106703014 資科三 陳家雋

1. Introduction

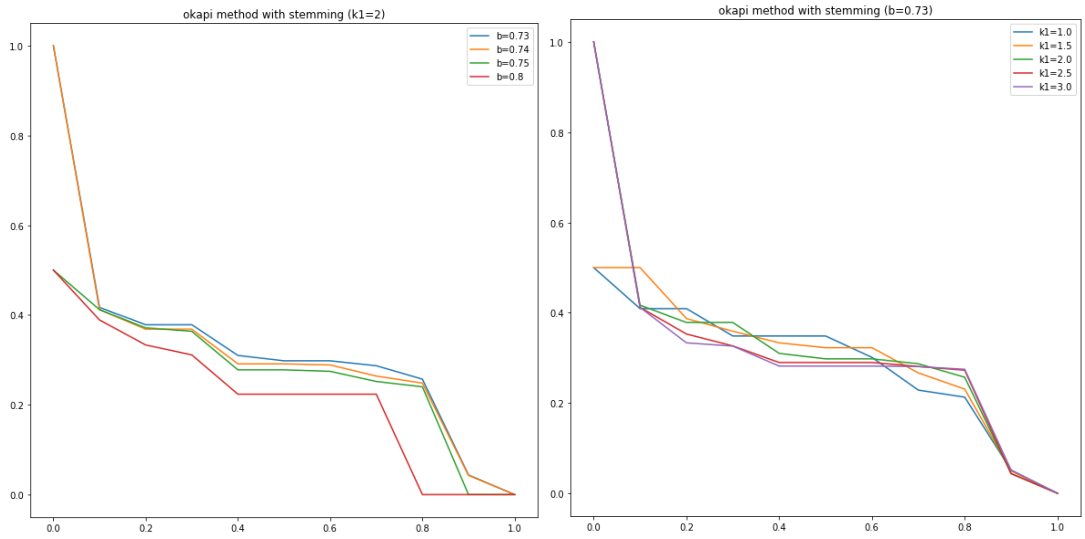
Information Retrieval is a daily work for most of the people in this era. A good search result can let the user find the correct answer. Thus, a good search engine is required. However, how to make a good search engine? A powerful retrieval method can give the user relevance answer. So, in this project, we will explore the difference of the retrieval methods, also compare the result of different retrieval methods. At first, we give a user's query, which is the word that the user wants to search. Then use different retrieval methods to assign a score to each document which was in a WT2G file. I use the Lemur toolkits as the tool to implement this experiment. In the following paragraph, I will introduce four retrieval methods.

2. Okapi BM25

Okapi BM25 is a ranking function used by search engines to estimate the relevance of documents to a given search query, it doesn't consider the proximity within the document. Following formula is okapi BM25's formula.

$$score(D,Q) = \sum_{i=0}^{n} IDF(\mathrm{qi}) * \frac{\mathrm{f(qi,D)} * (\mathrm{ki}+1)}{\mathrm{f(qi,D)} + \mathrm{k1} * (1 - \mathrm{b} + \mathrm{b} * \frac{|D|}{avgdl})}$$

$\mathrm{f(qi,D)}$ is qi's TF in the document, $|D|$ is the length of the document D, $avgdl$ is the average document length for all documents in the WT2G files. The above k1 and b both variables are free parameters, which are used to control the relevance of the result. Variable b need to fall on 0 and 1. The score of each document will fall on between 0 and 1. Whether 1 means the document is very relevance to the search query, 0 means not relevance. I respectively change the value of k1 and b to see which value is better for this query search.



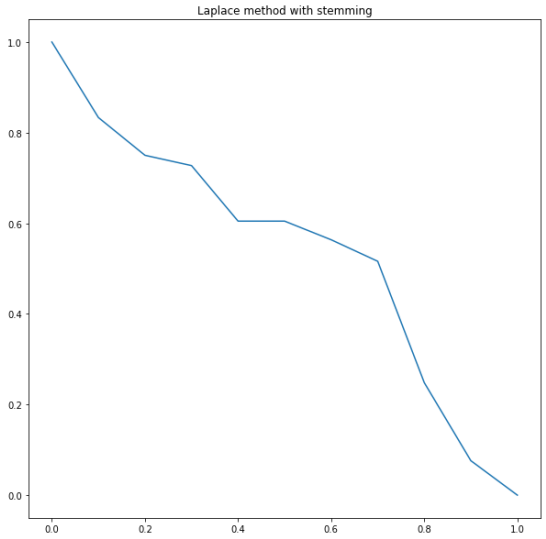
In the left graph above, I set k1 with 2, change the value of b to see the relevance result. We can see that as b growth up, the relevance will decrease more clipping. The result shows that at b=0.73, the top 10 documents is higher than b>0.73. Then, in the right graph above, I set b=0.73, change the variable k1 to see which will perform better. The graph shows that k1 will perform better for top 1 relevance document if k1>=2. But k1=1 or 1.5 will perform a little better than k>=2 for top2 to top10 relevance documents. So, I take k1=2, b=0.73 to be the best result for okapi BM25 retrieval method.

