Attributing Authorship to the Disputed Federalist Papers

Calvin Hulleman

Department of Data Science and Analytics, Bridgewater College

DSA 330: Data Warehousing

Dr. Ahmad Alqurneh

April 22, 2024

Introduction

Between October 1787 and May 1788, 85 essays were published in various New York newspapers. These papers, written by Alexander Hamilton, James Madison, and John Jay, urged the people of New York to ratify the already drafted United States Constitution. All 85 Essays were published anonymously under the pen name "Publius". It wasn't until Hamilton (in 1804) and Madison (in 1817) publicly claimed authorship that some of the authors were revealed. And it wasn't until 1818, when Jacob Gideon printed a full edition of the Federalist papers, that all Authors were identified together for the first time. However, Madison and Hamilton's lists, and their published revisions of the essays, contradicted each other, leaving us with 12 essays with disputed authorship.

These twelve essays were not the first, nor the last circumstance of disputed authorship in which scholars cannot completely agree on who wrote what. One of the most long-standing disputes of authorship is that of the Book of Mormon. Ever since it was published in 1830, there has been much dispute about whether or not Joseph Smith wrote the book in its entirety. The Book of Mormon, compared to the Federalists at least, is not disputed very heavily. Many studies, including the Larsen Study [4] on word prints have concluded that the Book of Mormon was not written in the authorship styles of any of the 19th-century candidates that they tested (including Joseph Smith). In 1990, Hilton [5] concluded "it is statistically indefensible to propose Joseph Smith or Oliver Cowdery or Solomon Spaulding as the author of the 30,000 words from the Book of Mormon attributed to Nephi and Alma." Now this poses the question, how do we decide you wrote a text and how can we improve on previous techniques in the case of Hamilton and Madison in the Federalist Papers?

Background

Authorship attribution is the process of identifying the authorship of a given text. It is a long-studied and, at times, controversial topic. There are many different methods and techniques used to identify and justify the identification of, the author of a text. Authorship identification of handwritten documents predates the 20th century, but now almost the entirety of data used to analyze text is digital. The creation of texts has also become somewhat of a digital playground, this means that there are now more modern methods to perform authorship attribution like the use of key-stroke biometrics. This is most applied to emails and online messaging.

Unfortunately, while modern techniques tend to be more efficient, and arguably more reliable, we do not have the luxury of knowing the keystroke biometrics of James Madison and Alexander Hamilton. The "old-fashioned" way of analyzing texts is to do so based on stylometric features, by the analysis of lexical, syntactic, or semantic properties of the document. For instance, in the 1960s, Mosteller and Wallace [1] used a simple statistical inference on the usage of a set of 165 specifically chosen words to attribute the authorship of the 12 disputed Federalist papers. The inference was based solely on the frequency of these words in the attributed Federalist papers and then compared to the disputed ones. For example, Hamilton used the word "upon" roughly 18 times more often than Madison. By using a most frequent word list, Mosteller and Wallace concluded that all twelve of the disputed papers should be attributed to Madison and not Hamilton. However, in an extended version [2] Mosteller and Wallace reference six separate study methods on how to assign authorship to the Federalist papers(a study involving a linear discriminant, the original MFW study, re-analysis of the primary study, a weight-rate analysis, a hand calculated Bayesian analysis, and a "three category analysis"). The first study ultimately convinced them of one method not to use, sentence length. It was not a great linear discriminator in the case of Hamilton and Madison because their mean sentence length (34.55 and 34.59 respectively) and standard deviation (19.2 and 20.3) were almost identical and thus totally unusable as a linear discriminator between the two authors.

In their main study, as outlined above, they used a set of "most frequent words" that were chosen based on their ability to successfully attribute authorship to undisputed papers outside of the test set. In this case, the test set was compiled from 94,000 written words by Hamilton (less than 10,000 from non-Federalist papers) and 114,000 written words by Madison (73,000 from non-Federalist papers). They started with a list of over 300 words that were then whittled down to 165 words that were determined to be able to differentiate between the two authors. Those 165 words were then cut down to 30 by testing on another set of texts that had not been used in the initial test set. With the set of 30 words, the Bayes Rule could be applied to statistical inferencing. Since they were dealing with only two outcomes, they were able to use log odds to simplify the Bayesian process:

$$\textit{Final log-odds} = \textit{initial log-odds} + \textit{log-likelihood}$$

Initially, Mosteller and Wallace worked with a Poisson and then a Negative binomial distribution for the marker words. However, after examining 90 non-marker words, they settled on using a Beta distribution. After deriving the posterior log-odds they found that they, simply put, had nearly irrefutable evidence in favor of Madison as the author of all 12 disputed documents.

Much like the first analysis performed by Mosteller and Wallace, the weight-rate analysis used linear discriminators. This analysis was mainly composed as a check on the "main" analysis. A pool of 117 words that had the potential to discriminate between the two authors was compiled. These words were then tested on 25 and 23 works by Madison and Hamilton respectively. Only the most effective 20 words were kept. A linear discriminant function was then calculated based on the weights of each of the 20 words. The function was then applied to a screening and calibrating set of Madison and Hamilton's works. Based on the calibrating set, only one (55) of the twelve disputed papers was within a 99% confidence interval for Hamilton, while all but two of the twelve were in a 90% confidence interval for Madison. This analysis provided an outcome much like that of the main Bayesian study.

The hand-calculated robust Bayesian analysis that Mosteller and Wallace performed gave almost the same results as the weight-rate analysis. Only showing evidence for number 55 to be written by Hamilton. Since Mosteller and Wallace relied heavily on Bayesian statistics to solve the issue of the Federalist papers, there have been many other attempts involving varying approaches.

