Your Name: Hailun Zhu

Your Andrew ID: hailunz

Homework 2

1 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)?
   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

2 Experiment 1: Baselines

	Ranked Boolean	BM25 BOW	Indri BOW
P@10	0.1500	0.3000	0.2300
P@20	0.1800	0.2950	0.2800
P@30	0.1667	0.2967	0.2900
MAP	0.0566	0.1304	0.1277
Time	00:15.5	00:16.78	00:16.89

3 Experiment 2: Queries with Synonyms and Phrases

3.1 Queries

List your queries.

10:#NEAR/3(#SYN(cheap economical inexpensive affordable) internet) internet.inlink
12:#SYN(djs dj #NEAR/1(disk jockey)) djs.keywords djs.url
26:#near/4(lower #near/1(heart #syn(rate pulse)))
29:#NEAR/1(#SYN(ps #NEAR/1(play station)) 2) games #near/1(ps.title 2.title)
33:elliptical #syn(trainer machine) #near/15(elliptical.title trainer.title)
52: avp.url #SYN(avp #NEAR/8(Association Volleyball Professional))
71: #near/1(living in) india
102:#NEAR/10(fickle creek) #syn(farm farmhouse field)
149: uplift at #NEAR/10(yellowstone #NEAR/2(national park))
190:#NEAR/1(brooks brothers) #SYN(clearance sale bargain discount) brooks.title brothers.title clearance.title

3.2 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.

10:#NEAR/3(#SYN(cheap economical inexpensive affordable) internet) internet.inlink
Use SYN NEAR/N and field. As cheap is an adjective, I could come up with several synonyms, such as inexpensive and affordable, and this word could compose of a phrase with internet, thus use NEAR.

12:#SYN(djs dj #NEAR/1(disk jockey)) djs.keywords djs.url
Use SYN, NEAR/N and field. As dj could be abbreviate of disk jockey, so use SYN to include this phrase and dj. Also use field.

26:#near/4(lower #near/1(heart #syn(rate pulse)))
Use NEAR/N as heart rate could be a phrase, thus use NEAR/1. And heart rate and heart pulse are synonyms.

29:#NEAR/1(#SYN(ps #NEAR/1(play station)) 2) games #near/1(ps.title 2.title)
Use SYN, NEAR/N and field. As ps could be abbreviate of play station, thus use NEAR/1 and SYN to include this phrase. And as ps 2 could be a phrase, thus use NEAR/1 to form this phrase.

33:elliptical #syn(trainer machine) #near/15(elliptical.title trainer.title)
Use SYN, NEAR/N and field. Use machine to be synonyms to trainer. And use title field.

52: avp.url #SYN(avp #NEAR/8(Association Volleyball Professional))
Use SYN, NEAR/N and field. As avp could be abbreviate of Association Volleyball Professional, so use NEAR/N and SYN to include it.

71: #near/1(living in) india
The 'living in' seems to be a phrase so that I use NEAR/1.

102:#NEAR/10(fickle creek) #syn(farm farmhouse field)
Use NEAR/n and SYN operator. These three words compose of a specified phrase so use NEAR/10. And use SYN to include synonyms of 'farm'.

149: uplift at #NEAR/10(yellowstone #NEAR/2(national park))
The Yellowstone national park seems to be a specified name. Meantime the national park could compose of a phrase, so compare the Yellowstone and national park separately. As those are the fixed phrase, I don't use other synonyms.

190:#NEAR/1(brooks brothers) #SYN(clearance sale bargain discount) brooks.title brothers.title clearance.title
The 'brooks brothers' is a brand, it is likely to be considered together, so use NEAR/1. And there are some synonyms could be used for clearance, including sale, discount.

3.3 Experimental Results

	Ranked Boolean	BM25 BOW	Indri BOW	Ranked Boolean Syn/Phr	BM25 Syn/Phr	Indri Syn/Phr
P@10	0.1500	0.3000	0.2300	0.2100	0.3500	0.3800
P@20	0.1800	0.2950	0.2800	0.2350	0.3600	0.4200
P@30	0.1667	0.2967	0.2900	0.2167	0.3567	0.4100
MAP	0.0566	0.1304	0.1277	0.0780	0.1702	0.1993
Time	00:15.51	00:16.78	00:16.89	00:08.54	00:08.16	00:09.74

3.4 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.

