Homework 2

1 Introduction

1.1 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

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

   No

3. Did you examine anyone else's software for this assignment (Yes or No)? It is not necessary to mention software provided by the instructor.

   No

4. 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

5. Are you the author of every word of your report (Yes or No)?

   YES

1.2 Instructions

2 Experiment 1: Baselines

2.1 Experimental Results

	Ranked Boolean AND (Exp-2.1a)	BM25 BOW (Exp-2.1b)	Indri BOW (Exp-2.1c)
P@10	0.4300	0.6000	0.6700
P@20	0.4400	0.5600	0.5900
P@30	0.4000	0.4800	0.5067
MAP	0.1060	0.1477	0.1607

3 Experiment 2: Indri Parameter Adjustment

3.1 Experimental Results

	μ (Note: λ=0.01)				
	1500 (Exp-3.1a)	**500 (Exp-3.1b)**	**1000 (Exp-3.1c)**	**2000 (Exp-3.1d)**	**2500 (Exp-3.1e)**
P@10	0.7000	0.6700	0.7100	0.6900	0.6800
P@20	0.6300	0.6050	0.6300	0.6300	0.6200
P@30	0.5300	0.5333	0.5367	0.5167	0.5200
MAP	0.1806	0.1742	0.1790	0.1800	0.1748

	λ (Note: μ=1500)				
	0.4 (Exp-3.2a)	**0.0 (Exp-3.2b)**	**0.2 (Exp-3.2c)**	**0.7 (Exp-3.2d)**	**1.0 (Exp-3.2e)**
P@10	0.6800	0.7000	0.6900	0.6300	0.0
P@20	0.6000	0.6300	0.6200	0.5850	0.0
P@30	0.5100	0.5300	0.5200	0.4667	0.0033
MAP	0.1682	0.1809	0.1748	0.1533	0.0000

3.2 Parameters

A grid search is performed on the mu and lambda space. I found out lambda around 0.01 and mu around 1500 performed well. So, I set lambda as 0.01 for the 3.1 experiments and set mu as 1500 for 3.2 experiments.

In order to explore the effect of mu under the setting of lambda being 0.01, I increased and decreased mu by interval of 500 and got 500, 1000, 2000 and 2500.

In order to explore the effect of lambda under the setting of mu being 1500, I chose lambda across its range. The range of lambda is from 0.0 to 1.0. So, other than 0.4, I took two points of each side in the range and also made sure that I have 0.0 and 1.0 included. The parameters I chose are 0.0, 0.2, 0.7 and 1.0.

3.3 Discussion

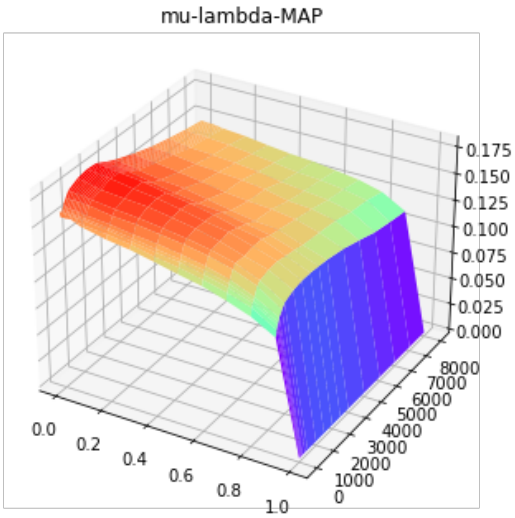
The score of Indri is calculated by $(\mathbf{1}-\lambda)\frac{tf+\mu P}{length\_d+\mu}+\lambda \boldsymbol{P}$, where $\boldsymbol{P}=\frac{ctf}{length\_c}$.

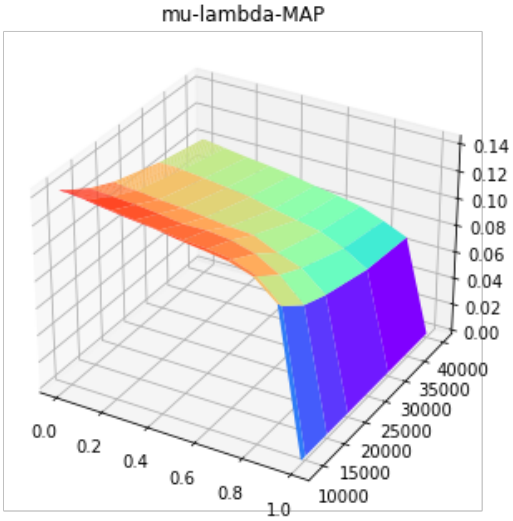
Analysis for lambda:

