Bryan Jensen
COMP-255
April 14, 2013 *Note: Missing any form of supervised learning (technical issues)

Comparison of Machine Learning Techniques For Authorship Attribution

As Applied to the Federalist Papers

Abstract:

In a standard example of authorship attribution, this report details the application of machine learning techniques in an attempt to solve the classic problem involving the Federalist Papers, wherein exist essays of authorship status both known and contested. The main method of Machine Learning implemented in this report was one from the category of unsupervised algorithms, Cluster Analysis. Two different inputs sets were used in an attempt to improve and cross-validate the results, along with various tuning parameters of the algorithms involved.

Introduction:

Authorship attribution is a very fundamental application of machine learning (ML). The process involves using a set of papers whose authors are known and undisputed as a “training set” for the ML algorithm to later apply to the “test set,” the body of works with unknown or contested authorship. This follows along with the algorithms of supervised learning quite well, however in this paper unsupervised learning was applied in an attempt to see what results could be obtained from simply analysis of the clustering of like papers.

The input data sets for these ML techniques are called “features,” represented as vectors of attributes of the data to be analyzed. Put more concretely, the features were a list (vector) of the counts of the words (the attributes). The attributes to cross-compare were narrowed down to the high-frequency words in all papers, as this has been shown to be the most effective for authorship attribution (Burrows 2002; Diederich et al., 2000; Grieve 2007; Hoover 2003a,b; Koppel et al., 2007; Martindale and McKenzie 1995; Uzuner and Katz 2005; Zhao and Zobel 2005; Yu 2008). There is much less work detailing the best ML method, and so the classifier is the main focus of this paper.

may
state
states
government
power
must
people
one
constitution
union
national
authority
powers
great
public
general
executive
united
might
federal
every
new
members
body
different
particular
time
senate
first
necessary
part
laws
made
without
subject
two
rights
foreign
much
number
less
cases
men
either
interests
well
several
liberty
system
york
influence
place
whole
common
political
among
found
congress
danger
principle
form
make
far
law
little
latter
governments
another
case
nature
war
interest
circumstances
necessity
former
proper
nations
yet
even
though
within
often
many
means
country
degree
situation
certain
article
still
others
important
administration
given
appear
equally
long
nothing
sufficient
perhaps
answer
whether
never
causes
right
appointment
supreme
magistrate
articles
side
commerce
confederacy
last
single
now
force
period
true
officers
give
like
also
independent
always
present
equal
years
happen
manner
parties
judges
local
established
considered
three
thing
point
community
nation
publius
experience
constitutional
treaties
however
britain
consideration
measures
become
let
best

Feature Sets:

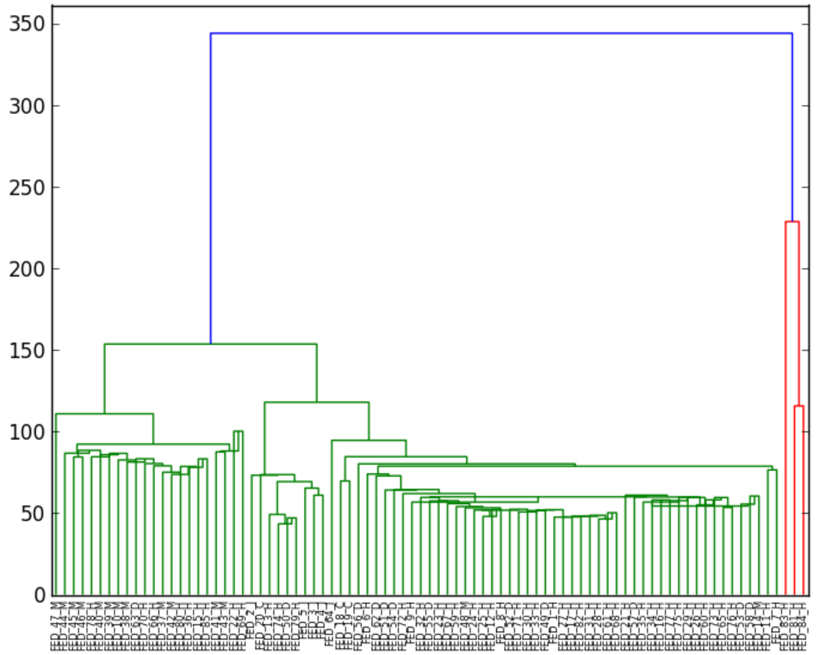
The feature sets input into the two different ML techniques were the "raw" (a slight misnomer) and the "preprocessed." The raw data set was the original data set with all words removed that weren't used by at least every author, at least once - meaning the only words kept were those used once by Madison, Hamilton *and* Jay, in any paper of theirs (note: in this case, the Disputed and Coauthored sections were also considered authors themselves). Further preprocessing consisted of removing anything below a certain overall mean frequency threshold of .05% - that is, if a word consisted of at least .05% of all the words across all the papers, then it was kept. This set was labeled as preprocessed. Both of these steps were done in accordance with previous studies showing that high frequency words give the most accurate results in authorship attribution (Burrows 2002; Diederich et al., 2000; Grieve 2007; Hoover 2003a,b; Koppel et al., 2007; Martindale and McKenzie 1995; Uzuner and Katz 2005; Zhao and Zobel 2005; Yu 2008).

