Diego Valdes

IST 707

Home Work 5

5/1/2019

**Introduction**

It is human nature to know. To know the answers to questions, to know the events of history, to know the behaviors of the animal kingdom, to know as much as possible about as much as possible. The nature for knowledge can be seen in many examples of society. Museums, documentaries, scientific journals, even religion are all ways that humans seek to gain knowledge from the unknown. A recent example is the unveiling of a picture of a black hole. The resources and technology to achieve creating a single image are beyond compare, but the payoff was knowledge. We know what it looks like now.

The Federalist Papers are comprised of eighty-five essays explaining to the citizens of New York why they should ratify the newly drafted United States Constitution. The writers included Alexander Hamilton, James Madison, and John Jay, but were published under the pen name “Publius”. Their authorship remained secret for thirty-one years. One problem arose from this reveal; a series of the essays, 11 total, had disputed authorship. Madison and Hamilton both claimed authorship of the works. It has been a mystery to historians the true authorship of the essays. They wanted to know.

Authorship attribution has become a more widely used technique. Simply stated, a writer’s use of words and the frequency of use is like a fingerprint. Recently, it was used to attempt to determine a similar conflict between Beetles songs written by either John Lennon or Paul McCartney. Like the Federalist papers, each song writer had written their own songs and they had collaborated on many works. This has been a common source of argument among fans, just as the authorship of the Federalist essays has been among historians.

**Analysis**

**About the Data**

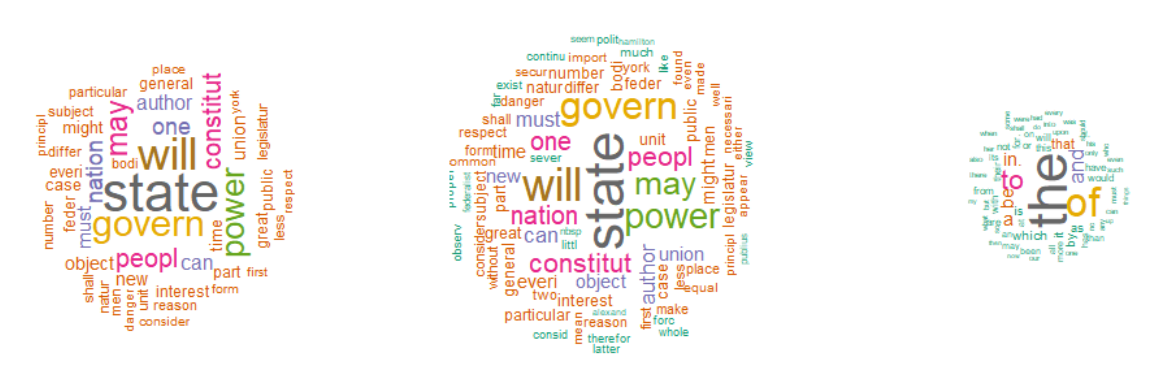
Three datasets generated from 85 essays (The Federalist Papers) with the authorship, Hamilton, Madison, Jay, HM (Hamilton and Madison) and dispt (unknown authorship that both Hamilton and Madison claim). The first two datasets were generated using the essays with the exclusion of function words (and, a, the, …). The data was formatted where each row is an essay, and the columns are the frequency of the word. 82 out of 4900 words were selected in this set based on the occurrence of that word across all the essays. These 82 words appear in 75% of the essays (fig.1.1)

Fig 1.1: word clouds for the 3 data sets

Each one of these was normalized differently; one based on the length of the essay (word frequency/essay length), the other scaled using max/min values (fig 1.2a,b).

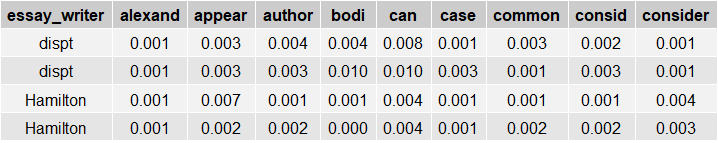


Fig 1.2a: scaled data based on essay length

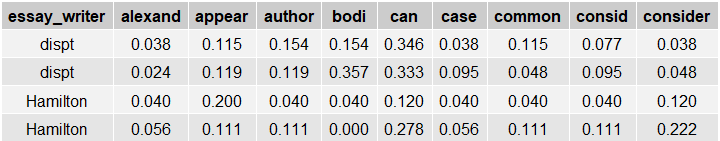


Fig 1.b2: scaled data based on max/min

The last dataset is formatted in the same manner as the others, normalized based on essay length, and contains 70 functional words as the column attributes (fig 1.3).