Another method of stylometrics used in the Federalist papers is the analysis of vocabulary richness. A multivariate approach to vocabulary richness was successfully applied to the Book of Mormon scripture by Holmes [6]. This approach uses five variables that all measure the vocabulary richness of the texts but hold constant text length. Using N as text length, the number of unique words in the text as V, and the number of words used r times as Vr:

$$R = (100 \log N)/(1 - (V1/V))$$

$$V2/V$$

$$K = 10{\wedge}4(\sum_{i=1}^{\infty} i^{2Vi} - N)/N{\wedge}2$$

and the final two parameters (α and θ) from the Sichel [7] distribution. These five parameters cover the entirety of an author's potential vocabulary richness distribution. In Holmes's analysis, he selected thirty-three papers to use for sampling. All the disputed text, all five of John Jay's papers, eight random Hamilton papers, five random Madison papers, and three of the joint Hamilton and Madison papers. Holmes used the Oxford Concordance program to create a word list for each text. The nouns in each word list were then filtered to combine singular and plural nouns into one and then used to create distributions of vocabulary for each text. The SICEHL [6] program was then used to compute vocabulary richness outputs. The first output of the richness variables was using average-linkage cluster analysis. This produced three clusters: 1) the Madison papers, two Jay papers, and one of Hamilton's, 2) two Jay papers and one Hamilton paper, 3) the rest of Hamilton's, one of Jay's, plus the three joint papers chosen prior.

A principal component analysis (PCA) was then performed on only the Madison and Hamilton papers. This PCA accounted for over 93% of the variation in the five dimensions. It also showed a continuing pattern, with Hamilton 72 as an outlier. Then the Jay and the Hamilton and Madison Joint papers were added as components in the analysis. This PCA accounted for 90% of the variation in the dimensions. A third PCA, of Hamilton, Madison, and the disputed papers, accounted for 89.9% of the dimensional variation. Hamilton 72 was repeatedly an outlier in the component analysis. Finally, a linear discriminant function and cross-validation were used on the disputed papers and the calibrating set. Using this method, eleven of thirteen calibrating papers were correctly assigned to their groups and all twelve disputed papers were attributed to Madison.

It must be noted here that one of the critical assumptions made by Mosteller and Wallace, Holmes, and which we will retain for simplicity's sake, is that the disputed Federalist papers only had one author. A somewhat newer prevailing argument and critique of Mosteller and Wallace's traditional and non-traditional approaches to the disputed Federalist papers is that they were joint efforts by Hamilton and Madison. Rudman [8] argued that there is no doubt that the Federalist Papers, overall, were a joint effort between the three men. We also know that there were at least three jointly written papers (Federalist 18, 19, 20). This alone is arguably enough to not immediately write off the possibility that the twelve disputed papers were also joint efforts. There are also a handful of traditional and non-traditional [9] studies that provide strong evidence for the case of multiple authorship for the twelve disputed Federalist Papers. Overall, the main issue with these studies is that they often disagree with each other. Since these studies fail to produce reliable or repeated results, we will stick to the assumption that either Hamilton or Madison wrote each paper.

There are several other problems with Mosteller and Wallace's analyses of the papers that many other non-traditional studies also obtain. Both main problems have to do with text sampling and control. To create the chunks of writing to base their analysis on, they go beyond the scope of the Federalist papers. This creates some serious doubts in their analysis. When doing any textual analysis, but especially non-traditional stylometric analysis, the starting text should be as close to the final text as possible. The further away they are the more errors you are likely to encounter. For Hamilton, they don't stray too far away (only 10,000 of 94,000 words were not from the Federalist papers). However, for Madison, the error is much worse. 73,000 of the 117,000 words used to produce the training set were from non-Federalist Papers. What makes this particularly egregious is that many of the non-Federalist papers were compiled from nearly 25 years of Madison's writing (this is a lot of time for writing style and vocabulary distribution to change) and, since the publication of Mosteller and Wallace's study, some of these texts have been shown to not even have been written by Madison. The second problem with the non-traditional approach is the use of any meaningful control, or rather lack thereof. The ‘training set’ used in their Bayesian analysis is not a valid control.

