Your Name : Aditya Gautam

Your Andrew ID : agautam1@andrew.cmu.edu

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 : Aditya Gautam

Your Andrew ID : agautam1@andrew.cmu.edu

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	0.2040	0.2220
P@20	0.2480	0.2680	0.2650
P@30	0.2573	0.2653	0.2897
MAP	0.1334	0.1462	0.1562

Document the parameter settings that were used to obtain these results.

Answer 1)

Parameters used :
BM25:k_1=1.2
BM25:b=0.75
BM25:k_3=0.0

Indri:mu=2500
Indri:lambda=0.4

These parameters were the default parameters used in the previous and this homework. I have used the above mentioned parameters for the comparison**.**

On running BM25, Indri BOW and Indri SDM, we can observe that the performance of BM25 was worst when compared with MAP score. Looking at top to documents, the best relevant documents where for Indri SDM and is expected also since it takes into consideration various combination of AND, OR and WINDOW operator to increase the score of document based on the relevance sentences present in the documents containing query terms.

The MAP of Indri BOW and Indri SDM is more or less same, however Indri SDM is better than Indri BOW model based on the MAP parameters.

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.

Answer 2:

The custom features that I used are:

1) Vector Space Model score i.e. lnc.ltc (field : Body)
2) Doc length/ No. of unique terms

The Reason for choosing the first feature is that it signifies the cosine similarity between the query term and the document i.e. the overlap between the terms and the document. This is something which has not been directly covered. We have Indri and BM25 score which takes into consideration the term frequency of the query term and the document length along with the smoothing parameters but vector space could be another very relevant features which would signifies the relevance of the document given a particular query. So, I believe that this feature would be significant and has less correlation with other features.

Second feature that I chose was the ration of the doc length/no. of unique terms. The reason for choosing this feature is look a document from a language model point of view. Word distribution in a good document would follow a particular pattern with respect words chosen and repetition of words. A poorly design document would have some different distribution of words. Like, a document on basketball would have lots of term related to win, loss, basketball, team names, main player's name etc. However, a spam or a poor document would have a different distribution w.r.t to repetition of words. So, the ratio of document length to the ratio of unique words is considered as one of the feature.

Computation complexity of this features is O(1) as we already have the values we need in the Index. Likewise, we just need to get the values, do smoothing and get the feature values. The higher feature values for the VectorSpace model score is considered good. High score of 2nd custom feature is considered as bad like a spam document or something. Extremely low value for the second feature is also considered as bad.

The MAP value has increased from 0.1875 to .1963 on adding these features. P@10 increased from .4000 to .4480.

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	0.1960	0.4040	0.4040	0.4480
P@20	0.2500	0.4107	0.4107	0.4140
P@30	0.2707	0.3747	0.3747	0.3907
MAP	0.1225	0.1246	0.1875	0.1963

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.

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.

Answer 3:

IR based features are the features based on the modelling used and various relevant parameters like term frequency, document length, present of word in different parts of the documents like body, inlink, field and title. So, this is something that can be easily seen and judged. Every Machine learning model is based on the relevance judgments. In this case the relevance judgment is done by the human and only the parameters like sentences, occurrences of the words etc. are taken into account. Likewise, we see that the value of IR based parameters are consistent with the model and we see a MAP value of .1225.

With Content based model, the MAP value is a more or less same as that of IR based feature since they both go in hand i.e. spam score is likely to rank low by the manual judgment and thus, SVM would classifier would classify them with some level of confidence. Both the content based features and IR based feature alone were not able to beat the baseline. The MAP and P@ values are lower than BM25 and Indri score alone. This behavior is expected since these feature alone wouldn't provide a good Machine learning model to classify the documents and are not very informative alone.

On combining all the features, we are adding more confidence bound on the SVM by adding more informative things. Likewise, the combined efficiency would go up and it went up from MAP value of .1225(IR based only) to .1875. This improvement is expected as we have more informative features which are not correlated. Combining these two sets of no correlated features would strengthen the model and add to a better classifier. Thus the MAP values increased due to re-ranking of the documents by the classifier.

All the base feature and my two custom features improved the overall MAP and P@ values. **The increase in MAP value if by 5.2% from .1875 to .1963. The improved in P@10 value is by 11.11% from .4040 to 4480.** This improvement was because of adding two features which were not very related with the old features and provide new information to the machine learning model. The features were adding information about language modelling of the words and cosine overlap of terms with frequency. Thus, the MAP values increase because of the re-rank of the document based on the SVM scores.