3. Laplace smoothing

Laplace smoothing is a technique used to smooth categorical data. As the name of smoothing, we can directly think that it smooth the result of relevance result. In okapi graph, we can see that top 1 relevance score is obviously higher than the others. Thus, Laplace smoothing can resolve this problem. Following is the formula:

$$\rho i = \frac{\mathrm{mi} + 1}{\mathrm{n} + k}$$

Where m = term frequency, n = number of terms in document, k = number of unique terms in corpus. The difference of Laplace smoothing and okapi is that Laplace doesn't have variable can control. You can only according to the document and the query and get the result.



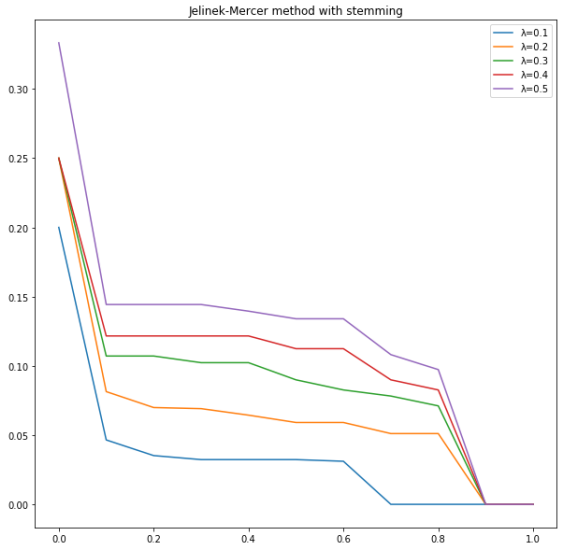
In the graph above, we can see that the top 10 relevance documents' score is decrease as the documents increase. However, it performed more linear than okapi BM25.

4. Jelinek-Mercer smoothing

Jelinek-Mercer smoothing is another method for smoothing. This method uses document probability and collection probability to smooth the score of each document. The formula for Jelinek-Mercer smoothing is:

$$\rho i = \lambda P + (1 - \lambda) Q$$

where P is the estimated probability from document, Q is the estimated probability from corpus. ρi is the score of each document about query relevance.



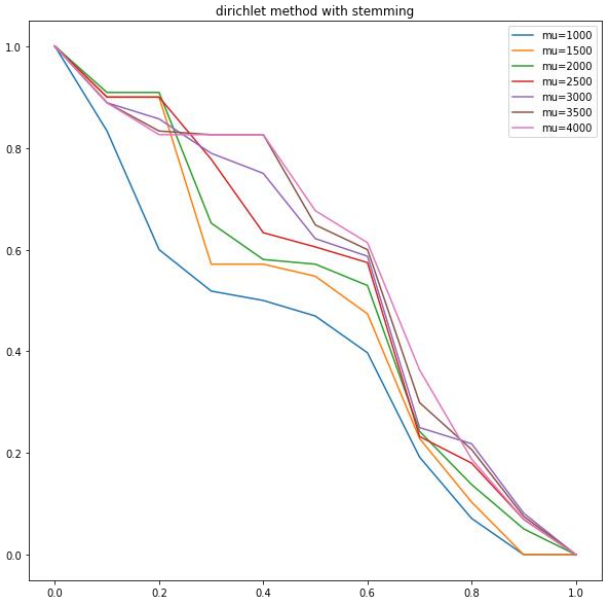
In the graph above, if the λ get higher, the result score is more concern about the document probability. And the graph shows that the higher λ the higher score the document is. So we consider λ =0.5 to be the best result.

