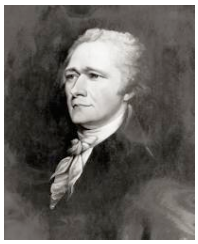
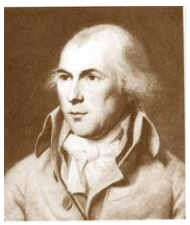
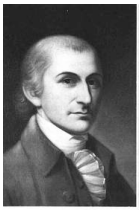
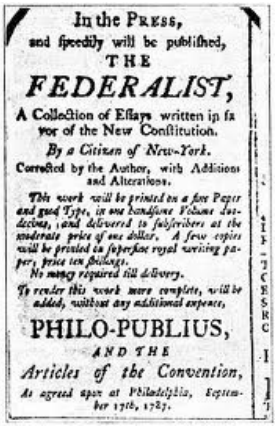
Authorship Attribution and the Federalist Papers


Hamilton


Madison


Jay

Summary:
You will replicate and extend the authorship attribution experiments of Jockers and Witten (2010) on the Federalist Papers. The **Federalist Papers** are "a series of 85 articles or essays advocating the ratification of the United States Constitution"(Wikipedia, 2011). In particular, you will use a number of machine learning techniques to gather evidence as to who *may have* written the twelve disputed papers and the three co-authored papers.


In the PRESS,
and speedily will be published,
THE
FEDERALIST,
A Collection of Eſſays written in fa vor of the New Conſtitution.
By a Citizen of New-York.
Corrected by the Author, with Additions and Alterations.
This work will be printed on a fine Paper and good Type, in one handſome Volume Duodecimo, and delivered to ſubſcribers at the moderate price of one dollar. A few copies will be printed on ſuperfine royal writing paper, price ten ſhillings.
No money required till delivery.
To render this work more complete, will be added, without any additional expence,
PHILO-PUBLIUS,
AND THE
Articles of the Convention,
As agreed upon at Philadelphia, September 17th, 1787.

(0) Read Matt Jockers and Daniela Witten's 2010 paper (see the starter kit for a .pdf of the paper). Which of the 85 papers gave some machine learning techniques a challenge? (that is, for papers where we *know* the author but the models often got them wrong?) Which author is most often attributed to the Disputed papers? You don't have to exactly replicate these results, but they form a "benchmark" set of results.

Methods

Workflow Part I – The Beginning Game

(1) Typically, you'd have to (a) find the digitized texts, (b) fix Optical Character Recognition (OCR) and/or other "errors", (c) parse the text, (d) count the words in each text, (e) save the proportion of counts in a dictionary, and (f) export the "matrix" to a .csv file. But "*hear yea, hear yea*" … this part has been done for you. See the files:

`lexos_1gram_inAll85_prop.csv` and
`lexos_2gram_top100_prop.csv`

Note: these files were produced with the *Lexos* web app. For each file, scrubbing settings were: whitespace (newlines, tabs, runs of spaces) denote "word" boundaries; convert everything to lowercase; remove digits; remove any punctuation *except* keep apostrophes, so for example we'd keep "`we'd`" as one token); hyphens were removed; no stop words were removed.

The initial file has proportional counts of "single" (1-gram) words that appear in all of the 85 texts. The second file has the proportional counts of "word pairs" (bigrams or 2-grams). This second file may contain some bi-grams that do not always appear in each text, that is, some zeros may appear if an author did not use that bi-gram in a specific text.

(2) **Refining the data:** The given datasets may not be completely ready for "prime time". As always, you should explore, check, and refine your starting data. For example, do you like the orientation of your dataframe? Do you want the data that shows the Totals and Averages? Should you even keep the Co-Authored texts? Remember, we are most interested in seeing if (a) our models can correctly predict authorship *for the authors we know* and if so, (b) what do our models predict for the Disputed and/or Co-Authored texts?

Regarding features (e.g., 1-gram words or bi-grams), what if a "token" is only used once in one text? Would it be better to keep only words that are used by each of the authors at least once within one of their texts? Or perhaps you want to be more stringent and only keep words that are used in all the texts at least once?

Whatever decisions you make, document your Methods by discussing any changes you make in your final notebook.

Workflow Part II – The "Middle Game"[1]

Using 1-grams initially and then using bi-grams:

(3) confirm the authorship of papers when we know the author using:
   (a) a clustering technique
   (b) a classification technique

(4) generate evidence of predicted authorship for the disputed and co-authored papers using:
   (a) a classification technique

Workflow Part III – Discussion

(5) Write up the results of your experiment in a Jupyter Notebook.

References

Federalist Papers. http://en.wikipedia.org/wiki/Federalist_papers Accessed: 03/08/2011

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

[1] When I lived in Australia during 2004-2005, I met Dr. John Burrow's who devised the Delta authorship attribution technique. John was very modest about his computational techniques, calling them *only* middle game techniques. That is, there is much scholarship to do at the start, we then perform computational experiments, but then there is still much scholarship to do afterwards.