WSM Project 2 Report

Charlene Chiang
National Cheng-Chi University, Bachelor of Management Information System
106306057@nccu.edu.tw)

Abstract: This is a project implementing four different retrieval methods based on using Anserini. Anserini, a powerful information retrieval toolkit, provides indexing, retrieving, and evaluating. In this report, you can see the different performances between the following methods: vector space model with BM25 weighting, language model with Laplace smoothing, language model with Jelinek-Mercer smoothing, and vector space model with BM25 weighting and pseudo-relevance feedback via RM3. Moreover, the advantage and disadvantage of stemming and IDF and the comparison between different smoothing techniques.

1 Model Introduction

Using WT2g data collection to construct a index with stemming and another one without stemming. Running 50 TREC queries against the collection with four different retrieval methods as below.

1.1 Vector space model with BM25 weighting

Vector space model is a model that represents queries and documents as vectors in order to calculate the similarity between them. With BM25 weighting, the model will also takes term frequency and document length into consideration. According to the formula below, k1 represents the tuning parameter controlling document term frequency scaling while b is the tuning parameter controlling scaling by document length. In this project, I set k1 as 2 and b as 0.75.

$$score(D,Q) = \sum_{i=1}^{n} IDF(q_i) \cdot \frac{f(q_i,D) \cdot (k_1+1)}{f(q_i,D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{avgd})}$$

1.2 Language model with Laplace smoothing

Language model is a model that calculates the probability of a sequence of words also known as query appearing in the document. With Laplace smoothing, also called add-one smoothing, which simply adds one to each count to avoid zeros.

$$P_{add-1}(w|D) = \frac{count(w,D)+1}{len(D)+|V|}$$

1.3 Language model with Jelinek-Mercer smoothing

With Jelinek-Mercer smoothing, this model can also referred to as a linear interpolation model, which mixes the probability of the query in the document and in the collection. Using λ to control the efficient of the model, which is 0.2 in this project.

$$p_\lambda(w \mid d) \;=\; (1-\lambda)\, p_{ml}(w \mid d) + \lambda\, p(w \mid C)$$

1.4 Vector space model with BM25 weighting and pseudo-relevance feedback via RM3

RM3 is a kind of query expansion strategy based on language modeling, which is used to create new queries by taking the terms in the top K results return from the original query as the potential expansion terms.

2 Model Evaluation

In this section, you can see the comparison result of two indexes with different methods. Including each of their mean average precision, precision at rank 10 documents, and also showing their non-interpolated and 11-point interpolated precision by using box plot.

2.1 Mean Average Precision

The mean of the average precision scores for each query.

bm25Stemmed	qldStemmed	qljmStemmed	bm25Rm3Stemmed
0.2385	0.2967	0.2423	0.3006

bm25Unstemmed	qldUnstemmed	qljmUnstemmed	bm25Rm3Unstemmed
0.0841	0.0988	0.0796	0.0966

The table above shows that the documents with stemming always perform better than them without stemming, while Vector space model with BM25 weighting and pseudo-relevance feedback via RM3 has the highest MAP.

2.2 Precision at Rank 10 Documents

The proportion of the relevant documents in the top 10 documents.

bm25Stemmed	qldStemmed	qljmStemmed	bm25Rm3Stemmed
0.4120	0.4440	0.3460	0.4620

bm25Unstemmed	qldUnstemmed	qljmUnstemmed	bm25Rm3Unstemmed
0.1592	0.1551	0.1265	0.1571

We can know that language model with Laplace smoothing and pseudo-relevance feedback via RM3 has the top two highest precision at rank 10 documents. It means that if the user cares more about the precision among the documents that have higher rank, they should use this method.

2.3 Precision (at top k documents)

We use the average precision at rank 5, 10, 15, 20, 30, 100, 200, 500, 1000 documents to draw the box plot. The result shows that documents with stemming perform better. Also, the larger data range means having a higher precision at top documents.

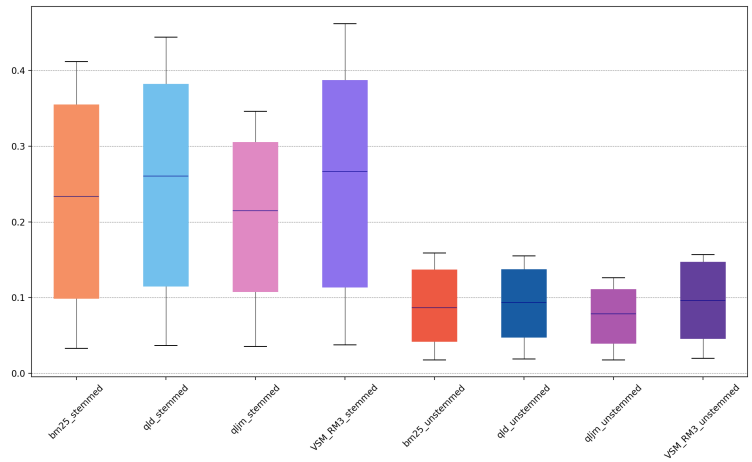

Figure 1: The precision at top k documents of the 2 indexes with different methods

2.4 Precision (11-point interpolated)

The 11-point interpolated shows the highest precision of the information retrieval system in 11 standard recall levels, which are, 0.0, 0.1, 0.2,..., 1.0. In the box plot below we can tell that it has the similar result with the box plot drawn by precision at top k documents.