P represents the frequency of term in the whole collection. $\frac{tf+\mu P}{length\_d+\mu}$ represents the frequency of term in the current document. Therefore, lambda balances these two effects. If lambda is 0, The score is only based on $\frac{tf+\mu P}{length\_d+\mu}$, if lambda is 1, the score will not be calculated by the term frequency in current document, which mean the score for a same term will be the same across all documents. As we can image, this will not help with search. And we can conjecture that lambda should be small, because P is only used to help smoothing the $\frac{tf+\mu P}{length\_d+\mu}$ but should not be the main factor for the score. As we can see from the chart, the map is relatively high when lambda is small.

Analysis for mu:

The effect of mu is also to help smoothing the score of term frequency. If term frequency in a document is zero and we do not have smoothing for it, the probability of the term in the document will be zero. We can think mu as the default word count of a document. Then length_d + mu is the total length of the document because we add some default words to the document. And mu*P is the occurrence of current term among those added default words. So. Tf + mu*P is the final occurrent of current term in the document. As we can image, if mu is way much larger than length, the effect of default words will dominate. So, there should be an optimal number for mu. According to the chart, we can see 1500 is a good number. And this is same for all choice of lambda. And MAP will not deteriorate dramatically according to mu.





4 Experiment 3: Indri Representations

4.1 Experimental Results

	Indri BOW (body) (Exp-4.1a)	**0.1 url 0.1 keywords 0.1 title 0.7 body (Exp-4.1b)**	**0.00 url 0.1 keywords 0.1 title 0.7 body (Exp-4.1c)**	**0.1 url 0.00 keywords 0.1 title 0.7 body (Exp-4.1d)**	**0.1 url 0.1 keywords 0.00 title 0.7 body (Exp-4.1e)**
P@10	0.7000	0.6800	0.6889	0.6800	0.6900
P@20	0.6300	0.6250	0.6278	0.6200	0.6400
P@30	0.5300	0.5200	0.5148	0.5267	0.5300
MAP	0.1806	0.1733	0.1803	0.1715	0.1808

4.2 Example Query

4.1a #AND(

#WSUM(0.0 Train.url 0.0 Train.keywords 0.0 Train.title 1.0 Train.body)

#WSUM(0.0 station.url 0.0 station.keywords 0.0 station.title 1.0 station.body)

#WSUM(0.0 security.url 0.0 security.keywords 0.0 security.title 1.0 security.body)

#WSUM(0.0 measures.url 0.0 measures.keywords 0.0 measures.title 1.0 measures.body))

4.1b #AND(

#WSUM(0.1 Train.url 0.1 Train.keywords 0.1 Train.title 0.7 Train.body)

#WSUM(0.1 station.url 0.1 station.keywords 0.1 station.title 0.7 station.body)

#WSUM(0.1 security.url 0.1 security.keywords 0.1 security.title 0.7 security.body)

#WSUM(0.1 measures.url 0.1 measures.keywords 0.1 measures.title 0.7 measures.body))

4.1c #AND(

#WSUM(0.0 Train.url 0.1 Train.keywords 0.1 Train.title 0.7 Train.body)

#WSUM(0.0 station.url 0.1 station.keywords 0.1 station.title 0.7 station.body)

#WSUM(0.0 security.url 0.1 security.keywords 0.1 security.title 0.7 security.body)

#WSUM(0.0 measures.url 0.1 measures.keywords 0.1 measures.title 0.7 measures.body))

4.1d #AND(

#WSUM(0.1 Train.url 0.0 Train.keywords 0.1 Train.title 0.7 Train.body)

#WSUM(0.1 station.url 0.0 station.keywords 0.1 station.title 0.7 station.body)

#WSUM(0.1 security.url 0.0 security.keywords 0.1 security.title 0.7 security.body)

#WSUM(0.1 measures.url 0.0 measures.keywords 0.1 measures.title 0.7 measures.body))

4.1e #AND(

#WSUM(0.1 Train.url 0.1 Train.keywords 0.0 Train.title 0.7 Train.body)

#WSUM(0.1 station.url 0.1 station.keywords 0.0 station.title 0.7 station.body)

#WSUM(0.1 security.url 0.1 security.keywords 0.0 security.title 0.7 security.body)

#WSUM(0.1 measures.url 0.1 measures.keywords 0.0 measures.title 0.7 measures.body))

4.3 Weights

My prior assumption on the weight is that body should be the most important field to match query and every other filed should be useful in the searching. Therefore, I assign the highest weight for body and assign equal small weight for url, keywords and title fields.