5. Dirichlet method

Dirichlet method is another method for smoothing. The major difference between Dirichlet and Jelinek-Mercer is that Dirichlet will punish long document. Which means that the longer the document, the less score it get. The following is the Dirichlet formula:

$$P(w|d) = \frac{c(w,d) + \mu p(w|C)}{|d| + \mu}$$

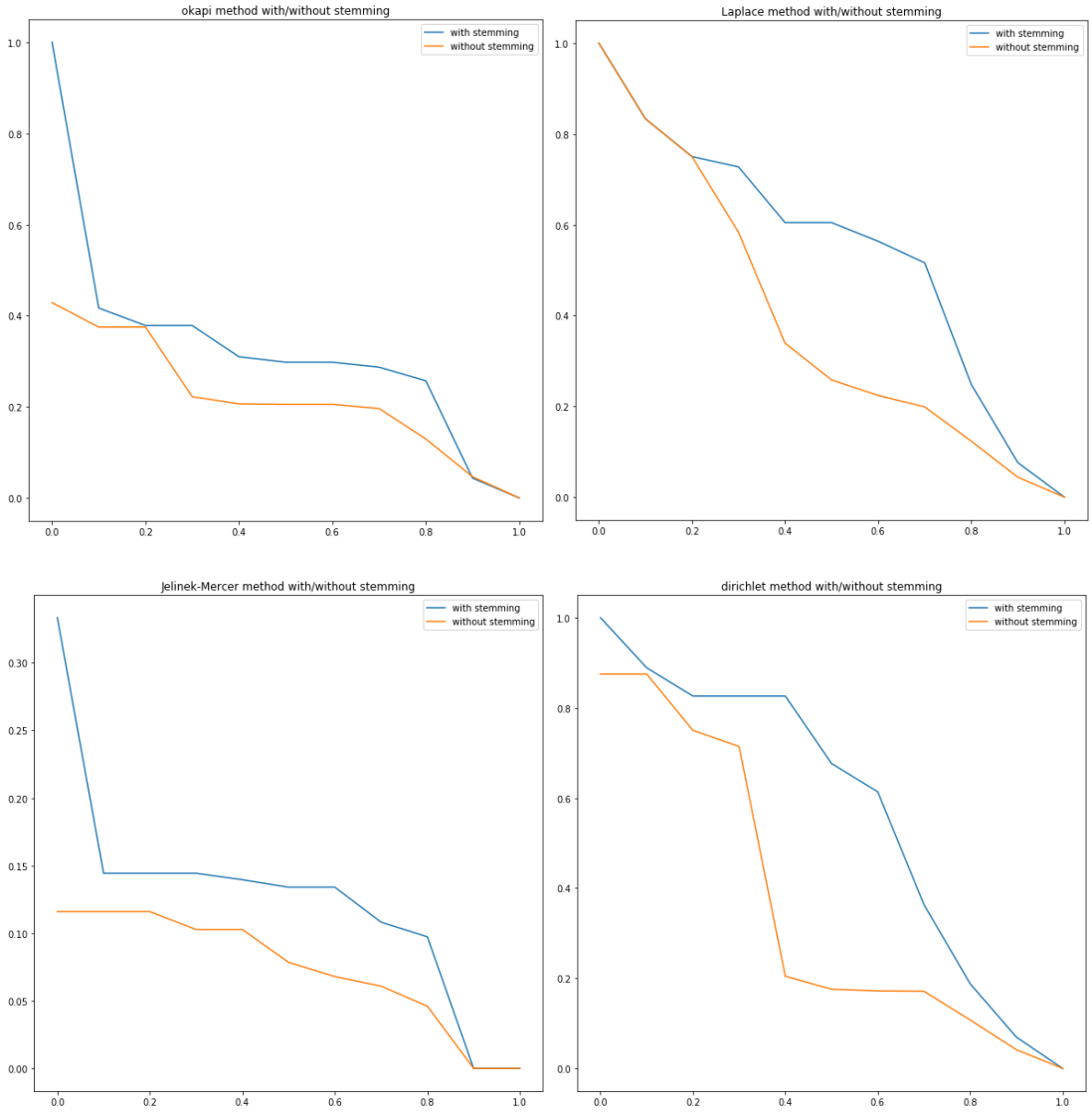
μ is a fixed value and is determined empirically. Typically m is set to maximize a retrieval metric like mean average precision for a set of queries and a collection of documents.



In the graph is the result of Dirichlet smoothing. We can see the score decrease as the document increase. For μ =1000 perform obviously worse than μ>1000. But as μ>1000, the score of each document are very similar. We take μ=4000 as the best result because the average score is the best.

6. Experiment

In the above four retrieval method, we use stemming to perform the relevance of query and documents. Stemming means removing the tail of the words. For example, playing, play, played, player. Stemming will see them as play. This step is to increase the relevance of query and documents. Following four graphs are the comparison of the above four methods for stemming and without stemming.



In the four graphs, we can clearly see that the blue lines (with stemming) are all perform higher score than the orange lines (without stemming).

7. Conclusion

As the result about four retrieval methods, we can conclude that Dirichlet smoothing method's result is better than other methods' result. But, if use different query, the best result might not be Dirichlet smoothing.