Your Name: Wenhan Lu

Your Andrew ID: wenhanl

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: Wenhan Lu

Your Andrew ID: wenhanl

Homework 2

1 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

2 Experiment 2: Queries with Synonyms and Phrases

2.1 Queries

10:#NEAR/1(#SYN(cheap inexpensive affordable) internet)
12:#SYN(djs dj #NEAR/1(disk jockey))
26:lower #NEAR/5(heart rate)
29:#NEAR/1(#SYN(ps #NEAR/1(play station)) 2) games
33:#SYN(#NEAR/3(elliptical trainer) #NEAR/3(elliptical machine))
52:#SYN(avp #NEAR/3(Association Volleyball Professional))
71:#SYN(living reside) in india
102:#NEAR/5(fickle creek farm)
149:uplift at #SYN(#NEAR/5(yellowstone national park) yellowstone)
190:#NEAR/1(brooks brothers) #SYN(clearance sale)

2.2 Query descriptions

10:#NEAR/1(#SYN(cheap inexpensive affordable) internet)
Cheap Internet should be a phase in information need. “cheap” can be replaced by synonyms.

12:#SYN(djs dj #NEAR/1(disk jockey))
Abbreviation expended and SYN with other close synonyms.

26:lower #NEAR/5(heart rate)
Not to much hack. “heart rate” is likely a phase so use NEAR on them.

29:#NEAR/1(#SYN(ps #NEAR/1(play station)) 2) games
Abbreviation expended as a synonym of “ps”.

33:#SYN(#NEAR/3(elliptical trainer) #NEAR/3(elliptical machine))
"elliptical trainer" and "elliptical machine" are close synonyms. And they need to put close as a phase.

52:#SYN(avp #NEAR/3(Association Volleyball Professional))
Abbreviation expended as a synonym of "avp".

71:#SYN(living reside) in india
Not too much hack. Close synonym added to "living".

102:#NEAR/5(fickle creek farm)
Can't do more on this. Confused about information need, so treat them as phase.

149:uplift at #SYN(#NEAR/5(yellowstone national park) yellowstone)
"yellowstone national park" is a phase, but sometimes only yellowstone is strong enough to identify. So put "yellowstone" as a synonym.

190:#NEAR/1(brooks brothers) #SYN(clearance sale)

"brooks brothers" is a brand name that should be place close. "sale" is a very close synonym to "clearance"

2.3 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	0.2700	0.3000	0.2600
P@20	0.1800	0.2950	0.2800	0.2950	0.3350	0.3400
P@30	0.1667	0.2967	0.2900	0.2800	0.3567	0.3667
MAP	0.0566	0.1304	0.1277	0.1227	0.1572	0.1625

2.4 Discussion

SYN (synonym)	Include fair number of synonyms will increase precision. SYN should only be used in case of meaning is very close and very replaceable. Otherwise it will be over-inference of information need and have negative effect. In case of query is like a abbreviation, expend it to full name and add it as a SYN is a good choice to increase precision.
NEAR (phrase)	For strong phrase, like brand name, we should use strong phrase operator, like #NEAR/1(brooks brothers). In case of phrase is not necessary get together, we should have loose restriction on distance on words, like #NEAR/5(fickle creek farm).

3 Experiment 3: BM25 Parameter Adjustment

3.1 k_1

	k_1							
	1.2	0.0	0.5	0.9	1.5	2.0	4.0	10.0
P@10	0.3000	0.0400	0.2800	0.3000	0.2900	0.2900	0.2800	0.2500
P@20	0.2950	0.0200	0.3200	0.2900	0.2950	0.2950	0.2850	0.2550
P@30	0.2967	0.0433	0.3067	0.3000	0.2933	0.3033	0.2867	0.2500
MAP	0.1304	0.0142	0.1281	0.1297	0.1298	0.1292	0.1279	0.1186

.

3.2 b

	b							
	0.75	0.0	0.2	0.4	0.5	0.6	0.9	1.0
P@10	0.3000	0.2400	0.2800	0.2400	0.2500	0.2600	0.2600	0.2200
P@20	0.2950	0.2800	0.3050	0.3050	0.2950	0.2900	0.3200	0.2950
P@30	0.2967	0.2733	0.2900	0.3133	0.3033	0.3167	0.3133	0.3067
MAP	0.1304	0.1076	0.1304	0.1313	0.1295	0.1302	0.1293	0.1205

3.3 Discussion

Variable selection:

Boundary variables are chosen extremely small and large. Variables in the middle are chosen evenly.

About k_1:

K_1 is parameter to control influence of term frequency. Suppose we have a constant parameter b, the higher the K_1, the higher the impact of a high term frequency is in final calculation. Documents with more occurrence of term are more likely to have high rank. In my result, set K_1 to 0 will give a very bad result, which means ignoring impact of tf is a bad idea. However, set K_1 too high is not best as well. The peak appears on about 1.2. In practice, we need to consider balance between influence of tf and idf.

About b:

b is parameter to control influence of document length. The larger the b is set, the higher impact of document length on result. Intuitively, shorter documents with same term frequency should have higher rank. So this parameter is set to find a balance to include impact from document length. (Set b to 0 is not too bad as observed, which means document length is as important as term frequency).

4 Indri Parameter Adjustment

4.1 μ

	μ							
	2500	0	100	500	1000	1500	5000	10000
P@10	0.2300	0.2600	0.2800	0.3200	0.2700	0.2300	0.2200	0.1900
P@20	0.2800	0.3000	0.3200	0.3100	0.3300	0.3250	0.2700	0.2950
P@30	0.2900	0.3100	0.3133	0.3167	0.3167	0.3133	0.2933	0.3000
MAP	0.1277	0.1254	0.1316	0.1346	0.1316	0.1315	0.1210	0.1163

4.2 λ

	λ							
	0.4	0.0	0.2	0.3	0.5	0.7	0.9	1.0
P@10	0.2300	0.2700	0.2500	0.2400	0.2300	0.1900	0.1500	0.0100
P@20	0.2800	0.3000	0.2950	0.2900	0.2850	0.2650	0.2150	0.0050
P@30	0.2900	0.3133	0.3000	0.2933	0.2833	0.2633	0.2500	0.0033
MAP	0.1277	0.1346	0.1318	0.1295	0.1267	0.1205	0.1093	0.0010

4.3 Discussion

Variable selection:

Mu variable selection is more like a power function. It's used to test mu change from small to large exponentially. For lambda, it's almost linear. It must be between 0 and 1, so we must select 0 and 1 as edge case, and choose the rest linearly.

About μ:

μ is used to smooth probability based on tf using some extent of idf. For short documents, probabilities are more granular, so large μ is important. For long document, probabilities are more smooth, so large μ is less important. We don't length distribution of our documents, but according to experiment result, I do get a peak MAP on about μ=500, which should be a balanced value for this document set.

About λ:

λ is used to balance "idf effect". Higher λ will give higher "idf effect" which means documents with rare words tend to have higher rank. But for short queries, usually every term must match. So rare words are less important. In my result, MAP decreases as λ increases, that's mainly because all our queries are short queries.