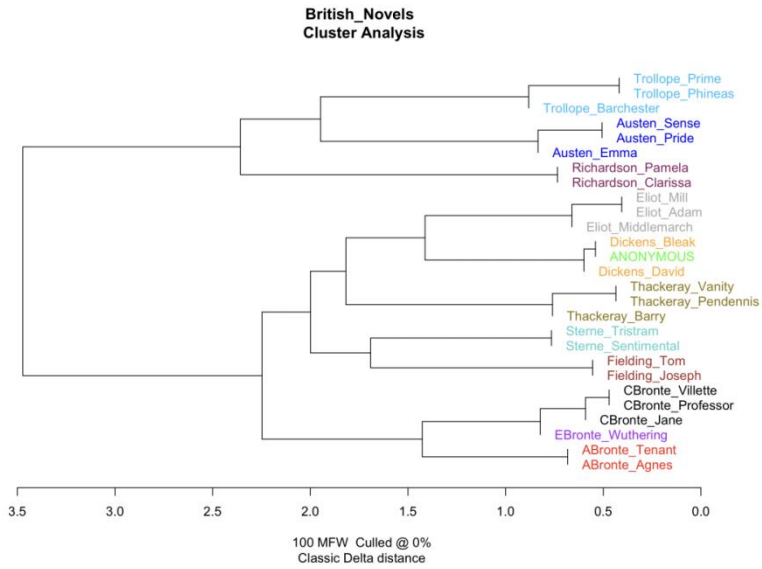
Assignment 3

Part A Collaborated with Spring Wan

We believe that the anonymous author here is Dickens. The least likely author is Richardson. Because it is the furthest from this branch, we consider them the least similar to each other. According to the structure of the tree diagram, the anonymous author is in the same branch as Dickens, similar to the branch of the author Eliot, and the structure of both is almost identical.

We run the first (cluster-) analysis, and set the parameters of our analysis: FEATURES: choose FEATURES: words, MFW SETTINGS:minimum:10; maximum:100; increment:10; start at freq.rank:1.



From the figure above, we can clearly see that a file named ‘ANONYMOUS.txt’ most likely is written by Dickens.

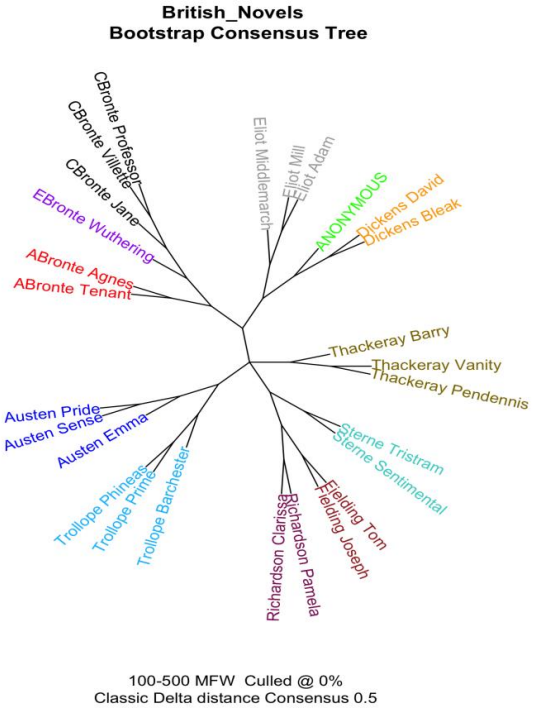
Firstly, let's look at it from an orientation perspective. This tree diagram grows horizontally from right to left, and on the left side, the clustering categories are divided into 2 major categories. Part 1 includes Trollope_Prime, Trollope_Phineas, Trollope_Barchester and so on, Part 2 includes Eliot_Mill, Eliot_Adam, Eliot_Middlemarch and so on. From the category where ANONYMOUS is located, we can easily exclude several authors in part 1.

Let's look again from the distance point of view, above we have narrowed it down to Part 2, the distance between Dickens_Bleak’s and ANONYMOUS is almost 0, while the distance between several other authors in Part 2 and ANONYMOUS is greater than 0. Therefore, we think the anonymous author is Dickens.