In order to test the effect of each field in url, keywords and title, I remove one of them each time while keeping other the same. The weights for the remaining fields are the same as before to ensure the relative ratios of them are unchanged. And I can compare the result from c, d, e to b.

4.4 Discussion

Comparing a with b

My assumption is that url, keywords and title are all useful for matching the query. However, the result shows that using all other three field equally might undermined the performance.

Comparing b with c, d, e

Adding url and keywords field helps with the performance. However, adding title filed undermined the performance. Moreover, while adding only url and keywords field, the performance is better than only using body fields, which gives the conclusion that url, keywords and title fields are useful while matching query, but their weight should be carefully selected.

Analysis of metric P@n

The MAP of Exp-e is higher than Exp-a, however, the P@10 is lower, which means that adding additional fields helps filter out irrelevant documents in a larger ranking scale while it might worsen the ranking of those most related documents.

This makes sense because adding more field to the query means the search engineer will look into more text of a document, so that the relevant information hidden in other fields could be retrieved too. The precision will increase. However, because the most important filed for ranking is the body field. Adding scores from other fields could influence the ranking only based on body.

5 Experiment 4: Sequential dependency models

5.1 Experimental Results

	Indri BOW (body) (Exp-5.1a)	**0.80 AND 0.10 NEAR 0.10 WINDOW (Exp-5.1b)**	**0.80 AND 0.10 NEAR 0.00 WINDOW (Exp-5.1c)**	**0.80 AND 0.00 NEAR 0.10 WINDOW (Exp-5.1d)**	**0.80 AND 0.00 NEAR 0.20 WINDOW (Exp-5.1e)**
P@10	0.7000	0.7000	0.6700	0.7600	0.7700
P@20	0.6300	0.6750	0.6750	0.6800	0.6600
P@30	0.5300	0.5733	0.5633	0.5700	0.5600
MAP	0.1806	0.1996	0.1974	0.2028	0.2054

5.2 Example Query

5.1a #and(Scottish Highland Games)

5.1b #wand(0.8 #and(Scottish Highland Games)

0.1 #and(#near/1(Scottish Highland Games) #near/1(Scottish Highland Games))

0.1 #and(#window/8(Scottish Highland Games) #window/8(Scottish Highland Games)))

5.1c #wand(0.8 #and(Scottish Highland Games)

0.1 #and(#near/1(Scottish Highland Games) #near/1(Scottish Highland Games))

0.0 #and(#window/8(Scottish Highland Games) #window/8(Scottish Highland Games)))

5.1d #wand(0.8 #and(Scottish Highland Games)

0.0 #and(#near/1(Scottish Highland Games) #near/1(Scottish Highland Games))

0.1 #and(#window/8(Scottish Highland Games) #window/8(Scottish Highland Games)))

5.1e #wand(0.8 #and(Scottish Highland Games)

0.0 #and(#near/1(Scottish Highland Games) #near/1(Scottish Highland Games))

0.2 #and(#window/8(Scottish Highland Games) #window/8(Scottish Highland Games)))

5.3 Weights

I assume the and operator should have the highest weight among the three operators. And I want to test whether NEAR or WINDOW is more important for the search. Therefore, I set up a baseline experiment where AND has the weight of 0.8, NEAR and WINDOW each has the weight of 0.1. In order to keep the relative ratio between different experiments. I made the 5.1c and 5.1d with weights of [0.8, 0.1, 0.0] and [0.8, 0.0, 0.1] instead of weights like [0.9, 1.0, 0.0]. After above experiments, I found out using WINDOW and AND operators without NEAR had a better result. Therefore, I try 5.1e with weights of [0.8, 0.0, 0.2] to test a different ratio of weight between WINDOW and AND.

5.4 Discussion

1. Comparing 5.1 and 5.2, we know that using sequential dependency is better. Because when user input the query, they are usually entering with a phrases or sentences.
2. Before the experiment, I though the NEAR operator should be weighted higher than WINDOW. Because NEAR has constrain on the order of its argument while WINDOW doesn't. It is harder to have a match for NEAR. Therefore, once NEAR matches, we should assign higher weight on it. However, 5.1b and 5.1c shows that WINDOW should be assigned higher weight rather than NEAR. Moreover, using WINDOW only instead of WINDOW and NEAR together give us a better result.