Machine Learning the Authors of the Federalist Papers

https://github.com/JSunde/FederalistPapers

Peter Giseburt
petergg@uw.edu

Jakob Sunde
jsunde@uw.edu

Abstract

The Federalist Papers were originally printed under a pseudonym, Publius. Since then it has been discovered that the papers were written by Hamilton, Madison, and John Jay. However, some of the papers are still of disputed authorship. To predict the authors of these paper, we will preform feature selection on frequently used words and then run the k-NN algorithm.

1 Project description:

We will take our data from Project Gutenberg, and then using the known authors we will convert the text into features and compare these features of known authors against the papers that are unknown or in dispute to attempt to determine the author. We can also use cross validation on known essays to test the accuracy of our model without using the test set.

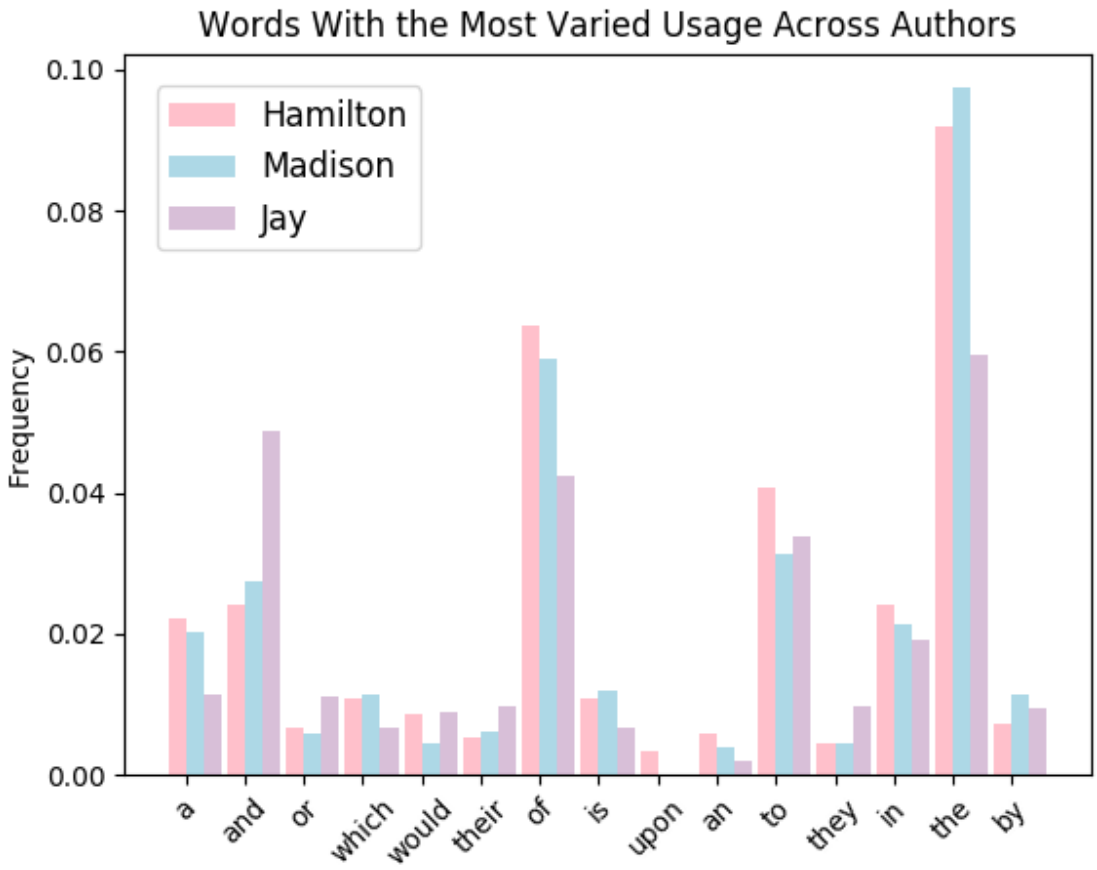
The idea of this task is that individual writers have individual styles, which can be represented in terms of how frequently certain words or phrases are used. From some research on the subject, certain "function words" are more indicative of a specific author, while other words, like nouns, can be highly influenced by subject matter and thus are not a good indicator of a specific writer [1].

Despite the main authors being Hamilton or Madison, John Jay is also credited with a few of the Papers. Due to this, we cannot use a simple binary classifier, as there are three potential authors. We chose to use k-NN as it allows for multiple different classifications, based on the sample created from the word frequency of the other authors. We chose words that we determined to have a high frequency across all documents, and then from those words used the ones that have the highest difference in frequency between authors. This is something we are uncertain about, and will look into other methods for choosing the words we consider in the future.

