Kristofer Hormoz

Stylometry & Text Analysis

Executive Summary

This project attempts to demonstrate the techniques and methods of stylometric analysis to correctly identify the authorship of unknown texts. In this project we will explain what stylometric analysis is as well as go over three key methods and discuss their advantages and disadvantages as related to the ability of identifying authors.

Our first method is regarded as “Mendenhall’s Characteristic Curves” and it is considered one of the first attempts in to identify authorship in the field of stylometric analysis. The method simply plots the frequency of individual words, and the theory is that each author has a unique characteristic curve that can then be compared to unknown or test data sets. In this project we present two characteristic curves, plot (1), of two known authors and then present six additional characteristic curves, in plot (2), of unknown works. The two known authors used are Alexander Dumas and Mary Wollstonecraft Shelly, and the texts written by these authors to be analyzed are “The Count of Monte Cristo” and “Frankenstein” respectively.

Our next two methods are known as the Adams Kilgruffs Chi-Squared method and John Burrow’s Delta Method and both these methods measure the distance between vocabularies employed by authors. However, unlike Kilgruffs Chi-squared method, Burrow’s Delta method can measure how an anonymous text and sets of texts written by different known authors all diverge from the average of all of them put together as well as give equal weight when

measuring different features to avoid the problems from the chi-squared method of giving more proportion of weight for more common words.

It was found that,

1. When interoperating the characteristic curves plot, Alexander Dumas's plot was like that of unknown texts 2,3,5, and 6 while Mary Shelly's plot was like that of unknown texts 1 and 4. However, unknown text 3, "Mathilda", belongs to Mary Shelly and not Alexander Dumas thus this method proves to not be overly accurate.
2. When analyzing the chi-squared method, we had conducted the test on both our training data texts and the test statistic proved that these two texts had greater distances between the vocabulary used between authors thus these two authors are in fact not the same individual. However, this method is not efficient when wanting to compare to multiple different texts to multiple different authors and it also puts more weight on more common words or phrases as compared to rarely used words that may distinguish two authors from each other.
3. When analyzing Burrow's Delta method, we had correctly identified each unknown texts as belonging to the correct author, and we were able to accomplish this more effectively as compared to the chi-squared method. Thus, this test statistic proved to be the most accurate in identifying authorship as compared to the three other methods.

Kristofer Hormoz

Introduction to Stylometric Analysis

From Polish philosopher Wincenty Lupolianski (1890) analyzing Plato's Dialogues to Rev. A. Q. Morton (1960) evaluating the fourteen Epistles of the New Testament, Stylometry data analysis has presented the tools needed to analyze components of texts to distinguish identity, authenticity, and other components that differentiate the style of writing between authors. Scholars, and researchers have used stylometry in order observe how "authors tend to write in relatively consistent, recognizable and unique ways" (Laramee, page 1). From vocabulary, use of vowels, size of sentences or paragraphs, and uses of semicolons or other punctuations no two authors will ever have an identical style of writing.

Early forms of stylometric analysis would plot how often words of different lengths appeared, believing that the plot curves "would look pretty much the same no matter what parts of the novel we had picked" (Laramee, page 6). The most popular of these methods were Mendenhall's Characteristic Curves of Composition. This method is considered very blunt and not as reliable as other stylometry analysis methods because it fails to consider the actual words in an author's vocabulary. However, due to the limitations of technology in Mendenhall's time (1880), this method was very easy to compile data as it was to be written in by hand.

Stylometry analysis has a vast variety of applications, from identifying authors, uses in forensic science, identifying plagiarized texts, etc. For example, in his article "Introduction to Stylometry with Python" Francois Laramee discuss a collection of 85 articles written between 1787 and 1788 known as the Federalist Papers. There has been a debate surrounding the identity of the authorship for these papers. Three alleged authors include Alexander Hamilton, James Madison, and John Jay. Although, the Federalist Papers were submitted under the pseudonym

“Publius” Francois Laramee believes that utilizing stylometry analysis it is possible to determine which of the 85 articles were written by either Hamilton, Madison, or Jay. Looking at cases where stylometry analysis was used in forensic science, in 1996 when comparing the Unabomber’s manifesto with letters written by the suspect, Ted Kaczynski, resulted in the apprehension and arrest of Kaczynski.

