Statistical Natural Language Processing
SS 2020
Prof. Dr. Dietrich Klakow

Final Project
Document Retrieval Model

Anar Amirli, 2581604, s8aramir@stud.uni-saarland.de
Ga Yeon Roß, 2568941, gayeonross1@gmail.com

Report

This report describes the comparison of baseline document retrieval model with a more advanced re-ranker method, Okapi BM25. In this study, we not only compared these two document retrieval methods, but also examined how the result of the BM25 model can be improved and how it's performance changes upon different corpus types. More specifically, we developed and evaluated a two-stage information retrieval model that given a query returns the n most relevant documents and then ranks the sentences within the documents. Initially, we implement a baseline document retriever with tf-idf features. Later, we improve over the baseline of the document retriever with a refined BM25 approach using Okapi BM25. Contrary to the TF-IDF approach which rewards term frequency and analyzes document frequency, the BM25 approach additionally accounts for term frequency saturation and document field-length normalization. Taking advantage of auxiliary parameters which ensure the aforementioned characteristics of the BM25, we later improved model performance by fine-tuning them in order to find the most relevant top-ranked sentences. Parameter search is done with a grid-search approach on the parameter space, using the precisions mean function as accuracy metrics. For evaluation of the performance of final models, we used the mean of precisions and mean reciprocal rank evaluation functions.

1. Evaluation

The following table shows the performance of the baseline and BM25 model using evaluation methods mean of precisions and mean reciprocal rank (MRR).

The results of the baseline are based on the top 50 retrieved documents. The BM25 model re-ranked the top 50 documents from the top 1000 documents retrieved by the baseline model. These top 50 documents are then splitted as sentences and ranked again, whereas each sentence is treated as one document.

When applying both evaluation methods, mean of precisions and MRR, the BM25 model on the top 50 documents performs slightly better than the

baseline but when re-ranking the sentences the evaluation results are getting worse compared to the baseline and BM25 on the top 50 documents.

	Baseline (top 50 documents)	BM25 (top 50 documents)	BM25 (top 50 sentences)
Mean of precisions	0.097	0.122	0.083
MRR	0.591	0.716	0.548

Table 1. *The comparison of results we achieved with different models upon two different evaluation methods.*

The BM25 function has two important parameters, namely k1 and b. As it described more in detail in the discussion part, we know that the parameter k1 controls how quickly an increase in term frequency results in term-frequency saturation. Lower values result in quicker saturation, and higher values in slower saturation. The parameter b controls how much effect field-length normalization should have. A value of 0.0 disables normalization completely, and a value of 1.0 normalizes fully. Based on this intuition, we decided to fine-tune these parameters. Above results (*Tab 1.*) for both BM25 models are the best results achieved by doing parameter tuning. Fine-tuning is done on the param space K1={0.05, 0.2, 0.5, 0.75, 1.5, 2.25, 2, 3, 4} and B = {0.05, 01, 0.25, 0.5, 0.75, 1} and mean precision is used as an evaluation metric. We fine-tuned parameters on both top 50 documents and top 50 sentences corpuses, then later did the ranking score calculation with best parameters, where best parameter found were k1=0.75 , b=0.05 and k1=0.5 , b=0.05 for 50 top documents and top 50 sentences corpuses respectively. More detailed overview of parameters search can be found in Fig *1.-2*.

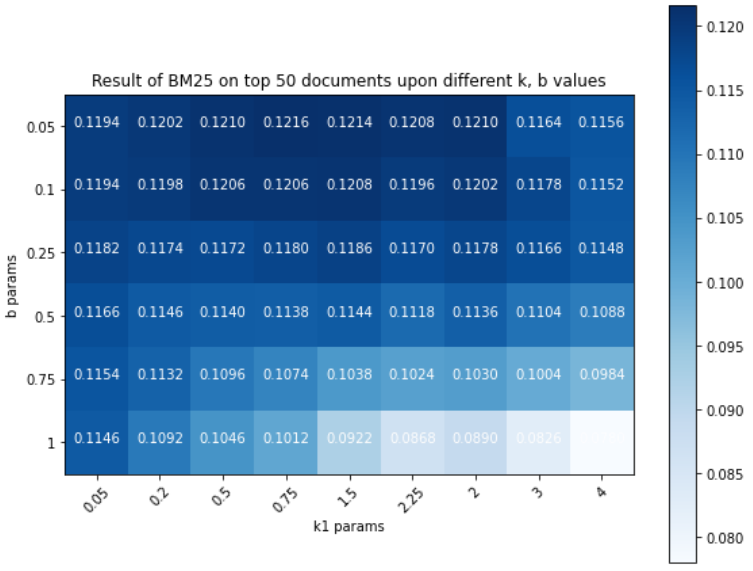

Figure 1. Show the result of parameter search on top 50 documents corpus.

