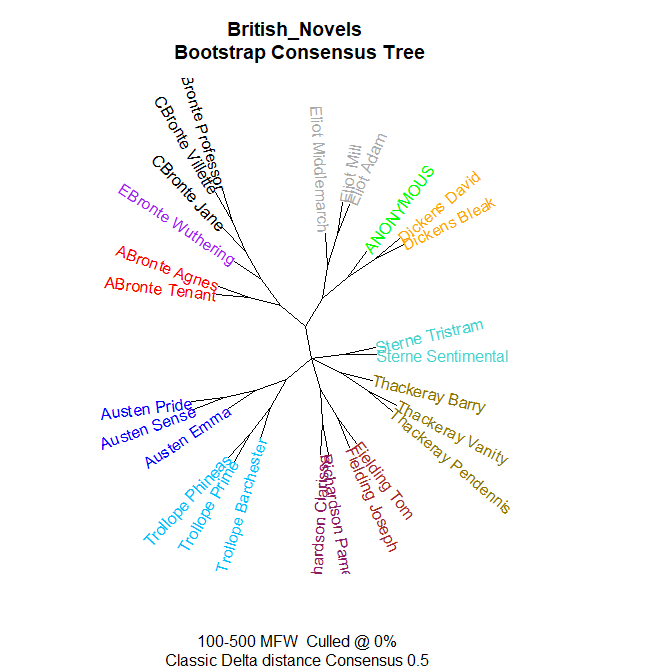
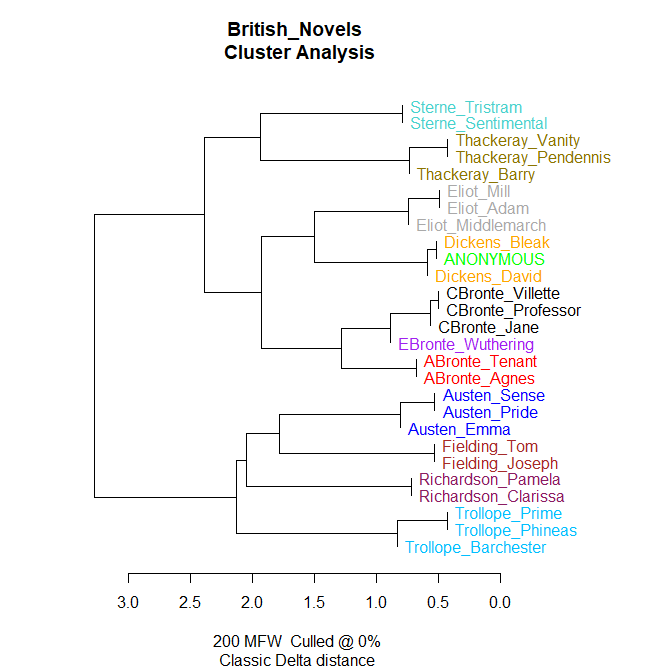
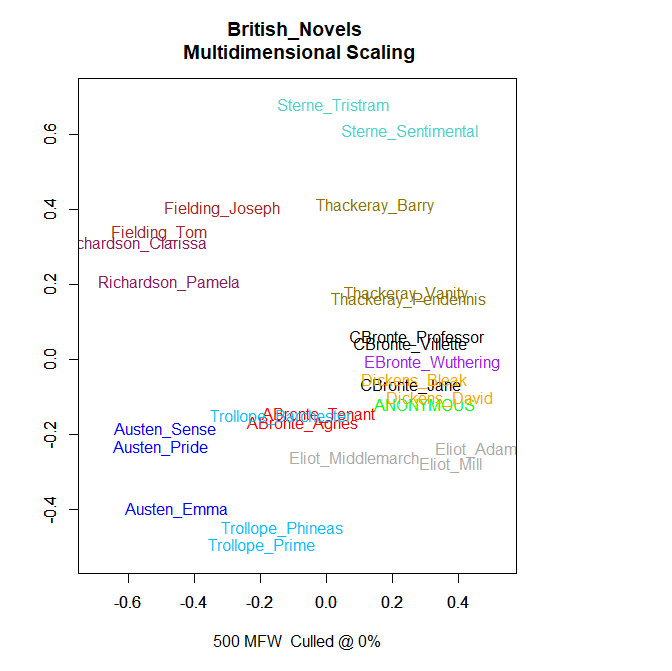
Assignment 3 Part A