The texts that we will be using for this assignment will be Mary Wollstonecraft (Godwin) Shelley’s “Frankenstein” and Alexandre Dumas “The Count of Monte Cristo”, as well as six additional “unknown” texts where three belong to Mary Shelly and three belong to Alexandre Dumas. Starting with Frankenstein, this is a horror science fiction novel written in 1818 by British author Mary Shelly. It consists of approximately 280 pages with 7316 lines of the original text. Looking towards The Count of Monte Cristo, this is an adventure novel written in 1844 by French author Alexander Dumas. It consists of approximately 928 pages with 61,056 lines of the original text. We can see that not only do both these novels have a different theme to them, Frankenstein is a horror science fiction novel where The Count of Monte Cristo is an adventure novel, they also are written by two authors from not only different geographical backgrounds but also a difference in time periods that would impact their writing style. The difference in each author’s writing style is what this project wishes to discover.

This project will examine the contents of these two texts (Frankenstein and The Count of Monte Cristo) to use stylometry analysis to develop a fingerprint for each authors writing style to identify any other texts as belonging to them. The following are the goals for this project.

1. Present how powerful and useful stylometry analysis can be, and to do this we will be examining three different stylometric approaches and explain the advantages as well as the disadvantages for each approach. We will begin with “Mendenhall’s Characteristic

Curves of Composition" (Laramee, page 8), then look at "Kilgariff's Chi-Squared Method" (Laramee, page 8), and lastly "John Burrow's Delta Method" (Laramee, page 8).

2. Implement the methods on unknown texts to identify which author wrote them

Preparing the Texts in Python

The main source of texts that were used came from http://www.gutenberg.org/, a website that allows users to download texts for free. However, these texts are not of the original form from when they were published. For example, in the downloaded text for "The Count of Monte Cristo" in addition to the original novel, translated to English, we can see that project Gutenberg had also included information on the website at both the start of the text file as well as the end of the text file. This pattern follows for all of project Gutenberg text files. Thus, the first challenge was to clean up the text files. To do this, we had to know at what lines do our original texts start and end and to solve this problem we used a multitude of python tools such as regular expression. Once we solved this problem, we had to now split our text files into individual words so that we may conduct some of our test statistics, such as chi-squared and Delta tests. Also, this will be needed when we are to plot the frequency of words for our first stylometric method.

To split the text files into individual words and count these words we had utilized not only regular expression and methods such as ".split()", but also the packages "collections" and "pandas". These two packages have functions like "Counter" which allows us to count the frequency of words in the text files, but before doing this we made sure that after we had split our texts files into words and had used the method "lower" to make sure that we ignore if a particular word is either lower case or upper case when counting them. The reasoning behind this is that for our test statistics we don't want a word such as "the" and "The" to be counted as two separate words.

Some of the challenges and problems that arose surrounding this area in the project was related to the uses of regular expression and the knowledge of when project Gutenberg had started the original text files and ended them. Although the convention was very similar, we had to explore the exact lines that began and ended each original text file to find the line count. Regarding the use of regular expression, while it is a powerful tool when working with text files it can also become challenging to understand when trying to use some of the methods such as "re.sub()" in succession to remove particular items in the text files.

Analysis & Results

(1) Thomas Mendenhall's Characteristic Curves

In 1887 Thomas Mendenhall had published "The Characteristic Curves of Composition" which was "one of the earliest attempts at stylometry, the quantitative analysis of writing style" (Jeremy Norman, page 1). Mendenhall was motivated by remarks made by mathematician Augustus de Morgan in 1851 and as such he proposed to analyze a composition by "forming what may be called a word spectrum, or characteristic curve" (Jeremy Norman, page 2). This characteristic curve would be a graphical representation of words in a text according to their length and frequency of occurrence. Whereas Mendenhall utilized these techniques to attempt to identify the true author of works usually attributed to William Shakespeare, we will use these techniques to identify techniques used by our two authors Mary Shelly and Alexander Dumas.

We will first examine the frequency of words and the frequency of different word lengths for each novel.

Table (1): Top 5 Most used Words for Frankenstein		
	Word	Frequency
1	“The”	4152
2	“and”	2958
3	“i”	2765
4	“of”	2638
5	“to”	2082

In table (1) we see that the most common word in Frankenstein is “the” with a frequency of 4152. In fact, we will soon find that this is a common trend where most texts have “the” as the most common phrase. It is interesting to note that the single use of the vowel “i’ is the most common use of a single word in Frankenstein whereas we will see how this will be different with The Count of Monte Cristo.