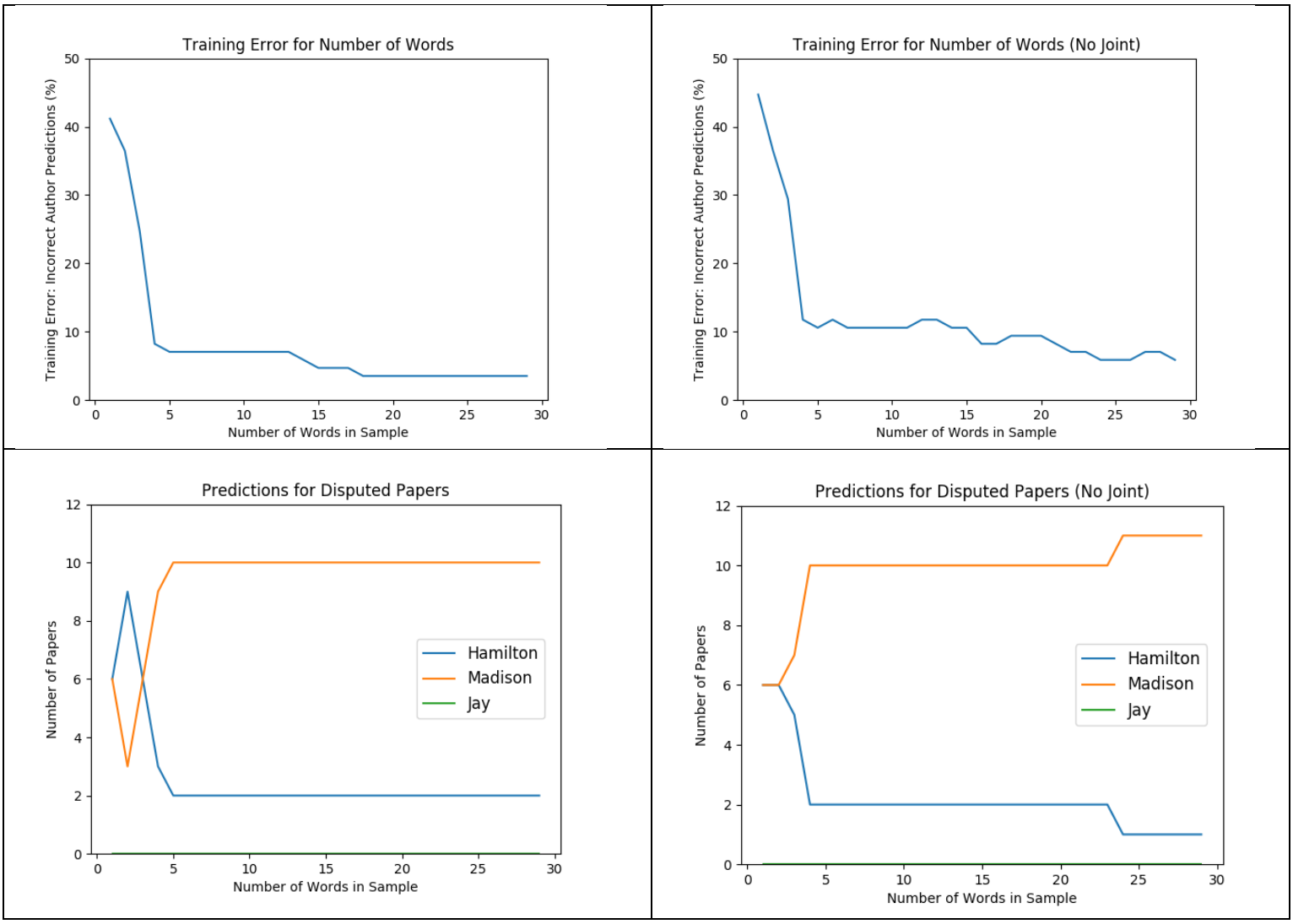
2 What we've done so far:

First of all, we gathered our data. We downloaded a .txt file of all the Federalist Papers from Project Gutenberg that contained all the papers, and wrote a parser that went through that file to collect the counts of the words used in each paper, which we converted into a frequency by the number of words in that paper. From knowing which authors wrote each paper, and the frequencies of words in each paper, we create an average word frequency vector for each author, as well as a global word frequency vector for all papers. We then calculated the distance from this overall average word frequency to each of the respective authors word frequencies, for each word. We summed this distance for all the authors to find the words that had a high variation from the mean frequency, which indicates a high difference from each other. Surprisingly, these words were mostly "function words," despite us not filtering for specific words at all. With these N (with N being the number of words that we use in our feature vector to classify texts) words, which we tested different numbers of N for, we had the features we could use to predict authors.

We then used k-NN to predict authors, both on the test data (unknown author papers) and the training data (known authors). We calculated the sample for each author by simply calculating the average word frequency for each of the chosen feature words over all of their known papers. We then tested the known papers against this sample for different N values, finding a large amount of error starting at N = 1, which sharply dropped until N = 5 and flattened out. For our test data, we found that it, after a shaky start, predicted Madison for 10 out of 12 papers from N=5 onwards. In Glenn Fung's paper, he talks about how his separating plane (with only 3 variables: to, upon, and would) predicated that all 12 disputed papers were Madison's, and other work on the subject agrees with him [1]. However, he used a separating plane, which would incorrectly classify John Jay's papers as either Hamilton or Madison and thus Jay's papers were ignored. All three of "to, upon, and would" appeared in the list of words that we highly valued.



Hamilton and Madison collaborated on some papers together, and as such should have an overlapping writing style for these papers, which may skew, the word frequency we expect from each of the authors. Upon removing the papers that both Hamilton and Madison collaborated on, we received similar, slightly mutated data. It had a slight increase in errors for the training data, however for high N's the model predicated 11 out of 12 to be Madison, which is close to the generally accepted result of all 12 being Madison's. The graphs titled with (No Joint) di not have the collaboration papers.



3 What We Plan to Do:

One of the things we plan to do is reexamine how to select the feature vector of words that we use to classify with. Both Fung and a paper he referenced predicated all 12 being Madison, despite each of them creating a separating plane with only 3 variables. We are looking into cross validation, among other methods, to select the words we use to determine the author for the future.

In addition, we will consider removing Jay from the data set. While it feels a bit like cheating, when trying to classify the disputed papers as Hamilton or Madison, Jay may be throwing our results off. Of course, its possible he wrote them, but neither our model or any other we have researched predicts him as the author of any. He also seemed to have a strong effect on the words chosen that have the highest difference between the authors, so maybe removing his data would make our results more clear.

Refrences

[1] Fung, Glen (2003) The disputed federalist papers: Svm feature selection via concave minimization. In: *Proceedings of the 2003 conference on Diversity in computing*, pp. 42–46. ACM Press, New York