Your Name: Leonardo Neves

Your Andrew ID : lneves

Homework 5

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: Leonardo Neves

Your Andrew ID: lneves

Homework 5

1 Experiment: Baselines

Provide information about the effectiveness of your system in three baseline configurations.

	BM25	Indri BOW	Indri SDM
P@10	0.2160	0.1680	0.2080
P@20	0.2480	0.2520	0.2320
P@30	0.2573	0.2560	0.2480
MAP	0.1334	0.1343	0.1236

Document the parameter settings that were used to obtain these results

BM25:k_1=1.2
BM25:b=0.75
BM25:k_3=0.0
Indri:mu=1000
Indri:lambda=0.4.

BOW = 0.2, NEAR = 0.4, Window = 0.4

2 Custom Features

Describe each of your custom features, including what information it uses and its computational complexity. Explain the intuitions behind your choices. This does not need to be a lengthy discussion, but you need to convince us that your features are reasonable hypotheses about what improves search accuracy, and not too computationally expensive to be practical.

My feature #17 is the title length. Less informative titles might be larger and comprehend more words, meaning the query terms that have matched might be less informative. My feature #18 is the maximum tf-idf among the query terms. Since tf-idf represents how relevant that term is to the file, having a larger tf-idf will show that the query term is relevant for the given document.

3 Experiment: Learning to Rank

Use your learning-to-rank software to train four models that use different groups of features.

	IR Fusion	Content-Based	Base	All
P@10	0.2280	0.2000	0.4080	0.4000
P@20	0.2420	0.2500	0.4000	0.4120
P@30	0.2667	0.2733	0.3747	0.3707
MAP	0.1218	0.1236	0.1870	0.1895

Discuss the trends that you observe; whether the learned retrieval models behaved as you expected; how the learned retrieval models compare to the baseline methods; and any other observations that you may have.

Clearly, the combination of features improved the baseline. It is interesting that, using the query term overlap made the precision at higher ranks worse, but improved the overall map, what might mean that it favors documents that are more comprehensive of all terms instead of ones that match one term very well. Another important observation is that the non-query related scores, like spam and pagerank, are of great importance to the rankings, improving considerably the previous scores on all rank levels.

Also, discuss the effectiveness of your custom features. This should be a separate discussion, and it should be more insightful than "They improved P@10 by 5%". Discuss the effect on your retrieval experiments, and if there is variation in the metrics that are affected (e.g., P@k, MAP), how those variations compared to your expectations.

My features showed good overall improvement, while making the precision at higher ranks to be smaller. I was not expecting that, since max tf-idf should be good for finding documents that are great matches to every of the terms, making it better at higher ranks. On the other hand, the title length feature seemed to be relevant. I have first tested it using the body length and it made the result worst.

4 Experiment: Features

Experiment with four different combinations of features.

	All (Baseline)	$Comb_1$ 1-4,17-18	$Comb_2$ All - 7,10,13,16	$Comb_3$ All - 17	$Comb_4$ All - 18
P@10	0.4000	0.4360	0.4360	0.4040	0.4000
P@20	0.4120	0.3960	0.4120	0.4100	0.4000
P@30	0.3707	0.3813	0.3893	0.3760	0.3693
MAP	0.1895	0.1861	0.1931	0.1892	0.1873

Describe each of your feature combinations, including its computational complexity. Explain the intuitions behind your choices. This does not need to be a lengthy discussion, but you need to convince

us that your combinations are investigating interesting hypotheses about what delivers good search accuracy. Were you able to get good effectiveness from a smaller set of features, or is the best result obtained by using all of the features? Why?

For my first combination, I wanted to investigate how the IR features would improve the results. Since they are the most expensive in query time and, thus, the most related to how long my search engine is going to take to produce results, I wanted to see if they would make such a difference that would require me to use them as features, or if I could just ignore them and my result would be "good enough". The truth is that they actually hurt the top rank precision, but in fact make the overall performance better and might be worth the cost, if we are looking for improvements on a MAP-like metric. My second combination aimed on getting the top overall results. Since I knew, from the previous question, that the query overlap features were not good. In fact, they proved to be bad features with this experiments and, indeed, I got the best results by ignoring them. Using these features should probably have been an attempt to normalize the results and give more weight to documents that match more query terms, but since some query terms on a bow model can be too general, the final result was hurt. My third and fourth combinations tried to prove that both my custom features are relevant and should be used together. From the last experiment 18, we can see that the title length is an ok feature, since it is not costly and could improve the MAP average, although hurting a bit of the precision, on the base value from question 3. Experiment 3, assessing performance of tf-idf, showed that this features is really important and, again, is not costly for the system since we are using values that were already computed for the indri computation. I imagined it would not be as effective since indri and BM25 have ways of accounting for document and term frequency, but the metric was better than expected.

5 Analysis

Examine the model files produced by SVM^{rank}. Discuss which features appear to be more useful and which features appear to be less useful. Support your observations with evidence from your experiments. Keep in mind that some of the features are highly correlated, which may affect the weights that were learned for those features.

Some of this discussion may overlap with your discussion of your experiments. However, in this section we are primarily interested in what information, if anything, you can get from the SVM^{rank} model files.

Looking at the magnitude of each feature, we can see that the spam scores and the "is from Wikipedia" features are the most relevant. In addition, the features related to the title and the tf-idf features are not as relevant, but have more influence than the other ones. I wouldn't be expecting this, but the pagerank scores are one of the lowest ones and so are the features related to the body. This might be because the pagerank and spam could have some correlation, since spam pages should not have a lot of other pages pointing to it and, thus, a smaller pagerank score, and the body information might be highly correlated to the title values or the tf-idf.