Answer

- **U6826541 yixi rao**

Q1

map: 0.13176310670592079
Rprec: 0.15703823953823953
recip_rank: 0.2803481871014636
P_5: 0.15333333333333335
P_10: 0.12000000000000002
P_15: 0.08888888888888892

Q2

map: 0.18124582815728202
Rprec: 0.1723953823953824
recip_rank: 0.44355440362883936
P_5: 0.20666666666666667
P_10: 0.1366666666666667
P_15: 0.11111111111111113

Q3

Q3.1

What modifications you made?

Apply the Stemming to the token t

Why you made them?

Because it can make the token list denser, and reduce the size of it by classifying a lot of derivative words into the same stem, therefore during the querying process, the other forms of the token can also be matched which will increase the precision and recall.

What the new performance is?

map: 0.234310295864282
Rprec: 0.20953823953823955
recip_rank: 0.4135098648814554
P_5: 0.21333333333333332
P_10: 0.16000000000000003
P_15: 0.12666666666666668

Why you think the modification did/did not work?

The Stemming turns tokens into stems, which are the same regardless of inflection and we may wish different forms of a root to match. For example, if a query contains the word “synchronisation”, then all the documents contained with synchronise,

synchronous, synchronised, synchronisation, synchrony, synchronizing will be investigated.

Q3.2
What modifications you made?
Apply the Lemmatization to the token t

Why you made them?
The lemmatization is a dictionary-based approach, and it can turn words into lemmas. So, lots of words which has the same meaning can be categorized into a same word which must appear in the dictionary. As a result, the IR system will be more effective and can reduce the size of the total token list.

What the new performance is?
map: 0.14641806300651367
Rprec: 0.17013347763347764
recip_rank: 0.32535950552303317
P_5: 0.17333333333333334
P_10: 0.13
P_15: 0.0977777777777778

Why you think the modification did/did not work?
Lemmatization can map some words with the same meaning but different forms to one lemma, and it can also keep the semantics of the word because it must come from the dictionary so it is good for text content analysis and querying. For example, a lot of verbs that its past tense and past participle are not ended will ‘ed’ can be identified by the lemmatization process and increase the precision and recall of the query. Such as ‘drive’, ‘drove’, ‘driven’ can be easily identified.

Q3.3
What modifications you made?
Apply the punctuation elimination to the token t

Why you made them?
It is possible that different documents have different writing styles, especially the use of the punctuation. Besides, the tokenization process may be not deleting all the punctuation which is useless in the information retrieving process. If we can find all the punctuation involved tokens, then we can reduce the size of token list and let it be effective.

What the new performance is?
map: 0.1546737468012121
Rprec: 0.15822871572871572
recip_rank: 0.31840601305485

P_5: 0.17333333333333334
P_10: 0.1366666666666667
P_15: 0.10444444444444446

Why you think the modification did/did not work?
During the tokenization process, some words which contains the comma or exclamation point or question mark or semicolon should be reconstruct by eliminating the punctuation, then the result word can map with the word in the query. For example, the uneliminated word token "subprime," is totally different with "subprime", after the modification, the comma should be clean and these two words can map.

Q3.4

Modification	MAP
Stemming	0.234310295864282
Lemmatization	0.14641806300651367
punctuation elimination	0.1546737468012121

The result is that "stemming" > "punctuation elimination" > "Lemmatization", the reason of choosing the MAP as the metric is that it can reflect all the recall levels. The best performance is the stemming modification as it is more proper for information retrieval than lemmatization as lemmatization emphasizes on the actual meaning of the word using it in the more granular and accurate text analysis and expression. On the other hands, the reason why the lemmatization does not perform well is that I did not specify the port of speech of the words when doing the lemmatization, it is important to specify the part of speech of the word, otherwise the reduction may not work well.

Q4

Welcoming
./gov/documents\31\G00-31-2565694
./gov/documents\42\G00-42-4180551
./gov/documents\85\G00-85-0255215
./gov/documents\86\G00-86-2161870
./gov/documents\86\G00-86-4087434
./gov/documents\97\G00-97-2878104
./gov/documents\98\G00-98-1962568
unwelcome OR sam
./gov/documents\25\G00-25-0827853
./gov/documents\52\G00-52-1913171
./gov/documents\83\G00-83-1047969
./gov/documents\92\G00-92-0698648
./gov/documents\99\G00-99-0915850
ducks AND water

./gov/documents\21\G00-21-2194623
./gov/documents\55\G00-55-3182412
./gov/documents\84\G00-84-2749066
plan AND water AND wage
./gov/documents\49\G00-49-1454872
./gov/documents\63\G00-63-2128846
./gov/documents\94\G00-94-2331424
./gov/documents\97\G00-97-3308777
plan OR record AND water AND wage
./gov/documents\49\G00-49-1454872
./gov/documents\61\G00-61-3739677
./gov/documents\63\G00-63-2128846
./gov/documents\94\G00-94-2331424
./gov/documents\97\G00-97-3308777
space AND engine OR football AND placement
./gov/documents\23\G00-23-0835873
./gov/documents\24\G00-24-3404638
./gov/documents\25\G00-25-1519978
./gov/documents\45\G00-45-0875203
./gov/documents\67\G00-67-4173730
./gov/documents\77\G00-77-3396578
./gov/documents\81\G00-81-0581460
./gov/documents\92\G00-92-0775281

Q5

(a)

I will evaluate it using the precision, recall, and F-Measure to evaluate the Boolean query system. In order to evaluate it, I need to create the contingency table which requires me to collect the data of a collection of documents, queries, and a set of relevance judgment. It is hard to get all the relevance judgement of all query-document pairs because we judge it by using the information need and it is hard and impossible. After the contingency table is built, we can then calculate the precision and recall, it is appropriate because we can use the precision to test whether our system can have the efficiency and accuracy of getting the relevant documents and do not retrieve the irrelevant documents. The recall is to test whether the system has the ability to retrieve all the relevant documents. Although there is no necessary relationship between precision and recall, in large-scale data sets, these two indicators are mutually restricted. So, in the case of the precision and recall are contradictory, we should use the comprehensive evaluation F-Measure, which is the weighted harmonic mean of precision(P) and recall(R). The higher the value, the better the effectiveness is.

(b)

To evaluate the ranked retrieval sets, I can use the Precision-Recall Curve metric and some single value metric such as MAP, Mean Reciprocal Rank, Precision at top 10, Recall at top 10. In order to conduct the test, I need to collect the data of a collection of documents, queries, and a set of relevance judgment. And we need to collect the data of the top 10 query results returned of each query. It is hard to collect the relevance judgment because it is hard to judge it by using the abstract information need, as a solution we can use other trustworthy ranked IR systems to help us tag the relevance of the query-document pairs. After we collect the data, we can plot the Precision-Recall Curve and can compare it with other ranked IR systems using the same data set, it is appropriate because we can check the tendency of the curve, if our system's PR-curve is more gently and less steep than others IR systems, then we can conclude that our system have better performance because the precision is high. For the single number metric Mean Average Precision (MAP), it is appropriate because it consider all the recall levels and the area size can reflect the precision of the top 10 result, for the Mean Reciprocal Rank (MRR), it is appropriate because it can reflect the first relevant document appeared in the results, it is possible that the system has better performance when the first document appears early.