Experiments and Discussion

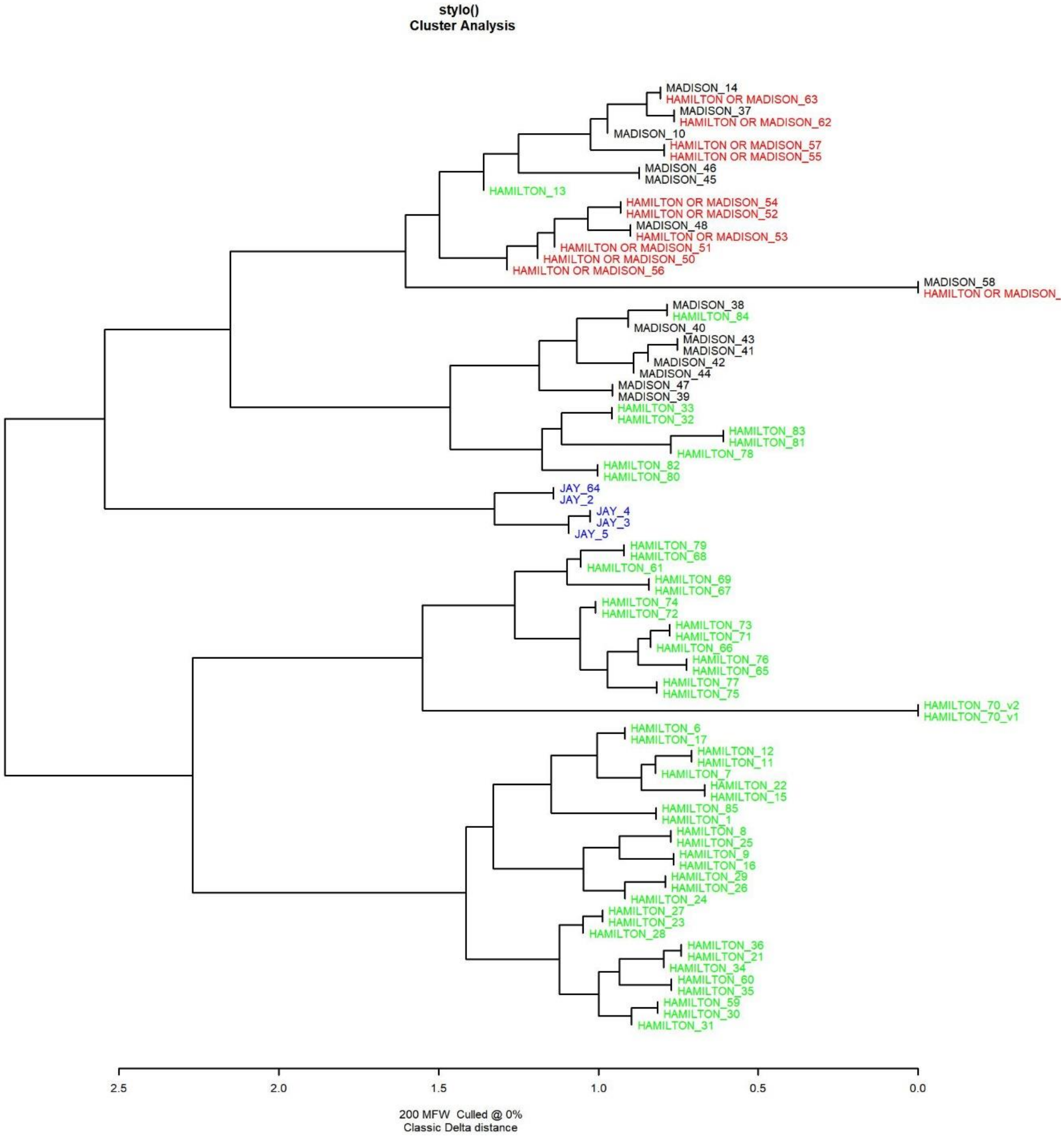
For this analysis, we will analyze and compare the outputs of only two distances, the classic delta [11] and Eder’s delta [12]. To justify this, we will first look at why we are not using the other options provided by Stylo. The Manhattan and Euclidean distances are notably poor when it comes to unevenly distributed objects, which most matrices of most frequent words are. Argamon’s Linear Delta [13] is particularly sensitive to the size of our corpus and is just a function of the Euclidean distance. The Canberra distance is particularly sensitive to rare vocabulary and does not adjust well to random noise like Eder’s delta does. We use the classic delta because it has a strong ability to handle uneven distributions and standardize them and we use Eder’s delta because it is an adjustment of the classic delta that rescales to suppress outliers in the word list.

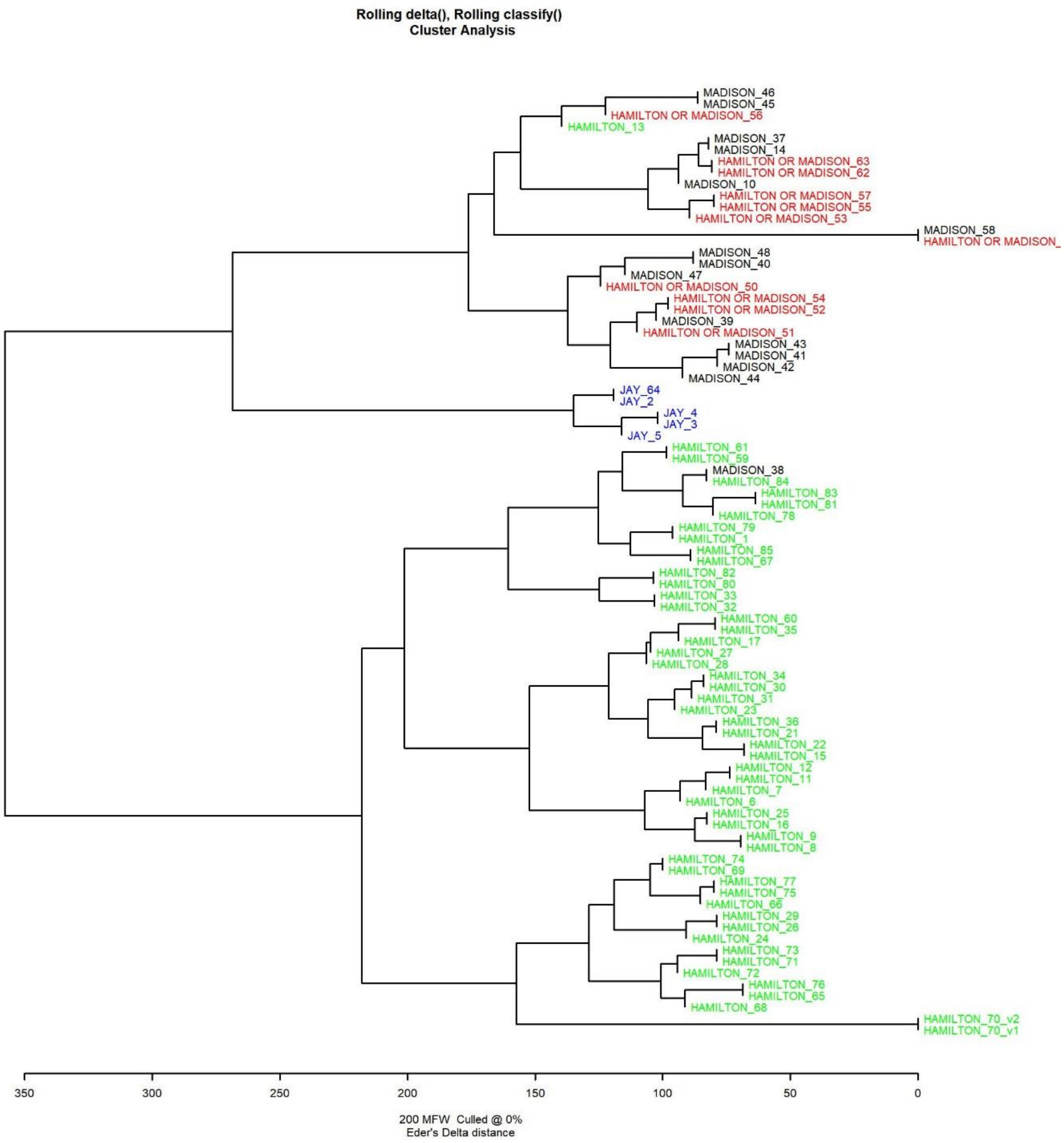
The entirety of this analysis is done using the “Stylo” package [10] in R studio. The first stage was done using the stylo() method, the second with the classify() method, and further information was provided by using the rolling.delta() method.

