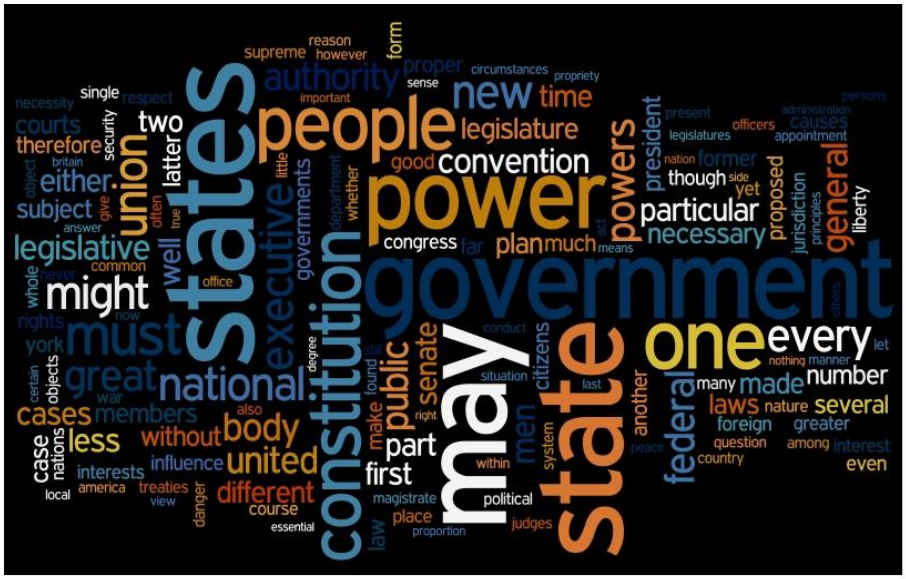
Artificial Intelligence Project 4

Text Mining

Code and Report by Caleb Pudvar



Introduction:

In authorship attribution analysis, writing styles of an unknown work are compared to styles of known authors to aid in determining the author of the unknown texts. These styles are categorized as vectors of 'features' for each paper. In this specific case, the features turned out to be frequencies of certain words. More detail will be provided on the formulation of these frequencies throughout. The authorship analysis of the Federalist Papers of Hamilton, Madison, and Jay is a classic homogeneous, closed data set for machine learning, and will be the topic of this report.

Text Preparation:

The text of the Federalist Papers was provided by Matthew Jockers, a professor at Stanford University. In order to split up the text properly, the Federalist Papers were initially a single text marked up into XML. Each paper was separated by "Div" tags, and important metadata was also accurately marked. Using python scripts, this XML document was split into a number of text files and relocated to directories categorizing papers by author. These categories are: 'Hamilton', 'Madison', 'Jay', 'Coauthored', and 'Disputed'. Text in each file was then lowercased and stripped of any unnecessary punctuation in order to determine word frequencies more accurately. The decision to solely use these frequencies as the features was done with the attempt to replicate the results found by Mathew Jockers and Daniela Witten on this same data set. Our results differ, because bigrams (sequential word pairings) were not counted in this analysis, whereas Jockers and Witten chose to include them.

Data Processing:

All results came from one of two versions of the data, which are to be known as 'raw data' and 'preprocessed data'. Raw and preprocessed data were computed as specified in Jockers and Witten (2010). The raw data set contains all words that Hamilton, Madison, and Jay each used at least once in one of their works. For example, say Hamilton and Madison both used 'government' in one of their works. If Jay never used 'government' in any of his writings, that word would be omitted from the raw data set. The resulting matrix for the raw set was of dimension 85 x 1211. The preprocessed data was a subset of the raw data with a mean relative

frequency of .05%. In other words, words in the preprocessed data must appear at least .05% of the time across the whole corpus, and must also be included in the raw data. The resulting matrix for preprocessed data was of dimension 85 x 208.

These matrices were then exported to spreadsheets where the data could be properly interpreted. First, a cluster analysis of both raw data and preprocessed data was done to visually represent similarities between texts. Then, the preprocessed data was run through CART (Classification And Regression Trees) which provided surprising results.

Results:

In order to see the results in a clear way, the dendrograms were produced and studied carefully. Due to their large size, the full dendrograms will be provided at the end of the report. There were some immediately noticeable features that were common to both dendrograms. Firstly, as shown in Figure 1, each of the Jay texts fell into order quite neatly without any error. This also shockingly occurs with the coauthored texts as well, represented in Figure 2. These patterns are apparent in both the raw and preprocessed sets.