Classifier Methods:

In the realm of unsupervised learning we used cluster analysis, the underlying idea of which is spatial proximity of the vectors in a graphical sense. As in the case of over 500 features, we cannot graphically represent such a high level of n-space and therefore we rely on computational analysis of numbers and the output thereof in graphical (dendrogram) form.

With the resulting dendrogram from the cluster analysis the ideal outcome is a clear grouping of the Hamilton papers separate from the Madison papers and the Jay papers, with the Coauthored and Disputed sorting themselves clearly into one of the groups. In order to better obtain an accurate and correct result the dendrogram has various linkage methods and methods of measuring the distance.

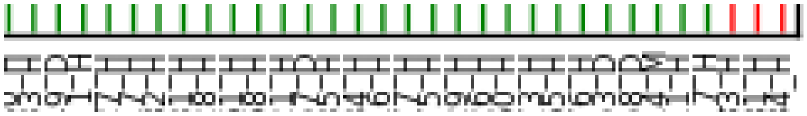
Example Dendrogram:



1.1: Dendrogram

Raw Data Set - Centroid Linkage - Euclidean Distance

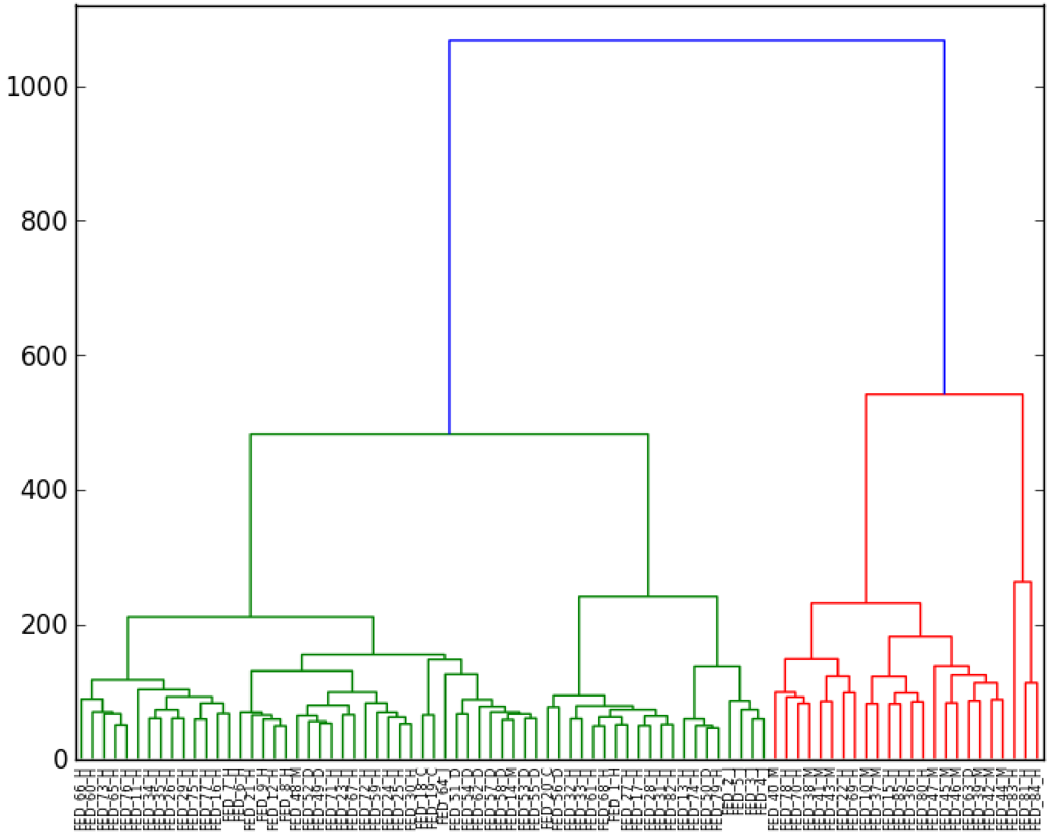
With the centroid linkage method it was clear (even though the labeling of the previous figure becomes near-impossible to make out) that not everything sorted out as we might've hoped. If we zoom in on the lower right:



1.2: Dendrogram 1.1 legend Zoomed-In

We can see some confusing results. For instance, there is a long string of Hamilton papers that are all very similar to each other and therefore link together first. And in the middle of that Hamilton run, there are a few Disputed papers, which indicates strongly the authorship of those papers was indeed written by Hamilton - however, right next to it is a Madison paper, one not contested at all. This throws some doubt on the results of this method, and other dendrogram linkage methods and distance metrics fare little better (all consulted dendrograms are included at the end).

The preprocessing of the feature sets combined with a different linking distance algorithm (Ward's) does indeed seem to have a positive effect on the resulting dendrogram:



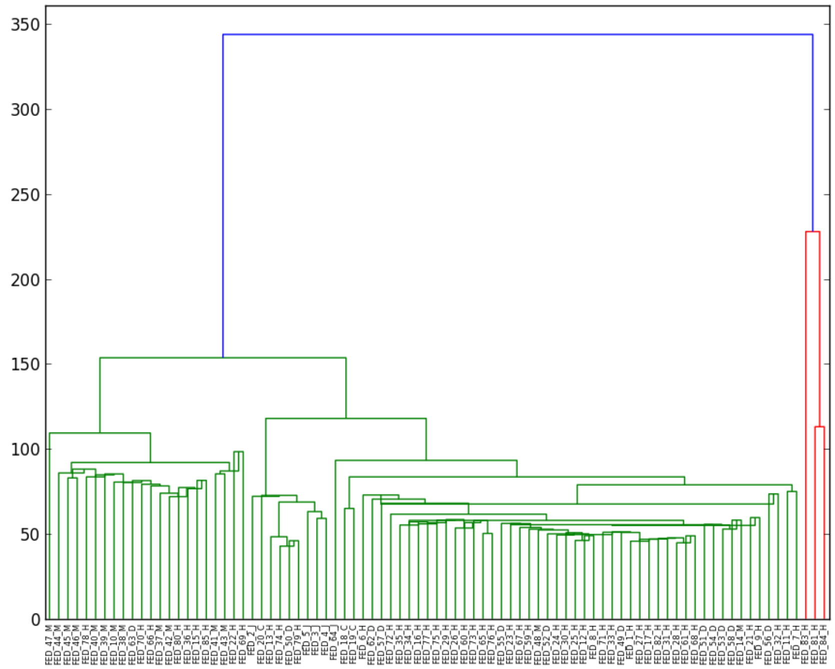
1.3: Dendrogram
Preprocessed Data Set - Ward Algorithm Linkage - Euclidean Distance

This dendrogram shows a much larger distance between the different papers and a rather clear sorting out of the two main authors. Also, an interesting phenomenon occurs where a majority of the disputed papers all are recognized as being very similar, slightly left of the middle of the graph. The presence of the Madison paper in the midst of this group gives hint to a possible similarity between the writing styles, but as there is only one from Madison, not much can be conclusively stated from that portion.