Explanatory Process

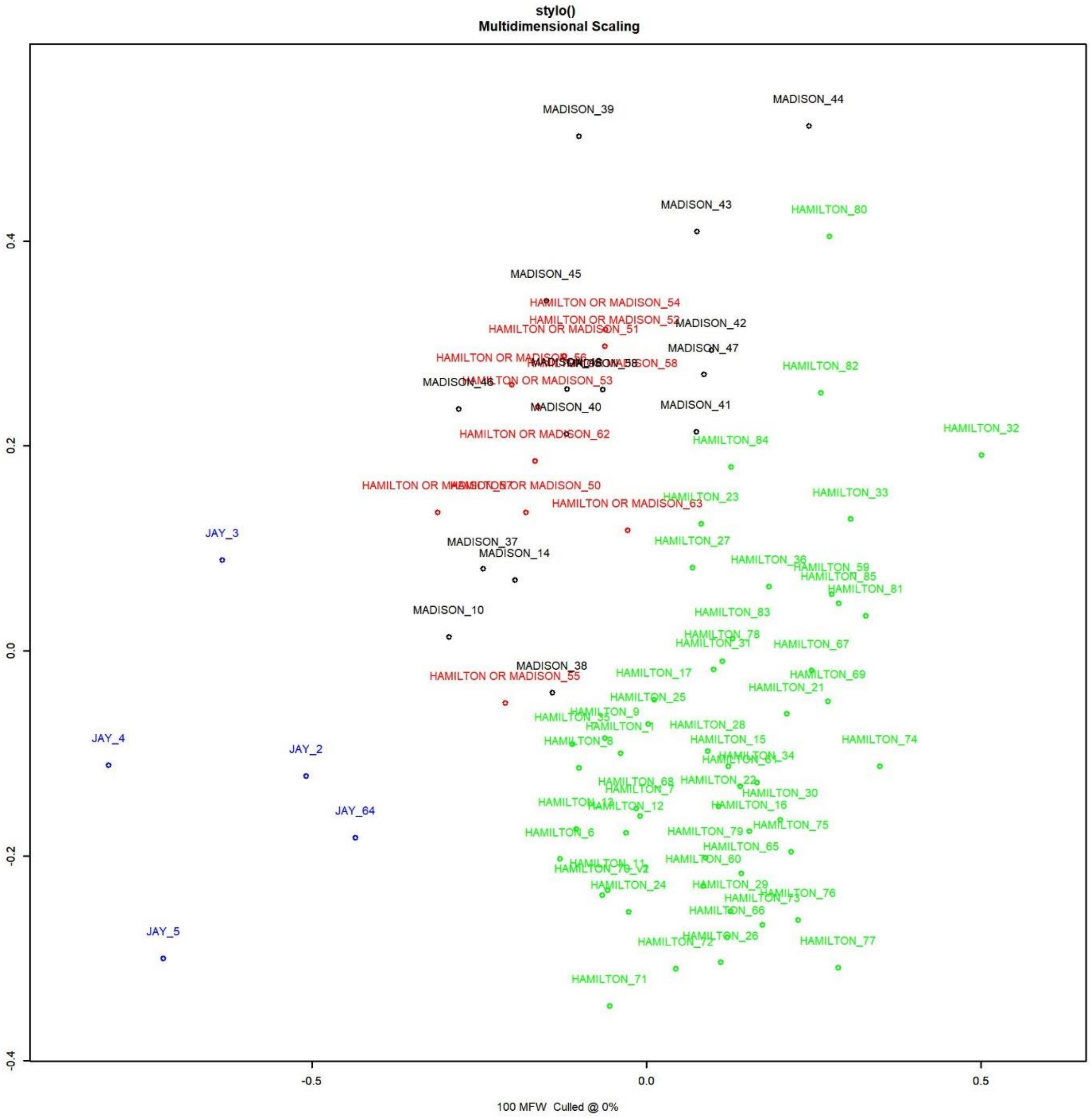
The Stylo function provides a most frequent word list for the entire set of texts (disputed and undisputed). Then it produces the word frequencies for each of the MFWs to create a matrix of words by text. It then normalizes the matrix, cuts down the words based on the analysis, and removes personal pronouns. The function then uses different statistical methods and distance calculations to determine the relationships between the texts and provide a visual representation.

