**Xiaoyu He**

**xiaoyuh1**

**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.

Yes, I asked Mouwu Lin, how to handle the case when TermVector is empty and when overlap score is zero.

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

No.

1. 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
2. Are you the author of every word of your report (Yes or No)?

Yes

**Xiaoyu He**

**xiaoyuh1**

**Homework 4**

# Instruction

# Experiment: Baselines

|  |  |  |  |
| --- | --- | --- | --- |
|  | **BM25** | **Indri**  **BOW** | **Indri**  **SDM** |
| **P@10** | 0.3200 | 0.2000 | 0.2167 |
| **P@20** | 0.2400 | 0.1940 | 0.2500 |
| **P@30** | 0.2173 | 0.1907 | 0.2250 |
| **MAP** | 0.1246 | 0.0865 | 0.1117 |

BM25:k\_1=1.2, BM25:b=0.75, BM25:k\_3=0, Indri:mu=2500, Indri:lambda=0.4

For sequential dependency model, I gave #and, #near, #window weight 0.3, 0.35, 0.3. An example query looks like:

102:#wand( 0.3 #and( fickle creek farm ) 0.35 #and( #near/1( creek farm ) #near/1( fickle creek ) ) 0.35 #and( #window/8( creek farm ) #window/8( fickle creek ) ) )

# 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.***

Feature 17 is Ranked Boolean score for body field. It can be calculated as the total sum of query term frequency in a document. The computation effort of computing Ranked Boolean is the cheap as compute overlap rate.

Though this model looks simple, I think it as can work as a complement of BM25 and Indri model, because this feature is only term frequency, without any normalization on the document length. I choose body field because I body field is relatively more important than other field (based on observation on the SVM weight). If I can add more costumed feature, I would also try other fields.

Feature 18 is VSM Score for body field. The idea of Vector space model is compute the similarity between query and documents. VSM score can be calculated in 2 steps. First, compute the dot product of query BOW vector and document BOW vector. To do this calculation, we will need TermVector of the document, dot product can be computed as the total frequency. Second, divide the dot product by the L2 norm of query vector and document vector. Those are not computationally very expansive.

My intuition behind this is, high cosine similarity means higher scores. I choose body field because I body field is relatively more important than other field (based on observation on the SVM weight). If I can add more costumed feature, I would also try other fields.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **IR**  **Fusion** | **Content-**  **Based** | **Base** | **All** |
| **P@10** | 0.2680 | 0.2800 | 0.2800 | 0.2840 |
| **P@20** | 0.2280 | 0.2320 | 0.2340 | 0.2340 |
| **P@30** | 0.2120 | 0.2133 | 0.2187 | 0.2147 |
| **MAP** | 0.1042 | 0.1113 | 0.1117 | 0.1123 |

***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.***

The general trend is, with more feature throw into the SVM, the performance gets better. Content based model performs nearly as good as baseline. Only use BM25 and Indri model performs comparably much worse than content based method. This means that, the term overlap score is can improve the system performance over only use BM25 and Indri model.

# Experiment: Features

Experiment with four different combinations of features.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **All**  **(Baseline)** | **Comb1** | **Comb2** | **Comb3** | **Comb4** |
| **P@10** | 0.2840 | 0.2840 | 0.2840 | 0.2840 | 0.0000 |
| **P@20** | 0.2340 | 0.2260 | 0.2280 | 0.2300 | 0.0000 |
| **P@30** | 0.2147 | 0.2107 | 0.2093 | 0.2147 | 0.0000 |
| **MAP** | 0.1123 | 0.1114 | 0.1121 | 0.1211 | 0.0000 |

***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?***

Since this SVM only use a linear model. Features that have higher weights are more important than features that have lower weights. First, I observed the weights of using all features, the sorted results are as follows:

No. 1, 5:0.506434 No. 2, 7:0.482350 No. 3, 10:0.379158

No. 4, 6:0.263601 No. 5, 8:0.258186 No. 6, 9:0.172078

No. 7, 1:0.146674 No. 8, 13:0.067030 No. 9, 2:0.043895

No.10, 16:0.042706 No.11, 11:0.016621 No.12, 14:0.012642

No.13, 17:-0.018357 No.14, 12:-0.021125 No.15, 18:-0.049446

No.16, 15:-0.076054 No.17, 4:-0.135203 No.18, 3:-0.153079

5,6,7,8,9,10 are top 6 most important features. They are BM25, Indri, overlap score of body field and title field.

Feature 1 Spam score ranks 7th, has weight(0. 146674) similar to Feature 3 Wiki score (-0.145181) and Feature 4 is PageRank(-0.135203). Theses three have comparably high absolute weight compare to the rest of features excluding the top 6.

Comb1: use body field 5,6,7, title field 8,9,10, spam 1, wiki 3. Ignore 2,4,11,12,13,14,15,16,17,18

The Comb1 only selected those features that have highest weights.

Comb1: use body field 5,6,7, title field 8,9,10, spam 1, wiki 3, costumed feature 17,18 Ignore 2,4,11,12,13,14,15,16

Added costumed features on Comb1.

Comb3: use body field 5,6,7, spam 1, wiki 3, costumed feature 17,18 Ignore 2,4,8,9,10,11,12,13,14,15,16.

Only used body field, deleted title field from Comb1. Here, I try to test how will the search result behaves if I only use the body field. The MAP score is even higher than the base line.

Comb3: use body field 5,6,7, 1,2,3,4, costumed feature 17,18 Ignore 8,9,10,11,12,13,14,15,16.

Further test the effectiveness of only use body field.

# Analysis

***Examine the model files produced by SVMrank. 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 SVMrank model files.***