STUDENT VERSION

Title:

Who Wrote That: Stylometry Investigations

Summary of the project

Who wrote Macbeth? Scholars agree that William Shakespeare (WS) is the author of Macbeth, but there is not agreement on the author of the play Thomas Lord Cromwell. In fact, there is near universal agreement that WS is not the author of the play Thomas Lord Cromwell. What evidence is available to help scholars reach such conclusions? "Stylometry" is a collection of techniques that are useful for discovering and matching patterns in samples of prose. What kinds of patterns? Things like the distribution of word length, i.e. three letter words, four letter words, etc. A bit more sophisticated would be the frequency distribution of specific words, i.e. road, castle, blood, shadow, etc.

Stylometry has been done by hand but computer applications make the cleaning of writing samples and the counting of words or word length easier, speedier, and more accurate.

This module will explain some of the techniques of stylometry and will take the user step by step through an example of using the language R and the package STYLOMETRY to investigate the connection of Samual Clemens (Mark Twain) to the Snodgrass letters which were published in 1861 in the New Orleans Daily Crescent. (Quintus Curtius Snodgrass). There is also an easier example using Minitab.

Instructions to students:

To prepare for this exercise read the brief Wikipedia article on stylometry [1]

In a short module it will not be possible to conduct an in-depth analysis of how stylometry is handled in Minitab or in R. We will make reliance on your professor to supply you with either a frequency distribution of word lengths found in two documents that you are comparing (Minitab example) or (stylo( ) example) to point you to two (or more) plain text samples of authors you wish to compare.

If doing a Minitab exercise investigate the meaning of the p-value that Minitab gives as output from they hypothesis test of similarity of the word length distribution in the two samples. Are the distributions too similar to be attributed to chance? What do you conclude about authorship of the two documents?

If doing a sytlo( ) example, follow your instructor's directions to start RStudio, add the stylo( ) package, and run stylo( ), selecting as output the Cluster Analysis. Reference [2] will help you learn how to interpret the resulting Cluster Analysis dendogram.

References:

[1] https://en.wikipedia.org/wiki/Stylometry

[2] For help on interpreting dendrograms the article linked below is very helpful.

https://wheatoncollege.edu/wp-content/uploads/2012/08/How-to-Read-a-Dendrogram-Web-Ready.pdf

[3] The package stylo

https://cloud.r-project.org/web/packages/stylo/stylo.pdf

[4] Mark Twain and the Quintus Curtius Snodgrass Letters: A Statistical Test of Authorship Author(s): Claude S. Brinegar Source: Journal of the American Statistical Association , Mar., 1963, Vol. 58, No. 301 (Mar., 1963), pp. 85-96 Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association Stable URL: https://www.jstor.org/stable/2282956

[5] A tutorial on using stylo( )

http://coltekin.net/cagri/courses/lingdiff/tutorial.html