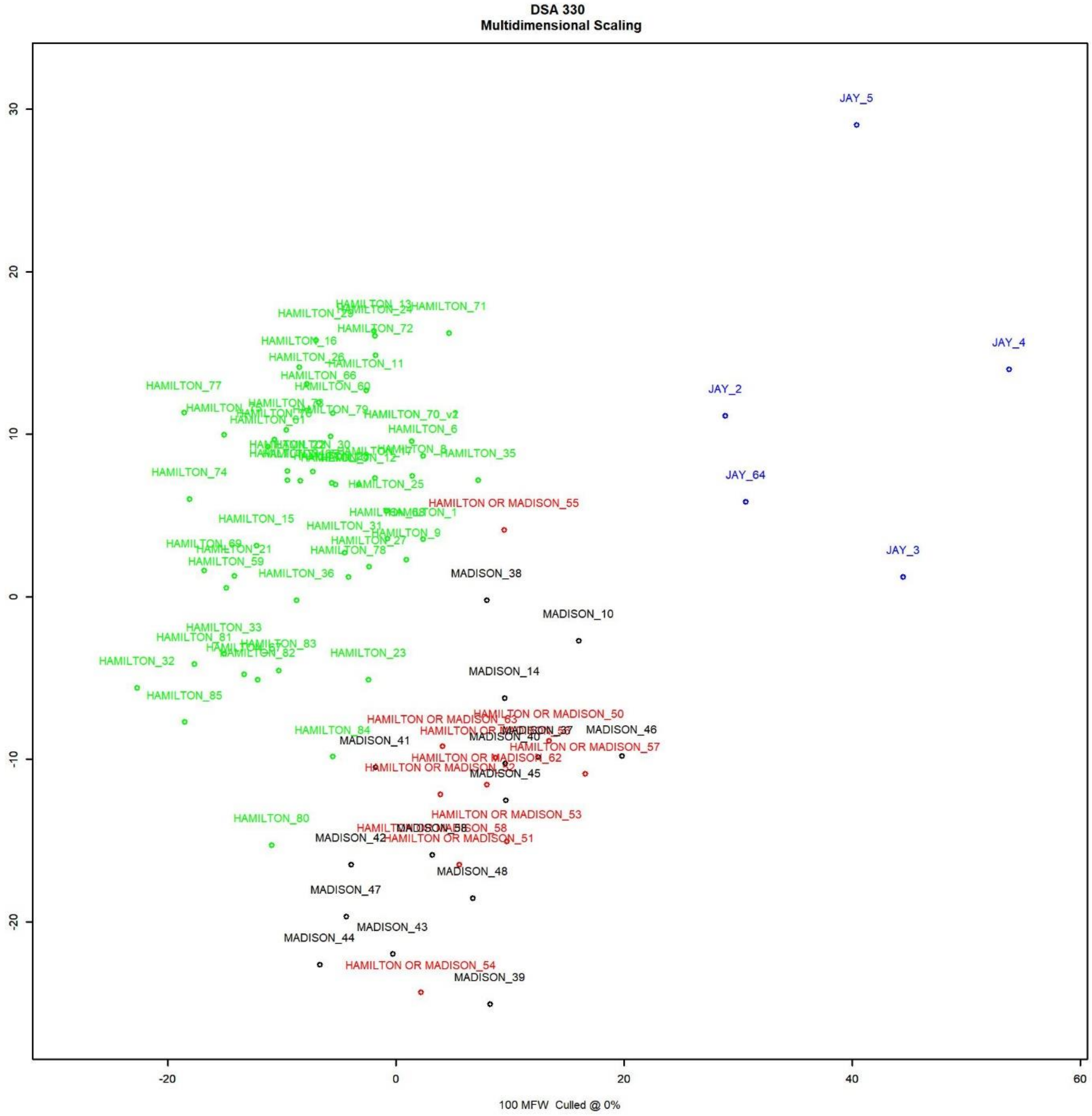
The three statistical outputs we will produce are a cluster analysis, a multidimensional scaling analysis, and a principal component analysis using a covariance matrix. The cluster analysis uses the distance matrix to sort the items into groups and produces a dendrogram of the results. The results for the classic delta and Eder's delta respectively are shown below:


stylo()
Cluster Analysis
MADISON_14
HAMILTON OR MADISON_63
MADISON_37
HAMILTON OR MADISON_62
MADISON_10
HAMILTON OR MADISON_57
HAMILTON OR MADISON_55
MADISON_46
MADISON_45
HAMILTON_13
HAMILTON OR MADISON_54
HAMILTON OR MADISON_52
MADISON_48
HAMILTON OR MADISON_53
HAMILTON OR MADISON_51
HAMILTON OR MADISON_50
HAMILTON OR MADISON_56
MADISON_58
HAMILTON OR MADISON_
MADISON_38
HAMILTON_84
MADISON_40
MADISON_43
MADISON_41
MADISON_42
MADISON_44
MADISON_47
MADISON_39
HAMILTON_33
HAMILTON_32
HAMILTON_83
HAMILTON_81
HAMILTON_78
HAMILTON_82
HAMILTON_80
JAY_64
JAY_2
JAY_4
JAY_3
JAY_5
HAMILTON_79
HAMILTON_68
HAMILTON_61
HAMILTON_69
HAMILTON_67
HAMILTON_74
HAMILTON_72
HAMILTON_73
HAMILTON_71
HAMILTON_66
HAMILTON_76
HAMILTON_65
HAMILTON_77
HAMILTON_75
HAMILTON_70_v2
HAMILTON_70_v1
HAMILTON_6
HAMILTON_17
HAMILTON_12
HAMILTON_11
HAMILTON_7
HAMILTON_22
HAMILTON_15
HAMILTON_85
HAMILTON_1
HAMILTON_8
HAMILTON_25
HAMILTON_9
HAMILTON_16
HAMILTON_29
HAMILTON_26
HAMILTON_24
HAMILTON_27
HAMILTON_23
HAMILTON_28
HAMILTON_36
HAMILTON_21
HAMILTON_34
HAMILTON_60
HAMILTON_35
HAMILTON_59
HAMILTON_30
HAMILTON_31
2.5
2.0
1.5
1.0
0.5
0.0
200 MFW Culled @ 0%
Classic Delta distance

