Your Name: Shuang Xu

Your Andrew ID: sxu1

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: Shuang Xu

Your Andrew ID: sxu1

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: BM25 Parameter Adjustment

2.1 k_1

	k_1							
	1.2	0	1	2	3	4	100	20
P@10	0.4200	0.1100	0.4200	0.4100	0.4000	0.4000	0.2700	0.3100
P@20	0.3500	0.1350	0.3450	0.3450	0.3600	0.3550	0.2400	0.3150
P@30	0.3667	0.1533	0.3700	0.3667	0.3633	0.3600	0.2933	0.3267
MAP	0.1985	0.0665	0.1996	0.1874	0.1746	0.1686	0.1085	0.1422

2.2 b

	b							
	0.75	0	1	0.5	0.25	0.125	0.2	0.1
P@10	0.4200	0.3900	0.3900	0.4200	0.4600	0.4100	0.4100	0.4200
P@20	0.3500	0.4450	0.3600	0.4000	0.4400	0.4500	0.4450	0.4650
P@30	0.3667	0.4600	0.3500	0.4067	0.4267	0.4467	0.4233	0.4667
MAP	0.1985	0.1977	0.1703	0.2118	0.2179	0.2163	0.2157	0.2148

2.3 Parameters

K:
First let k1=0 to check if k1 has an impact on precision and recall. It does have affects.
Assume k1 within a small range will produce more accurate results. So test range [1,4] and find the similar precision results.

Guess there would be a threshold, if k1 is larger than the threshold, the precision would go down greatly. So test k1=100 and k1=20.

B:
First try the boundary to check if b has an impact on precision and recall. It seems no big affects. And then try binary search to find optimal b.

2.4 Discussion

If the document is relatively longer, it has more chance to include the query term, so it tends to get higher score. The parameter b is used to smooth document length. But in this experiment, it seems have little affect on the results. Larger b would greatly punish on long document. Parameter k will scale up the punishment.

b=0.25 produces highest P@10. When k=0, we totally ignore effects of long documents and make tf weight equals 0, which means document term frequency has no effect and idf dominate, it produces worst result. Small k, from 1-4 produces identical results but decreasing MAP. For this corpus, small punishment on long document and big rewards document term frequency would get better results.

3 Experiment 3: Indri Parameter Adjustment

3.1 μ

	μ							
	2500	0	1000	10000	100000	1000000	10000000	100000000
P@10	0.4000	0.4300	0.4200	0.3700	0.3000	0.2800	0.2700	0.2700
P@20	0.4700	0.4100	0.4400	0.4750	0.4700	0.4350	0.4050	0.4050
P@30	0.4233	0.3933	0.4267	0.5033	0.4600	0.4633	0.4600	0.4633
MAP	0.2057	0.1952	0.2143	0.1805	0.1470	0.1328	0.1297	0.1292

3.2 λ

	λ							
	0.4	0	1	0.5	0.25	0.1	0.3	0.75
P@10	0.4000	0.4000	0.0000	0.4000	0.3900	0.3900	0.3900	0.3700
P@20	0.4700	0.4750	0.0150	0.4700	0.4700	0.4750	0.4750	0.4300
P@30	0.4233	0.4233	0.0200	0.4167	0.4267	0.4233	0.4233	0.4000
MAP	0.2057	0.2142	0.0020	0.2012	0.2097	0.2132	0.2132	0.1862

3.3 Parameters

μ:
Set μ=0 to see if it has on impact on precision and recall. And it seems μ=0 generates better results. Then try larger and larger μ to see if the results would vary a lot.

λ:
First try the boundary to check if λ has an impact on precision and recall. It seems smaller λ is better. But actually the results for λ in [0,0.5] are quite similar. So check a number close to 1 but not 1 to see if gets the same result.

3.4 Discussion

The effect of λ is like that of idf. If λ is larger, it rewards the frequent terms in the whole collection, and if it is smaller, it rewards frequent terms in the document. In this experiment, small λ would produce better results. That's because the query set here contains short queries and for short queries 'idf weighting' is less important and small λ is good.

μ makes smoothing by introducing a small sample. It helps avoid 0 frequency and explain unseen words. In this case, small μ would produce better results as the documents are long and probabilities are more smooth and small μ is enough.

The results for small μ and λ are quite stable.

