P3 Indexing Report

1. The indexing occurs from line 12 to line 250. The evaluation occurs from line 266 to line 324

2. I used two maps to store the reverted list. The first map is Map<String, HashMap<String, ArrayList<Integer>>>, the key is term and the value is also a map which key is doc id and value is the list of its positions. The second map is Map<String, HashSet<String>>, the key is term and the value a set of the play id. Using these two maps, I can easily get the scene(s) of play(s) where a term occurs. For phrase-based queries, I implemented a recursive function to find the scene. For each recursion, I pick-up the first term of the phrase and return the rest phrase and a map which only contains the position neighbor to the last term. Finally, return the remain scenes. At the beginning, I used Map<String, ArrayList <Posting>> to store the reversed list, but found it hard to pick-up values. So I decided to build another type of map.

3.
    1. java.io: read file and write file to txt
    2. java.util: create Maps, Set and the max value of Integer
    3. org.json.simple: read file fro Json

4. Because counts does not contains the position detail of each term. For instance, scenes “ABBA” and scenes “AABB” has the same word counts be they are not the same. I can fix this by recording the position of each term.

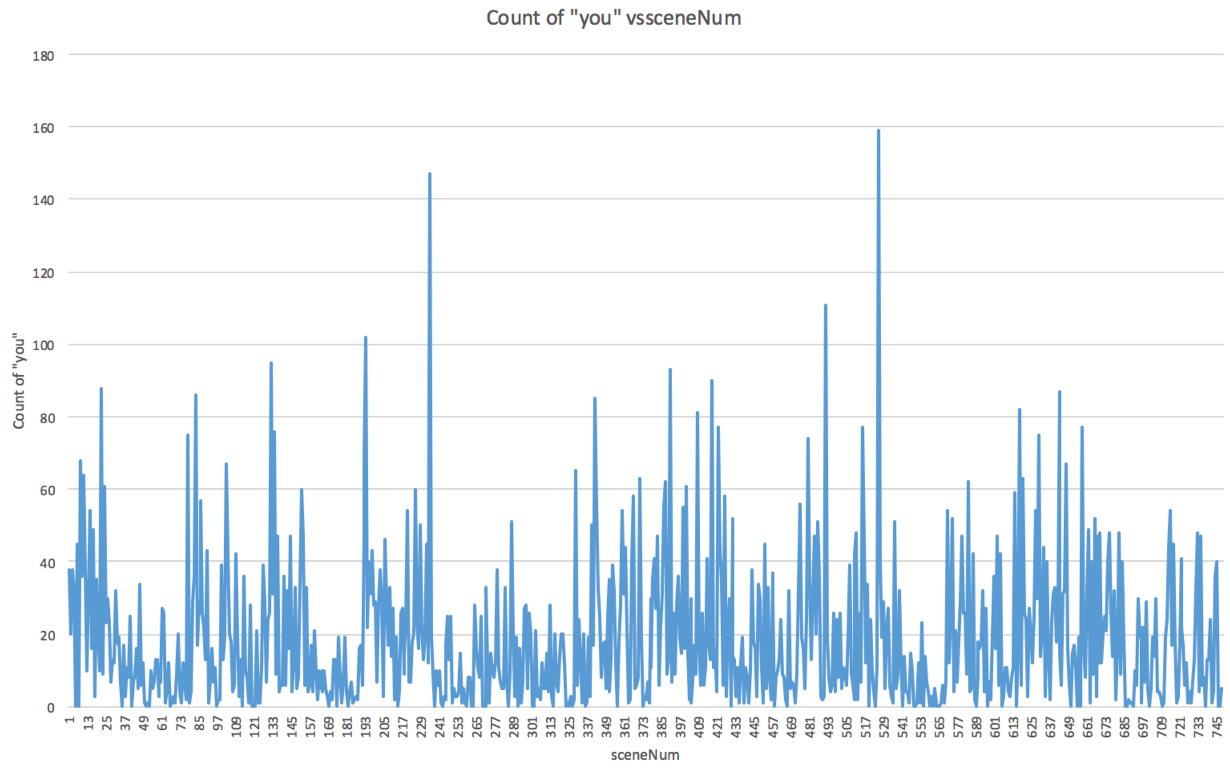
5. Term0: 2096520 ns
   Term1: 453174 ns
   Term2: 18936 ns
   Term3: 79639 ns
   Phrase0: 978357 ns
   Phrase1: 3460151 ns
   Phrase2: 200705 ns

