ASSIGNMENT 3

Jane Jiang S5356970

Part A. Author attribution

Li Yang S5315573, Jane Jiang S5356970

We used stylo to identify an 'anonymous' text by comparing it to several texts written by lots of potential candidate authors. First, we used the stylo software to analyze the 100 most frequently occurring words to produce the first tree diagram. By analyzing the 100 most frequent words, stylo divided the 26 known authors into 17 branches. As for the authors of the 'anonymous' text, based on the similarity between the 100 most frequent words, the tree diagram shows that stylo divided the text of ANONYMOUS into a branch with Dickens Bleak and Dickens David. Moreover, in this one branch, the 'anonymous' text is closer to Dickens Bleak. Thus, according to stylo's analysis, we would think that the author of the 'anonymous' text is likely to be Dickens Bleak under the division of the 100 most frequently occurring words.

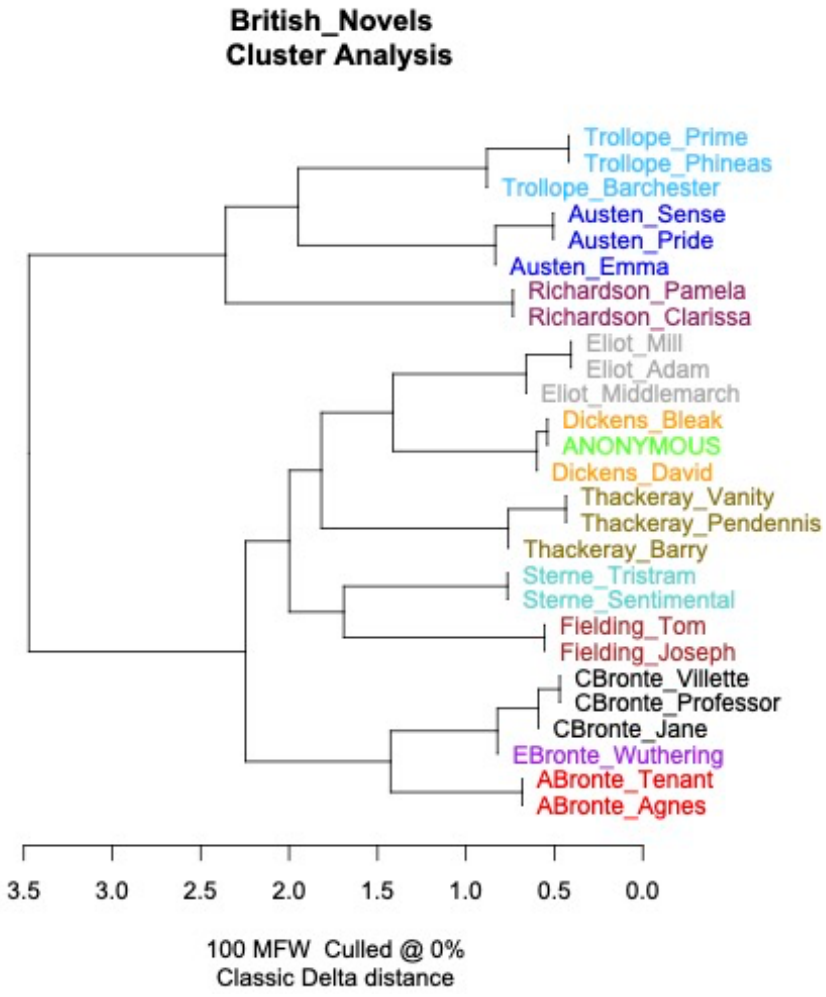

Figure 1. 100 Most Frequent Words

In the second step, we used stylo to analyze the 100 most frequent 2-grams in each text. The 100 most frequently used 2-grams by each author were analyzed to compare the stylistic differences between the texts written by different authors. The following tree diagram was derived from the stylo analysis. Through this tree diagram, we can see that, unlike the first step of comparing the 100 most frequent words, the authors were reclassified into 16 branches in the analysis of the 100 most frequent 2-grams. In the analysis of the most frequent 2-grams, the guesses about the authors of ANONYMOUS texts do not differ much from the results of the first graph. stylo analysis, the authors who wrote ANONYMOUS texts use 2-gram habits still closer to Dickens Bleak and Dickens David, and in this branch, the usage habits of authors of ANONYMOUS texts are closer to Dickens Bleak. Therefore, after using stylo analysis, we believe that the author of the 'anonymous' text is most likely to be Dickens Bleak based on the division of the 100 most frequent 2-grams.