Table (2): Top 5 Most used Words for The Count of Monte Cristo		
	Word	Frequency
1	“The”	27779
2	“of”	12637
3	“to”	12628
4	“and”	11598
5	“a”	9118

In table (2) we see that the most common words used in The Count of Monte Cristo are very similar to that of Frankenstein. In fact, four out of the five most common words used in this text are equivalent to that of Frankenstein however, we see that apart from the word “the” all the other four words are listed differently. Furthermore, it appears that Alexander Dumas appears to favor the phrase “a” a single letter word more than other single letter words. Although we have only begun to examine certain aspects of each novel, it is interesting to see how this method already shows dissimilarities and similarities between each author.

We will now examine the frequency of each word length, from length one to twenty-three for each novel.

Table (3): Frequency of Different Word Lengths

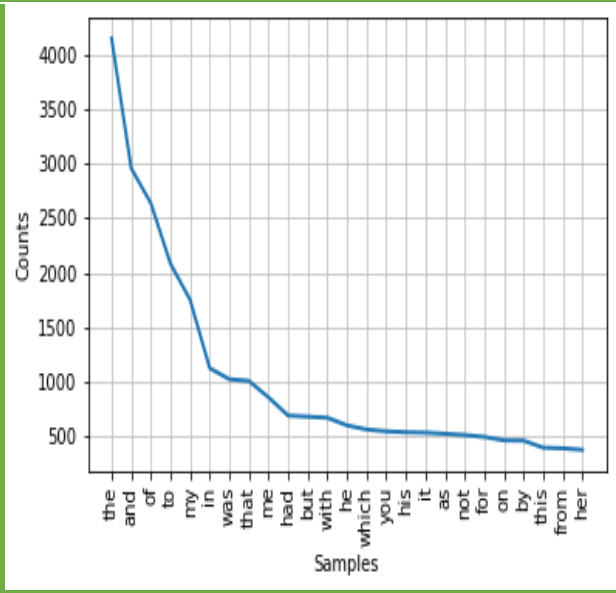
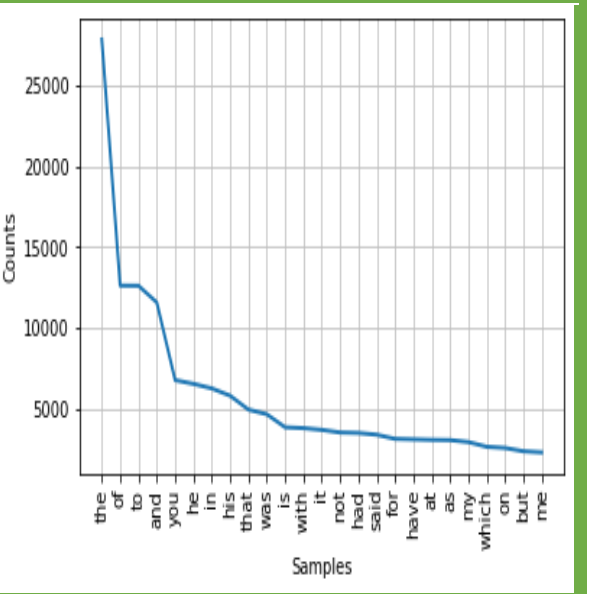
Frankenstein	The Count of Monte Cristo
words of length 1 = 5.6%	words of length 1 = 3.7%
words of length 2 = 18.2%	words of length 2 = 17.1%
words of length 3 = 21.5%	words of length 3 = 22.7%
words of length 4 = 15.6%	words of length 4 = 17.4%
words of length 5 = 10.3%	words of length 5 = 11.0%
words of length 6 = 8.3%	words of length 6 = 8.5%
words of length 7 = 7.0%	words of length 7 = 6.7%
words of length 8 = 5.0%	words of length 8 = 4.9%
words of length 9 = 4.1%	words of length 9 = 3.6%
words of length 10 = 2.2%	words of length 10 = 2.1%
words of length 11 = 1.3%	words of length 11 = 1.1%
words of length 12 = 0.6%	words of length 12 = 0.6%
words of length 13 = 0.2%	words of length 13 = 0.3%
words of length 14 = 0.1%	words of length 14 = 0.1%
words of length 15 = 0.0%	words of length 15 = 0.1%
words of length 16 = 0.0%	words of length 16 = 0.0%
words of length 17 = 0.0%	words of length 17 = 0.0%
words of length 18 = 0.0%	words of length 18 = 0.0%
words of length 19 = 0.0%	words of length 19 = 0.0%
words of length 20 = 0.0%	words of length 20 = 0.0%
words of length 21 = 0.0%	words of length 21 = 0.0%

words of length 22 = 0.0% words of length 23 = 0.0%	words of length 22 = 0.0% words of length 23 = 0.0%

