Your Name: ADITYA GAUTAM

Your Andrew ID: agautam1@andrew.cmu.edu

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.
   Ans) 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)?
   Ans) 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.
   Ans) 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)?
   Ans) 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: Aditya Gautam

Your Andrew ID: agautam1@andrew.cmu.edu

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/3(sewing #SYN(instruction guidelines)) #NEAR/6(#SYN(instruction guidelines) sewing)

79:voyager.body voyager.url

84:#NEAR/3(#SYN(tectonic continental) plates) #NEAR/3(plates #SYN(tectonic continental))

89:#NEAR/120(Obsessive compulsive disorder) ocd.body

108:#NEAR/1(ralph owen brewster) #NEAR/3(ralph brewster) #NEAR/1(owen brewster)

141:#NEAR/40(#SYN(va virginia) #SYN(dmv driving motor vehicle) registration) #NEAR/35(virginia motor registration) #NEAR/35(va motor registration) #NEAR/35(virginia dvm registration) #NEAR/30(va dvm registration)

146:#NEAR/1(sherwood regional library)

153:pocono.body #NEAR/1(pocono.title mountains.title) #NEAR/1(pocono mountains)

171:#NEAR/1(ron howard) #NEAR/1(ron.title howard.title)

197:#NEAR/40(idaho state flower) #NEAR/1(Philadelphus lewisii) #NEAR/1(Philadelphus.title lewisii.title) Syringa.body Syringa.title #NEAR/1(idaho.title state.title flower.title)

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/3(sewing #SYN(instruction guidelines)) #NEAR/6(#SYN(instruction guidelines) sewing)

Here, the information need is to get the instructions of sewing. So, I used the corresponding synonyms words like guidelines along with instructions. With this, since the 'instructions' is supposed to be near by the sewing, so the distance of near operator is kept around 3 in forward and 6 in reverse.

79:voyager.body voyager.url

Since this is the single word query, All the results in the body or URL would be relevant and thus fetched. Tried with title and keywords also, but somehow that reduced the MAP, not sure why.

84:#NEAR/3(#SYN(tectonic continental) plates) #NEAR/3(plates #SYN(tectonic continental))

Since the information need to look for continental plates, I have added a SYN with tectonic, since it is the widely used replacement of continental and occurs frequenctly across web pages. Also, both the context of possible combination i.e. forward and reverse has been taken into consideration.

89:#NEAR/120(Obsessive compulsive disorder) ocd.body

Full form expansion has been done for the OCD. This will help in reducing the ambiguity and making the query better. Distance of 120 is used after various experimentation and tuning and since the possibility of the words can be within couple of sentences. So, 120 suits the match.

108:#NEAR/1(ralph owen brewster) #NEAR/3(ralph brewster) #NEAR/1(owen brewster)

Search is for the a specific person, so all the words needs to be next to each other. However, one of the three word can be skipped, so other relevant queries related to this was added to cover all cases.

141:#NEAR/40(#SYN(va virginia) #SYN(dmv driving motor vehicle) registration) #NEAR/35(virginia motor registration) #NEAR/35(va motor registration) #NEAR/35(virginia dvm registration) #NEAR/30(va dvm registration)

Word expansion of VA is used along with various possible combination of dmv since any/all of the words can be present in the sentence. Word registration is necessary since not adding this word could give us irrelevant results. So, various near distance is used based on experiments and heuristics.

146:#NEAR/1(sherwood regional library)

Information need is for a specific query thus words needs to be next to each other. So, NEAR operator is used with a distance of 1.

153:pocono.body #NEAR/1(pocono.title mountains.title) #NEAR/1(pocono mountains)

Pocono is a place which is very famous for mountains. So, It is very likely that a usual single word query is related to that famous thing associated with it. Likewise, single Pocono word in the body is coupled with the Pocono mountains present in the body or the title of the web page, next to each other.

