Homework 2

1 Introduction

1.1 Collaboration and Originality

Your report must include answers to the following questions:

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
   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 <u>of any kind</u> 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. 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 <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
   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.

5. Are you the author of <u>every word</u> 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.

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)				
	1500 (Exp-3.1a)	**1000 (Exp-3.1b)**	**500 (Exp-3.1c)**	**2000 (Exp-3.1d)**	**1400 (Exp-3.1e)**
P@10	0.7000	0.7100	0.6700	0.6900	0.7000
P@20	0.6300	0.6667	0.6050	0.6300	0.6450
P@30	0.5300	0.6300	0.5333	0.5167	0.5333
MAP	0.1809	0.1791	0.1745	0.1801	0.1814

	λ (Note: μ=0)				
	0.4 (Exp-3.2a)	**0.3 (Exp-3.2b)**	**0.5 (Exp-3.2c)**	**0.2 (Exp-3.2d)**	**0.1 (Exp-3.2e)**
P@10	0.4500	0.4600	0.4500	0.4600	0.4700
P@20	0.4850	0.4850	0.4760	0.4800	0.4850
P@30	0.4533	0.4633	0.4367	0.4600	0.4667
MAP	0.1409	0.1454	0.1360	0.1484	0.1531

3.2 Parameters

When lambda is fixed to be 0, and test for mu, I used a gap of 500 to test. I first run baseline 1500 and then 1000, 500 since I have no idea how would the score change. After I found the MAP keep decreasing, I test 2000(base+500), and found MAP decrease as well. The last value I put 1400, since I think if there is an optimal value for lambda, it will be more likely to locate between 1500 and 1000, and closer to 1500.

When mu is fixed to be 0, and test for lambda, I used a gap of 0.1 to test, since I know lambda would have to be in between 0 and 1. I first test the baseline 0.4 and +- 0.1, since with experience above, it seems that there will be a local optimum. The results favors smaller lambda, and therefore I test 0.2. Observing an improvement, so I further decrease lambda and test 0.1.

3.3 Discussion

When lambda is fixed to be 0, and test for mu. We know that larger mu put less weight to individual document term frequency relative to the whole collection. The result confirms this idea, that as we decrease mu, top score documents tend to match better as when we decrease from 1500 to 1000, P@10, P@20, P@30 all increase. However, the overall Map score might decrease, this may due to the fact that we enforced a tighter restriction on selected documents and some weakly relevant documents are lost.

When mu is fixed to be 0, and test for lambda. We know lambda plays a similar role to mu in terms of smoothing, except that it will leverage the whole score rather than just tf and doc length. Since (1-lambda) is the associated coefficient. Larger lambda will put less weight on individual doc score, too. The results also confirm this idea, that as we decrease lambda, both the MAP and top P@ score are improved.

4 Experiment 3: Indri Representations

4.1 Experimental Results

	Indri BOW (body) (Exp-4.1a)	0.2 url 0.00 keywords 0.3 title 0.5 body (Exp-4.1b)	0.3 url 0.00 keywords 0.5 title 0.2 body (Exp-4.1c)	0.1 url 0.00 keywords 0.1 title 0.8 body (Exp-4.1d)	0.1 url 0.1 keywords 0.1 title 0.7 body (Exp-4.1e)
P@10	0.6700	0.6200	0.5900	0.6700	0.6600
P@20	0.5900	0.5450	0.4850	0.5850	0.5950
P@30	0.5067	0.4700	0.4033	0.5000	0.4967
MAP	0.1607	0.1489	0.1331	0.1570	0.1591

4.2 Example Query

4.1b #AND(#WSUM(0.2 Train.url 0.3 Train.title 0.5 Train.body)#WSUM(0.2 station.url 0.3 station.title 0.5 station.body)#WSUM(0.2 security.url 0.3 security.title 0.5 security.body)#WSUM(0.2 measures.url 0.3 measures.title 0.5 measures.body))

4.1c #AND(#WSUM(0.3 Train.url 0.5 Train.title 0.2 Train.body)#WSUM(0.3 station.url 0.5 station.title 0.2 station.body)#WSUM(0.3 security.url 0.5 security.title 0.2 security.body)#WSUM(0.3 measures.url 0.5 measures.title 0.2 measures.body))

4.1d #AND(#WSUM(0.1 Train.url 0.1 Train.title 0.8 Train.body)#WSUM(0.1 station.url 0.1 station.title 0.8 station.body)#WSUM(0.1 security.url 0.1 security.title 0.8 security.body)#WSUM(0.1 measures.url 0.1 measures.title 0.8 measures.body))

4.1e #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))

4.3 Weights

All Indri models use mu 2500 and lambda 0.4, output length 100

The baseline is giving body weight 1. And for 4.1b and 4.1c, I tried to put more weights on the title and url in attempt to reward matching in url and title fields, since intuitively, matches in those two fields normally signifies high relevance between the content and our query. The reason that I did not use keyword field is because when I do hw1, keywords seems to not work at all.