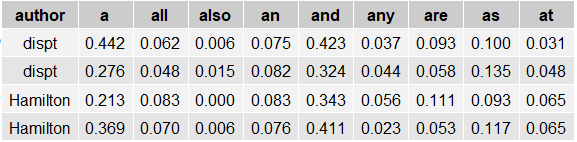


Fig 1.3: data with functional words

Neither dataset contained NA or missing values. To prepare it for analysis, the function word columns were converted to numeric values and the cols with ‘author’ and ‘filename’ converted to factors.

Authorship is not equally represented in the data (fig 1.4). Hamilton has more than triple the samples than the next author, Madison. This may become a problem in creating a model, as the datasets are not balanced with authorship. This may lead to HM and Jay being much less likely to be able to be recognized among the authors with greater representation.

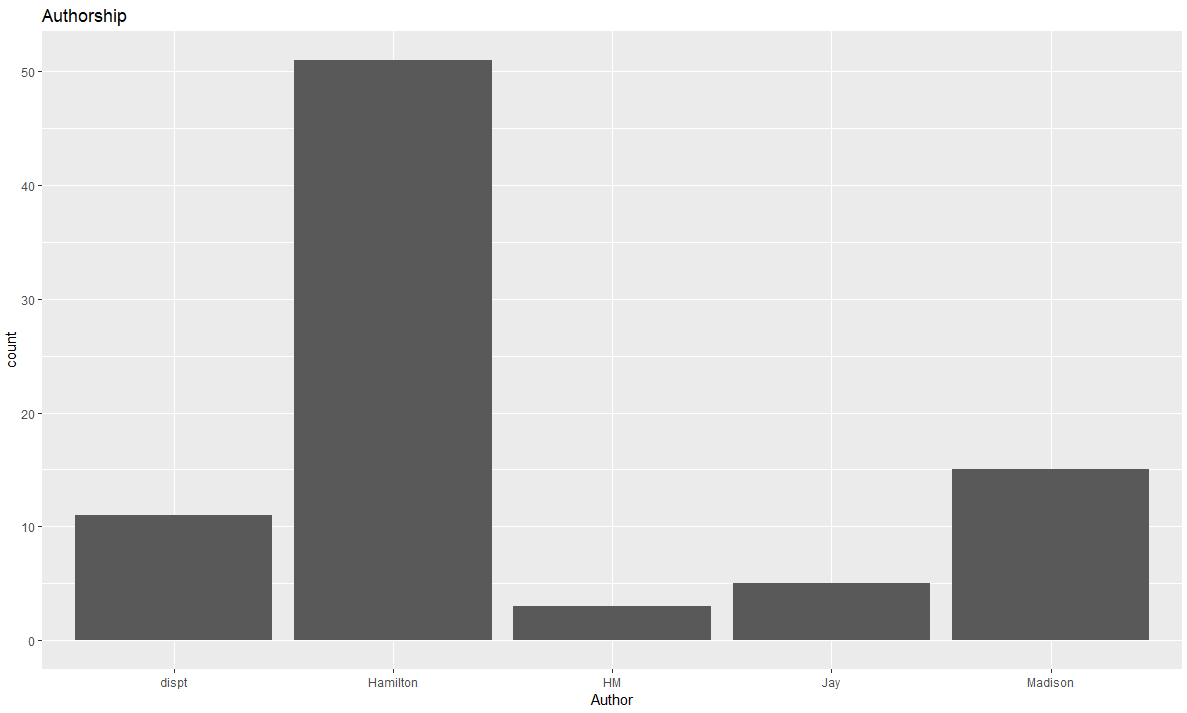


Fig 1.4: distribution of authorship in all 3 datasets

Visually examining the difference between the three datasets (fig. 1.5) shows that there is more variance within the two that are not using functional words. Although the scale is different for these two datasets, they seem to have many similarities in the words that are ‘hottest’ among the authors. The functional words: ‘the’, ‘of’, ‘to’, ‘in’, ‘and’, and ‘a’ are the standouts in the 3rd dataset.

