Nicholas Lichtsinn

IST 707

Homework 5

**Introduction**

The Federalist Papers were a collection of eighty-five essays pushing New Yorkers to ratify the newly formed United States Constitution. The pieces were first published anonymously in New York newspapers in 1787 and 1788 under the pen name “Publius,” and were written by Alexander Hamilton, James Madison, and John Jay. The Federalist Papers are regarded as one of the most important sources for analyzing and comprehending the Constitution’s original intent.

Seventy-four of the eighty-five articles have an identified author, however there are eleven articles that have a disputed author. Both Alexander Hamilton and James Madison claimed to have written these documents and the truth is unknown. This had led to the use of a variety of machine learning and natural language processing techniques to determine the correct author of the disputed articles.

K-Means and Hierarchical Clustering techniques have previously been used to figure out the appropriate author for the eleven disputed documents. This paper will utilize decision tree models to determine the author of the eleven disputed papers.

**Analysis**

Data Preprocessing:

The data files consisted of one csv file that contained the eighty-five text files term frequency of the 72 most frequent words found in the Federalist Papers. The dataset was loaded into R as a dataframe.

This dataset was already pre-cleaned so not much pre-processing was needed to start to work on the data. However due to the number of words the top 40% of words were taken to use for the models.

Decision Tree models were then created using training data from all the authors and from only articles written by Hamilton and Madison alone as these will help the model determine the author more clearly. The models were trained with a constant 65% mix of Hamilton to Madison papers as Hamilton wrote more and an uneven balance in the training set would skew results.

**Results**

The first unpruned model was created by taking out the disputed articles and using all articles to train the model and evaluate its accuracy on the training set. All models were created using cross validation with rpart and this prunes the tree back to the point of information gain perfectly.

A picture containing text, clock, screenshot

Description automatically generatedThis gave “Root node error: 14/47 = 0.29787”. Here is the plotted decision tree:

This model had a lot of noise, and the articles were not written by Jay or Hamilton and Madison together, so another model was created using only the articles written by Hamilton and Madison alone.

Diagram

Description automatically generatedThis unpruned model gave a “Root node error: 9/42 = 0.21429” and is shown below:

This model has a lower root node error and shows more clearly the model’s decision-making process for each author. The prediction table shows how the model split the articles correctly identifying all of Hamilton’s articles but misclassifying two of Madison’s.

Text

Description automatically generated with medium confidence

Text

Description automatically generated with low confidenceDiagram

Description automatically generatedA fully pruned model was created without using rpart.control to classify the minbucket and minsplit parameters giving the same Root node error: 9/42 = 0.21429 but correctly guessing all articles written by Hamilton and Madison article.

Using the unpruned and pruned tree model to predict the eleven disputed articles both models performed the same and predicted all eleven articles to be written by James Madison.

Graphical user interface, text, application

Description automatically generated

**Conclusions**

Using decision tree models to determine the author of the eleven disputed Federalist Papers found that the author for all eleven articles was James Madison. This finding is the same when using different clustering models. It is possible that there is some crossover between the two authors however as even the pruned model went back and forth on predicting Hamilton and Madison’s articles correctly so there is a lot of similarity in some of the texts.

More investigation could be done using a Random Forest model to see if there is any change in the prediction. The data could also be processed differently, and this might train the models to a different outcome, by using a 65% split of the data this study ensured the correct distribution, but different approaches might give alternative outcomes.

Natural Language Processing and alternative sentiment analysis techniques are becoming more powerful as technology progresses there will be more intuitive ways to look at the Federalist Papers and these new techniques might give a different result than was found here.