Jeff Levesque

Professor Gates

IST 565

Homework #5

**Introduction**

Three founding fathers, Alexander Hamilton, James Madison, and John Jay greatly contributed to the ratification for the Constitution of the United States of America. In the papers they authored, The Federalist Papers, a consistent message was delivered in favor for signing the Constitution.

These founding fathers wrote the collection of 85 papers. However, 11 were disputed to have authorship of either Alexander Hamilton, or James Madison. This controversy motivated many to perform various analysis. In a previous result, using kmeans clustering, it was found that the disputed authorship belonged to Alexander Hamilton. However, it was also noted in the result, the analysis was likely flawed. Instead of discretizing the author, the removal would have produced better clustering results.

In this study, decision trees will be used to verify whether previous results were accurate. This will be done, by generating a train set using all papers except those of disputed authorship. Once a model is generated, a prediction will be made on the disputed authorship. Various error values, and tuning methods will be attempted to verify the integrity of the decision trees.

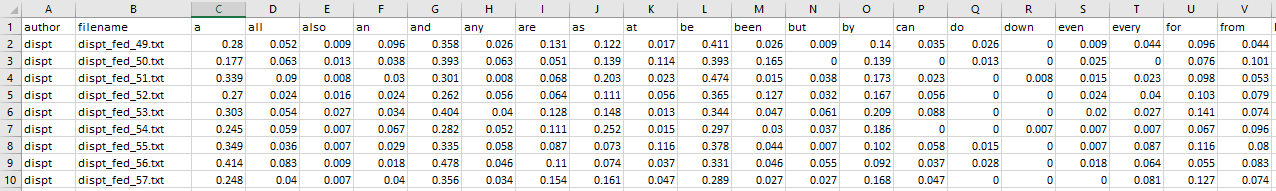
**Analysis**

Data Preparation:

A csv dataset representing instances authorship:

* Alexander Hamilton (**Hamilton**)
* James Madison (**Madison**)
* John Jay (**Jay**)
* Alexander Hamilton and James Madison (**HM**)
* Alexander Hamilton or James Madison (**dispt**)

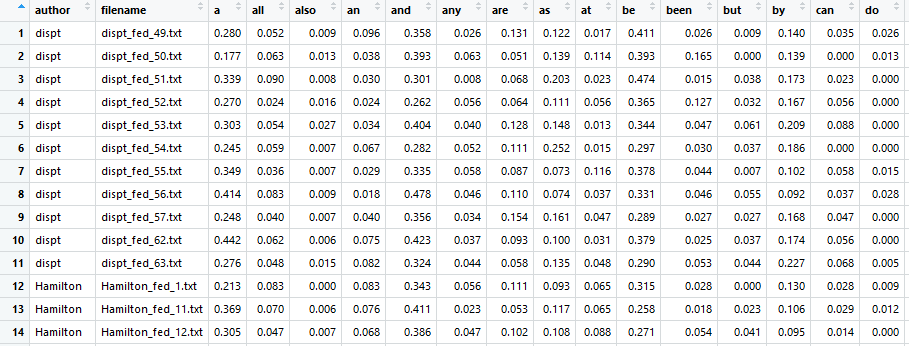
Each having a normalized frequency measure of 70 words:



The second column titled filename, corresponds to a supplied textfile, containing the content for the respective paper, with respect to the given author.

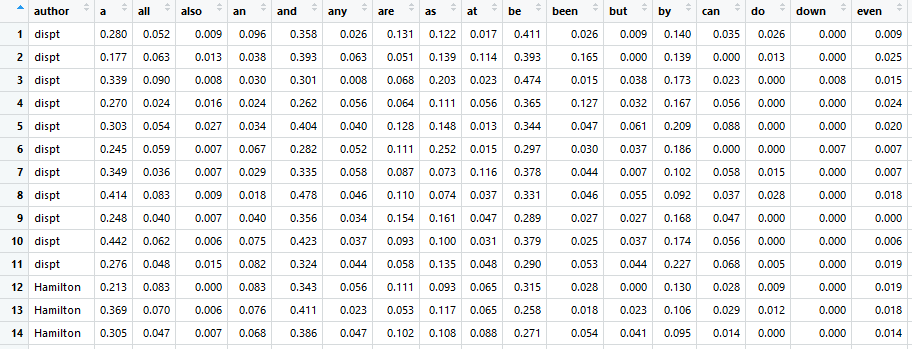
Processing:

The provided dataset was loaded into R as a dataframe, using standard read.csv():



Once loaded, some initial preprocessing was implemented. Specifically, the filename column was removed, since the supplied fedPapers85.csv, contained the same redundant file information as normalized values:

df = df[, -c(2)]



Once the needed dataframe were defined, the train set was defined as any row instance of the dataframe not containing dispt. Conversely