Analyzing table (3), we can see that both authors tended to utilize three letter words more often than words of any other lengths. However, we can quickly point out stark dissimilarities from table (3) that would indicate that these two authors are not alike. For instance, whereas Mary Shelly prefers words of length two to be her second most used words in her novel, Alexander Dumas prefers words of length four. Furthermore, it appears that Alexander Dumas has vocabulary of words with a high degree of length because almost 7.9% of the words he had used in The Count of Monte Cristo had a length of 10 or greater whereas Mary shelly only accounts for 4.4% of words in Frankenstein to be of length 10 or greater.

Thus, Alexander Dumas is 1.79 times more likely to use words of length 10 or greater as compared to Mary Shelly, which would help use in determining a characteristic attributed to Alexander Dumas. But, to compare these two authors using Thomas Mendenhall method we must now plot these frequencies to create our “characteristic curve”. As we recall, the characteristic curve plots words based on their frequencies of usage, with the assumption that an author’s vocabulary does not change significantly over time. Thus, if two characteristic curves appear to be similar, we may conclude that the vocabular between these two novels are significantly similar and thus may be written by the same author.

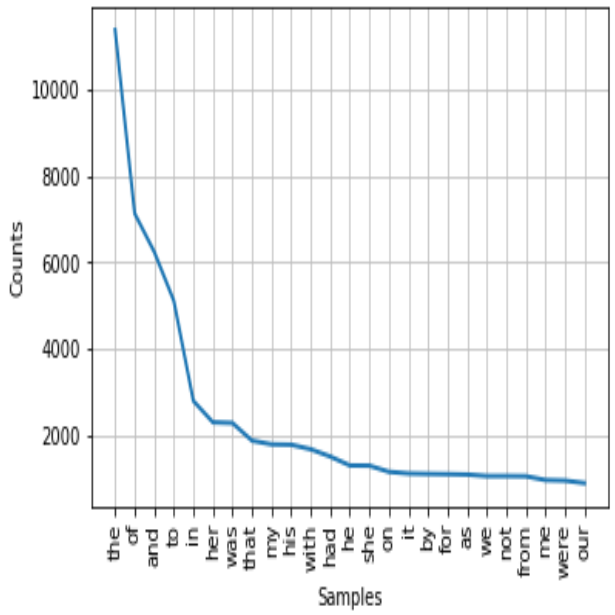
Plot (1) Characteristic Curves

Frankenstein	The Count of Monte Cristo
	

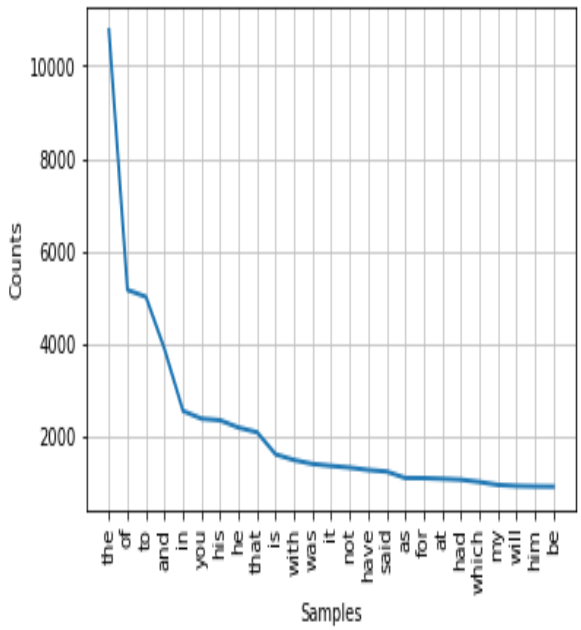
We see that in fact both authors do indeed have very different characteristics curves, and what this method theorizes is that each author's characteristic curve of a word length usage "would be so precise that it would be consistent over the lifetime of the author" (Laramee, page 10). Looking at Frankenstein, we see that Mary Shelly had a smoother inverse exponential decline in the frequency of words she most used, while Alexander Dumas has a "bump" in his characteristic curve between the second and fourth most used word. We will now look at the characteristic's curves of our six unknown text files to compare their curves with the known authors in our project. We will specifically look for curves that are either a smooth inverse exponential decline like Frankenstein or have a "bump" in its curve like The Count of Monte Cristo.