Conclusion:

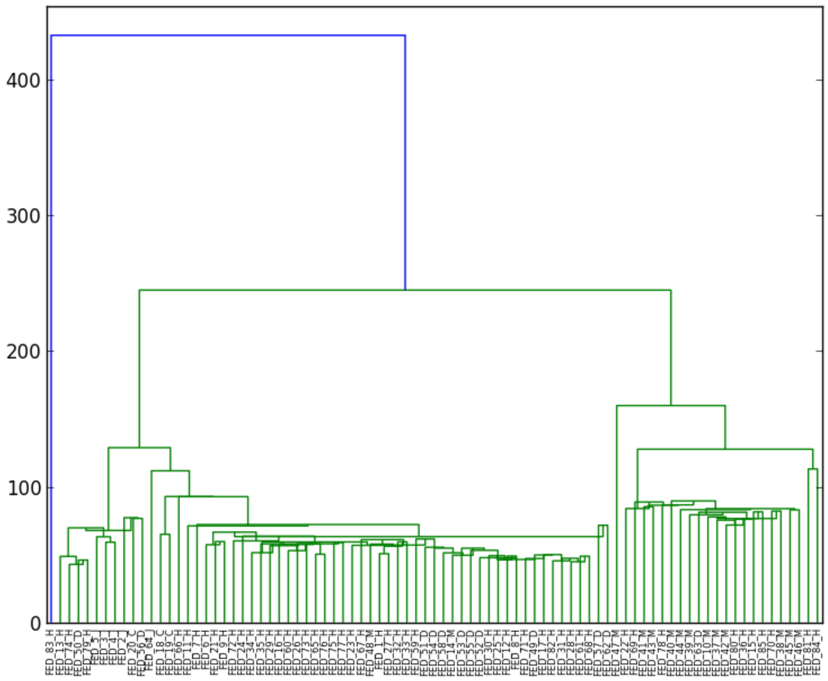
From the data presented purely in dendrogram form, there can be no clear consensus as to whether the majority of the Disputed papers actually belong to Madison as has been indicated by other such studies (Jockers and Whitten, 2010). With no supervision present in this technique, it is impossible to cross-validate and therefore get a better sense of whether or not any tuning changes made actually produced a better result from the data itself, or simply because we know the results and can alter the parameters based off of that. This can be no clear consensus from this paper alone as to where the majority of the disputed paper's true origins lie, but from previous experiments it seems the popular consensus is that Madison is the true author of those papers claimed by both Madison and Hamilton.

Consulted Dendrograms:



1.4: Dendrogram

Preprocessed Data Set - Centroid Linkage - Euclidean Distance



1.5: Dendrogram

Preprocessed Data Set - Median Linkage - Euclidean Distance

References:

Burrows, J. (2002). 'Delta': a measure of stylistic difference and a guide to likely authorship. *Literary Linguistic Comput: J Assoc Literary Linguistic Comput*, 17(3): 267–87.

Diederich, J., Kindermann, J., Leopold, E., and Paass, G. (2000). Authorship attribution with support vector machines*. Appl Intell*, 19(1–2): 109–23.

Grieve, J. (2007). Quantitative authorship attribution: an evaluation of techniques. *Literary Linguistic Comput: J Assoc Literary Linguistic Comput*, 22(3): 251–70.

Hoover, D. L. (2003a). Another perspective on vocabulary richness. *Comput Humanities*, 37(2): 151–78.

Hoover, D. L. (2003b). Multivariate analysis and the study of style variation*. Literary Linguist Comput: J Assoc Literary Linguist Comput*, 18(4): 341–60.

Jockers, M. L., & Whitten, D. M. (2010). A comparative study of machine learning methods for authorship attribution. *Oxford Journals*, *25*(2), 215-223. Retrieved April 15, 2013.

Koppel, M., Schler, J., and Bonchek-Dokow, E. (2007). Measuring differentiability: unmasking pseudonymous authors. *J Mach Learn Res*, 8: 1261–76.

Martindale, C. and Mckenzie, D.(1995). On the utility of content analysis in author attribution: The Federalist. *Comput Humanities*, 29(4): 259–70.

Uzuner, O. and Katz, B. (2005). A comparative study of language models for book and author recognition. *Lecture Notes in Computer Science*. Berlin: Springer.

Yu, B. (2008). An evaluation of text classification methods for literary study. *Literary Linguist Comput*, 23: 327–43.

Zhao, Y. and Zobel, J. (2005). Effective and scalable authorship attribution using function words. *Lecture Notes in Computer Science*. Berlin: Springer.