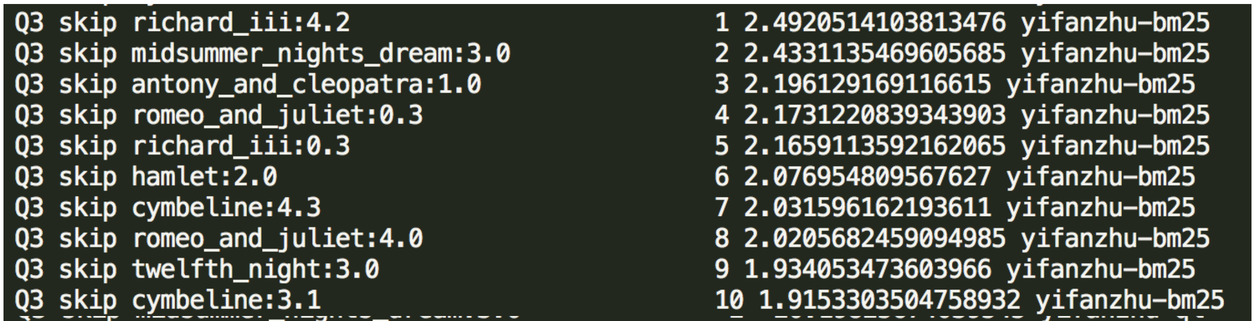
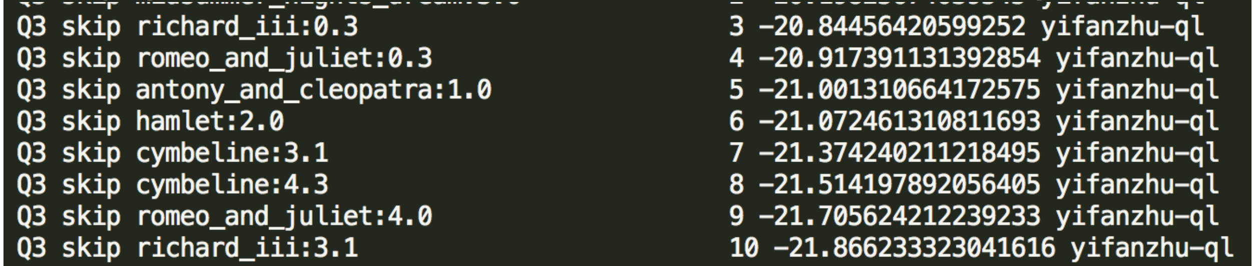
P4 Retrieval Report

1. The BM25 occurs from line 258 to line 314. The QL occurs from line 316 to line 363. The evaluation occurs from line 368 to line 389.
2. The system is basically the same as P3 Indexing. All the parameters in BM25 and QL can be retrieved from the map that I constructed in last project. All I need to do is create a for loop to calculate the rank score of each term in the query and all them them. One Question I had is the output is hard to read because the write function of java.io cannot align the text. I solve this by create a new string builder and add different spaces based on the length of scene.
3.
    1. java.io: read file and write file to txt
    2. java.util: create Maps, Set, the max value of Integer and calculate the log
    3. org.jason.simple: read file from JSON
4. The results for Q6 is bad. Because “to be or not to be” is a sentence. The meaning of it will totally change if we separate it. The ranking score we calculate is based on separate words of it. In the case, documents contain the sentence “to be or not to be” may have lower ranking score than some documents contains more “to”, “be”, “or” or “not”.
5. I cannot say the results for “setting the scene” is bad or good. Because it is relatively bad compare to Q1-Q5 but it is also relatively good compare to Q6. In this query, “the” is a very common word and “scene” also appears in every document exactly once. Which mean these two words provide limited information. Based on the feature of BM25 and QL, once a term appears too many times, it will has limited influence on the rank score. Overall, I think it is relatively bad.
6.



```
Q3 skip richard_iii:4.2                          1 2.4920514103813476 yifanzhu-bm25
Q3 skip midsummer_nights_dream:3.0               2 2.4331135469605685 yifanzhu-bm25
Q3 skip antony_and_cleopatra:1.0                 3 2.196129169116615 yifanzhu-bm25
Q3 skip romeo_and_juliet:0.3                     4 2.1731220839343903 yifanzhu-bm25
Q3 skip richard_iii:0.3                          5 2.1659113592162065 yifanzhu-bm25
Q3 skip hamlet:2.0                               6 2.076954809567627 yifanzhu-bm25
Q3 skip cymbeline:4.3                            7 2.031596162193611 yifanzhu-bm25
Q3 skip romeo_and_juliet:4.0                     8 2.0205682459094985 yifanzhu-bm25
Q3 skip twelfth_night:3.0                        9 1.934053473603966 yifanzhu-bm25
Q3 skip cymbeline:3.1                            10 1.9153303504758932 yifanzhu-bm25
```



```
Q3 skip richard_iii:0.3                          3 -20.84456420599252 yifanzhu-ql
Q3 skip romeo_and_juliet:0.3                     4 -20.917391131392854 yifanzhu-ql
Q3 skip antony_and_cleopatra:1.0                 5 -21.001310664172575 yifanzhu-ql
Q3 skip hamlet:2.0                               6 -21.072461310811693 yifanzhu-ql
Q3 skip cymbeline:3.1                            7 -21.374240211218495 yifanzhu-ql
Q3 skip cymbeline:4.3                            8 -21.514197892056405 yifanzhu-ql
Q3 skip romeo_and_juliet:4.0                     9 -21.705624212239233 yifanzhu-ql
Q3 skip richard_iii:3.1                          10 -21.866233323041616 yifanzhu-ql
```

Q3: hope dream sleep

Top 10 of BM25 Q3

Scene Id	Count(hope, dream, sleep)
richard_iii:4.2	1,12,11
midsummer_nights_dream:3.0	0,9,2
antony_and_cleopatra:1.0	2,1,1
romeo_and_juliet:0.3	0,12,4
richard_iii:0.3	2,4,5
hamlet:2.0	1,2,5
cymbeline:4.3	0,3,6
romeo_and_juliet:4.0	0,3,2
twelfth_night:3.0	0,2,1
cymbeline:3.1	3,4,4

Top 10 of QL Q3

Scene Id	Count(hope, dream, sleep)
richard_iii:4.2	1,12,11
midsummer_nights_dream:3.0	0,9,2
richard_iii:0.3	2,4,5
romeo_and_juliet:0.3	0,12,4
antony_and_cleopatra:1.0	2,1,1
hamlet:2.0	1,2,5
cymbeline:3.1	3,4,4
cymbeline:4.3	0,3,6
romeo_and_juliet:4.0	0,3,2
richard_iii:3.1	1,1,2

The system do pretty well on Q3. QL appears to be better. I calculate the counts of hope, dream and sleep. The result shows that the rank of QL is more related to the query. Further more, I read all the scene text and rated them from not relevant to relevant. The result also shows that the rank of QL is more correspond to the relevant rate.