Plot (2) Six Characteristic Curves for Six Texts with Unknown Authors

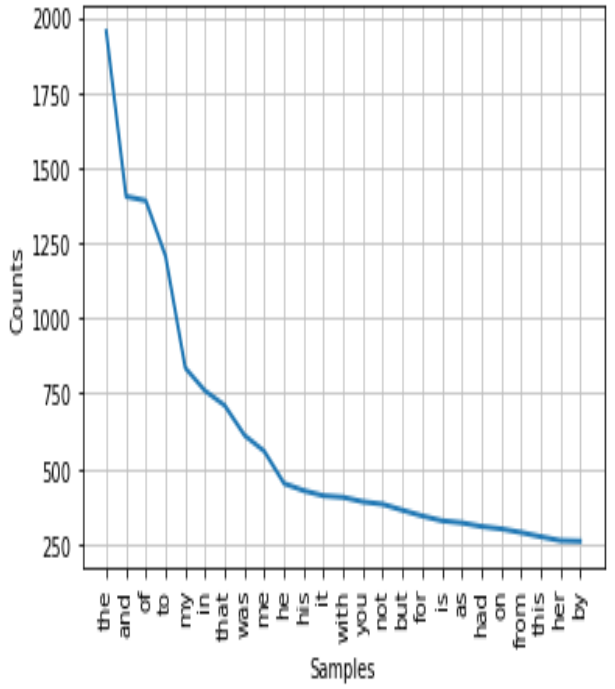
Unknown Text (1)



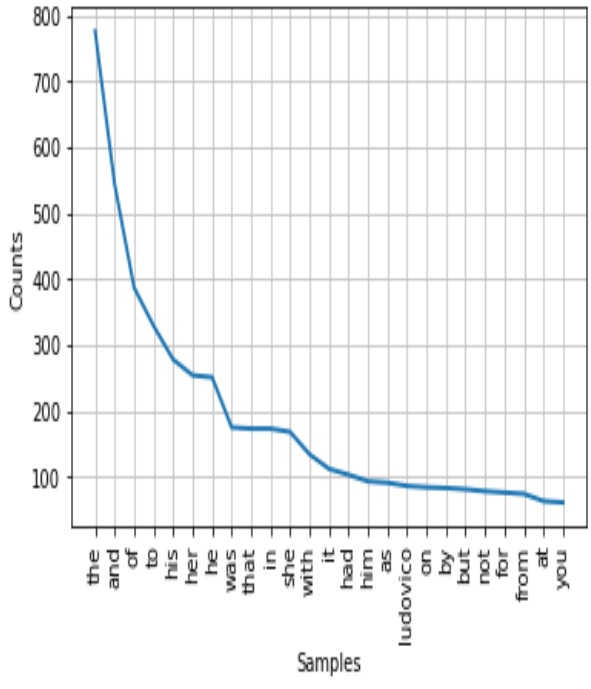
Unknown Text (2)

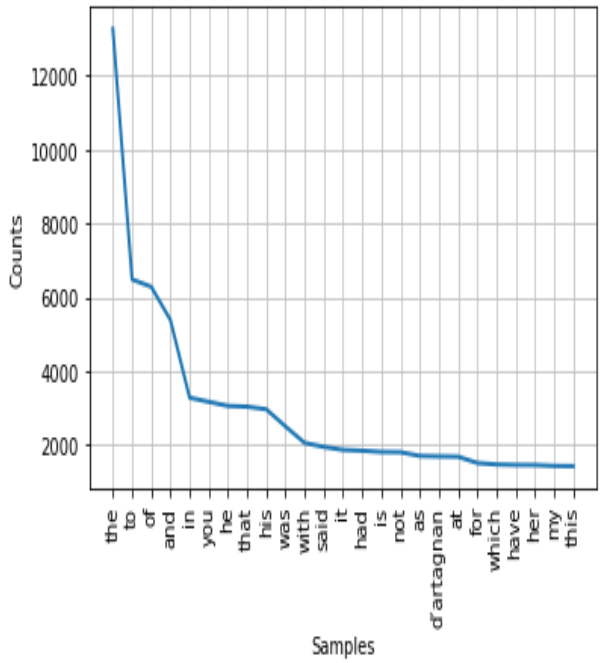
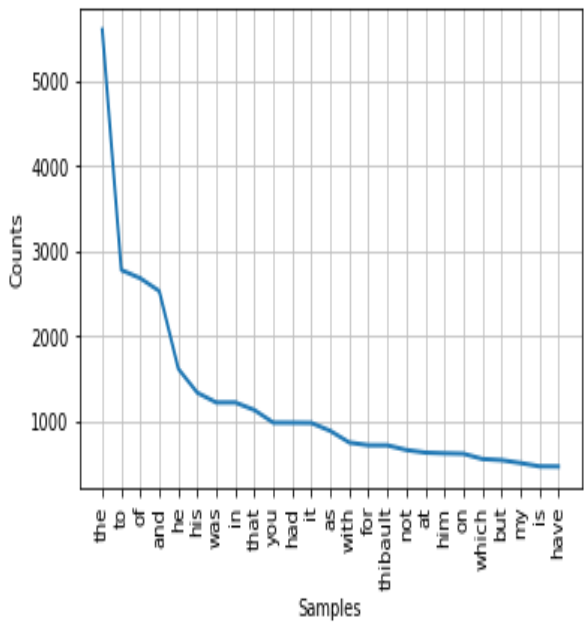