171:#NEAR/1(ron howard) #NEAR/1(ron.title howard.title)

Information need is for a particular person with a specific name. Likewise NEAR operator with distance 1 would give us the best match as words needs to be next to each other. With body, the other field which would be of interest would be the title field in the webpage since any title with the person name would be very relevant.

197:#NEAR/40(idaho state flower) #NEAR/1(Philadelphus lewisii) #NEAR/1(Philadelphus.title lewisii.title) Syringa.body Syringa.title #NEAR/1(idaho.title state.title flower.title)

Information need is for a specific state flower of Idaho which is also famous/called by other names. Thus all the other famous name of the flower was also added to the query. All the words are essentials and needs to be there in the web page else we might get generalized web pages. So, I have used the operator NEAR with the distance 40. Distance of 40 is used since the words like 'Idaho' and 'State flower' could be at a distance within a sentence or a paragraph and would be a relevant result.

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	0.5500	0.6100	0.5400
P@20	0.2800	0.3500	0.4700	0.6550	0.5850	0.6050
P@30	0.3367	0.3667	0.4233	0.6333	0.5400	0.6000
MAP	0.1071	0.1985	0.2057	0.3022	0.2964	0.3206

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.

Using synonyms improve the overall MAP and P@n as the relevant pages with other similar words would also be fetched and ranked higher. Performance of Indri is slightly better than BM25 for the same structured query because of probabilistic modelling and 2 stage smoothing. For the specific words like known person, the NEAR operator with least distance works the best as expected, however for some queries, the higher distance gives better performance rather than short distance like for OCD. This is a bit different from expectation. As per my understanding, the distance of 1 or 2 in NEAR operator should work the best, but is it not the case with the given data set.

Performance of Indri seems to be the best among all the three ranking system and is expected also since it will rank based on the probabilistic model with 2 stage smoothing so it will consider the docs in which any of the query word occurs. With Ranked Boolean AND, we consider only the document in which all the query words are present but in Indri, we consider the probabilistic AND operation, thus better performance of results.

In MB25 and Indri, changing different parameters like smoothing etc. would have different effect on the MAP and the most optimized value is like a global maxima, increasing of decreasing the value results in lower MAP, P@n etc. Like for Indri, the best optimal value for mu seems to be 2500 and 0.4 for lambda. Changing any parameters either side lowers MAP. This is expected also, since these parameters gives the proper balance to the prior information and current term occurrence.

Another strange thing to note is that, sometimes the NEAR operator distance gives the best results when with large distance values even though it should be less as per the intuition like for 'Idaho state flower', the result is best with distance of 40 as compared to 5-10(with institution that words needs to be near)

3 Experiment 3: BM25 Parameter Adjustment

3.1 k_1

	k_1							
	1.2	0	0.1	0.4	0.8	1	1.5	5
P@10	0.6100	0.4900	0.5600	0.5600	0.5600	0.5600	0.5700	0.5500
P@20	0.5850	0.4500	0.5700	0.5700	0.5750	0.5700	0.5650	0.5400
P@30	0.5400	0.5333	0.5367	0.5367	0.5333	0.5367	0.5267	0.5233
MAP	0.2964	0.2184	0.2941	0.2949	0.2917	0.2926	0.2933	0.2905

.

3.2 b

	b							
	0.75	0	0.1	0.2	0.4	0.6	0.8	1.0
P@10	0.6100	0.5600	0.5500	0.5200	0.5200	0.5500	0.5500	0.5700
P@20	0.5850	0.5950	0.5900	0.5750	0.5400	0.5550	0.5600	0.5550
P@30	0.5400	0.5900	0.5800	0.5700	0.5467	0.5500	0.5300	0.5300
MAP	0.2964	0.2964	0.3048	0.3013	0.3008	0.3003	0.2903	0.2737

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.

