Your Name Mengyu Yang

Your Andrew ID mengyuy

Homework 3

Collaboration and Originality

1. Did you receive help <u>of any kind</u> 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

2. Did you give help <u>of any kind</u> to anyone in developing their software for this assignment (Yes or No)?

   No

3. Are you the author of <u>every line</u> of source code submitted for this assignment (Yes or No)? It is not necessary to mention software provided by the instructor.

   Yes

4. Are you the author of <u>every word</u> of your report (Yes or No)?

   Yes

Your Name Mengyu Yang

Your Andrew ID mengyuy

Homework 3

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

BM25: k_1 = 1.2, b = 0.75, k_3 = 0.

Indri: mu = 2500, lambda = 0.4.

2. Experiment 2: Different representations

	Indri BOW (body)	0.2 url 0.2 keywords 0.2 title 0.2 body 0.2 inlink	0.05 url 0.05 keywords 0.3 title 0.4 body 0.2 inlink	0.01 url 0.01 keywords 0.37 title 0.6 body 0.01 inlink	0.3 url 0.3 keywords 0.05 title 0.05 body 0.3 inlink	0.25 url 0.25 keywords 0.24 title 0.01 body 0.25 inlink
P@10	0.2300	0.0900	0.1000	0.1300	0.1200	0.1100
P@20	0.2800	0.1150	0.1100	0.1600	0.1250	0.1250
P@30	0.2900	0.0933	0.1000	0.1700	0.1133	0.1167
MAP	0.1277	0.0465	0.0527	0.0815	0.0395	0.0350

My Strategy for setting the weights is to emphasize some fields by setting high weights and reduce other fields' influence by giving negligible weight in each experiment run. Different combinations are tried so that the effect of different fields can be compared. For certain significant field (such as body), set weights of different levels so that the trend of performance with weighting can be analyzed.

Basically for this query set, all of the five multiple representations performed worse than BOW. And the less the body field weight, the lower the accuracy. The reason behind this might be that some relevant documents do not have the query terms in title or url or other fields except for body.

In the first experiment, I distributed equal weight to five fields (0.2 in this case), so that this group can be a reference to later runs. From the chart we can see, the performance of equal distributed weight among fields performed badly for both MAP and P@10/20/30, which means that equal weighting is detrimental for both precision and recall. In the second run, three fields [title, body, inlink] are emphasized and [url, keywords] are given very little weights, both precision and recall are improved. The reason of the improvement might be that body and title contains most of the words of a document. In the third run, the weight of [body, title] are further increased. As expected, both precision and recall are improved a lot. The opposite weighting strategy of run 3 is also tried, which is to put most weights on [url, keywords, inlink] and negligible weight on [body, title]. As can be seen from the chart, the MAP of run 4 is even worse than run 1 (equal weight), but the precision of top 30 documents are a little bit improved. The precision is query-based. For query #26, #33, #190, the precision of top 30 documents are improved a lot, but for others precision is similar between run 3 and run 4 or even dropped a little bit in run 4. A observation after four runs is that the less the body field weight, the worse the MAP is. In order to confirm this guess, run 5 is conducted with almost equal weight on [url, keyword, inlink, title] and negligible weight 0.01 on [body]. As expected, run 5 gives the worst MAP 0.035 among all the experiments. The precision of top documents is a bit higher than equal weight (run 1) but not higher than run 3 (body weights 0.6).

The time of running the query set for BOW is 15s while the time for different five weighting are 24s ~ 27s . The reason why different representation is slower might be that the search engine needs to check all five fields and calculates scores instead of just considering body fields in BOW.

3. Experiment 3: Sequential dependency models

Example Query: Provide your structured query for query "fickle creek farm".