Unknown Text (3)



Unknown Text (4)



Unknown Text (5)	Unknown Text (6)
	

In Plot (2), we are given six characteristic curves for six texts with unknown authors. When comparing these curves to that of Plot (1) we can see that for unknown texts 2,3,5,6 all has similar characteristics curve patterns to that of “The Count of Monte Cristo”, while unknown texts 1 and 4 have similar characteristics curve patterns to that of “Frankenstein”. Thus, by Mendenhall’s method, we would conclude that Alexander Dumas is the author of texts 2,3,5,6 and Mary Shelly is the author of texts 1 and 4.

This, in theory, would allow us to distinguish a pattern that is unique to an author’s writing style and allow us to identify an authors work as belonging to them. However, Thomas Mendenhall’s method did not consider the actual words in an author’s vocabulary nor the genre differences when comparing two or more texts. This is obviously problematic considering the goal is to be able to identify characteristics of a writing style for authors that would aid in or investigation into the identity of authors for any texts. Furthermore, we will soon find that the

conclusions made regarding the true authorship of the unknown text files was in fact wrong when we use the Delta method later in the project. We must therefore disregard this method as being non trustworthy for our stylometry analysis and therefore find other methods which would be more accurate.

(2) Adam Kilgarriff's Chi-squared Method Test

The chi-squared test statistic method is usually used when wanting to compare the probability distribution between two sets of observations to test whether the two sets have equal probability distributions or pattern. However, Adam Kilgarriff in 2001 proposed using this test statistic to determine authorship of texts. He stated that the test statistic will allow us to "measure the distance between vocabularies employed by authors in two sets of text" (Laramee, page 14). The hypothesis is that the more similar the vocabularies would indicate a higher likelihood of the text being written by the same author. There are some assumptions with this method however, namely that we assume the vocabulary word usage pattern of an individual would remain relatively constant. Right away we see the advantage of this method when compared the Mendenhall's characteristics curves method because Kilgariff's method accounts for the range of an author's vocabulary.

To conduct the chi-squared test we must first have a frequency count of all words in a text but, we must be carful not to count lower case and capital case words separately rather we count them as one word. To not run into the risk of giving undue importance to a word that occurs very rarely, this method suggests to only compare between 100 to 1000 of the most common words between two authors. We will use the 500 of the most common words in a text in

our test. We next merge the two sets of 500 most common words between he two authors into one set in order to conduct our chi-squared test. Note the test follows the following formula

$$\chi^2 = \sum_i \frac{(c_i - E_i)^2}{E_i}$$

where c_i represents the frequency of word I in the merged set and E_i represents the frequency of the word I, in the separate set belonging to the individual author's text. In theory, the smaller the chi-squared value, the more similar the two sets of words are thus the higher the likely-hood that these two authors are one in the same in terms of writing style.

Table (4): Frequency of Top 500 Words

	Frankenstein Frequency	Monte Cristo Frequency
1	4152	27779
2	2958	12637
3	2765	12628
4	2638	11598
5	2082	9118
…	…	…
496	17	55
497	17	51
498	17	187
499	17	7
500	17	38

In Table (4) we see the frequency of the top 500 words that will be used to compare distance between words from our two texts. If the distance is small, i.e., the writing style is similar, we should expect a smaller test statistic. Our null hypothesis and alternative hypotheses are,

$$H_o: two\ two\ texts\ were\ written\ by\ the\ same\ author$$

$$vs.$$

$$H_a: the\ two\ texts\ were\ not\ written\ by\ the\ same\ authors$$

after calculating our chi-squared test statistic, we find that $\chi^2 = 625156$. Because the test statistic is so large, we can conclude that vocabulary between the two texts is so dis-similar that they could not have been written by the same author, so we reject our null hypothesis.

