Department of Statistics

MASTER'S THESIS PRESENTATION

JASMINE TAN

Department of Statistics
The University of Chicago

Text Mining and Authorship Classification: The Federalist Papers Revisited

FRIDAY, October 23, 2015, at 9:00 AM
Eckhart 110, 5734 S. University Avenue

ABSTRACT

This study revisits the problem about the authorship of the Federalist Papers. The purpose of this paper is twofold: to verify the authorship allocation of the twelve disputed Federalist Papers and share of authorship of the three joint papers as presented in previous historical and statistical papers; and to propose optimal datasets of words that can be used toward solving future authorship disputes. Rates of words are the variable of choice used for classification, where the selection of the type of word is of particular concern here. In particular, we use "function" words, which are invariant under different topics, to distinguish between authors. By using the Federalist Papers of known authorship as the test set, we establish elastic-net models to predict the authorship of the disputed Federalist Papers. After analysis, we find that high-frequency "function" words are optimal for distinguishing between authors. We also find compelling evidence to suggest that Madison was the author of all the disputed papers and the main executor of the joint papers.

For information about building access for persons with disabilities, please contact Laura Rigazzi at 773.702-0541 or send an email to lrigazzi@galton.uchicago.edu. If you wish to subscribe to our email list, please visit the following web site: https://lists.uchicago.edu/web/arc/statseminars.