Rohini Shrivastava

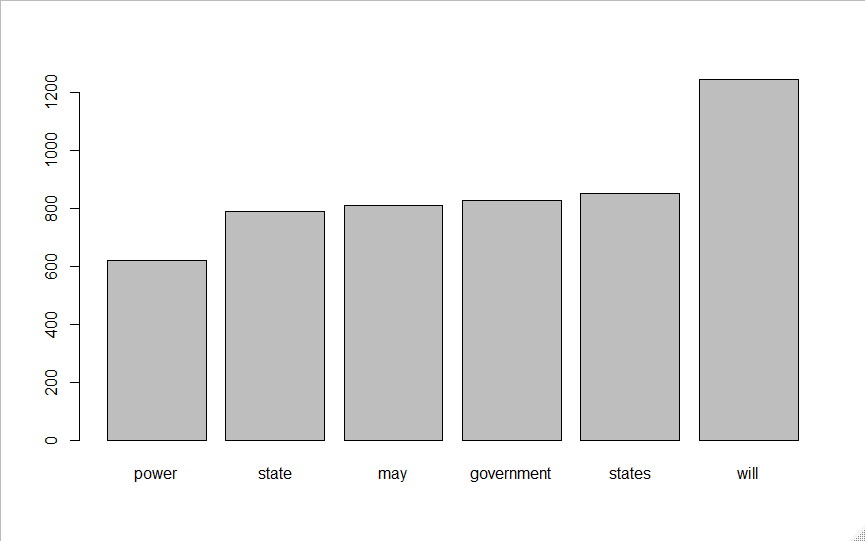
IST707

Homework 5

The Federalist Papers is a collection of 85 essays written by Alexander Hamilton, James Madison, and John Jay to promote the ratification of the United States Constitution. The papers were all published under one pseudonym making it hard to understand who wrote what. Over the years people have discovered the authors of the papers, however there are still 11 papers that are disputed between the four authors.

**SECTION 1: DATA PREPERATION**

First the Federalist papers were downloaded from the Library of Congress. There were 85 papers, each as an individual text file. The text files were imported into an R script as a corpus. The words for all the documents were vectorized and the five most used words were found for all the documents (*Figure 1).*

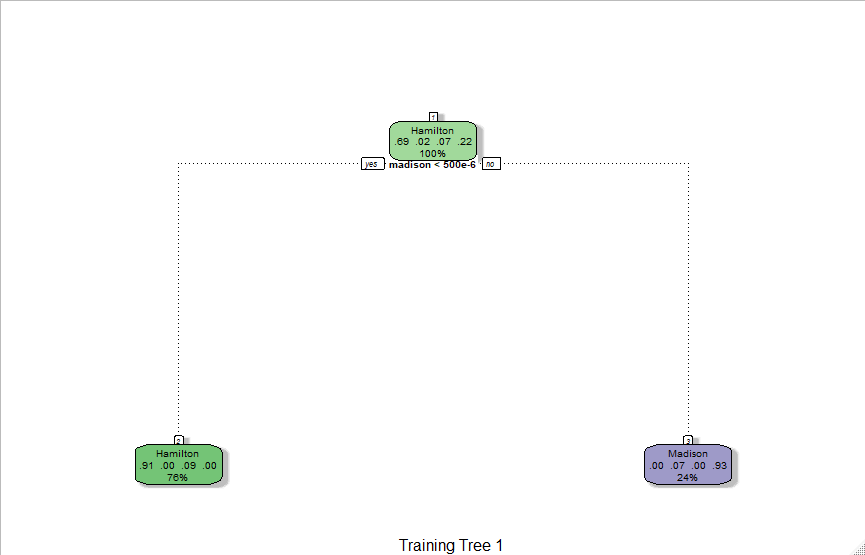


*Figure 1: Top 5 Most Used Words*

The corpus that was imported was converted into a dataframe. This dataframe had all the English stopwords removed. The disputed papers were removed when creating the training and testing data set as it would potentially throw the classifications off since the models are trying to determine who wrote the disputed papers. A training and test set were created using the dataframe without the disputed papers. 80% of the data was set for the training, and 20% was set for the test set. Once the decision trees have been tested with out the disputed papers, the best model will be run only on the disputed papers.

