Instructor Notes

Disciplines/courses suitable for this project: This project is an appropriate supplement to any introductory course in history or literature or political science, could be used in Religion or Theology.

Degree of difficulty: Depending on the amount of setup and post-project discussion on interpretation of the results, this project is classified as M (moderately difficult), or **A** (advanced.)

Software: The code given in this set of notes is an R Script. It uses the package "Stylometry" and can easily be ran by copying the included script into the RStudio Script window. The connection to computer programming can be near zero if you explain to a class what the computer is doing and present them with the output, then spend time deciding what the output is telling them. A more hands on approach would to be sure they have access to RStudio and give them the script. Let them build a corpus of text and select the output they desire. Because I cannot put my hands on the full text original Quintus Curtius Snodgrass letters, but do have access to counts of word size in both a subset of the Quintus Curtius Snodgrass papers and an uncontested sample of Mark Twains writings as a first example I have used this data and Minitab to do a quick chi square goodness of fit test on the hypothesis "Distribution of Word Length is the Same."

Expanding the Study: An interesting second step would be to compare undisputed work of Mark Twain to letters of "Thomas Jefferson Snodgrass" which are three letters contributed as travel letters to the Keokuk Post (Keokuk Iowa) on November 1, November 29, 1856 and April 10, 1857.

About running individual or group project: This project can easily be presented either way. I would lean toward teams of two doing research on the output from the stylometry package and interpreting the results.

Discussion on duration of the project: This module will take at least two class periods. To fit it into two class periods I suggest assigning readings about stylometry and about cluster analysis before starting the project.

Particular notes for Faculty:

First Example: Frequency Counts for the first three QCS letters and a sample from MT

length	QS-A	MT-A	
2	997	349	
3	1026	456	
4	856	374	
5	565	212	
6	366	127	
7	318	107	
8	258	84	
9	186	45	
10	96	27	
11	63	13	
12	42	8	
13	25	9	

What Minitab Does with these counts.

The MT-A counts were converted to relative frequencies and the Total Counts of QS-A were multiplied by the relative frequencies to obtain Expected Counts. Then a Chi Squared "goodness of fit" test was applied to the Expected Counts and Observed Counts which are the counts in MT-A.

In Minitab-20 the menu selections needed to do this analysis follows:

STAT > TABLES > Chi-Square goodness of fit (one variable) > and in the resulting menu supply these values: Observed Counts (QS-A), Category Names (length), Proportions Specified by Historic Counts (MT-A)

Among the possible outputs that can be selected are:

Observed and Expected Counts

Category	Observed	Historical Counts	Test Proportion	Expected	Contribution to Chi-Square
2	997	0.192711	0.192711	924.63	5.6646
3	1026	0.251795	0.251795	1208.11	27.4513
4	856	0.206516	0.206516	990.86	18.3556
5	565	0.117062	0.117062	561.67	0.0198
6	366	0.070127	0.070127	336.47	2.5918
7	318	0.059083	0.059083	283.48	4.2030
8	258	0.046383	0.046383	222.55	5.6480
9	186	0.024848	0.024848	119.22	37.4042
10	96	0.014909	0.014909	71.53	8.3688
11	63	0.007178	0.007178	34.44	23.6798
12	42	0.004417	0.004417	21.19	20.4224
13	25	0.004970	0.004970	23.84	0.0560

Chi-Square Test

N	DF	Chi-Sq	P-Value
4798	11	153.865	0.000

The Null Hypothesis was that the word length proportions are equal, suggesting that the two works were by the same author. The test has a p-value of 0.0000 indicating that the Null Hypothesis is to be rejected and we conclude that the analysis does not suggest common authorship between CS-A and MT-A, i.e. Mark Twain is likely not the author of the Quintus Curtius Snodgrass letters.

The language R and the package "stylo" can conduct a more sophisticated analysis using the frequency of words, not just the count of word length. As an example consider a paper by a Mathematician and another paper by a Theologian. The math paper may have a high count of the word "equal" and the theologian may repeatedly use the word "agape." If you are <u>only</u> counting word length, the use of these signature words will not help you distinguish authorship because both are being counted as examples of five letter words.

As an example of R and the stylo( ) package I installed stylo( ) and attached it to my copy of RStudio with the library(stylo) command. The only other preparation step is to build a sub directory in your working directory, it is by default called "corpus" and in that sub-directory put plain text versions of the works you intend to compare. For illustration I have built my corpus from the memoirs of U. S. Grant and W. T. Sherman. In this example I am not investigating authorship but rather looking at how much similar education and experiences influenced the writing of these two generals.

The one line command, "RESULTS<-stylo( )" , breaks the two works into numerous samples, prepares a list of most frequent words and counts their frequencies, and identifies the samples that are closest to each other (clusters). Many of us will recognize these steps from earlier less powerful programs in which we were required to perform each step manually.