Sometimes if we use improper synonyms, the performance could be worse. For example, for No.10 query, I have tested [10:#syn(cheap #near/1(low priced)) internet.inlink]. However the 'low priced' turns out to be an improper synonyms, thus the performance decreased. Cases are the same with No.26 query, when I use #syn(lower decrease).

But if we choose proper synonyms, there could be a impressive improvement on the results. That is the case with No.10 query, (cheap, inexpensive, affordable), No.33 query, (trainer machine), No.102 query, (farm, farmhouse, field), and No.190 query, (clearance, bargain, sale, discount). Those terms are easy to find synonyms, whereas some are difficult.

A special case for the use of synonyms is to include the full complete phrase for those abbreviates. Examples are No.12, No.29 and No.52. By giving more detailed information, the result could be improved.

As for those terms that are very likely to be phrases, using Near operator to form phrases could be a good idea. For instance, ‘national park’, ‘heart rate’, ‘brooks brothers’, ‘disk jockey’ and ‘living in’ are all fixed phrases, thus to use NEAR/1 is a good choice. However for others, like adjective and nouns, we could not be sure about the distance between them, in this situation, it would be better to use larger distance.

In conclusion, when using synonyms and phrases, we should be very careful. Using right terms and phrases could help, but it is also likely to impair the results.

4 Experiment 3: BM25 Parameter Adjustment

4.1 k_1

	k_1							
	1.2	0.9	0.1	0.0	1.5	3.0	10.0	300.0
P@10	0.3000	0.3000	0.2800	0.0900	0.2900	0.2900	0.2500	0.2000
P@20	0.2950	0.2900	0.3000	0.0500	0.2950	0.2900	0.2550	0.1950
P@30	0.2967	0.3000	0.3100	0.0600	0.2933	0.2933	0.2500	0.1900
MAP	0.1304	0.1297	0.1265	0.0176	0.1298	0.1290	0.1186	0.0947
Time	00:16.00	00:16.18	00:15.85	00:15.99	00:15.66	00:15.80	00:15.41	00:16.30

.

4.2 b

	b							
	0.75	0.40	0.00	0.10	0.70	0.80	0.90	1.00
P@10	0.3000	0.2400	0.2400	0.2600	0.2700	0.3000	0.2600	0.2200
P@20	0.2950	0.3050	0.2800	0.2900	0.2950	0.3000	0.3200	0.2950
P@30	0.2967	0.3133	0.2733	0.2967	0.2967	0.3033	0.3133	0.3067
MAP	0.1304	0.1313	0.1076	0.1243	0.1285	0.1301	0.1293	0.1205
Time	00:16.00	00:15.47	00:15.78	00:15.88	00:15.67	00:15.39	00:15.83	00:15.74

4.3 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.

As for my understanding, the bigger k1 is, the more influence that the document length has. And when the document length(doclen) has fewer influence, the term frequency(tf) will have more influence. The bigger b is, the stronger penalty is for long documents. Thus document with bigger length will have fewer scores.

4.3.1 For K1.

I choose default value 1.2 as a threshold, and then I choose 3 values that are less than it, and 4 values that are larger than it.

For the values that are less than 1.2, I first choose 0.9. As stated in the lecture slide, for large collections, 0.9 is often chosen as k1. For large collections, there may be lots of relevant documents, thus we should slightly reduce the influence of the document length. The result shows that for the corpus we use, the performance of 0.9 is slightly worse than 1.2.

I secondly choose 0, in this corner case, the doclen and tf will have no effect. The performance dramatically drops compared to the default performance. This indicates that the tf is a very important factor to represent relevance. And this makes sense.

Thirdly I choose 0.1, which is slightly bigger than 0. The performance is much better than the previous one. This also indicates that tf plays a significant role in this model.

