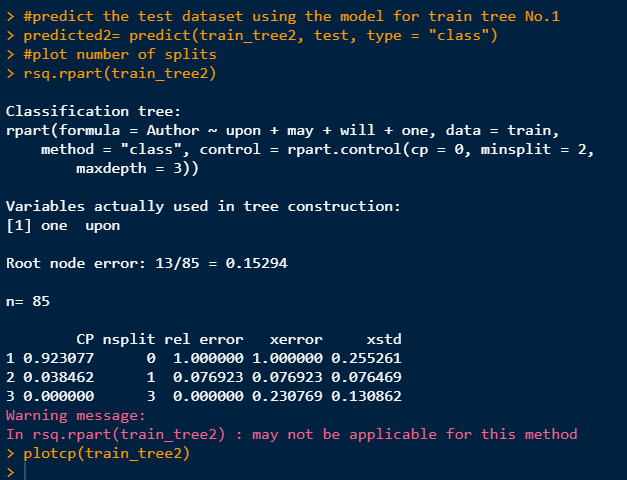
Description automatically generated

Text

Description automatically generated with medium confidence







Chart, line chart

Description automatically generated

Diagram

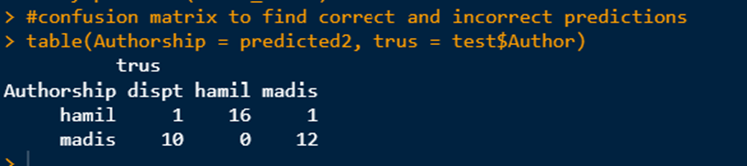
Description automatically generated

Text

Description automatically generated

A picture containing text, clock

Description automatically generated



**Conclusion**

Per HAC clustering and decision trees, Madison was the author of the disputed documents of the Federalist Papers.