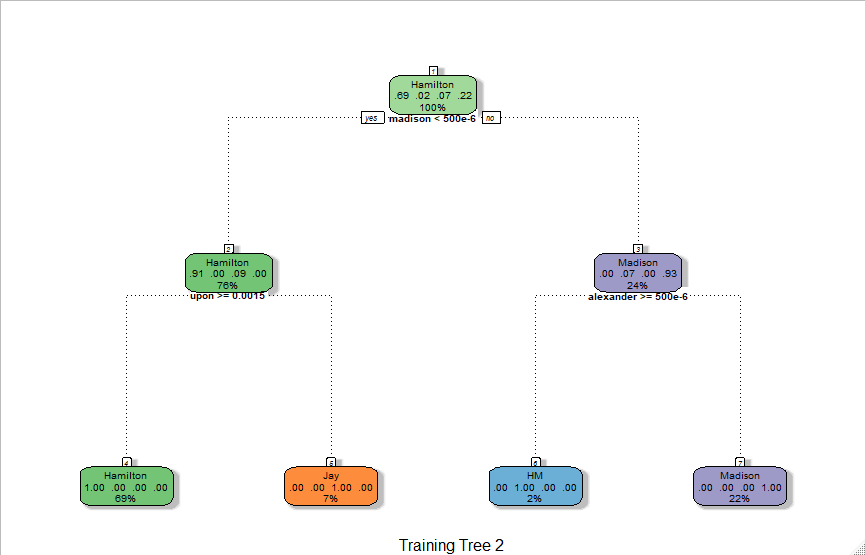
**SECTION 2: MODELS- DECISION TREES:**

Three different decision trees were created. The first one was a basic one with no additional parameters. This model just took the data and ran it with a “class” method. There was no fine tuning of the maximum or minimum depth. In Figure 2, the data only looks at Hamilton and Madison papers. There is nothing written by Jay or HM. 76% of the texts were categorized as Hamilton and 24% were categorized as Madison.



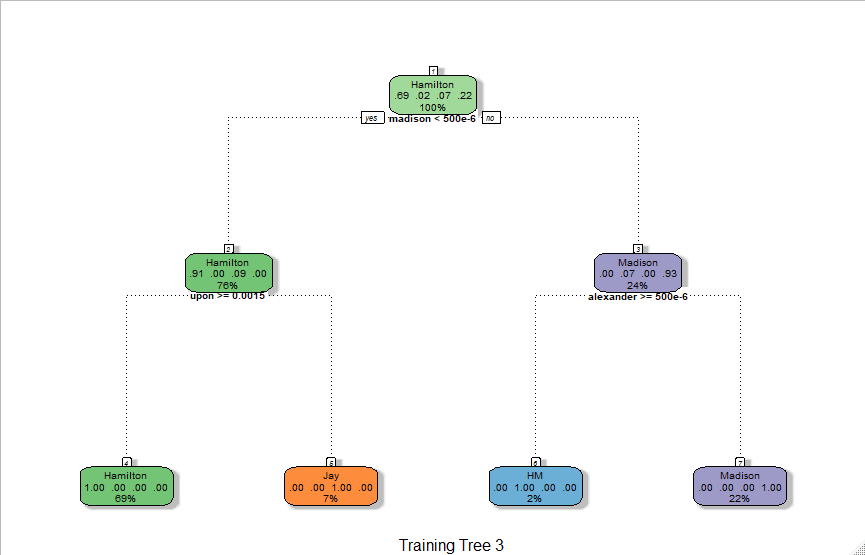
*Figure 2: Basic Decision Tree*

In the second decision tree that was run, there was minsplit of 2 and a maxdepth of 3. In this situation there needs to be a minimum of two observations in the node before a split is done, and the tree will go 3 nodes deep maximum (*Figure 3)*



*Figure 3: Second Decision Tree*

The third model had at least 1 bucket, with the minimum split being 7 and maximum depth as 8. The minimum bucket sets the minimum number of observations in the final node.

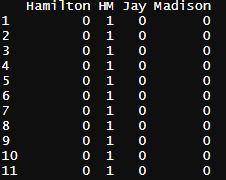
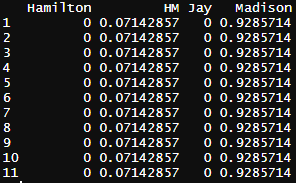


*Figure 3: Third Decision Tree*

From the three models, the third model had the highest accuracy at 100%. The first model had the lowest accuracy at 80%. The third model will be used to predict the disputed papers.

**SECTION THREE: PREDICTION**

The third decision tree was used to find the author of disputed papers. According to the model, the 11 papers were written by HM (both Madison and Hamilton) (*Figure 4).* The third model had an accuracy of 100%, so there is most likely some form of overfitting happening that we are unable to determine. In order to combat this, we will also run the second decision tree model to see how the two compare. In the second model, all 11 papers are pointing to Madison writing them (*Figure 5).*

*Figure 4: Model 3 prediction* *Figure 5: Model 2 Prediction*

In the previous clustering models, it was determined that the texts are closer to Hamilton’s writing style than Madison’s. The HAC clustering technique was more ambiguous than the K-Means model. The K-Means model was clearly pointing the disputed texts were Hamilton’s, however the HAC was left more to interpretation based off the dendric tree. In the decision tree, it is determining that the papers are either written by Madison or both of them. The decision tree model did not reach the same conclusion that the clustering algorithms did.