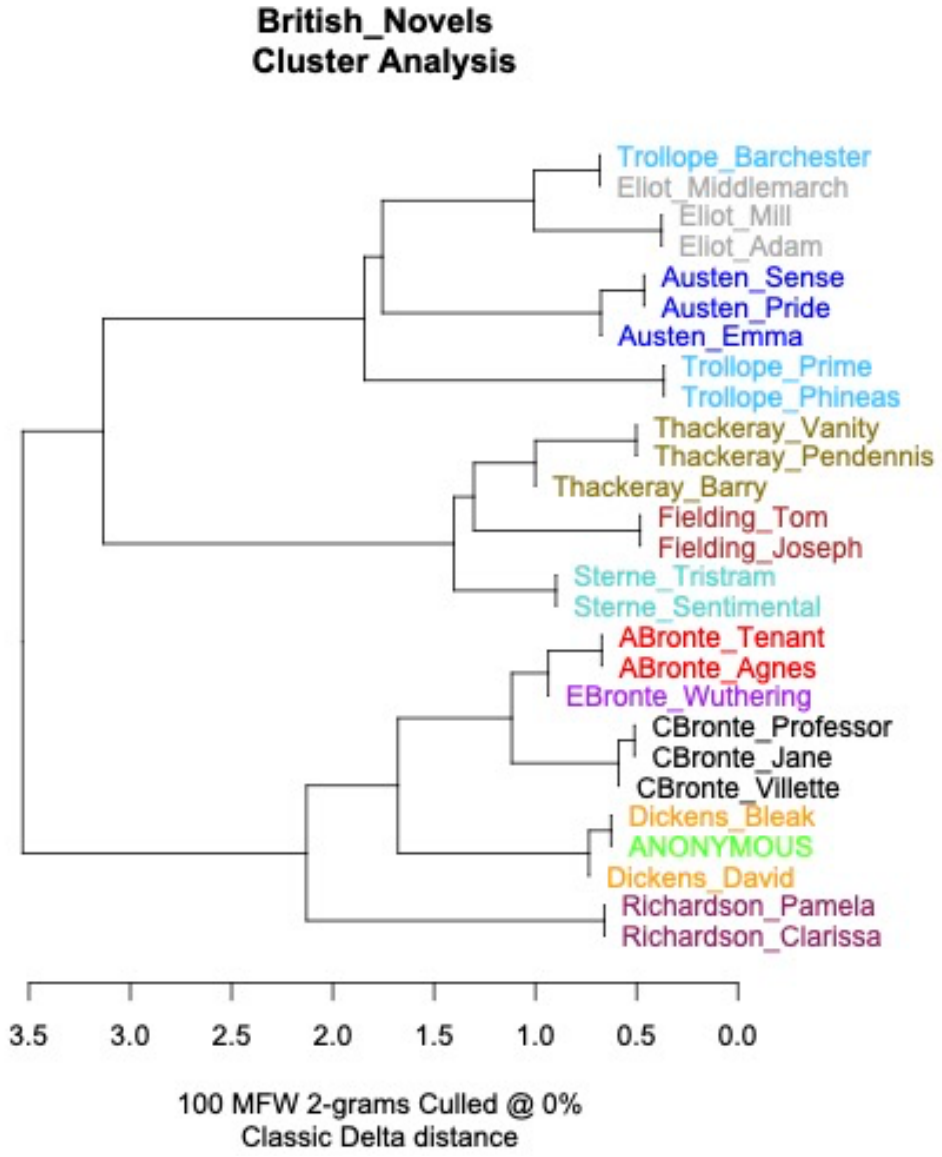

Figure 2. 100 Most Frequent 2-grams

In the third step, we used stylo to analyze the 100 most frequently used CHARACTER 4-grams. This time the results of the author's speculation about the ANONYMOUS text differed from the previous two analyses of the 100 most frequently occurring words and the 100 occurring 2-grams answers. The tree diagram derived from the analysis of the 100 most frequent CHARACTER 4-grams by stylo has changed this time. Although the 26 known text authors were still divided into 17 branches this time, the classification according to the 100 most frequently used character 4-grams showed that the ANONYMOUS text authors still wrote texts in style closer to the writing habits of Dickens Bleak and Dickens David. However, the difference in the analysis of the writing habits of the 100 most frequently used character 4-grams is that this time the ANONYMOUS author's distance is closest to that of Dickens David, indicating that authorship, in this case, is more biased toward Dickens David. This is different from the previous two graphs, which are biased toward Dickens Bleak is different.

