Your Name: Hao Wang

Your Andrew ID: haow2

Homework 2

Collaboration and Originality

Your report must include answers to the following questions:

1. Did you receive help of any kind from anyone in developing your software for this assignment (Yes or No)? It is not necessary to describe discussions with the instructor or TAs.
   No
   If you answered Yes, provide the name(s) of anyone who provided help, and describe the type of help that you received.

2. Did you give help of any kind to anyone in developing their software for this assignment (Yes or No)?
   No
   If you answered Yes, provide the name(s) of anyone that you helped, and describe the type of help that you provided.

3. Are you the author of every line of source code submitted for this assignment (Yes or No)? It is not necessary to mention software provided by the instructor.
   Yes
   If you answered No:
   a. identify the software that you did not write,
   b. explain where it came from, and
   c. explain why you used it.

4. Are you the author of every word of your report (Yes or No)?
   Yes
   If you answered No:
   a. identify the text that you did not write,
   b. explain where it came from, and
   c. explain why you used it.

Your Name:

Your Andrew ID: haow2

Homework 2

1 Experiment 1: Baselines

	Ranked Boolean	BM25 BOW	Indri BOW
P@10	0.1700	0.4200	0.4000
P@20	0.2800	0.3500	0.4700
P@30	0.3367	0.3667	0.4233
MAP	0.1071	0.1985	0.2057

2 Experiment 2: Queries with Synonyms and Phrases

2.1 Queries

69:#NEAR/2(sewing instructions)

79:#SYN(voyager.keywords)

84:#NEAR/2(continental plates)

89:#SYN(ocd.keywords)

108:#NEAR/2(#SYN(ralph owen) brewster)

141:#NEAR/3(va #NEAR/2(dmv registration))

146:#NEAR/3(sherwood #NEAR/2(regional library))

153:#SYN(pocono.keywords)

171:#NEAR/2(ron howard)

197:#NEAR/5(#NEAR/3(idaho state) flower)

2.2 Query descriptions

For each query, provide a brief (1-2 sentences) description that identifies which strategy was used for that query, any important deviations from your default strategies, and your intent, i.e., why you thought that particular structure was a good choice.

69:#NEAR/2(sewing instructions)

I think sewing and instructions are very likely to be part of a phrase, so I use NEAR/2 for structured queries to get a better result.

79:#SYN(voyager.keywords)

I think voyager should appear in keywords, which is more possible to make the result more relevant than any other fields.

84:#NEAR/2(continental plates)

I think continental and plates are very likely to be part of a phrase, so I use NEAR/2 for structured queries to get a better result.

89:#SYN(ocd.keywords)

I think ocd should appear in keywords, which is more possible to make the result more relevant than any other fields.

108:#NEAR/2(#SYN(ralph owen) brewster)

I think the query should aim at finding two brewsters whose name are ralph and owen. So I used SYN to two names and near/2 for brewster to make it a good query.

141: #NEAR/3(va #NEAR/2(dmv registration))

I think the query should aim at finding dmv registration places or websites or phone numbers in Virginia. So dmv and registration should be a phrase so I use NEAR/2 to find them, and va should be referred nearby in the document, So I used NEAR/3.

146: #NEAR/3(sherwood #NEAR/2(regional library))

I think the query should aim at finding a regional library in sherwood. So regional library should be a phraseso I use NEAR/2 to them, and sherwood should be referred nearby in the document, so I used NEAR/3.

153:#SYN(pocono.keywords)

I think pocono should appear in keywords, which is more possible to make the result more relevant than any other fields.

171:#NEAR/2(ron howard)

I think ron and howard are very likely to be part of a person's name, so I use NEAR/2 for structured queries to get a better result.

197: #NEAR/5(#NEAR/3(idaho state) flower)

I think idaho means the state in USA, and the query aims at finding some information about flower in that state, so I used NEAR/3 idaho and state, because flower is not directly related to the phrase, but should appear nearby, so I used NEAR/5 to them.

2.3 Experimental Results

	Ranked Boolean	BM25 BOW	Indri BOW	Ranked Boolean Syn/Phr	BM25 Syn/Phr	Indri Syn/Phr
P@10	0.1700	0.4200	0.4000	0.4111	0.4556	0.4667
P@20	0.2800	0.3500	0.4700	0.5056	0.4556	0.5000
P@30	0.3367	0.3667	0.4233	0.5185	0.4185	0.4741
MAP	0.1071	0.1985	0.2057	0.2329	0.2250	0.2449

2.4 Discussion

Discuss any trends that you observe; whether the use of synonyms and phrases behaved as you expected; and any other observations that you may have.

According to my observation, it is explicit that the improvement is very significant. In P@10, P@20, P@30 and MAP, the result of RankedBoolean, BM25, and Indri all get significant improvement. But according to the table, here is no enough information to know the improvement in recall, for only precision is considered in the table. The increase of P@10, P@20, P@30 and MAP can fully convinced that the precision is improved.

The reason of the improvement is in two expects. The first one is the difference of NEAR/n and AND to retrieval phrases. In unstructured queries, there are no NEAR/n, but AND. So it is impossible for it to take the distance of words into consideration, which is not a good way to retrieval phrases. However, NEAR/n cannot perform very well when doing retrieval of words that are not phrases. So in order to have a higher precision, I used NEAR/n for phrases, and SYN for words that are not phrases. In structured queries, I used NEAR/n to retrieval possible phrases and SYN to combine them together, so the result improved a lot.

The other reason is that I modified the fields of some queries to get a better result. The default field is body, which is very general, especially to SYN and AND. Because they didn't take the position of words in documents into consideration and there are too many words in body filed, so the result can be much worse.

