Information Retrieval HW4

System model: unigram model  
mean average precision of original query: 0.34927217

Review

In this assignment, I chose unigram as my model base, and used Lidstone smoothing and interpolation with corpus to avoid zero probability issue. Overall, the programming is not so hard if I have well understanding every part of this model. The reason I chose unigram instead of bigram is because the result of unigram model is well enough by experience. For the smoothing technique, I originally wanted to use Laplace’s law to do the smoothing. However, considering the effect of the adjustment, it would be more appropriate to lower the weight of additional count given.

From my experiment, I set smoothing value as 0.00001 and interpolation weight as 0.5. Under this setting, the result of the short query retrieval on unigram model is the best, which is 0.34927217. Nonetheless, this setting didn’t work well on the training query set, in which the query is much longer than the former one. One reason I think might be the value of Lidstone method. The original function of the smoothing only took the query term with zero probability in document model into consideration. That is, the more words in query don’t show in document model, the more additional count will be added to the frequency.

Although it seems my model is not a good one and suitable for any kind of query, yet at least it does well on short query. I hope that I can find way to raise the effectiveness of this model on long query after conducting more experiments.