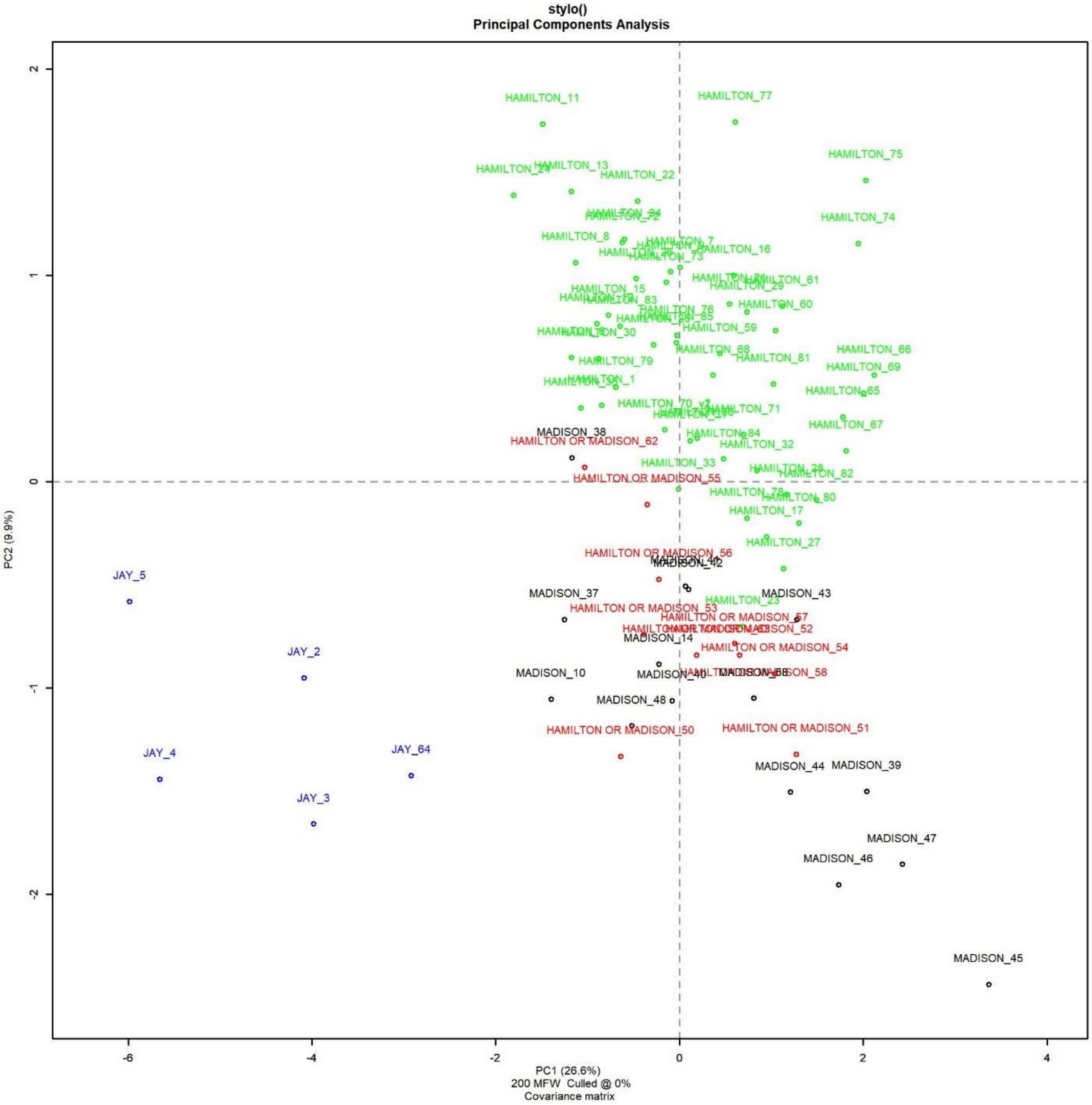


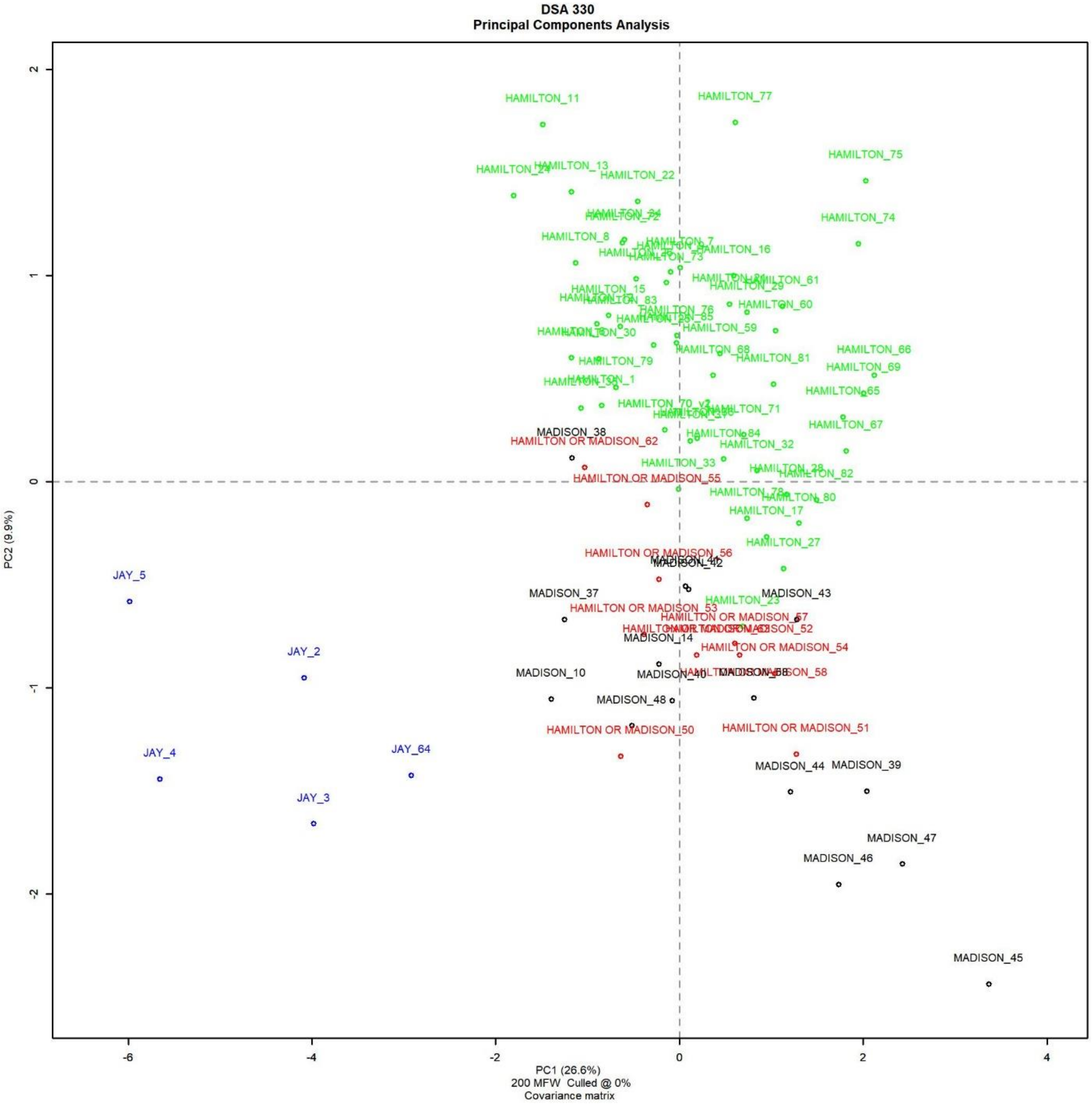
The dendrograms have very similar groupings. Both have the Jay papers separate from the others and most of the Hamilton papers are far from the Madison or disputed papers. Both also have Hamilton 13 as an outlier and grouped with Madison. Both charts also clearly attribute all twelve disputed papers to Madison. However, the classic delta distance struggles to distinguish between a small group of 8 of Madison and 8 of Hamilton's undisputed papers. The multidimensional scaling of these shows a similar result while separating the undisputed papers more effectively and without Hamilton 13 as an outlier.





