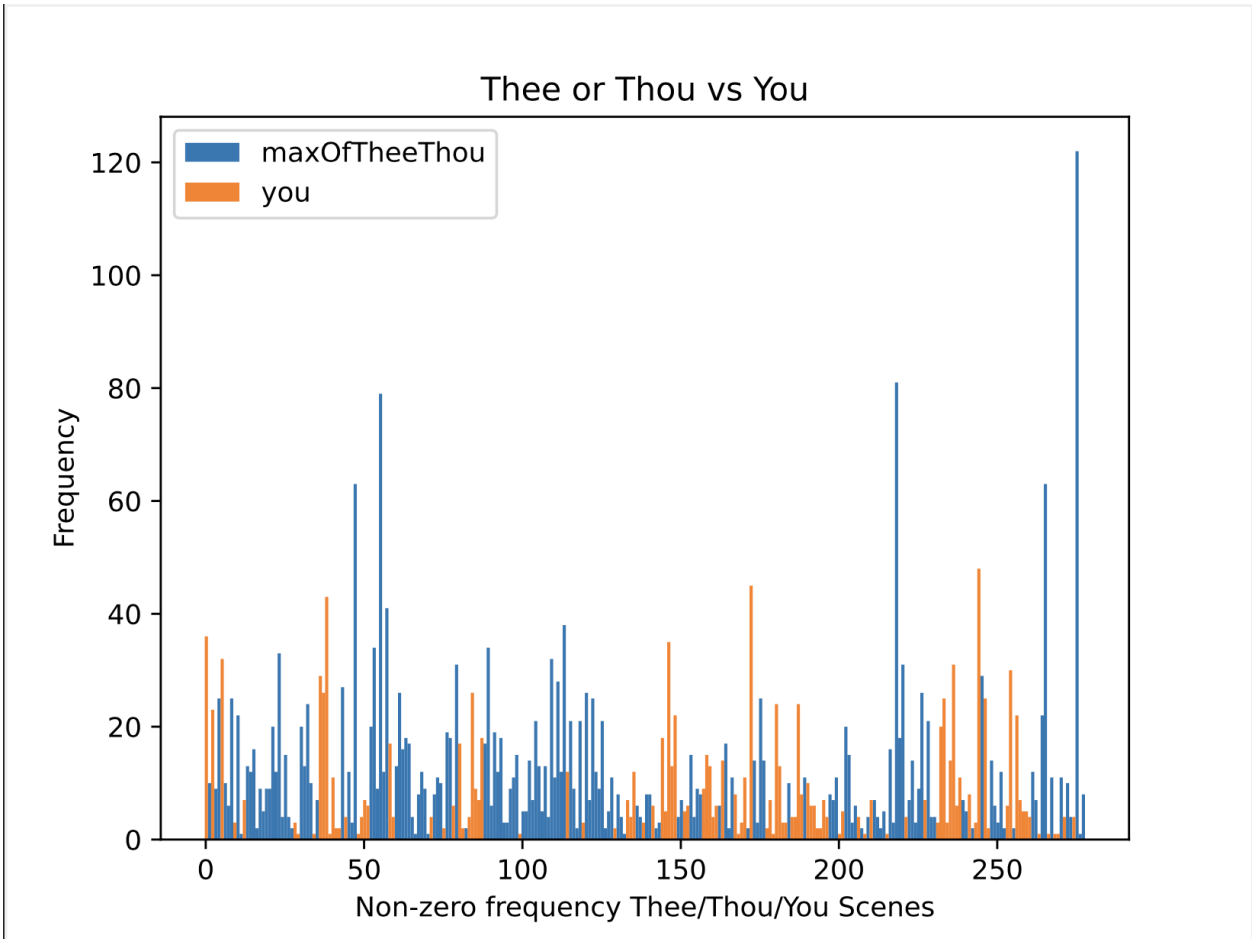
1. I indicated in the code where the Indexing and evaluation exists.
2. My indexing process was based on the given pseudo code. In order to figure out how to compare thee and thou to you, I created a compare function that does so. Then I had to check the frequencies and then check for any matching. I had trouble with the plot at first trying to figure out how I actually wanted it to look like, but then decided to do it like this in question 7.
3. In my P3.py file, I imported 3 software libraries: gzip, json, and matplotlib. I imported gzip and json so that I could open the shakespeare-scenes.json.gz file. Then I imported matplotlib so that I could create a plot of the first query.
4. The counts might be misleading features for comparing different scenes because the length of some scenes will be longer than others so the counts might be misleading. In other words, the count will be greater in the scenes that are longer. A way to fix this would be by trying to level out the larger scene into the same scale as a shorter scene so that the counts could be more accurate.
5. The query that took the longest to execute was comparing thee, thou, and you. These words were used way more often than other words which made it execute much longer. Since these words appear a lot more, the program will take longer to execute.
6. The average length of a scene is around 1199. The longest scene is loves_labors_lost:4.1 and the shortest scene is anthony_and_cleopatra:2.8. The longest play is Hamlet and the shortest play is comedy_of_errors.

Graph below

7.


Thee or Thou vs You
maxOfTheeThou
you
Frequency
0
20
40
60
80
100
120
0
50
100
150
200
250
Non-zero frequency Thee/Thou/You Scenes