Your Name Yiding Ou

Your Andrew ID yidingo

Homework 4

Collaboration and Originality

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 Yiding Ou

Your Andrew ID yidingo

Homework 4

Instruction

1 Experiment: Baselines

	BM25	Indri BOW	Indri SDM
P@10	0.1680	0.1720	0.1880
P@20	0.2040	0.1980	0.2040
P@30	0.2053	0.1893c	0.2147
MAP	0.0924	0. 1146	0.1180

For BM25, I use k1 = 1.2, b = 0.75, k3 = 0.

For indri, I used lambda = 0.2, mu = 2500.

2 Custom Features

The first feature is calculating the length of inlink field, the time complexity should be O(n). the intuition behind this that I think a page with more inlink is kind of a high priority. And it is likely a popular page.

The second feature is calculating the length of title, since title contains the most important information in its body, so longer title has higher possibility to be a relevant document. The complexity is O(n) n for the number of documents.

Both features are not expensive on complexity, and has a potential good possibility to retrieve a relevant document.

3 Experiment: Learning to Rank

	IR Fusion	Content-Based	Base	All
P@10	0.0160	0.0320	0.0160	0.0160
P@20	0.0480	0.0480	0.0140	0.0140
P@30	0.0547	0.0480	0.0253	0.0253
MAP	0.0199	0.0177	0.0080	0.0080

3.1 Parameters

The parameter for MB25 is k1 = 1.2, b = 0.75, k3 = 0.

For indri is: lambda = 0.2, mu = 2500

The SVC rank value of c is 0.001

3.2 Discussion

From MAP perspective, the Information Retrieval fusion has the best performance, this is because these two IR approaches seeking to retrieve more relevant documents, so IR fusion has the best MAP score. However the content base has the better p@n score, I think the reason behind this is that is the term overlap feature are more focus on retrieve the relevant document at top of the retrieves sets. The base and all retrieve documents will 0 score, this is probably the first four feature's reason. They are the most influential features, and some of them results in the result list are all 0 score.

4 Experiment: Features

Experiment with four different combinations of features.

	All (Baseline)	**1,2,3,4**	**BM25(5,8, 11,14)**	**Indri(6, 9,12,15)**	**My Features(17,18)**
P@10	0.0160	0.0160	0.0120	0.0040	0.0320
P@20	0.0140	0.0140	0.0360	0.0200	0.0340
P@30	0.0253	0.0253	0.0507	0.0227	0.0373
MAP	0.0080	0. 0080	0.0179	0.0106	0.0108

4.1 Parameters

The first combo is the first four feature, because of the finding in experiment1, I think the first four feature will only return documents with 0 score.

The second combo is testing the BM25 features, combine the body, title, inlink, and url fields, the corresponding features are 5, 8, 11, 14.

My third combo is testing the Indri features, to compare with the BM25, and it is also combined with body, tiltle, url and inlink fields. The corresponding features are 6, 9, 12, 15.

My last combo is test my features, to prove my hypothesis.

4.2 Discussion

The first one complexity is O(n), I choose this combination to figure out my finding from baseline experiment, I think the first four make the retrieved set document with 0 score. And from this experiment, I think I proved my finding, the reason behind this is probably because some document do not have the field that the feature is testing.

In order to find out which feature cause the problem, I also run some sub experiments, with feature 1,2 and 3,4. I found feature 1,2 can product nonzero scores, and 3,4 produced 0 scores. So, I run 2,4 and find that feature 3, which is the Wikipedia feature causes the score becomes 0. The result probably is that not many documents in this collection is wiki page.

The second experiment complexity is O(n), since it just iterate all documents once, I run this experiment to compare with the third combo. I want to see the BM25 feature and Indri feature's detail, since in the baseline experiment only test them as a whole.

As for the third combo, I found the BM25 has better performance both in p@n and MAP than Indri, however, compare to the IR fusion, which is the combination of BM25 and Indri, merely use Indri feature is worse than the IR fusion, I think this proves the combination of Indri and BM25 features has the better performance in both MAP and p@n results.

For the fourth combo, I want to figure out does my features is better than the baseline test, and the result is positive. Which proves documents with more inlink and longer title are likely relevant with the information need.

All in all, I think for this corpus we should disable the first four features, and for now the best combination is the IR fusion.

5 Analysis

1:0.81086707

2:-1.5883787

3: NA

4:-0.32302964

5:-0.30610749

6:-0.59844077

7:1.2202779

8:1.3193432

9:-0.33790779

10:-0.023688884

11:1.3377178

12:-0.5191533

13:-0.40780509

14:0.96725547

15:-0.66426867

16:-0.48575038

17:-0.067760386

18:-0.0024084777

From above we can see that feature 1, 7, 8, 11, 14 are the positive features, and 8, 11, and 14 are BM25. So I think my experiment proves that BM25 has a good performance compares to others. The corpus does not have many wiki pages so the third feature got all 0 for the output file. My features are trivial since the score is near 0. The worst performance feature is the second feature, I think we should reverse this feature to get a better performance, this also proved by my 17 feature, which is the length of inlink field, which means longer url with more inlink should be rank lower than other pages. Probably is because that this kind of pages tend to contain less information. For the body field, the term overlap feature beats two other features. This proves that when we consider measure with body field, we should use term overlap feature, and the reason probably is that in the body field, the structure of sentences and keywords are similar. For the title, url, and inlink field, BM25 has a better score. The reason should be that BM25 is a best matching IR model, fields like url, title and inlink are short, the more matches shows their relationship.