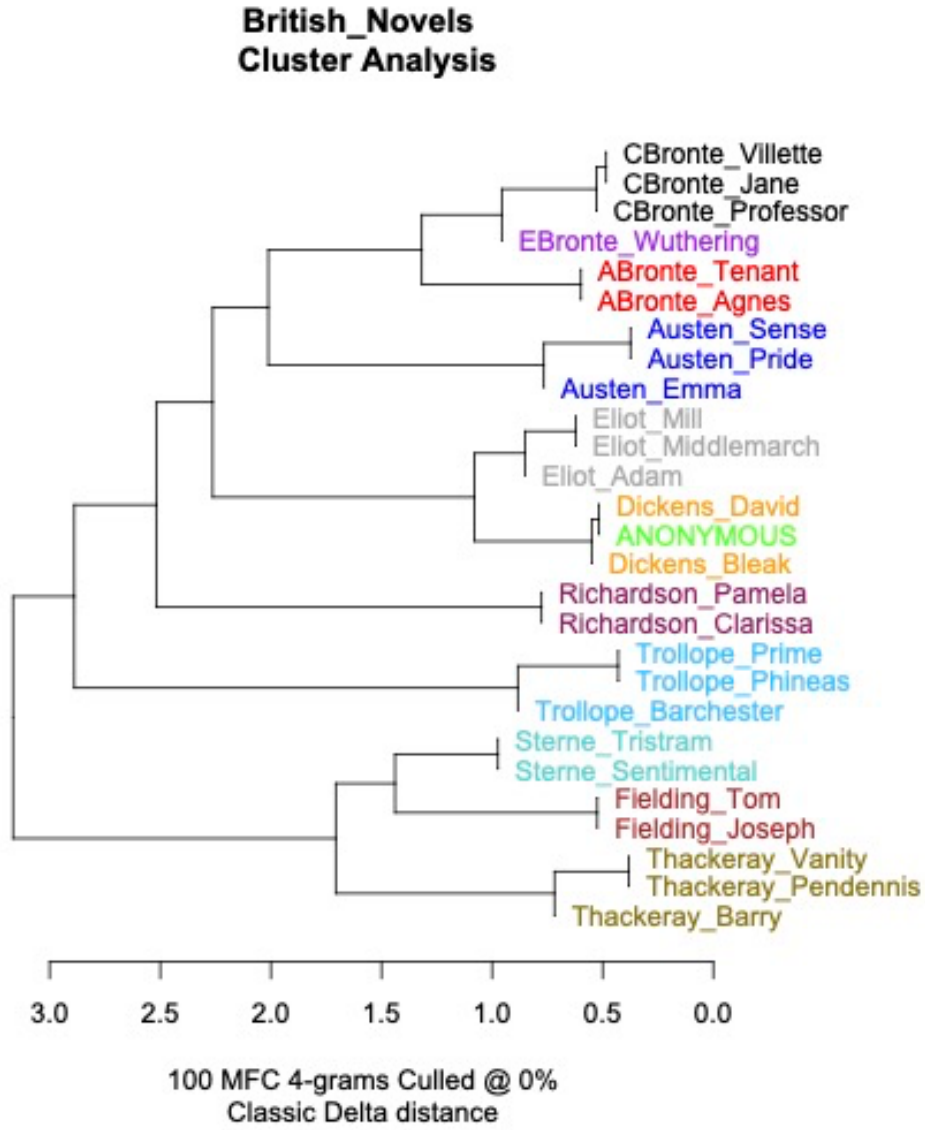

Figure 3. 100 most frequent character 4-grams

Based on the results of the three dendrograms, we finally concluded that "Dickens Bleak" is the most likely author of the 'ANONYMOUS' text.

Part B. Stylometric analysis of gender or genre

In this assignment, I chose to analyse the word preferences of male and female authors respectively in a corpus of British fiction. The corpus of texts for female authors was set to Preferred and the corpus for male authors was set to Avoided, and the 70 words that were used most frequently by male and female authors respectively were presented by calculating craig's Zeta scores through stylo.

As can be seen from the graph, the x-axis represents the ranking, with the frequency of words decreasing from left to right. The y-axis shows the frequency of the words. The farther away from the dotted line in the middle means that the word appears more frequently. Above the dotted line are the 70 most frequent words in the text base of women authors of British fiction, while below the dotted line are the 70 most frequent words in the text base of men authors of British fiction.

From Figure 4, we can see that four of the most frequent 10 words used by female authors are variations of feel, and a total of five words are used in relation to sensation, suggesting that female authors tend to use words related to experience and feeling in their writing. In contrast, the 10 most frequent words used by male authors are more often words related to honor, or tendency to power, suggesting that British male authors tend to use more power-related words in their writing. And it is not just the 10 most frequent words that seem to be the basis of our speculation after a closer look at the words in Figure 4.