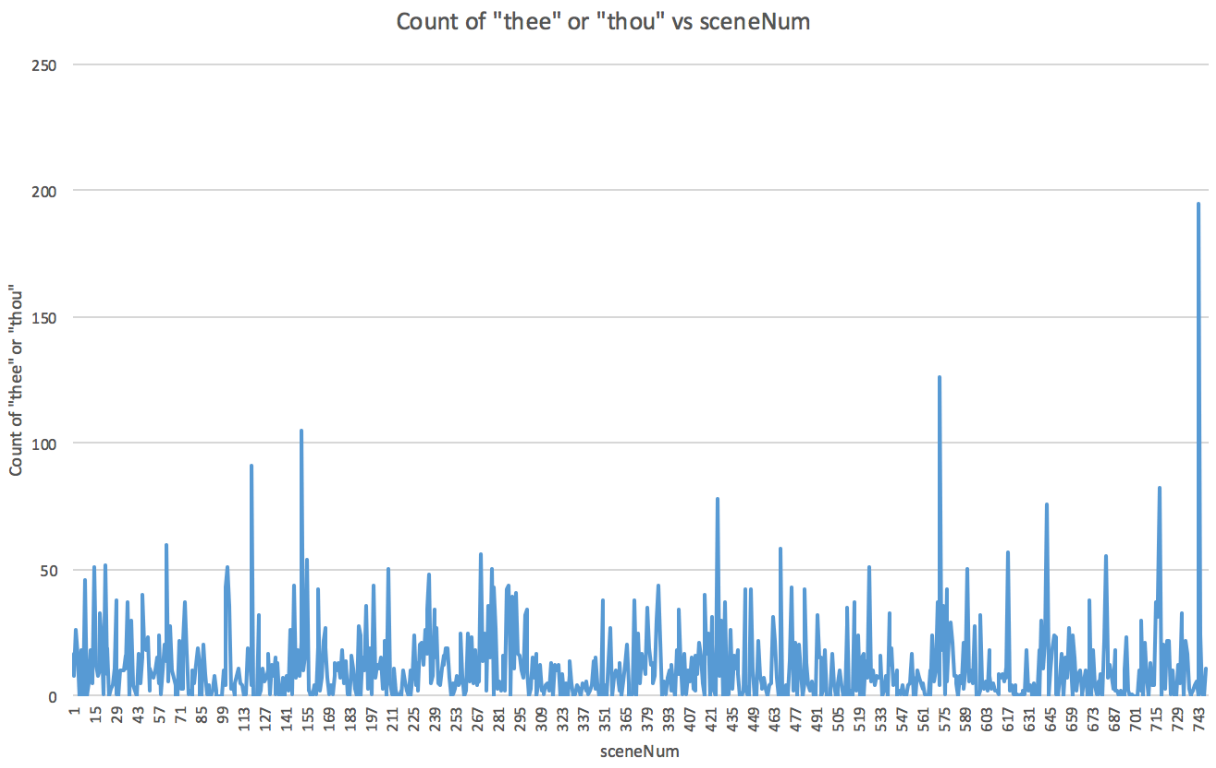
   Phrase1 “a rose by any other name” took the longest to execute. Because it is the longest phrase which contains more terms than others. So the time of visiting inverted list is the longest.

6. Scene count = 748
   Terms count = 897268
   Average length = 897268/748 = 1200

Shortest scene: antony_and_cleopatra:2.8	Count: 47
Long play: hamlet	Count: 32867
Shortest play: comedy_of_errors	Count: 16415

7.


Count of "you" vssceneNum
Count of "you"
sceneNum
0
20
40
60
80
100
120
140
160
180


Count of "thee" or "thou" vs sceneNum
Count of "thee" or "thou"
sceneNum
0
50
100
150
200
250