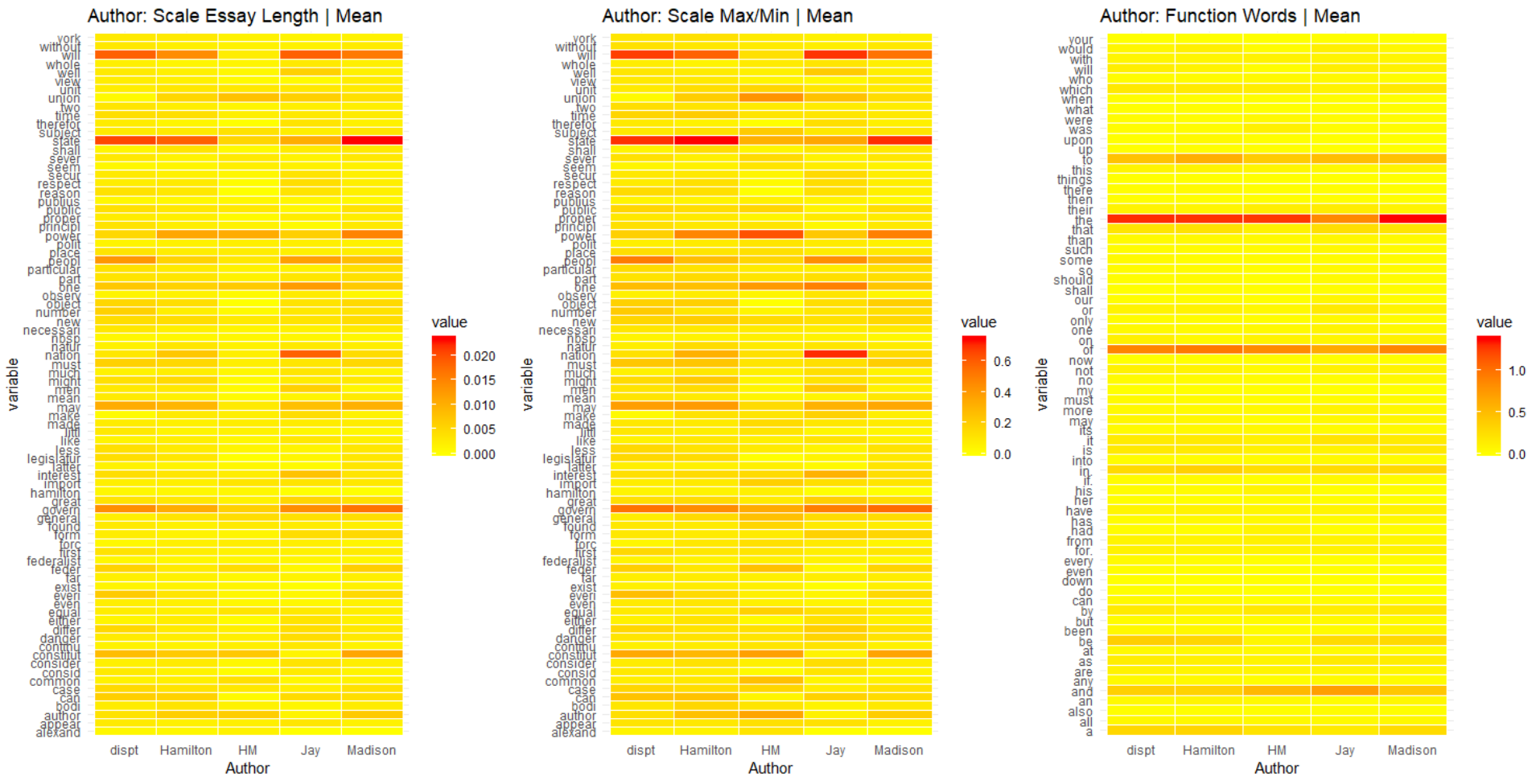


Fig 1.5: Heat maps of word frequencies by authorship

While both nonfictional examples have similar heat spots, they are hottest in words like ‘state’, ‘govern’, ‘power’, which are all in line with the type of vocabulary expected to be found in a persuasive piece favoring the Constitution. There are words that stand out per author. Jay seems to have liked the using ‘nation’ over his counterparts, and Hamilton and Madison had an affinity for ‘power’.