The comparison of three retrieval models shows that the overall precision of BM25 and Indri are similar and they are both better than RankedBoolean. Although RankedBoolean have a higher precision in structured queries, but in most structured queries and unstructured queries it is not so good as BM25 and Indri. And in unstructured queries, difference of the precision of RankedBoolean and BM25, Indri are too huge. The reason is explicit, for RankedBoolean model only considers tf to calculate scores, while Indri

and BM25 considers many more variables, including idf, df, document length, average document length etc., and change the weight of different of different parts of value to get a better result.

3 Experiment 3: BM25 Parameter Adjustment

3.1 k_1

	k_1							
	1.2	0.0	0.5	1.0	4.0	8.0	16.0	100.0
P@10	0.4200	0.1100	0.4200	0.4200	0.4000	0.3600	0.3400	0.2700
P@20	0.3500	0.1350	0.3400	0.3450	0.3550	0.3250	0.3050	0.2400
P@30	0.3667	0.1533	0.3733	0.3700	0.3600	0.3500	0.3267	0.2933
MAP	0.1985	0.0665	0.2056	0.1996	0.1686	0.1575	0.1468	0.1085

.

3.2 b

	b							
	0.75	0.0	0.2	0.4	0.6	0.8	0.9	1.0
P@10	0.4200	0.3900	0.4100	0.4500	0.4000	0.4100	0.3900	0.3900
P@20	0.3500	0.4450	0.4450	0.4150	0.3650	0.3550	0.3650	0.3600
P@30	0.3667	0.4600	0.4233	0.4033	0.3867	0.3733	0.3400	0.3500
MAP	0.1985	0.1977	0.2157	0.2144	0.2082	0.1966	0.1826	0.1394

3.3 Discussion

Explain your reasons for choosing the values that you tested, and how those reasons are related to how BM25 works. Discuss any changes in retrieval performance that you observed, and the significance of any trends that you observed.

k1:

K1 has to be larger or equal to 0, so I let the value of k1 double each time and also set a very large value, 100.0 to test its effect of a larger value. I can observe the effect of k1 by looking at P@10, P@20, P@30, and MAP. When k1 equals to 0, the value of P@10, P@20, P@30 and MAP are the global lowest value. Then when k1 increases, the value of P@10, P@20, P@30 and MAP increase. When k1 increases, the precision first becomes better then becomes worse. According to my observation, the best value for k1 should in the range of 0.5 and 4.0, because other values of k1 performs much worse than theirs, in both P@10, P@20, P@30 and MAP. If I narrow the range, it should be from 0.5 to 1.0.

b:

b has to be larger or equal to 0.0 and less or equal to 1.0, so I set the value of b quite evenly and tested the situation when b equals to 0.0 and 1.0. I can observe the effect of b by looking at the P@10, P@20, P@30 and MAP. When b equals to 1, the precision of BM25 is worst. Then when b increases from 0 to 1. The precision just has quite a little change. But when b is bigger than 0.9, the precision decreases much faster.

According to my observation, the best value of b should be in 0.2 to 0.6. There is no adequate evidence for me to narrow the range. Actually, when b ranges from 0 to 0.8, it doesn't make much difference to precision.

4 Indri Parameter Adjustment

4.1 μ

	μ							
	2500	0	1000	2000	4000	7000	10000	25000
P@10	0.4000	0.4300	0.4200	0.3600	0.3400	0.3600	0.3700	0.3200
P@20	0.4700	0.4100	0.4400	0.4400	0.4550	0.4800	0.4750	0.4550
P@30	0.4233	0.3933	0.4267	0.4200	0.4733	0.5000	0.5033	0.4800
MAP	0.2057	0.1952	0.2143	0.2069	0.1987	0.1882	0.1805	0.1609

4.2 λ

	λ							
	0.4	0.0	0.1	0.3	0.5	0.7	0.9	1.0
P@10	0.4000	0.4000	0.3900	0.3900	0.4000	0.3800	0.3500	0.0000
P@20	0.4700	0.4750	0.4750	0.4700	0.4700	0.4400	0.3950	0.0150
P@30	0.4233	0.4233	0.4233	0.4233	0.4167	0.4000	0.3867	0.0200
MAP	0.2057	0.2142	0.2132	0.2085	0.2012	0.1902	0.1641	0.0020

4.3 Discussion

Explain your reasons for choosing the values that you tested, and how those reasons are related to how Indri works. Discuss any changes in retrieval performance that you observed, and the significance of any trends that you observed.

μ:

The value of μ should be larger or equal to 0. So my set the begin value as 0 then gradually increase the value. I also tried two very large value to test its effect on extreme situations. Other values are nearly the double of its former value. I can observe the effect of b in terms of precision by looking at the P@10, P@20, P@30 and MAP. When μ equals to 0, the precision is not so bad, for MAP is still similar to 0.2. But when μ is very large, after 7000, the precision decreases quite fast, but its precision is still acceptable. I think the precision of μ doesn't change much when its value is larger than 0 and smaller than 4000. I think the best value of μ should be in the range of 1000 to 4000, but there is no full evidence for me to narrow the range.

λ:

The value of λ should be larger or equal to 0 and less or equal to 1. So I tested the situation when λ is 0 and 1. Then I tested 0.1 and 0.9 and see how does the precision change when the value of λ approaches extreme value. When the value of λ is 0 and 0.1, it doesn't make much difference in terms of precision. But λ changes a lot in precision, when the value of λ ranges from 0.9 to 1.0.

Surprisingly, the precision decreases by 98% in such a small range. The precision doesn't change much when λ ranges from 0 to 0.5. So I think the best value of λ should be in this range, but there is no full evidence for me to narrow the range.