Figure 1: Note that all papers by Jay have clustered together to form their own 'finger' of the dendrogram (source: preprocessed_dendrogram.png)

Figure 2: The three texts co-authored by Hamilton and Madison find each other beautifully

Of course, the main goal of the analysis was to determine the authorship of the disputed papers. As shown in Figure 3, a section of disputed papers filters into a branch Madison papers, indicating there are distinct similarities between the writing styles of Madison and those select disputed papers.

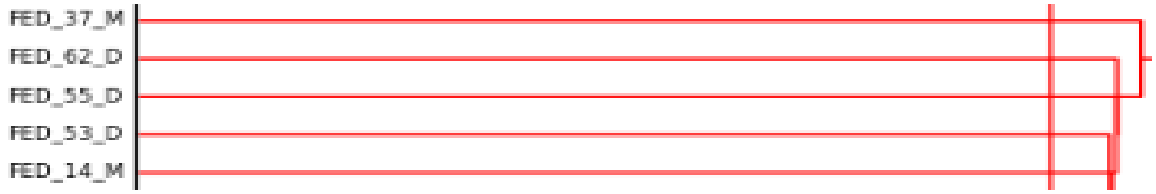


Figure 3: A selection of disputed papers clusters with two Madison papers, indicating similarities.

Unfortunately, the dendrograms were not successful in separating a large portion of the Madison texts from the Hamilton. This in turn makes further hypotheses to disputed authorship in these dendrograms inaccurate. However, when CART presented its results, the associations became astonishingly clear. As shown in Figure 4, a tree is presented. For each section, the texts included are represented by slashes. The order is as follows:

- Coauthored/Disputed/Hamilton/Jay/Madison

For each of these, the dominant author(s) is represented by a letter. To read this, take a look at the top branch breaking to the left. This branch is saying that if 'upon' is apparent in a text more than .125% of the time, then 48 out of 48 times that paper will be written by Hamilton. This chart is surprisingly able to break down each text to an author based on frequencies of only the words 'upon', 'union', 'people', and 'be' with great accuracy. One noticeable separation is on the first branch, where all but one Hamilton paper breaks from all others. This is indicative of the presumption that most of the disputed papers were written by Madison (Wikipedia). This is further suggested on the next branch, where the majority of the Madison papers and disputed papers break to the left on 'union'.