**Haozhe Bai S5171326**

**Findings of the research**





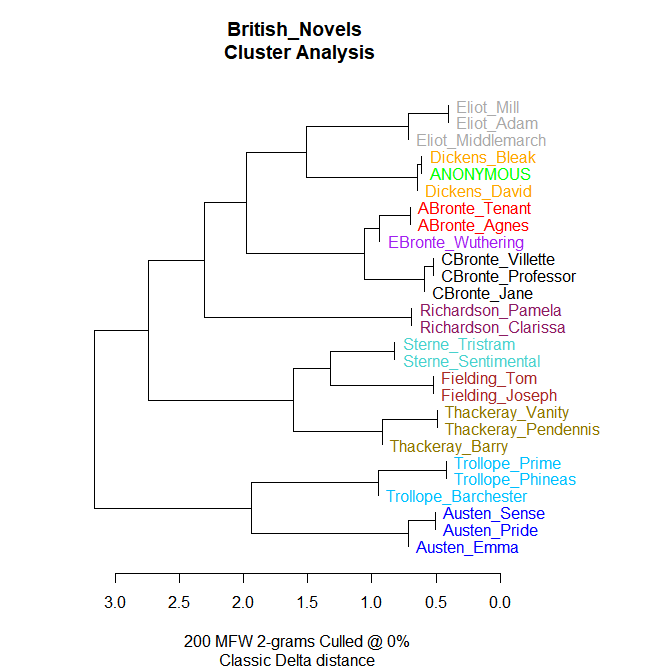
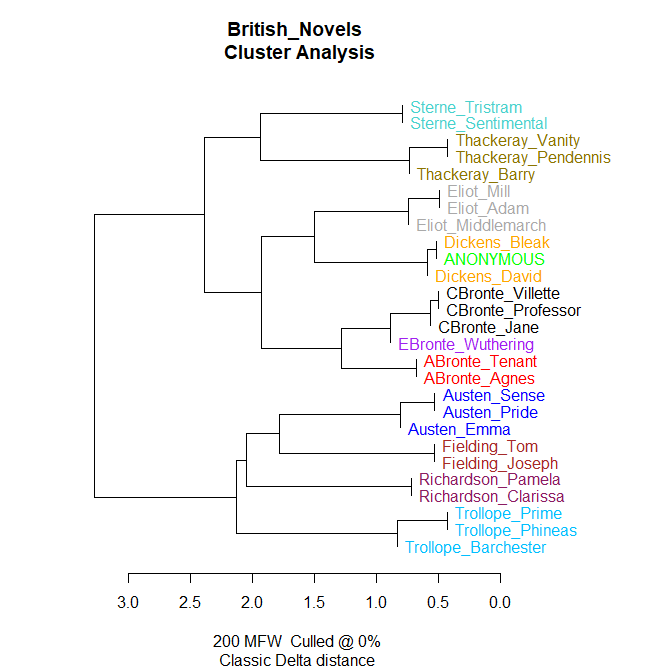
The author should be Chares Dickens

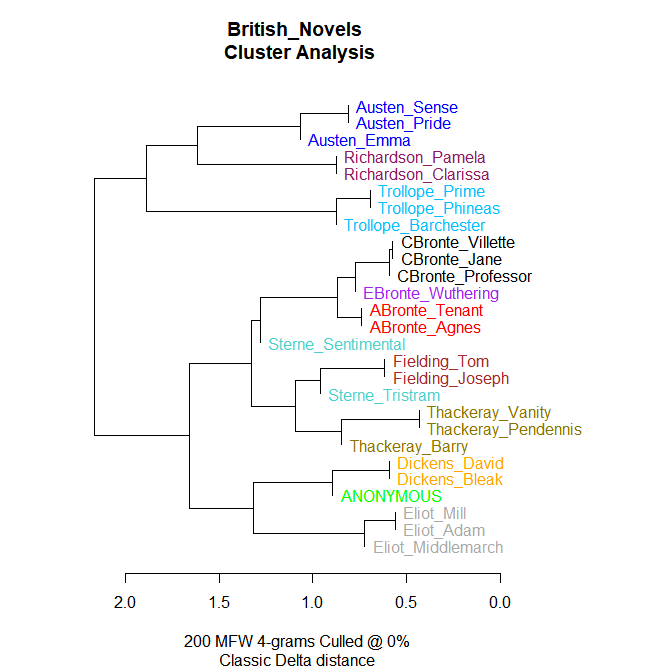
I used the Cluster Analysis to analyse the text content in the files. At last the photo shows that compared to other authors and books that anonymous one is much more familiar with the work of Dickens BLEAK HOUSE and DAVID COPPERFIELD. Just like the knowledge that we had learnt in the class the distance between two documents = the absolute mean of the differences between the frequencies of a set of features (F₁, F₂, F₃, F₄) in those documents. The graph shows the difference between authors clearly. Besides the connection with Dickens, we still can find that the novel written by Dickens is similar to Eliot’s work.Not only the Cluster Analysis, but also the Bootstrap Consensus Tree put two authors’ work in the same branch. In terms of the Bootstrap Consensus Tree, The anonymous work is in the same branch with BLEAK HOUSE and DAVID COPPERFIELD. They have more samilitaries than other works. And according to the Multidimensional Scaling this work around the Dickens work.