The number of variables in the data set equals the dimension of the data. We want to minimize it as having too many dimensions makes our work more difficult and adds needless effort. We can do a PCA analysis to lower the dimensions. To minimize the dimension, PCA essentially projects the observed data onto a separate axis (PCs). To put it another way, PCA analysis identifies the key variables in the dataset that account for most of the variances in the data and enables us to eliminate the unnecessary variables. Every numerical variable is subjected to PCA, and before doing so, the data is scaled to have a unit variance, which is consistent with one of the PCA assumptions. Differently Scaled variables may dominate and cause disruptions to our analysis. PCA using classic delta and Eder's delta present similarly to our MDS and cluster analysis:





The one strong difference between the output obtained from the PCA and the first two outputs is that Federalist 62 is attributed to Hamilton int the PCA as compared to all twelve being attributed to Madison in the cluster analysis and MDS.

The classify() function is a supervised learning function. This method divides the analysis into two phases. In the first stage, a classifier—a set of rules for identifying an author's "uniqueness" in style—is created based on the traceable differences between samples. The machine uses the trained classifier to assign additional text samples to the authorial classes that the classifier has established. Any disputed or anonymous samples will also be assigned to one of the classes if the classification is typically based on probabilistic criteria. This second step is predictive in nature.

Classification

Of the ten methods used (Classic delta, Eder's delta: delta, k-NN, SVM, NaiveBayes, NSC) only 5 produced general attribute success (GAS) rates of 90% or higher. All methods used an MFW of 100 and were not culled. A support vector machine SVM is an algorithm that determines boundaries between data points based on predefined classes, labels, and outputs. Using the classic delta and an SVM, a GAS rate of 93.1% was achieved (it incorrectly identified Hamilton's 79 and 84 as Madison's). This method identified eleven of the disputed papers as belonging to Madison and one belonging to Hamilton (Federalist 55). An SVM with Eder's delta produced the same results.

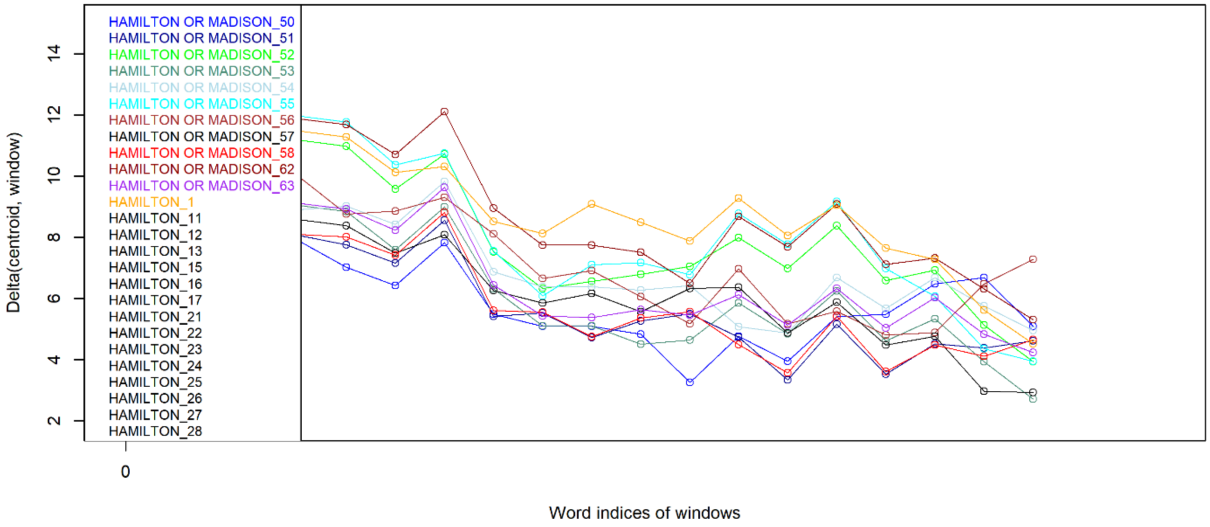
Nearest Shrunken centroid classification (NSC) computes the centroids for every class and uses soft thresholding to "shrink" the centroids toward 0. The class that has the smallest distance between an observation and the (shrunken) centroid is subsequently given the assignment for subsequent observations. Using an NSC algorithm and a classic delta, a GAS rate of 100% was produced (it correctly identified every undisputed text) and attributed all but Federalist 55 of the disputed texts to Madison. The same approach, but with Eder's delta, achieved the same results.