4 Experiment: Features

Experiment with four different combinations of features.

	All (Baseline)	Comb_1	Comb_2	Comb_3	Comb_4
P@10	0.4040	0.3440	0.3360	0.3880	0.2480
P@20	0.4107	0.3100	0.3360	0.3560	0.3060
P@30	0.3747	0.3147	0.3187	0.3227	0.3053
MAP	0.1875	0.1470	0.1461	0.1493	0.2480

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?

Answer 4)

Following are the description of the combinations used.

Combination 1: This consists of all the **BM25 related features i.e. features#5,8,11 and 14.** This model will imply the impact of only BM25 on the search ranking relevance. Also, it can be seen that the performance is much better than the baseline of BM25. Map value has increase from .1412 to .1470 with a huge improvement in the top documents i.e. P@10 from .2240 to .3440. Likewise, we can see that a weighted combination of different field with same model is much better than just using the model itself for one particular field.

Time : 62 sec

Combination 2: This consists of all the **Indri related features i.e. features#6,9,12 and 15.** This model implies the impact of Indri feature on the search ranking relevance w.r.t to the baseline of Indri model. The performance of the weighted indri model (taking into consideration various fields) is more or less same as that of weighted combination of Indri applied to different fields. The MAP value is more or less same.

Time : 63 secs

Combination 3: This consists of all the **term overlap related features i.e. features#7,10,13 and 16**. This model implies the impact of term overlap on all the fields. I wanted to see the impact of smoothing and other variable and things added to Indri and BM25. Surprisingly, the MAP value for this combination is really high i.e. .1493 which is extremely good, taking into consideration that only term overlap are considered among the documents.

Time : 62 sec

Combination 4: All the body related features are considered in this. **Features#5,6 and 7** are related to the body of the document and were evaluated together to see **the impact of only body field** in the document. It doesn't take into account any other field in the document. The Map value of this combination is surprisingly the highest i.e. .1533 implies that the importance of the body is extremely high among all the fields.

Time : 63 sec

Time complexity of the set of features are more or less same. It is little less than the time complexity when all the features are considered.

The best results where using the combination of all the features as all the features are adding additional information for the document, thus making the classifier much better or have much better understanding of the document features which are used for ranking. Using the sets or combinations of features defines a part or particular features of document, thus MAP observed was lower and was expected, as compared to all the features combined. With all features, we are actually defining a documents distinguishing characteristics in a more refined manner. Thus, SVM would have more feature and weights vector to better define the position of the documents from the support vector boundary. Likewise, it is expected to see the higher MAP value if all the features are considered, provided that all the features are informative and identify the documents with good bound.

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.

Answer 5) SVM defines the weight of features based on the training data. It has a regularization associated with it and the weights vector are the significance of the feature. Higher the weights given to a particular feature, higher is the importance of that feature in classifying the document. So, the features which are not very information would be given less importance and would have lower weigh associated with it. Likewise, the contribution of a feature is proportional to the weights defined to it.

The most important feature is **Wikipedia(feature#3), followed by Spam score(feature#1)**. This is expected as these two are very significant in determining the importance of the document. If document is from Wikipedia, we should give it the first importance and if it is spam, the document should be neglected irrespective of the score of Indri, BM25 and other IR related scores. These two features are extremely significant in determining the relevance of the document. This is as per the intuition and as per the SVM classifier weights. This is followed by the features of Indri and BM25.

Surprisingly, the features related to BM25 body score and Indri body score should be highly important but their weights are not that significant. **The weight of feature#10 is comparatively high. This means this feature is important in making a good classifier and classifying the document or re-ranking them. This is evident from the combination#3 in previous experiment, the MAP value of the combination 3 was very high as compared to the combination of IR models Indri and BM25. This is in correlation with the features weights present below. The value of feature#10 is high and likewise the result of combination#3 in previous question is also good.**

Feature values of SVM after training:

1 1:0.42561281 2:-0.065501206 3:0.7464264 4:-0.024974447 5:0.050052676 6:-0.032252461
7:0.0014127479 8:0.22688623 9:0.065221541 10:0.200247 11:0.12436787 12:0.10417015
13:0.091182724 14:-0.016245818 15:0.0060333572 16:-0.038878668 17:0.018872045 18:-
0.3770031 #