<u>An R Script</u>

```
# Stylometry June 2021
Install.packages("stylo")
library(stylo)
RESULTS<-stylo()
names(RESULTS)
RESULTS$features
```

RESULTS, a word of my choosing, is where stylo( ) stores its results.

name(RESULTS) reveals what is stored in the variable RESULTS

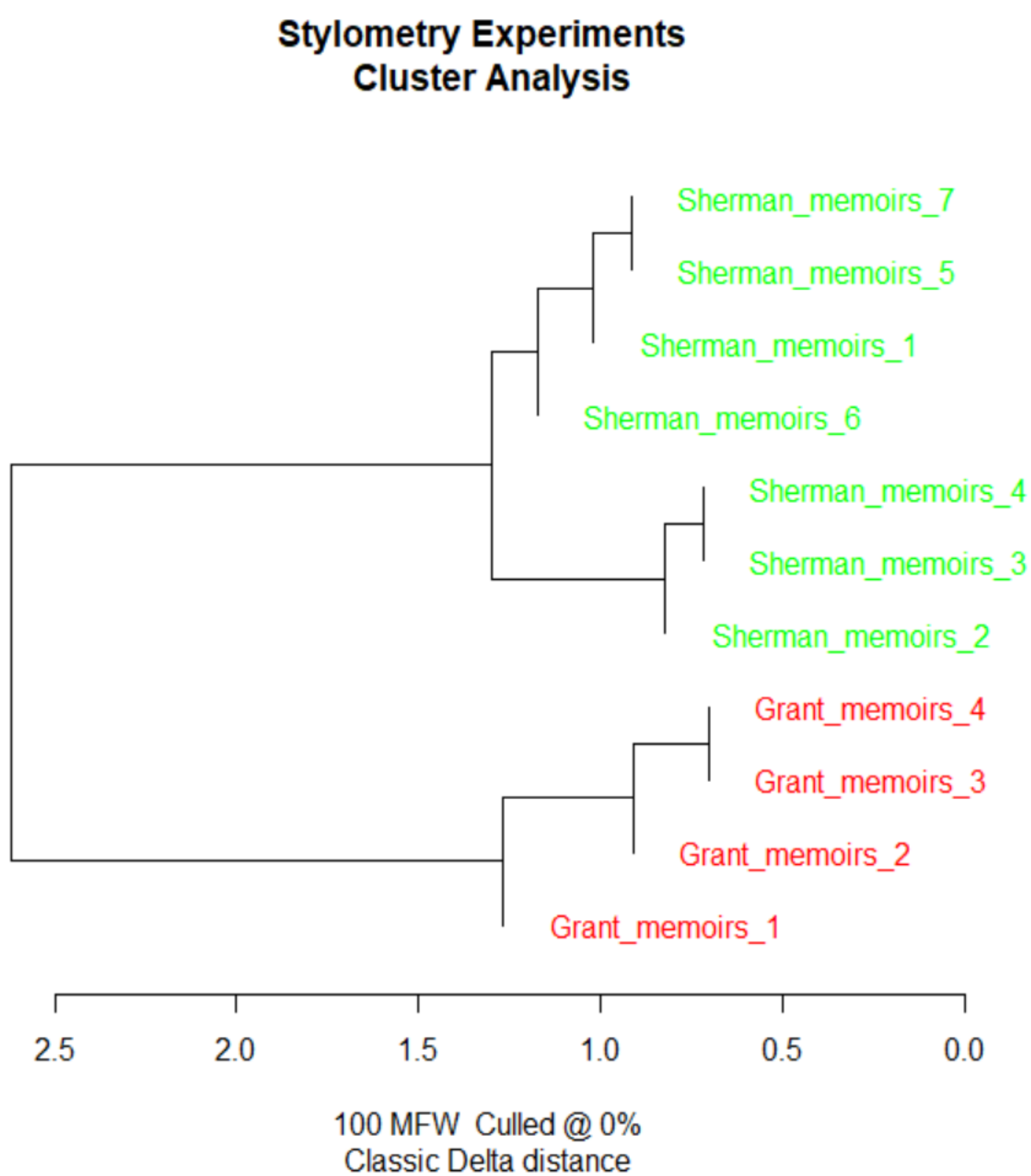
```
> names(RESULTS) # discover what is stored in RESULTS

[1] "distance.table"         "features"               "features.actually.used"
[4] "frequencies.0.culling"  "list.of.edges"          "table.with.all.freqs"
[7] "table.with.all.zscores" "call"                   "name"
```

These are components that will be used to produce a selected output. It is neither possible or necessary to detail the mathematics of what stylo( ) is doing as it produces output.

RESULTS$features, features being one object stored in RESULTS, reveals what is stored in "features." What is stored is a long list of the most frequent words.

In this example I have elected to break the samples up into sets of 10,000 words. Then stylo( ) decides which of the samples are the "closest" to each other. The user has the choice of distance metrics used, the default is "Classic Delta Distance." The output is also at the user's discretion. In this example I have used Cluster Analysis applied to the samples of 10,000 words using the 100 most frequent words in each sample. A dendrogram of the clusters is the output that I selected which I reproduce below.



Interpretation of Results:

Briefly the graph illustrates the samples (seven from Sherman, 10,000 words each and four from Grant, also 10,000 words each) that are closest to each other and moves outward until ultimately, we see that all of the Sherman samples are closest to other Sherman samples and Grant Samples are closest to Grant samples, leading to the conclusion that each general, even given the similarities of education and life, have developed their own unique writing style and word pattern.

Further Studies:

The methods illustrated here could of course be used to investigate disputed authorship in cases where you have several thousand words of prose from both authors. It could also be used to investigate the drift of vocabulary in an author's work from early writings to later writings.

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