Your Name: LIU Xi

Your Andrew ID: xiliu1

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: LIU Xi

Your Andrew ID: xiliu1

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

BM25 achieved best performance while Ranked Boolean the worst.

2 Experiment 2: Queries with Synonyms and Phrases

2.1 Queries

List your queries.

Synonyms: I use ***#SYN*** for ***synonyms***. When constructing queries, I looked into HW1 FAQ to identify the information needs and extract synonyms.

Phrase: I use ***#NEAR/n*** for ***phrases***, for example #NEAR/1(heart rate) represents the phrase "heart rate".

Original Query	**My Query with #SYN and #NEAR/n**
cheap internet	#NEAR/1(#SYN(cheap inexpensive affordable low-cost) internet)
djs	#SYN(djs dj #NEAR/1(radio disc jockey) #NEAR/1(disc jockey))
lower heart rate	lower #NEAR/1(heart #SYN(beat rate))
ps 2 games	#NEAR/1(#SYN(ps playstation) 2 games)
elliptical trainer	#SYN(#NEAR/1(elliptical trainer) #NEAR/1(elliptical machine))
avp	#SYN(avp #NEAR/2(Association of Volleyball Professionals) #NEAR/2(Alien vs Predator) #NEAR/1(Wilkes-Barre Scranton International Airport))
living in india	#NEAR/2(#SYN(living housing travel life) in india)
fickle creek farm	#NEAR/1(fickle creek farm)
uplift at yellowstone national park	#SYN(uplift elevation eruption) at #NEAR/1(Yellowstone national park)
brooks brothers clearance	#NEAR/1(brooks brothers) #SYN(clearance sale discount coupons outlet)

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.

Strategy:

1. Based on information needs, identify synonyms for each term and add them into #SYN to expand the query.
2. Based on information needs, identify phrases and use #NEAR/1 narrow the searching scope to exactly the phrase instead of separate words.
3. For phrases contains frequent words, for example, "of", adjust #NEAR/1 to #NEAR/2 or #NEAR/3.
4. #SYN could be used inside #NEAR/n, and vice versa.

Descriptions:

#NEAR/1(#SYN(cheap inexpensive affordable low-cost) internet)

Strategy:

1. Use #SYN and add ***inexpensive affordable low-cost*** to expand term "cheap".
2. Use #NEAR/1 to search by phrase instead of separate word "cheap" and "Internet".

Deviation:

1. add ***inexpensive affordable low-cost*** to expand term "cheap".

#SYN(djs dj #NEAR/1(radio disc jockey) #NEAR/1(disc jockey))

Strategy:

1. Use #SYN and add ***"dj", "radio disc jockey"*** and ***"disc jockey"*** to expand term "djs".
2. Use #NEAR/1 to represent by phrase ***"radio disc jockey"*** and ***"disc jockey"*** instead of separate words.

Deviation:

1. add ***"dj", "radio disc jockey"*** and ***"disc jockey"*** to expand term "djs"

lower #NEAR/1(heart #SYN(beat rate))

Strategy:

1. Use #SYN and add ***"beat"*** to expand term "rate".
2. Use #NEAR/1 to represent by phrase ***"heart beat"*** and ***"heart rate"*** instead of separate words.

Deviation:

1. add ***"beat"*** to expand phrase " heart rate" by "***heart beat***".

#NEAR/1(#SYN(ps playstation) 2 games)

Strategy:

1. Use #SYN and add ***"playstation"*** to expand term "ps".
2. Use #NEAR/1 to represent by phrase ***"ps 2 games"*** and ***"playstation 2 games"*** instead of separate words.

Deviation:

1. add ***"playstation"*** to expand term "ps"

#SYN(#NEAR/1(elliptical trainer) #NEAR/1(elliptical machine))

Strategy:

1. Use #SYN and add **"elliptical machine*"*** to expand term "**elliptical trainer**".
2. Use #NEAR/1 to represent by phrase ***"elliptical machine"*** and ***"elliptical trainer"*** instead of separate words.

Deviation:

1. add **"elliptical machine*"*** to expand term "elliptical trainer"

#SYN(avp #NEAR/2(Association of Volleyball Professionals) #NEAR/2(Alien vs Predator))

Strategy:

1. Use #SYN and add ***"Association of Volleyball Professionals"*** to expand term "***Alien vs. Predator***".
2. Use #NEAR/2 to represent by phrase ***"Association of Volleyball Professionals"*** and "***Alien vs. Predator"*** instead of separate words.
3. Adjust #NEAR/1 to #NEAR/2 because "of " and "vs" are frequent words and will be removed in query parsing.

Deviation:

1. Add ***"Association of Volleyball Professionals"*** to expand term "Alien vs. Predator".
2. Remove "." In "***Alien vs. Predator"*** to avoid confliction with structure query syntax.
3. Adjust #NEAR/***1*** to #NEAR/***2*** because "of " and "vs" are frequent words and will be removed in query parsing.