In addition, we also make another Bootstrap Consensus Tree using different

settings: Minimum: 100, Maximum: 500, Increment: 50. Based on the principle of similar category concentration, we can still draw the same conclusion from the above figure.

Combining the two sets of experiments with different parameters and analyses, we strongly believe that Dickens would be the most likely author based on these results.



For example, when using visual analysis of the most famous ancient Chinese novel, Dream of the Red Chamber, there is no question of the authenticity of Dream of the Red Chamber, but mainly the study of whether the chapters were written by the same author. As early as the 20th century, someone used a word-frequency-based analysis to point out that the wording habits of the first and second halves differed greatly, corroborating to some extent that the second half was continued by someone else. For this reason, we believe that if the first eighty and the second forty pages of Dream of the Red Chamber were written by the same author, then his diction and style of writing must be uniform. On the contrary, if there are obvious differences in the wording, it proves that there is a suspicion of forgery in the second forty returns.

Part B

In the female perspective of the British fiction corpus, we find that the verbs are used more frequently, from which we can infer that female writers describe most of the dialogues. In addition, the top 5 words with the highest vocabulary use are felt, feelings, feel, seem, sense, and 3 words with the lowest vocabulary usage are pain, calm, idea. In the male perspective, we find that nouns are used more frequently, so we can speculate that male writers are more inclined to use descriptive statements. In addition, pronouns such as lady, sir, gentleman, dear, and men are used more frequently. Male writers prefer to use pronouns to refer to fictional characters, while female writers use pronouns less frequently, so we speculate that female writers prefer to refer to fictional characters by their first names.

Nevertheless, the oppose-ability function's to reveal the full range of vocabulary used in a corpus is limited. This is explained by the fact that the function only compares two sets of texts and identifies the words that appear more frequently in one set than the other. As a result, this function only takes into account the most frequent words in the texts and may miss the nuances of less frequent words or phrases. For example, it could obscure aspects of language use such as sentence structure, syntax, and grammar.

Furthermore, the oppose-function obscures the vocabulary that is used in both sets of texts equally or similarly. This implies that the function may not highlight words that appear frequently in both sets of texts. The oppose-function visualizations may provide some insight into the language used in “male" and “female" styles, while they may reveal some linguistic differences between the two categories, they also have the potential to perpetuate gender stereotypes and oversimplify the complexities of language use. Furthermore, any differences observed may be influenced by factors other than gender, such as a genre, authorship, and audience.

The Sapir-Whorf Hypothesis argues that language is not merely a product of society, but can in turn influence the construction of the human mind and spirit. Language differences have both social and cultural origins, but there are also differences that cannot be simply reduced to socio-cultural causes. We need to focus not only on the linguistic differences between men and women, but also on the linguistic commonalities between men and women, so that we can place the variable of gender in a larger linguistic context and explore the relationship between gender and language in a comprehensive and objective way.

British_Novels
Craig's Zeta
Score
-1.0
-0.5
0.0
0.5
1.0
Rank of the item
0
10
20
30
40
50
60
70
honour
lord
favour
friend
although
lady
upon
occasion
honest
sir
tis
thou
story
gentleman
dear
letter
therefore
lovelace
dearest
madam
pen
sex
write
letters
harlowe
worthy
london
creature
wretch
says
fellow
whom
knows
virtue
beloved
clarissa
order
men
howe
thus
account
written
servant
persons
unhappy
guilty
company
manner
indeed
thousand
reputation
noble
thee
thy
o
fortune
bosom
art
jack
permit
captain
condition
received
belford
name
wrote
nevertheless
parliament
lady^s
inn
felt
feelings
feel
tone
feeling
seemed
sense
oh
rose
quiet
yes
chapter
cold
glance
minutes
entered
deep
smile
added
silence
looked
scarcely
strong
turned
its
paused
quite
seated
suddenly
while
effort
round
really
speaking
consciousness
i^ve
longer
sort
evening
wanted
towards
looking
something
clear
quietly
speak
jane
immediately
continued
beyond
change
dark
light
pale
instead
closed
slight
pause
bit
attention
voice
exclaimed
pleasant
painful
white
surprise
sat
idea
calm
pain
Avoided
Preferred