Two attributes, ‘Alexander’ and ‘Hamilton’ were removed prior to analysis from the datasets generated from the federalist papers. The guide of words appearing in over 70% of the essays included the author of the works.

**Models**

**Decision Tree (RandomForest)**

Each data set was used to create one decision tree model. Preparation of the training data included removing all disputed authorship essays. The remaining 73 observations were randomized and 70% selected as training data. The remaining 30% would be used to test the model’s accuracy.

Along with an error rate of the training data and an accuracy rating on test data, the model would also produce an importance rating for each word in the model, a confidence interval to determine the range of the true mean of the accuracy rating, and a Kappa score. Each model was run multiple times to record differences in all these measures. The final run of the models is what is being reported.

**Results**

Model 1 used the dataset scaled by essay length, model 2 used dataset scaled using max/min, and model 3 used the function word dataset. Models were ran multiple times and each time may have produced slightly different results. Final results for the run of each model are listed (fig 2.1)

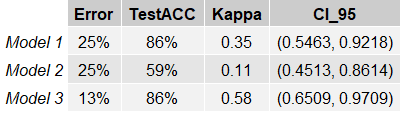


Fig 2.1: final results from last training run of 3 models using randomForest()

Originally, there was no plan to run each model more than once. After recording initial results, to create a different plot, the models were ran again. The results for model 1 where around the same as before, but model 2 and 3 produced different error and test accuracy. Due to this observation, it became of interest to run each model multiple times and record the differences in results. Although error and test accuracy could change drastically within each model, Kappa would only vary slightly. Other factors that wouldn’t change with each run are the importance of the words in each model (fig 2.2) and the confidence interval measured on the confusion matrix (fig 2.3) of the test set accuracy. Based on the Kappa score and confidence interval, model 3 performed the best.

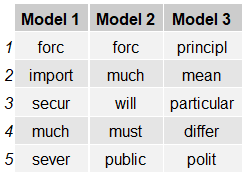
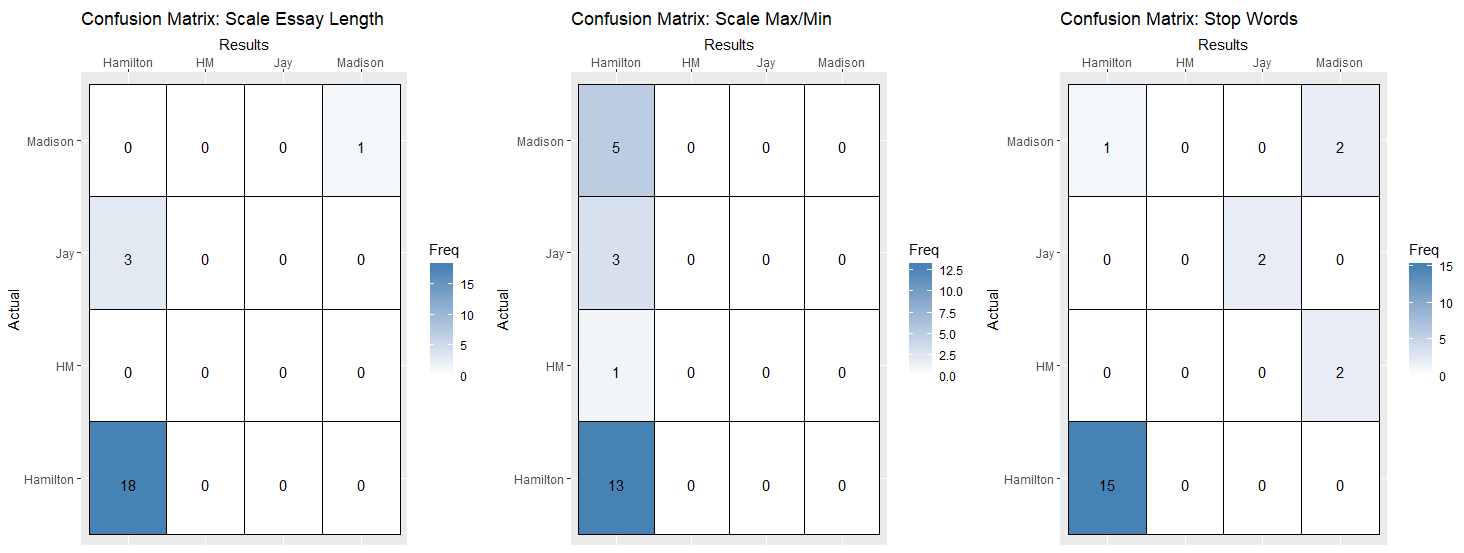
 