For values that are larger than 1.2, I first choose 1.5, the difference between this value and 1.2 is 0.3, which is the same as the difference between 0.9 and 1.2. The result is very similar to 0.9's performance. In fact, I have also tested 1.0, 1.1, 1.3, 1.4, those performance are very similar to 1.2 and are better than 0.9's and 1.5's.

I have also tested 3.0, 10, 300. The results shows a descending trend, which implies that assigning more importance on document length than on tf does no good. That is to say, it could yield better results when considering doclen and tf, but tf is more important.

4.3.2 For b.

By comparing results of 0 and 1, I could conclude that doclen truly has effect on document retrieval.

By comparing results of 0 and 0.1, 0.9 and 1, using b to adjust doclen bias could improve the performance.

By comparing results of 0.7, 0.75 and 0.8, 0.9, I could say that putting more penalties on longer document will not always be better. There exist some good values in between.

As 0.4 is for large collection usually, so I test it as well. This parameter yields the best MAP value among all the parameters that I have test. I have also tested 0.3, 0.5 and 0.6, there is no obvious trend by increasing b.

To sum up, using b to do document length normalization could improve the performance.

5 Indri Parameter Adjustment

5.1 μ

	μ							
	2500	0	1000	1500	2000	500	100	5000
P@10	0.2300	0.2600	0.2700	0.2300	0.2200	0.3200	0.2800	0.2200
P@20	0.2800	0.3000	0.3300	0.3250	0.3050	0.3100	0.3200	0.2700
P@30	0.2900	0.3100	0.3167	0.3133	0.2900	0.3167	0.3133	0.2933
MAP	0.1277	0.1254	0.1316	0.1315	0.1304	0.1346	0.1316	0.1210
Time	00:16.47	00:15.71	00:15.37	00:15.28	00:15.15	00:15.14	00:14.97	00:15.66

5.2 λ

	λ							
	0.4	0.0	0.1	0.2	0.3	0.6	0.7	0.9
P@10	0.2300	0.2700	0.2700	0.2500	0.2400	0.2000	0.1900	0.1500
P@20	0.2800	0.3000	0.3000	0.2950	0.2900	0.2750	0.2650	0.2150
P@30	0.2900	0.3133	0.3067	0.3000	0.2933	0.2700	0.2633	0.2500
MAP	0.1277	0.1346	0.1334	0.1318	0.1295	0.1241	0.1205	0.1093
Time	00:16.47	00:15.80	00:15.36	00:16.11	00:15.71	00:15.88	00:16.02	00:14.48

5.3 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.

5.3.1 For μ

As stated in the lecture slide, μ is the parameter used in Bayesian smoothing with Dirichlet Priors. When μ is 0, then there is no Bayesian smoothing.

So I first tested 0, to see what it looks like to only have Mixture Model Smoothing. Actually the result surprises me. It seems that not having this μ is better than $\mu=2500$ for P@n, and the MAP value is only a little bit worse.

Then I tested values in the range 1000~2000, because the lecture says usually best μ is within this range. However the results are not so satisfying, though the results are better than 2500's. And I noticed that the performance is in a descending trend when increasing μ. So that I next tested the $\mu=500$. The result of 500 yields a significant improvement. Then I have tested values around 500, like 100, 200, 600, 800, and I find out that 500 may be a peak in this range and is better for our queries and documents.

To ensure that there is a descending trend when increasing μ beyond 500, I tested 5000. The result shows that it is a descending trend.

The results indicate that larger μ sometimes could have negative effect. And as stated in the slide, for long documents larger μ is less important, and as our queries are short, bigger idf-like effect is not adorable. In this case, a smaller μ is more important.

5.3.2 For λ

I first tested the corner cast λ=0, the result also surprises me because it is much better than λ=0.4. Then I tested the following values in an ascending manner. And the results show the performance is in a descending trend. This indicates that in our experiment, not having λ will yield better results. When λ=0, there is only Bayesian smoothing with Dirichlet Priors. The reason why adding Jelinek-Mercer smoothing will do more harm maybe because our queries are too short. Thus every query term must match, so that when using Jelinek-Mercer smoothing idf weighting becomes important and makes some terms less important.

To sum up, larger λ will be better when there are long queries. For short queries, smaller λ is preferred.