#NEAR/2(#SYN(living housing travel life) in india)

Strategy:

1. Use #SYN and add *"**housing", "travel", "life"*** to expand term "**living**".

2. Use #NEAR/2 to represent by phrase ***“living in india”*** instead of separate words.

Deviation:

1. Adjust #NEAR/***1*** to #NEAR/***2*** because “in” is frequent word and will be removed in query parsing.

#NEAR/1(fickle creek farm)

Strategy:

1. Use #NEAR/1 to represent by phrase ***“fickle creek farm”*** instead of separate words.

#SYN(uplift elevation eruption) #NEAR/1(Yellowstone national park)

Strategy:

1. Use #SYN and add ***“elevation”, “eruption”*** to expand term “uplift”.
2. Use #NEAR/1 to represent by phrase **“Yellowstone national park”** instead of separate words.

Deviation:

1. add ***“elevation”, “eruption”*** to expand term “uplift”.

#NEAR/1(brooks brothers) #SYN(clearance sale discount coupons outlet)

Strategy:

1. Use #SYN and add ***“sale”, “discount”, “coupons”, “outlet”*** to expand term “clearance”.
2. Use #NEAR/1 to represent by phrase “brooks brothers” instead of separate words.

Deviation:

1. add ***“sale”, “discount”, “coupons”, “outlet”*** to expand term “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	0.2600	0.3400	0.3100
P@20	0.1800	0.2950	0.2800	0.2800	0.3500	0.3800
P@30	0.1667	0.2967	0.2900	0.2667	0.3433	0.3767
MAP	0.0566	0.1304	0.1277	0.1173	0.1750	0.1839

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.

Trend:

Trend 1: the use of synonyms and phrases largely improves the P@10, P@20, P@30, and MAP.

Compared with Ranked Boolean BOW, the use of synonyms and phrases on Ranked Boolean largely improves the P@10, P@20, P@30, and MAP.

Compared with BM25 BOW, the use of synonyms and phrases on BM25 largely improves the P@10, P@20, P@30, and MAP.

Compared with Indri BOW, the use of synonyms and phrases on Indri largely improves the P@10, P@20, P@30, and MAP.

Trend 2: BM25 and Indri outperform in terms of P@10, P@20, P@30, and MAP.

Compared with Ranked Boolean BOW, BM25 BOW and Indri BOW outperform with higher P@10, P@20, P@30, and MAP.

Compared with Ranked Boolean with phrases and synonyms, BM25 with phrases and synonyms and Indri with phrases and synonyms outperform with higher P@10, P@20, P@30, and MAP.

Behavior:

Yes, the use of synonyms and phrases behaved as I expected, that is, largely improved P@10, P@20, P@30, and MAP on Ranked Boolean, BM25, and Indri.

Other Observation:

Without phrases and synonyms, BM25 performs best; while with phrases and synonyms, Indri performs best.

The precision at each position is relatively close to each other for the same retrieval model with the same query.

For the models with phrases and synonyms, Indri should be chosen if we emphasis on the ***overall*** precision, while BM25 should be chosen if we emphasis on the precision for the ***top documents***.

Experiment 3: BM25 Parameter Adjustment

2.5 k_1

	k_1							
	1.2	0.2	0.4	0.8	1.5	3	8	120
P@10	0.3	0.28	0.28	0.3	0.29	0.29	0.26	0.2
P@20	0.295	0.305	0.31	0.3	0.295	0.29	0.265	0.205
P@30	0.2967	0.3067	0.31	0.3	0.2933	0.2933	0.2567	0.2
MAP	0.1304	0.1266	0.1275	0.1303	0.1298	0.129	0.1216	0.099

2.6 b

	b							
	0.75	0.95	0.85	0.65	0.55	0.45	0.35	0.25
P@10	0.3	0.24	0.27	0.26	0.25	0.24	0.22	0.25
P@20	0.295	0.305	0.315	0.295	0.285	0.3	0.305	0.305
P@30	0.2967	0.32	0.3067	0.3167	0.3133	0.31	0.32	0.3033
MAP	0.1304	0.1256	0.1299	0.1287	0.1303	0.1307	0.1309	0.1302

2.7 Discussion

Explain your reasons for choosing the values that you tested, and how those reasons are related to how BM25 works.

k1:

By definition, k1>0, so I chose 0.2, 0.4, 0.8, 1.5, 3, 8, 120 to experiment values from very small value (1/6 of the default setting) to very large value (100 times of the default setting).

According to the formula, when k1 is very large, the influence of (1-b) would be more significant while the influence of term frequency in the documents would be less significant, that is, the term frequency within current document is not impacting the score. On the other hand, if k1 is very small, larger term frequency leads to smaller document score.