Fig 2.2: most important words per model Fig 2.3: confusion matrix, correct results on lower left to upper right line.

Although model 3 correctly identifies Jay’s writing in the final run, it was rare to find his and HM writings correctly identified. It was also more common to have all of Hamilton and Madison’s writings to be correctly identified, the errors being HM and Jay’s writings being included with them.

The last and most important factor that didn’t change with each run was the results when the model was used to predict the disputed papers. Both Model 1 and 2 predicted Hamilton as the author of most of the disputed essays, only 1 occurrence of Madison in model 1. Model 3 attributed authorship mainly to Madison, with only 2 essays being assessed to Hamilton (fig 2.4).

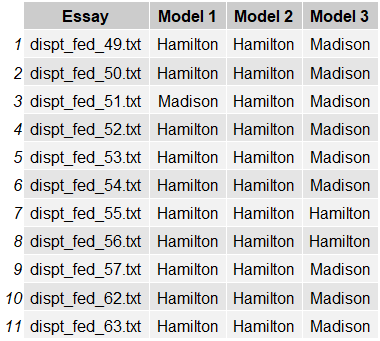


Fig 2.4: predicted results of models

Due to the data being unbalanced based on authorship, the same models were ran on 3 datasets with Jay and HM (Model A), Madison and Hamilton (Model B), and a balanced set with all writers (Model C).

Model A always had an error rate of 0, test accuracy of 1, and Kappa of 1. The prediction for disputed papers wasn’t considered, as there was no Hamilton and Madison representation, and Jay was predicted multiple times and historically he is not one of the authors considered to be a writer. Results with Model B were in line with results from model 3, but had 10 predictions for Hamilton and 1 for Madison. Model C performed the worst (Kappa: .14, 95% CI: (0.0063, 0.8059). This is most likely due to the small dataset, 16 observations.

Due to the findings from training these models, results from this analysis are not considered for part of the conclusion.

**Conclusion**

It is often said that writers have a voice. There is a way they sound, command the language, and use punctuation to control the beats and flow of their work. There is also an art in word choice, as a writer looking to persuade an audience to his/her view, will choose language to bring the reader to a certain point of view. Word choice, however, is more than mere hamper stance; for a writer it can be as identifying as a fingerprint.

Word choice, however, is not the only measure. Functional words, the words often overlooked in works are important as well. The usage of these words may be a better measure on their own versus non-functional words. These are the words that are not subjective to writer, they have to be used in order to communicate. The frequency of their use is just as much a part of an author’s voice as the usage of complex words that can be found by scavenging through a thesaurus.

Although there were 3 models, model 1 and 2 were constructed in similar fashions and had similar performance. When comparing them to model 3, it would appear that using the functional words is a more reliable approach to determining authorship. This is a flawed idea. Perhaps the issue with the model is in the approach, separating functional and nonfunctional words.

As it stands, model 3 is performing better, but it might be of consequence to further develop the model using nonfunctional words. This may be the key to providing a more accurate determination of the authorship of the Federalist Papers.