Using a simple classic delta approach produced a 93.1% GAS rate (misattributing Hamilton 67 to Madison and Jay 64 to Hamilton) and attributed Federalists 53, 55, and 56 to

Hamilton with the rest to Madison. A k-NN approach using Eder’s distance achieved a 93.1% GAS rate and attributed Federalists 55 and 56 to Hamilton with the rest to Madison.

Further Justification for Single Authorship

To further justify why we performed authorship attribution based on the assumption that only one (either Madison or Hamilton) wrote each of the twelve disputed papers, we can use the rolling.delta() function. This function follows a windowing procedure for each reference text. Each text is divided into successive, equal-sized samples, or "windows," in the first step of the process. There can be some partial overlap between the samples. Like Delta, this method makes use of the relative frequencies of a group of n terms that were the most common across all reference texts in the collection. Next, we calculate a representative centroid for every reference text, which is comprised of the standard deviation and the mean relative frequency for every n words in the windows that were taken out of the text. We next go on to the test text analysis. Additionally, we divide it into windows, and we compute the style difference (the Delta) between each test window and each reference centroid sequentially. The output of the rolling delta on our disputed papers is below:



As you can see in the graph, the word indices of the windows are mostly stable throughout the disputed papers. While there is arguably a large enough change to be of note at the third marker, this is consistent with the graphs for both Madison and Hamilton so is not of significance.

Conclusion

When looking at all the analyses overall, there are a few consensuses that we have come out with. First, it is nearly impossible to find any evidence that Hamilton wrote any of the disputed papers except Federalists 55 and 56. It seems abundantly clear based not only on the work done in this study but also based on the work of many others that Madison wrote Federalists 50-54, 57, 58, 62, and 63. Based on the explanatory process (the stylo() function) we used, our results would agree with that of Mosteller and Wallace. However, the problems with their study and newer studies like it remain the same. They did not have accurate and meaningful training or control data and their methods are not as mathematically sound as the traditional approach that we took in the classification section of this study. By only using the Federalist Papers to train and test out algorithms, we were more effectively able to grasp the distribution of the essays. By using the classic delta and Eder’s delta we also ensured that the potentially drastically different vocabulary distributions were normalized, and high variation was accounted for. Finally, by graphing the word indices in their respective windows we were able to show strong evidence that each of the disputed papers was not a joint effort and was written only by either Hamilton or Madison, not both. Based on these factors, we can safely attribute the authorship of Federalists 50-54, 57, 58, 62, and 63 to James Madison, and Federalist 55 to Alexander Hamilton. This leaves Federalist 55 with a significant level of uncertainty.

Bibliography

1. Mosteller, Frederick and David L. Wallace (1963). “Inference in an Authorship Problem”. Journal of the American Statistical Association 58.302, pp. 275–309
2. Mosteller, F. and Wallace, D. L. (1984). Applied Bayesian and Classical Inference: The Case of the Federalist Papers. Addison-Wesley, Reading, MA.
3. What can stylometry tell us about Book of Mormon authorship? (2020a). Retrieved from https://knowhy.bookofmormoncentral.org/knowhy/what-can-stylometry-tell-us-about-book-of-mormon-authorship#footnote18_qf40xr0
4. Larsen, W. A., Rencher, A. C., & Layton, T. (1980). Who Wrote the Book of Mormon? An Analysis of Wordprints. *BYU Studies, 20*(3), 225–251. Reprinted in Reynolds, N. B. (Ed.). (1982). *Book of Mormon Authorship: New Light on Ancient Origins* (pp. 157–188). Provo, UT: Religious Studies Center, Brigham Young University.
5. Hilton, J. L. (1992). Wordprints and the Book of Mormon. In Welch, J. W. (Ed.), *Reexploring the Book of Mormon: A Decade of New Research* (pp. 221–226). Salt Lake City and Provo, UT: Deseret Book and FARMS
6. Holmes, D. I. (1991). Vocabulary Richness and the Prophetic Voice. *Literary and Linguistic Computing,* 6(4): 259-268. (1992). A Stylometric Analysis of Mormon Scripture and Related Texts. *Journal of the Royal Statistical Society Series A,* 155(1): 91-120.
7. Sichel, H. S. (1975). On a Distribution Law for Word Frequencies. *Journal of the American Statistical Association, 70*(342), 542-547.

8. Rudman, J. (2012). The Twelve Disputed 'Federalist' Papers: A Case for Collaboration. *Digital Humanities 2012*.
9. Collins, J., Kaufer, D., Vlachos, P., Butler, B., & Ishizaki, S. (2004). Detecting Collaborations in Text: Comparing the Authors' Rhetorical Language Choices in the Federalist Papers. *Computers and the Humanities*, *38*(1), 15–36.
10. Eder, M., Rybicki, J., Kestemont, M., & Pielstroem, S. (2016). Stylometric Multivariate Analyses. *Digital Humanities 2012*. Retrieved from https://journal.rproject.org/archive/2016/RJ-2016-007/index.html
11. Burrows, J. (2002). Delta: A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing, 17*(3), 267–287.
12. Eder, M. (2011). Style-markers in authorship attribution: A cross-language study of the authorial fingerprint. *Studies in Polish Linguistics, 6*, 99–114. Retrieved from http://www.wuj.pl/page,art,artid,1923.html
13. Argamon, S. (2008). Interpreting Burrows's delta: geometric and probabilistic foundations. Literary and Linguistic Computing, 23(2): 131–47