Among the experimental k1s, k1=1.2 outperforms with higher MAP and precision@n. However, k1=1.5 achieved nearly same performance too.

b:

By definition, 0<b<1, so I chose 0.95, 0.85, 0.75, 0.65, 0.55, 0.45, 0.35, 0.25 to experiment values from very small value to large value within the scope. The values are uniformly distributed in the selection.

According to the formula, when b is very small, the influence of k1 would be more significant; meanwhile the influence of document length in the documents would be less significant, that is, long documents and short documents are treated similarly and documents receives larger score if the term frequency is large. On the other hand, if b is close to 1, the length of documents brings more impact on the score, that is, longer documents receive smaller score.

Among the experimental bs, b=0.35 outperforms with higher MAP and precision@n. However, b=0.45 achieved nearly same performance too.

Discuss any changes in retrieval performance that you observed, and the significance of any trends that you observed.

k1:

When 0<k1<1.2, the performances, in terms of precision and MAP are close but increases with the increasing of k1; while 1.2<k1<3 the performances are close but decrease with the increasing of k1; and after that the performance decreases with the increasing of k1.

The best setting of k1 for this corpus should lies in 0.8<k1<1.5, and current optimal setting is 1.2.

The trend is not significant within 0<k1<3, but turns to be significant when k1>3.

b:

P@10, P@20, P@30 MAP remains relatively stable with change of b within 0<b<1. They slightly increase with the increasing of b within 0.25<b<0.45, and slightly decreases with the increasing of b within 0.45<b<1.

The best setting of k1 for this corpus should lies in 0.25<b<0.45, and current optimal setting is 0.35.

The trend is not significant within 0<b<1.

3 Indri Parameter Adjustment

3.1 μ

	μ							
	2500	4500	3500	1500	800	500	50	5
P@10	0.23	0.22	0.22	0.23	0.29	0.32	0.26	0.26
P@20	0.28	0.27	0.28	0.325	0.325	0.31	0.325	0.3
P@30	0.29	0.2833	0.2833	0.3133	0.31	0.3167	0.3067	0.3033
MAP	0.1277	0.1225	0.1248	0.1315	0.1311	0.1346	0.1284	0.126

3.2 λ

	λ							
	0.4	0.7	0.6	0.5	0.3	0.2	0.1	0.05
P@10	0.23	0.19	0.2	0.23	0.24	0.25	0.27	0.27
P@20	0.28	0.265	0.275	0.285	0.29	0.295	0.3	0.3
P@30	0.29	0.2633	0.27	0.2833	0.2933	0.3	0.3067	0.31
MAP	0.1277	0.1205	0.1241	0.1267	0.1295	0.1318	0.1334	0.1341

3.3 Discussion

Explain your reasons for choosing the values that you tested, and how those reasons are related to how Indri works.

μ:

By definition, μ >0, so I chose 4500, 3500, 2500,1500, 800, 500, 50 to experiment values from very small value (1/50 of the default setting) to very large value (1.8 times of the default setting).

According to the formula, when μ is very small, the influence of λ would be more significant; while the influence of document length would be more significant, that is, longer documents get smaller score for a certain term. On the other hand, if μ is very large, p_{MLE} gets more impact so that documents with more frequent terms receives larger document score.

Among the experimental μ s, μ =500 outperforms with higher MAP and precision@n. However, μ =1500 achieved nearly same performance too.

λ:

By definition, 0< λ <1, so I chose 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1 to experiment values from very small value to large value within the scope. The values are uniformly distributed in the selection.

According to the formula, when λ is very small, the influence of p_{MLE} would be more significant; meanwhile the influence of μ and document length in the documents would be less significant, that is, long documents and short documents are treated similarly and documents receives larger score if it contains terms appears frequently in the collection. On the other hand, if λ is close to 1, document length and term frequency brings more impact on the score that is, longer documents receive smaller score, and documents containing frequent terms get larger score.

Among the experimental λ s, λ =0.05 outperforms with higher MAP and precision@n. However, λ =0.1 achieved nearly same performance too.

Discuss any changes in retrieval performance that you observed, and the significance of any trends that you observed.

μ:

MAP remains relatively stable with change of μ. P@10 and P@30 increase with the increasing of b within 5<b<500, and slightly decreases with the increasing of μ within **μ** >800.

The best setting of μ for this corpus should lies in 50<k1<2500, and current optimal setting is 500.

For P@n, the trend is significant within 500< μ <2500, but turns to be less significant when μ >2500. For MAP, it remains stable.

λ:

P@10, P@20, P@30 decrease with the increasing of λ within 0.05<b<0.7, the trend is not significant till $\lambda > 0.3$. MAP slightly decrease with the increasing of λ, relatively stable though.

The best setting of λ for this corpus should lies in 0< λ <0.1, and current optimal setting is 0.05.

The trend for MAP is not significant within 0< λ <1 while for P@n the decreasing turns to more significant till $\lambda > 0.3$.