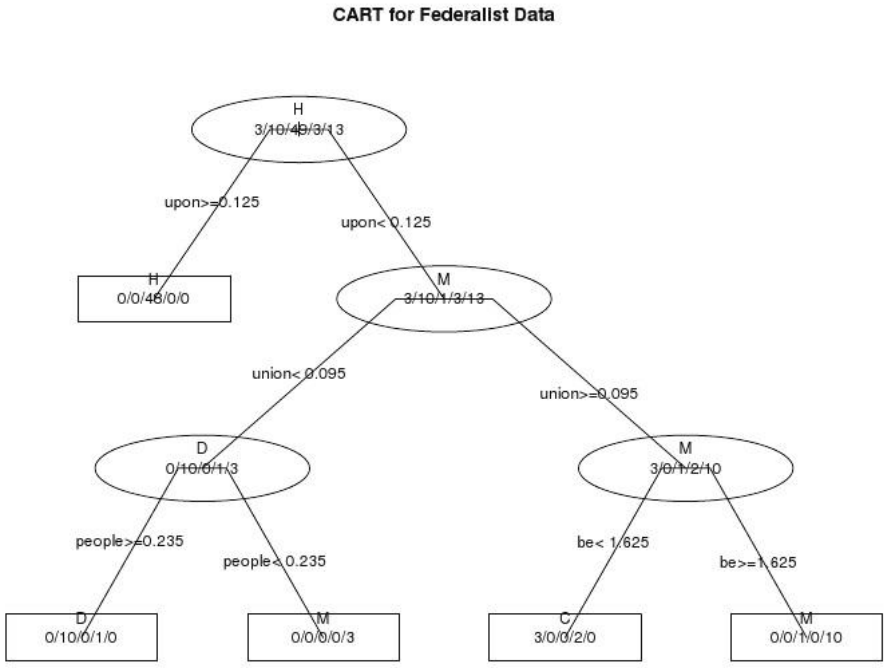
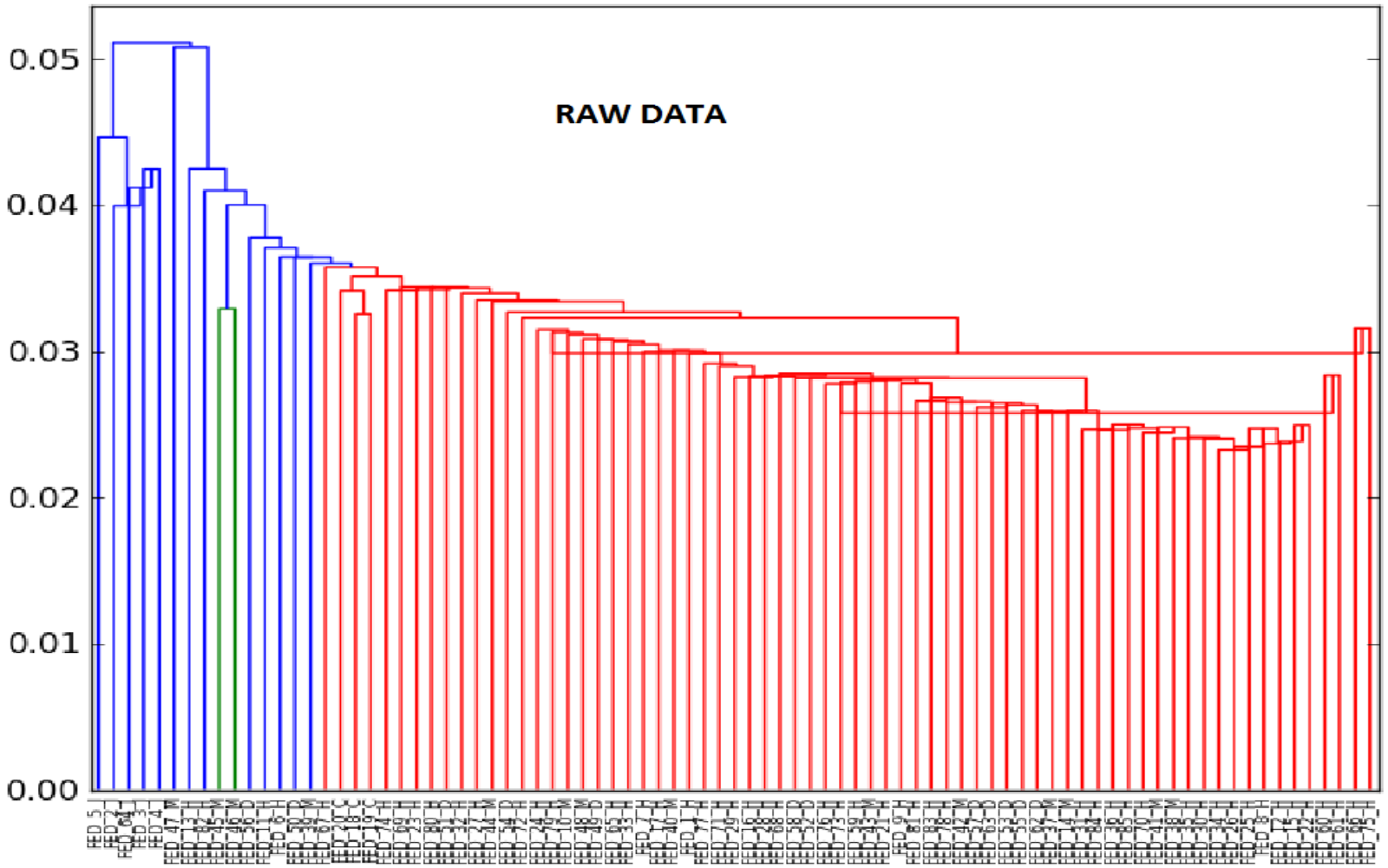