The value chosen for both the parameters is incremental increasing as I wanted to see the impact of doc length weighting parameters and the term frequency weighting. How slowly increasing one of the parameter (keeping other constant) will change the overall MAP, and at what value would I get the most

optimized performance w.r.t one parameter. BM25 considers the probability of the word occurrence in the index terms of a particular document for the specified field of the doc. Likewise, when we reduce the b k1 i.e reducing the importance of term frequency, we see the MAP increases to some extend. With K1 set as zero, we see that no importance is given to term frequency, thus the retrieved results would be very less significant thus the MAP is lowest in BM25 ranking method.

However, on **increasing K1**, the value of MAP increases to some extend and then starts decreasing. This implies that weightage of different parameters will give best model and taking any value to extreme value would decrease the overall effectiveness of the results as each term has its importance. Parameter K1 controls the weightage given to the term frequency and document length w.r.t average length.

Parameter b determine the importance of doc length in the ranking method. With this set to zero, we are only considering the term frequency irrespective of the length of doc thus we are expected to get not so good result. Likewise, the MAP value is lower. I kept on increasing the value by 0.1 and 0.2 later to see the impact of the document length. As expected, the MAP values increases to some extend and then starts decreasing. At the extreme value of 1, we are getting lowest MAP as too much importance/weightage has been given to the doc length.

Likewise BM25 works best for some chosen parameter based on experiments/heuristics and give significance to the doc length, term frequency and the idf.BM25 performance was much better then unranked Boolean and the ranked Boolean with default query. However, structured query of Ranked Boolean performed better than BM25.

4 Indri Parameter Adjustment

4.1 μ

	μ							
	2500	0	500	1000	1500	2000	3000	5000
P@10	0.5400	0.5100	0.5900	0.5600	0.5700	0.5700	0.5600	0.5900
P@20	0.6050	0.5400	0.5450	0.5700	0.5900	0.5900	0.6100	0.6200
P@30	0.6000	0.5200	0.5433	0.5567	0.5800	0.5900	0.6033	0.6067
MAP	0.3206	0.2831	0.3117	0.3165	0.3203	0.3198	0.3187	0.3197

4.2 λ

	λ							
	0.4	0	0.1	0.2	0.5	0.6	0.8	1
P@10	0.5400	0.5400	0.5400	0.5400	0.5400	0.5400	0.5300	0.5700
P@20	0.6050	0.6050	0.6050	0.6050	0.6050	0.6000	0.5950	0.5200
P@30	0.6000	0.6000	0.6000	0.6000	0.6000	0.6000	0.6000	0.4933
MAP	0.3206	0.3204	0.3204	0.3205	0.3200	0.3199	0.3197	0.2455

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.

The two main parameters for Indri model are Mu and lambda, which defines the smoothing and the weightage given to the prior probability of occurrence of the term in the corpus. I played around with lambda in incremental fashion starting from 0 to 1 with the step size of 0.1 and 0.2 sometime. This is to see the impact of prior probability in the overall ranking of the document. In the same fashion keeping lambda constant, I increased the mu from 0 to 5000 with step size of 500/1000 as I wanted to see the impact of this smoothing of the information ranking.

Keeping **lambda to value between 0.2-0.4 gives the best result** as we give less weightage to the prior and more to the term frequency occurrence. As we keep on increasing lambda, we see that give much more importance to the prior, thereby reducing the effect of the term occurrence. So, the result was worse as expected. **At an extreme value of 1, we are only considering the prior probability of the term**. This means that no relevance is given to the term frequency and document length. So, we got **worst possible result with MAP of .2455.**

On increasing the mu from 0 to 5000, we see that best/most optimal performance is at a value of 2500 as it will define proper importance to the smoothing and the term frequency along with document length. At mu=0 and extremely high values, we see the worst performance because of unbalances weightage given to prior and the actual term occurrence/doc length. Also, changing the value by 500 wouldn't have a very significant impact into the ranking. At higher value, MAP is decreasing very slowly with increase in lambda.

Results of Indri was much better than BM25 and Ranked Boolean. Indri performance is the best among all the methods experiment yet as it has a good impact of two stage smoothing, would fetch documents even with one or more term absent and provide a good balance in prior and term frequency. None of these things are considered in a single model and thus Indri performed best.