It turns out that both b and c perform worse than only use body field. Then I decide to change my strategy. In stead of giving huge reward for matches in other fields, I put a small weight on fields other than body, which means we are actually score based on body, and those without matching titles and urls get punished a little comparatively. That is why I put 0.1, 0, 0.1, 0.8.

In my last attempt, I was thinking that all other fields are only given a small weight and will not influence the decision so much, maybe it’s time to try keywords field. Surprisingly, the score went up a little bit.

4.4 Discussion

Results from b and c clearly shows that it is not a good idea to focus on fields other than body. It appears that the less weight we put on body, the worse the score we get. Result from d can work as a proof of our previous observation, do not distribute too much weight on other fields.

The result from e is interesting. By giving keywords 0.1 weight and keeping the total weight the same, we decrease the relative weight of Body compare to that of Url and Title. Which means, if keyword is entirely useless, we should get a lower score, according to my previous observation, since we shift out weight from body. Therefore, it’s possible that there are docs where keywords are set appropriately to be utilized by our search engine. Although we still didn’t beat the baseline, this observation is somehow useful.

5 Experiment 4: Sequential dependency models

5.1 Experimental Results

	Indri BOW (body) (Exp-5.1a)	0.7 AND 0.2 NEAR 0.1 WINDOW (Exp-5.1b)	0.55 AND 0.2 NEAR 0.25 WINDOW (Exp-5.1c)	0.65 AND 0.1 NEAR 0.25 WINDOW (Exp-5.1d)	0.9 AND 0.1 NEAR 0.5 WINDOW (Exp-5.1e)
P@10	0.6700	0.6500	0.6600	0.6900	0.7400
P@20	0.5900	0.6300	0.6300	0.6300	0.6400
P@30	0.5067	0.5433	0.5467	0.5600	0.5633
MAP	0.1607	0.1822	0.1896	0.1962	0.1978

5.2 Example Query

5.1b #WAND(0.7 #AND (Scottish Highland Games) 0.2 #AND (#near/1 (Scottish Highland) #near/1 (Highland Games)) 0.1 #AND (#window/10 (Scottish Highland) #window/10 (Highland Games)))

5.1c #WAND(0.55 #AND (Scottish Highland Games) 0.2 #AND (#near/1 (Scottish Highland) #near/1 (Highland Games)) 0.25 #AND (#window/10 (Scottish Highland) #window/10 (Highland Games)))

5.1d #WAND(0.65 #AND (Scottish Highland Games) 0.1 #AND (#near/1 (Scottish Highland) #near/1 (Highland Games)) 0.25 #AND (#window/10 (Scottish Highland) #window/10 (Highland Games)))

5.1 e #WAND(0.9 #AND (Scottish Highland Games) 0.1 #AND (#near/1 (Scottish Highland) #near/1 (Highland Games)) 0.5 #AND (#window/10 (Scottish Highland) #window/10 (Highland Games)))

5.3 Weights

All Indri models use mu 2500 and lambda 0.4, output length 100, near distance 1, window size 10.

For 5.1b I used the weights that shown on the lecture slides to get a first glance at the performance of SDM model. I'm glad to finally see that the result is better than the baseline.

Then I think P@30 value is a bit low than the other two, so I wanted to lift it up. For 5.1c, I decide to put more weight on window. The MAP score did increase. However, I later found out that it was just an coincidence. By raising the weight on WINDOW and decrease the weight on AND, I actually tighten the restriction and will tend to include less relevant docs for P@ larger numbers.

In order to loosen the constrain, I reduce the weight for NEAR for 5.1d from 0.2 to 0.1. NEAR is obviously the most informative constrain and the most strict as well.

For 5.1e, I still want to loosen my constrain, but further reduce weight for NEAR may eliminate its function in SDM, so I increase both AND and WINDOW weights by 0.25 such that the weight of NEAR is slightly alleviated.

5.4 Discussion

AND is to ensure that we include all docs that have some sort of relevancy to our query. Guarantee the recall of our search to a certain degree.

NEAR is one of the most insightful part of the SDM model, that if we found a match in NEAR, it probably indicates relevancy of that doc, so we should definitely make use of it. However, we should not put too much weight on it which might tighten the restrictions and give us high precision on top results with low precision in a larger scale, resulting in low MAP score.

WINDOW is another informative part, but it’s softer restriction than NEAR, therefore we could put a little more weight on it. By doing so, we can balance the tradeoff of peaking most relevant docs and not losing less relevant docs.

The performance may not seem that ideal. I think there are two reasons. First, we did not tune the best parameter for mu and lambda, 2500 and 0.4 may not be a good weight pair. Second, near distance 1, and window size 10 might still be too small. I believe increase the length parameter of either would give us a better MAP score, but that’s not the focus of this experiment.