We still can find something interesting in the three graphs. The Bronte family authors have a very similar writing style. Their works are shown in the same branch in both Cluster Analysis and Bootstrap Consensus Tree. As a family they have similar education and they can learn from each other as well as they can influence each other. Based on that their works have a high degree of similarity.George Eliot and Charles Dickens are both considered realist literary authors in a sense, and their works depict real-life tragedies. Due to the similarity of the subject matter, their works also show a high degree of similarity. And both Sterne's and Thackeray's works lean toward romanticism. This characteristic is well reflected in Sterne's work The Life and Opinions of Tristram Shandy, Gentleman. This is a stream-of-consciousness novel. The author enjoys great freedom and flexibility in the narration of the story.

By the way we still can find that different analysis methods can show different results. From the Cluster Analysis we can find that Fielding’s work has more similarity with Austen than Richardson. But when we change the index, using the Bootstrap Consensus Tree, the result is a little bit different. It shows the Fielding and Richardson are in the same branch. Austen and Trollope are in the same branch. So we can find out that if we change the methods the outcome will be changed as well. If we want to the the specific answer we have to do more analysis but we can not just rely one single method.

**Change parameters of n-gram size**





When we adjust the value of the n-gram size, we will get a different result. Although, this does not affect the correct result. It is interesting to compare the differences between them and this can help us to enhance our understanding of stylo.

First of all, we focus on the relationship between ANONYMOUS and two other Dickens' works. When we set the n-gram size to 1 and 2, the program considers ANONYMOUS to be more similar to Bleak and less similar to David. But when I set the n-gram size to 4, the program thought that Bleak and David were more similar and that ANONYMOUS was not very similar to either work. Overall, the program was able to classify works by the same author together regardless of whether the value of n-gram size was 1, 2 or 4. The only exceptions are Steme's Tristram and Sentimental novels. When we set the n-gram size to 4, the program considered the novels by Tristram and Fielding to be more similar, and the novels by Sentimental and CBronte, EBronte and ABronte to be more similar. This suggests that something is wrong with the setting of the parameters.

**Research  Expansion：A Comparative Study of Science Fiction Writers**

[I wish to use cluster analysis to examine science fiction by some of the leading contemporary science fiction authors. This is also the research topic of our group assignment. I have downloaded the works of 7 famous science fiction authors from 19c-20c from the website Project Gutenberg, a total of 27 novels. Since the website has very limited textual resources. Therefore, I was not able to collect all the works of the 7 authors. Dataset is a txt file and I have named the files using the initials\_title\_date\_of\_publication. Apart from the comparison of authors, we can also compare these novels in terms of year of publication.](https://www.gutenberg.org/)

[](https://www.gutenberg.org/)



In these two images we can find two works by H. G. Wells that show a high degree of similarity to Isaac Asimov's novels. I think this is an interesting phenomenon. These two novels are not science fiction in the traditional sense; The red room is a novel with elements of horror. The red room is a novel with elements of horror, while Anticipations is a novel with a large section devoted to social science and philosophical issues. I think these different novels can be associated with the work of Isaac Asimov, one of the greatest science fiction writers of all time. This event illustrates the development and progress of science fiction in the decades between the late 19c and mid-20c. Another author of note is Mary Shelley, whose work bears a strong resemblance to Arthur Clark's The clipper ship era and Ephraim Douglass Adams' Great Britain and the American Civil War.