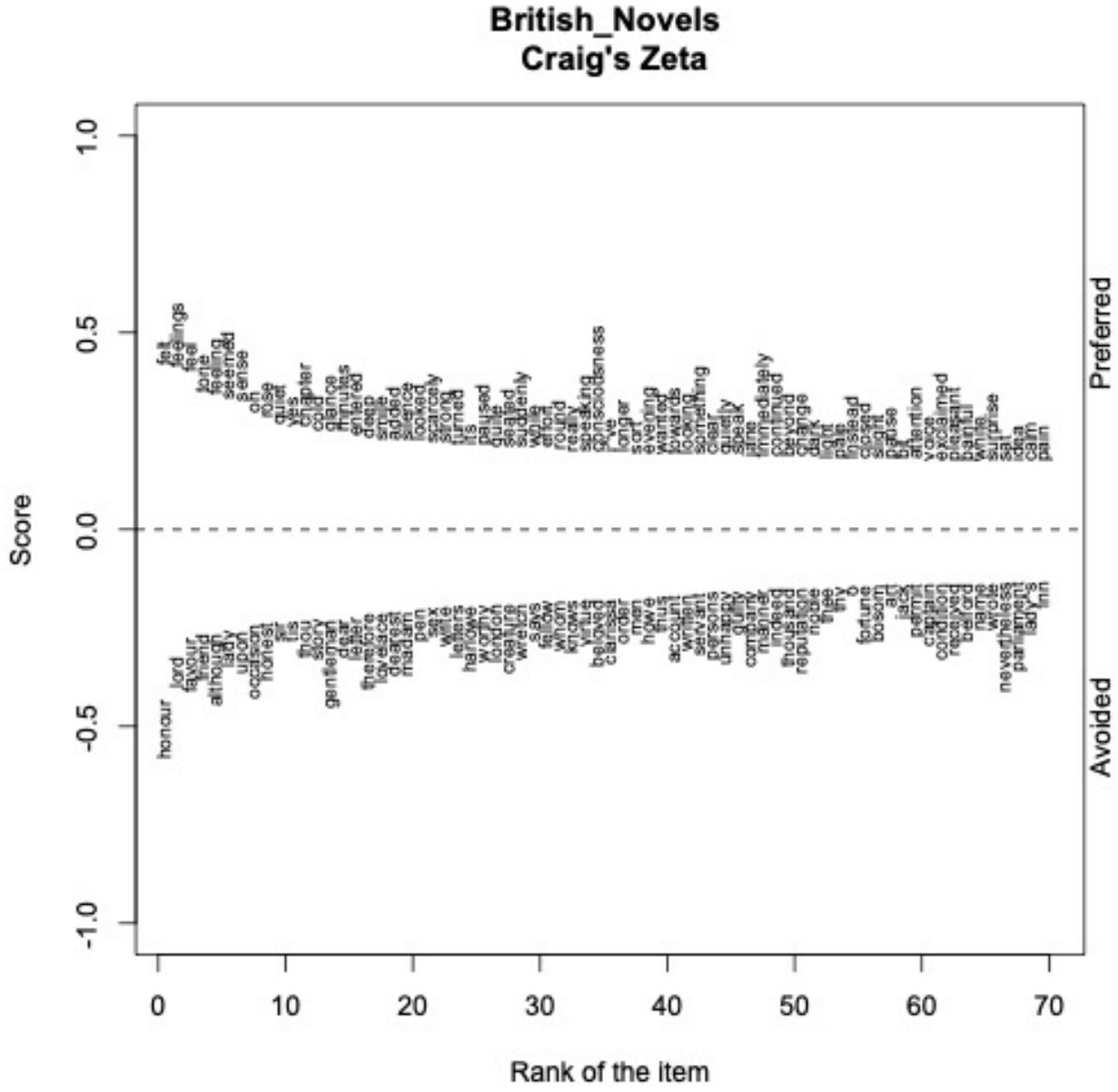

Figure 4

As mentioned above, we can easily obtain the 70 most frequently used words from Figure 4 and further analyse them with annotations and statistics, but the opposition function has some limitations. First of all, due to a flaw in the function itself, the graph generated is horizontal rather than vertical, which makes it inconvenient to read. Also, although the function reflects the most frequently used words in each of the two text corpora, the specific frequency of the words is not shown. At the same time, only the difference in word frequency between the same text corpus is intuitive, and the comparison of word frequency distributions between different text corpora is not obvious enough: the specific distance between words and the middle 0 is difficult to determine with human eyes. This makes it difficult to compare the frequency of the same word in different text collections.

However, despite some limitations of the contrastive function, we can draw some conclusions: it is certain that the 'masculine' style tends more towards power, while the 'feminine' style

tends more towards emotions expression. It is also possible to formulate some hypotheses based on this stylistic difference: Is and could this difference apply to all writers and not just novelists? Perhaps, on a larger scale, does this mean that women are more focused on feelings while men prefer power? These are all questions that we might find in this chart of the difference in frequency of words used by male and female authors.

Furthermore, we can also analyse these 70 words for lexical style and study the differences in the use of words between female and male authors. Alternatively, an active or passive analysis of these words could be carried out to examine the preferences of female and male authors in terms of the active and passive use of words, and further research could be done on these words to discover more stylistic differences between female and male authors.

In conclusion, from the results of the opposite-function we can clearly obtain lists of the most frequently used words for each text. We can also draw some conclusions and inspirations by roughly determining the difference between the two. However, due to the lack of specific frequencies and the way in which they are presented visually, the comparison of frequencies between the different lists is not as obvious, which makes it difficult to compare the frequencies of the same ranked words or the frequencies of the same words used in different text corpora.