Although this test statistic showed that the two texts do indeed belong to two separate authors, which we had already known, there are still questions surrounding the accuracy of this test. Namely, words that appear very frequently tend to carry a disproportionate amount of weight for the test statistic, this can cause the test statistic to overlook writing styles that have subtle differences by the ways in which authors may use uncommon words or phrases. Furthermore, this method is not efficient when trying to compare anonymous texts to many different authors at the same time. Thus, we will now introduce a new stylometric analysis method that considers the faults of the previous two methods and resolves them.

(3) John Burrow's Delta Method

John Burrow's Delta method, like Kilgariff's chi-squared method, "measures the distance between a text whose authorship we want to ascertain" (Laramee, page 26). Unlike the chi-squared method, the Delta method is designed to compare texts whose authorships are

unknown to many known works by different authors. More precisely “the Delta method measures how the anonymous texts and sets of texts written by different known authors all diverge from the average of all of them put together” (Laramee, page 26). Also, the Delta method gives equal weight when measuring different features in order to avoid the problems of from the chi-squared method of giving different proportions of weight for more common words.

The statistic for Burrow’s Delta method is as follows.

$$\Delta_c = \sum_i \frac{|Z_{c(i)} - Z_{t(i)}|}{n}$$

Where $Z_{c(i)}$ is the z-score for the word “i” in candidate “c” and $Z_{t(i)}$ is the z-score for the word “i” in the test case. The Symbol n represents the most frequent words in the training data set to be used. To calculate the z-score we use the following equation.

$$Z_i = \frac{C_i - \mu_i}{\sigma_i}$$

Where C_i is the observed frequency for the word “i”, and μ_i, σ_i represent the mean and standard deviation of the word “i” respectively in the dataset. It should be stated that for Burrows Delta method, if the statistic provides a high number such as three, this will imply that the two authors writing styles are too dissimilar to be the same author. Thus, the lower the Delta statistic means the greater the likelihood of authorship to a particular author.

To conduct this test, we will be following Thomas woods approach to Burrows Delta method, where he provides the python library “faststylometry” which allows us to compare authors based on their writing style utilizing Burrows Delta method.

To conduct this test, we will be implementing six additional texts to our data set, where three texts belong to Alexander Dumas and three belong to Mary Shelly. We will compare how to test statistics compare using both a cleaned-up version of the texts, note that this means no

additional information was added to the original novel in the txt file, to that of the original txt file downloaded from project Gutenberg.

Table (5): Burrows Delta Method on Clean up Texts (1) & (2)

Authors	Unknown Text (1)	Unknown Text (2)
Alexander Dumas	1.987	1.272
Mary Shelly	1.635	1.655

Table (6): Burrows Delta Method on Project Gutenberg Texts (1) & (2)

Authors	Unknown Text (1)	Unknown Text (2)
Alexander Dumas	1.928	1.356
Mary Shelly	1.644	1.734

Based on the Burrows Delta test statistic it appears that the unknown text (1) belongs to Mary Shelly, and the unknown text (2) belongs to Alexander Dumas. This is in fact true because the unknown text (1) is in fact “Mathilda” written by Mary Shelly and the unknown text (2) was “The Three Musketeers” written by Alexander Dumas. Furthermore, it appears that the extra information provided in the original text by project Gutenberg had increased the test statistic for the authors to be the true authors of the texts and decreased the test statistics for the authors that are known to not be the true authors of the text. The reasoning behind the test statistic decreasing could be attributed to the similarities in the addition of extra content by project Gutenberg for both the training data sets and the test data sets.

We will now continue our test by providing the Delta test statistic on our next set of unknown texts.