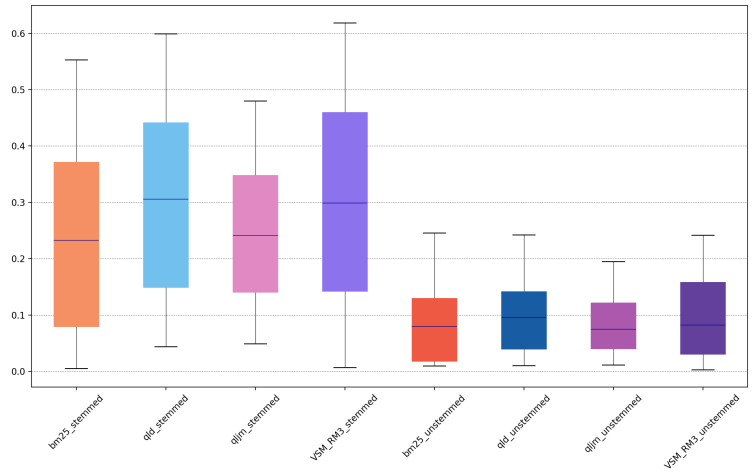

Figure 2: The 11-point interpolated precision of the 2 indexes with different methods

We can find out that language model with Laplace smoothing (qld) and vector space model with BM25 weighting and pseudo-relevance feedback via RM3 (VsmRM3) both have excellent performance, so now let's take a closer look on these two models .

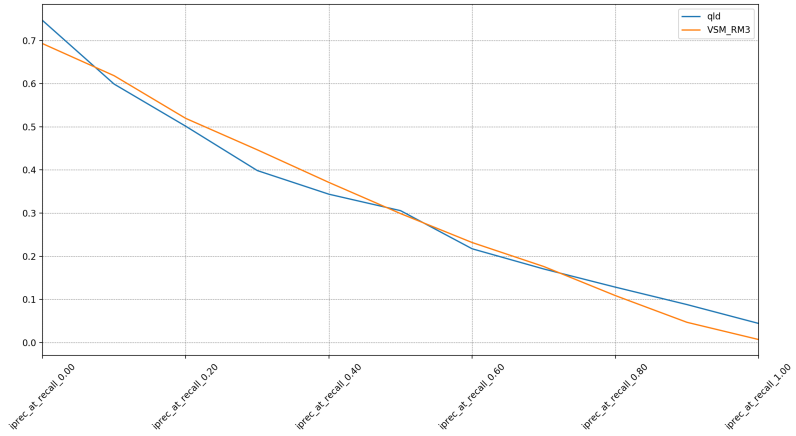

Figure 3: The performance of 11-point interpolated precision of qld and VsmRM3

According to the line chart above, we can consider both model have their own strengths. For users that only retrieve a few documents, vector space model with BM25 weighting and pseudo-relevance feedback via RM3 is more suitable for them. On the other hand, language model with Laplace smoothing will perform better when users care more about the precision on a larger amount of documents retrieved.

3 The comparison of stemming and different smoothing techniques

In the first part of the section we will discuss about the advantage and disadvantage of stemming and IDF. Then, we will compare the two different smoothing techniques, Laplace smoothing and Jelinek-Mercer smoothing.

3.1 Stemming and IDF

From section 2 we can know that index with stemming always performs better than index without stemming. Stemming is to process the word back to its word stem, so the word won't be known as different word when it is in different tenses or different part of speech. The disadvantage is that we can't be sure that the words can always be stemmed to their right root word. As for IDF, it measures how important a term is, so its advantage is that it can scale up the rare ones while weigh down the frequent terms.

3.2 Smoothing Techniques

The below line chart is the comparison of the precision at top k documents on different smoothing techniques using index with stemming, it is obvious to see from the result that qld performs better, no matter to the top documents or to the entire collection. Therefore, we can say that in this collection, it is more suitable to use Laplace smoothing instead of Jelinek-Mercer smoothing.

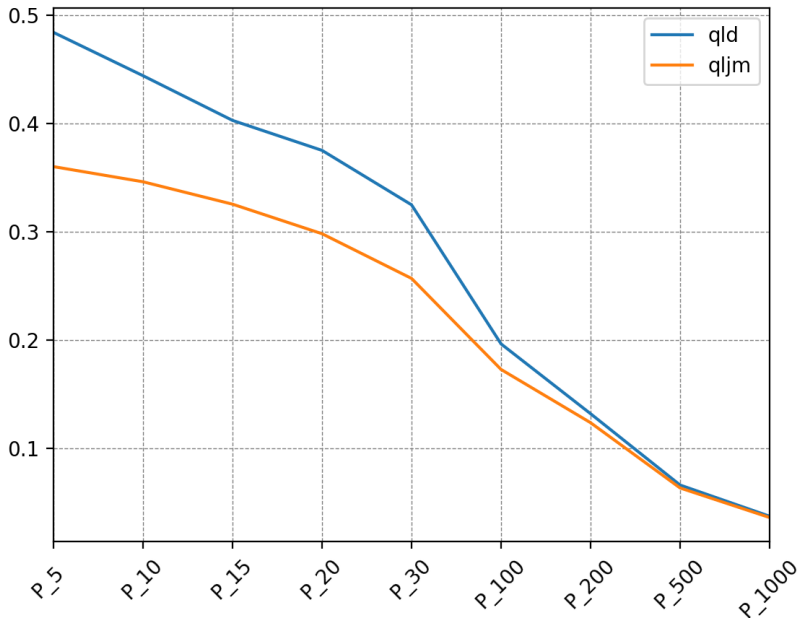

Figure 4: The precision of top k documents (with stemming) with two smoothing methods

4 Conclusion

In this project we use Anserini to perform information retrieval, by using simple shell scripts to implement the methods and the parameters we want, we can get the results efficiency. Due to the result from the above sections, there is no doubt that the outstanding performance of BM25 model. Therefore, we can say that if you really care a lot about your model precision, taking term frequency and document length into consideration is necessary .