4 Experiment 4: Different representations

4.1 Example Query

```
#AND (
  #WSUM(0.2 sherwood.url  0.2 sherwood.keywords  0.2 sherwood.title  0.2 sherwood.body  0.2 sherwood.inlink)
  #WSUM(0.2 regional.url  0.2 regional.keywords  0.2 regional.title  0.2 regional.body 0.2 regional.inlink  )
  #WSUM(0.2 library.url  0.2 library.keywords   0.2 library.title    0.2 library.body  0.2 library.inlink))

#AND (
  #WSUM(0.1 sherwood.url  0.1 sherwood.keywords  0.2 sherwood.title  0.5 sherwood.body  0.1 sherwood.inlink)
  #WSUM(0.1 regional.url  0.1 regional.keywords  0.2 regional.title  0.5 regional.body 0.1 regional.inlink  )
  #WSUM(0.1 library.url  0.1 library.keywords   0.2 library.title    0.5 library.body  0.1 library.inlink))


#AND (
  #WSUM(0.1 sherwood.url  0.1 sherwood.keywords  0.1 sherwood.title  0.5 sherwood.body  0.2 sherwood.inlink)
  #WSUM(0.1 regional.url  0.1 regional.keywords  0.1 regional.title  0.5 regional.body 0.2 regional.inlink  )
  #WSUM(0.1 library.url  0.1 library.keywords   0.1 library.title    0.5 library.body  0.2 library.inlink))


#AND (
  #WSUM(0.1 sherwood.url  0.2 sherwood.keywords  0.1 sherwood.title  0.5 sherwood.body  0.1 sherwood.inlink)
  #WSUM(0.1 regional.url  0.2 regional.keywords  0.1 regional.title  0.5 regional.body 0.1 regional.inlink  )
  #WSUM(0.1 library.url  0.2 library.keywords   0.1 library.title    0.5 library.body  0.1 library.inlink))

#AND (
  #WSUM(0.1 sherwood.url  0.2 sherwood.keywords  0.2 sherwood.title  0.4 sherwood.body  0.1 sherwood.inlink)
  #WSUM(0.1 regional.url  0.2 regional.keywords  0.2 regional.title  0.4 regional.body 0.1 regional.inlink  )
  #WSUM(0.1 library.url  0.2 library.keywords   0.2 library.title    0.4 library.body  0.1 library.inlink))
```

4.2 Results

	Indri BOW (body)	0.20 url 0.20 keywords 0.20 title 0.20 body 0.20 inlink	0.10 url 0.10 keywords 0.20 title 0.50 body 0.10 inlink	0.10 url 0.10 keywords 0.10 title 0.50 body 0.20 inlink	0.10 url 0.20 keywords 0.10 title 0.50 body 0.10 inlink	0.10 url 0.20 keywords 0.20 title 0.40 body 0.10 inlink
P@10	0.2000	0.0000	0.2000	0.1000	0.2000	0.2000
P@20	0.1500	0.1000	0.1000	0.1000	0.1000	0.1000
P@30	0.2000	0.0667	0.1333	0.1333	0.1667	0.1000
MAP	0.1234	0.0161	0.0833	0.0754	0.0833	0.0621

4.3 Weights

Test the same weight for all fields.
Assume body gets the highest weight and arrange a relatively higher weight for only one field while holding other fields weights the same to see which field is more important.
Give higher weights for the most two important fields to see if the results get better.

4.4 Discussion

Keywords and title gives more information than inlink and url for P@10 but have identical performance for P@20, P@30 and MAP. Higher weight on keywords field produces best results as the query is perfectly suitable to be used as keywords. Although different queries would prefer different fields, I believe title and keywords filed would produce better accuracy as they themselves represent highly frequent terms of the document and they're usually longer than inlink or url (counting effective words).

5 Experiment 5: Sequential dependency models

5.1 Example Query

#wand(0.5 #and(sherwood regional library) 0.25 #and(#near/1(regional library) #near/1(sherwood regional)) 0.25 #and(#window/8(regional library) #window/8(sherwood regional)))

#wand(0.25 #and(sherwood regional library) 0.5 #and(#near/1(regional library) #near/1(sherwood regional)) 0.25 #and(#window/8(regional library) #window/8(sherwood regional)))

#wand(0.25 #and(sherwood regional library) 0.25 #and(#near/1(regional library) #near/1(sherwood regional)) 0.5 #and(#window/8(regional library) #window/8(sherwood regional)))

#wand(0.2 #and(sherwood regional library) 0.4 #and(#near/1(regional library) #near/1(sherwood regional)) 0.4 #and(#window/8(regional library) #window/8(sherwood regional)))

#wand(0.02 #and(sherwood regional library) 0.49 #and(#near/1(regional library) #near/1(sherwood regional)) 0.49 #and(#window/8(regional library) #window/8(sherwood regional)))

5.2 Results

	Indri BOW (body)	**0.50 AND 0.25 NEAR 0.25 WINDOW**	**0.25 AND 0.50 NEAR 0.25 WINDOW**	**0.25 AND 0.25 NEAR 0.50 WINDOW**	**0.20 AND 0.40 NEAR 0.40 WINDOW**	**0.02 AND 0.49 NEAR 0.49 WINDOW**
P@10	0.2000	0.5000	0.6000	0.6000	0.6000	0.6000
P@20	0.1500	0.4000	0.4000	0.4000	0.4000	0.4000
P@30	0.2000	0.3333	0.3333	0.3333	0.3333	0.3333
MAP	0.1234	0.3520	0.3823	0.3805	0.3823	0.3823

5.3 Weights

Holding two operators same weights and another highest weight to see if that operator has biggest effect. Find #AND has smallest effect so try to lower the weight of #AND and see if the results would get better.

5.4 Discussion

The results are stable and much better than original query. It proves that sequential dependency queries help a lot in precision in this case. #WINDOW and #NEAR have similar performance because the three words in query tend to be very close in their locations.