Table (7): Burrows Delta Method on Clean up Texts (3) & (4)		
Authors	Unknown Text (3)	Unknown Text (4)
Alexander Dumas	2.624	1.759
Mary Shelly	1.973	1.871

Table (8): Burrows Delta Method on Project Gutenberg Texts (3) & (4)		
Authors	Unknown Text (3)	Unknown Text (4)
Alexander Dumas	2.361	1.615
Mary Shelly	1.906	1.765

Based on the Delta statistic it appears that the unknown text (3) belongs to Mary Shelly and unknown text (4) belongs to Alexander Dumas. Again, this is in fact true because unknown text (3) is “The Heir of Mondolfo” written by Mary Shelly and text (4) is “The Wolf Leader” written by Alexander Dumas. Furthermore, it again seems to be the case that the texts that are not cleaned up have a lower test statistic as compared to the texts that are cleaned up, which again can be contributed to the writing style of project Gutenberg being similar throughout all the texts that they add to their website, and this distorts the true test statistic when applying it to the original texts.

Now we will look at the last two sets of unknown texts that we have downloaded from project Gutenberg.

Table (9): Burrows Delta Method on Clean up Texts (5) & (6)

Authors	Unknown Text (5)	Unknown Text (6)
Alexander Dumas	0.986	2.013
Mary Shelly	1.656	1.264

Table (10): Burrows Delta Method on Project Gutenberg Texts (5) & (6)

Authors	Unknown Text (5)	Unknown Text (6)
Alexander Dumas	1.039	1.963
Mary Shelly	1.630	1.252

Looking at the unknown text (5), we see that Alexander Dumas is the author based on the Delta statistic and this is in fact true because unknown text (5) is “The Man in the Iron Mask”. Unknown text (6) is determined to be written by Mary Shelly by the Delta statistic and this is in fact true because unknown text (6) is in fact “The Last Man”. When looking at the test statistics between the cleaned-up texts and the downloaded texts from project Gutenberg, we have smaller test statistics for all the cleaned-up texts as compared to the downloaded texts from project Gutenberg. This behavior is unique when compared to Tables (5) and (6) as well as Tables (7) and (8). It appears that the added writing from project Gutenberg is too dissimilar between the

training data sets containing the novels “Frankenstein” and “The Count of Monte Cristo” and the test data set.

To conclude, we see that utilizing Burrows Delta test statistic allowed us to compare multiple unknown texts with multiple authors more efficiently as compared to the chi-squared method and the test statistic does not put disproportional weight on more common words as compared to uncommon words which was another disadvantage of the chi-squared method.

Conclusion

Stylometric analysis, although it is still in its infancy of becoming an established tool of data science, has the potential to assist in fields of identifying texts, notes, or any papers as belonging to individuals with a certain degree of error. Historical texts, forensic evidence, claims of plagiarism, all can be supported by this field of data science.

In this project, we had looked over the earliest methods of stylometric analysis such as Mendenhall’s characteristic curve to that of more advanced and recent methods like Burrow’s Delta method. Furthermore, when comparing the accuracy of the characteristic curve’s method to the Delta method we can see the faults in the assumptions of Mendenhall’s method. Mendenhall’s method relies on the interpretation of the individual conducting the test to state whether a curve plot resembles that of another author’s curve plot and during this project we had conclude that a text, which does not belong to an author, was written by that particular author based on their curve plot. Although Mendenhall’s method was inaccurate as compared to the Delta method, it was this method that laid the foundation of current stylometric analysis techniques.

What I would like to have done, if given more time, is to further explore more accurate methods that may be available or build upon the current Delta method and apply this to historical texts that are currently being argued over the identity of authorship. Also, pertaining to the Delta method, I would have like to have presented the actual probability of authorship to unknown texts in addition to presenting the Delta statistic.

Sources

1. Laramée, F. D. (2018, April 21). *Introduction to stylometry with python*. Programming Historian. Retrieved November 29, 2021, from https://programminghistorian.org/en/lessons/introduction-to-stylometry-with-python.
2. *Fast stylometry tutorial - freelance data scientist: Thomas Wood*. Freelance Data Scientist | Thomas Wood. (2021, April 9). Retrieved November 29, 2021, from https://freelancedatascientist.net/fast-stylometry-tutorial/.
3. *Thomas C. Mendenhall issues one of the earliest attempts at stylometry*. Thomas C. Mendenhall Issues One of the Earliest Attempts at Stylometry : History of Information. (n.d.). Retrieved November 29, 2021, from https://www.historyofinformation.com/detail.php?id=4120.
4. GoTrained. (2019, January 10). *NLTK corpus*. GoTrained Python Tutorials. Retrieved November 29, 2021, from https://python.gotrained.com/nltk-corupus/.
5. *Free ebooks | project gutenberg*. (n.d.). Retrieved November 29, 2021, from http://www.gutenberg.org/.