#WAND (0.7 #and (fickle creek farm) 0.2 #and (#near/1(fickle creek) #near/1(creek farm)) 0.1 #and (#window/8 (fickle creek) #window/8 (creek farm)))

	Indri BOW (body)	**0.7 AND 0.2 NEAR 0.1 WINDOW**	**0.5AND 0.4 NEAR 0.1 WINDOW**	**0.33 AND 0.33 NEAR 0.34 WINDOW**	**0.2AND 0.4 NEAR 0.4 WINDOW**	**0.1 AND 0.4 NEAR 0.5 WINDOW**
P@10	0.2300	0.2800	0.3300	0.3400	0.3500	0.3500
P@20	0.2800	0.2850	0.3100	0.3050	0.3200	0.3200
P@30	0.2900	0.2967	0.3033	0.3000	0.3033	0.3067
MAP	0.1277	0.1411	0.1541	0.1690	0.1683	0.1676

I start with [0.7 0.2 0.1] for three components[AND NEAR WINDOW], and then I gradually decrease the weight for AND and increase the weight for NEAR and WINDOW to find the best combination.

All five SDM has better performance than BOW on both MAP and Precision@10/20/30. By comparison between BOW and the first run[0.7 0.2 0.1] we can see that introducing bigrams and short window operators improved both precision and recall a lot. The reason behind that is that SDM gives a boost to documents that contain query terms close together, which is more likely to be relevant. After that, with decreasing weight on #AND and increasing weight on #WINDOW and #NEAR, the performance become better and better and reach the peak when #AND, #NEAR, #WINDOW has equal weight. The best MAP is up to 0.169 and the Precisions at top documents are more than 0.30. After that, when I keep increasing the weight of #NEAR and #WINDOW, the performance is a little bit drop but pretty much stable at ~0.16.

The time of running BOW query set is 15s while running SDM query set is 21-23s. The reason why SDM is slow is that SDM not only process default operator but also process bigrams and short window operators. Since the accuracy improved about 33%, the increase of computational cost is acceptable when accuracy is more important than time.

4. Experiment 4: Multiple representations + SDMs

Example Query: Provide your structured query for query “fickle creek farm”.

#WAND(0.5 #AND (#WSUM(0.01 fickle.url 0.01 fickle.keywords 0.37 fickle.title 0.6 fickle.body 0.01 fickle.inlink) #WSUM(0.01 creek.url 0.01 creek.keywords 0.37 creek.title 0.6 creek.body 0.01 creek.inlink) #WSUM(0.01 farm.url 0.01 farm.keywords 0.37 farm.title 0.6 farm.body 0.01 farm.inlink)） 0.5 #WAND (0.33 #and (fickle creek farm) 0.33 #and (#near/1(fickle creek) #near/1(creek farm)) 0.34 #and (#window/8 (fickle creek) #window/8 (creek farm))))

	Indri BOW (body)	**w=1.0 (Exp 2)**	**w1=0.8**	**w2=0.6**	**w3=0.5**	**w4=0.3**	**w5=0.1**	**w=0.0 (Exp 3)**
P@10	0.2300	0.1300	0.1600	0.1800	0.1900	0.2700	0.3200	0.3400
P@20	0.2800	0.1600	0.1850	0.2000	0.2100	0.2550	0.2950	0.3050
P@30	0.2900	0.1700	0.1967	0.2200	0.2333	0.2667	0.2900	0.3000
MAP	0.1277	0.0815	0.0913	0.0936	0.0985	0.1350	0.1650	0.1690

The query structure is defined to take the best of both multiple representations and SDMs to improve the precision and MAP values. However, from the Exp2, the best result of multiple representation is stills much lower than BOW. Therefore in the mixed model, as expected, the more weight on multiple representation, the worse the accuracy in final result. With w decreasing (the weight on SDMs increasing), the accuracy become better and better and closed to the best result given by only SDM. The results of 5 runs of mixed model are between the result of only multiple representations and the result of only SDMs.

The running time for BOW, multiple representations, and SDM is 15s, 24-27s, and 21-23s respectively. And the running time for multiple representations+SDM is 27-28s, which is higher than BOW, but pretty much similar to the running time for multiple representations. Since the accuracy is not improved by adding multiple representations to SDMs, it doesn't worth to cost extra time to run the multiple representation part. Using only SDM part will cost less time and give better performance at least on this query set.