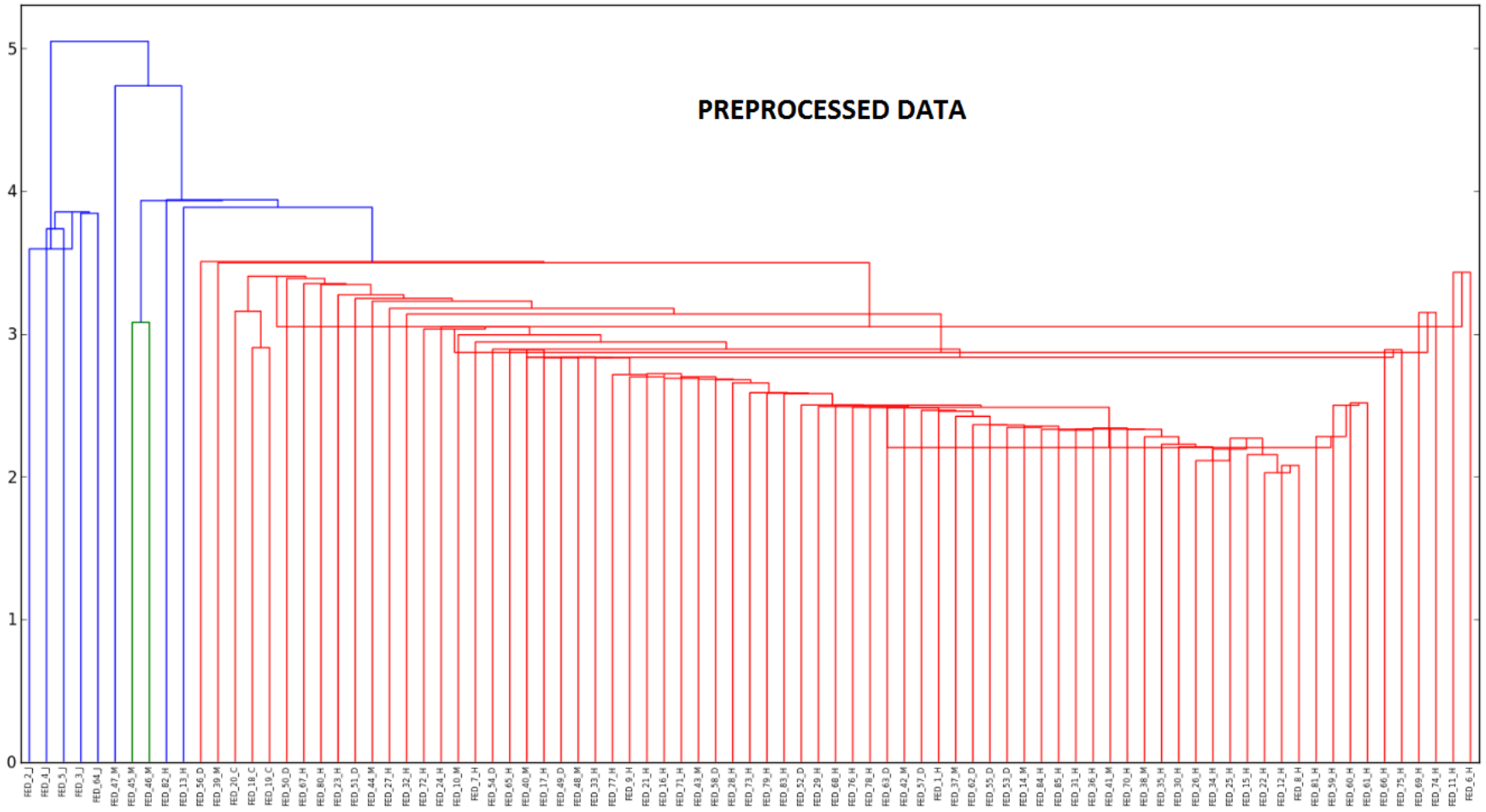
Figure 4: CART Tree separating papers by word frequencies. Note that Hamilton papers break away from all others, and Madison papers tend to track with the Disputed.

References

Federalist Papers. http://en.wikipedia.org/wiki/Federalist_Papers Accessed: 04/12/2013

Jockers, M.L. and Witten, D.M.(2010). A comparative study of machine learning methods for authorship attribution. *Literacy and Linguistic Computing,* v25(2), p215-223.


RAW DATA
0.05
0.04
0.03
0.02
0.01
0.00


PREPROCESSED DATA
5
4
3
2
1
0
FED_2_J
FED_4_J
FED_5_J
FED_3_J
FED_64_J
FED_47_M
FED_45_M
FED_46_M
FED_82_H
FED_13_H
FED_56_D
FED_30_M
FED_20_C
FED_18_C
FED_10_C
FED_50_D
FED_67_H
FED_80_H
FED_23_H
FED_51_D
FED_44_M
FED_27_H
FED_32_H
FED_72_H
FED_24_H
FED_10_M
FED_7_H
FED_54_D
FED_65_H
FED_40_M
FED_17_H
FED_49_D
FED_48_M
FED_33_H
FED_77_H
FED_9_H
FED_21_H
FED_16_H
FED_71_H
FED_43_M
FED_58_D
FED_28_H
FED_73_H
FED_79_H
FED_83_H
FED_52_D
FED_29_H
FED_68_H
FED_76_H
FED_78_H
FED_63_D
FED_42_M
FED_57_D
FED_1_H
FED_37_M
FED_62_D
FED_55_D
FED_53_D
FED_14_M
FED_84_H
FED_85_H
FED_31_H
FED_36_H
FED_41_M
FED_70_H
FED_38_M
FED_35_H
FED_30_H
FED_26_H
FED_34_H
FED_25_H
FED_15_H
FED_22_H
FED_12_H
FED_8_H
FED_81_H
FED_59_H
FED_60_H
FED_61_H
FED_66_H
FED_75_H
FED_69_H
FED_74_H
FED_11_H
FED_6_H