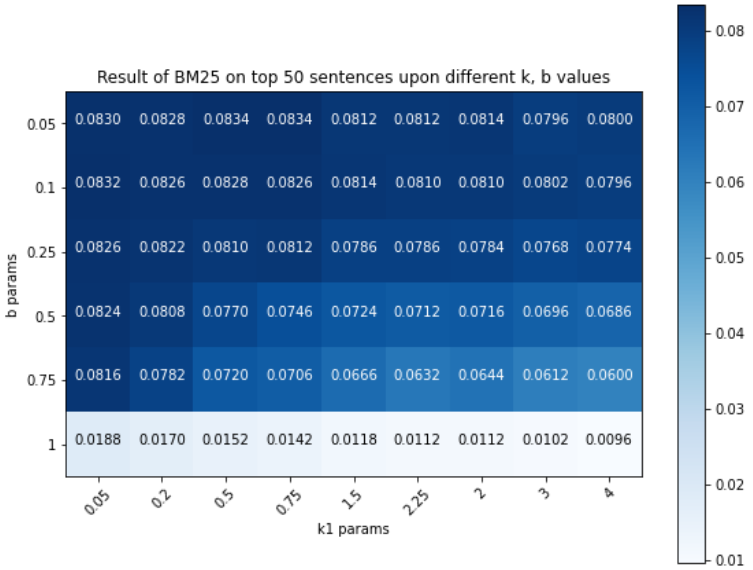

Figure 2. Show the result of parameter search on top 50 sentences corpus.

2. Discussion

2.1 Why does BM25 rank better than the baseline?

In summary, a simple TF-IDF model rewards term frequency and analyzes document frequency, as well as taking document length into consideration. BM25 score equation (1), however, is the modified variant of TF-IDF which additionally accounts for term frequency saturation and document length with different approaches.

$$score(D,\ Q)\ =\ \sum_{t \in Q} \frac{f_{t,D} \cdot (k_1 + 1)}{f_{t,D} + k_1 \cdot (1 - b + b \cdot \frac{|D|}{avgdl})} \cdot log \frac{N - n_t + 0.5}{n_t + 0.5} \quad (1)$$

The first part of the formula is the modified TF such that ft, D is the term frequency of ti in document D, where *k1* is used for frequency saturation and *b* is used to adjust saturation curve based on average document length avgdl. The second part of the formula responds as a probabilistic IDF derived from Binary Independence Model subject to Robertson-Spärck Jones weight (0.5), where n_t is the number of documents that contains term ti.

Contrary to normal TF, term saturation enables TF not to have a significant impact on the score when the document is already saturated with the occurrences of a term. In other terms, the further occurrences of the term don't increase the likelihood of relevance significantly when there are already enough mentions. In a mathematical sense it's something that behaves like a square root equation; it increases fast when the TF is small but changes more slowly when it's small. This

intuition is achieved by the TF/(TF+k) approach, which behaves like a square root equation and furthermore enables us to adjust the saturation curve. Saturation term consequently rewards complete matches over the partial ones. For example, imagine the query contains terms “w1” and “w2” and have the same IDF value. When we do information retrieval for complete query “w1 w2", it’s more desired to have a document which contains at least one instance of each term over the document that contains two or more instances of only one of the terms but not the others. In these kinds of cases, a document that contains both “w1” and “w2” terms is ranked higher than let's say the document which contains only two “w1”. It is because each term contributes to the score more when each occurs at least once than “w1" contributes when it occurs twice. The parameter b here is used to adjust satsaturation term for different sized documents, becaused number of occurrences in order to reach the saturation level varies on the document size. To summarise, *k1* controls how quickly an increase in term frequency results in term-frequency saturation and how much it contributes to the score, the lower the value is the quicker the saturation is. *b* controls how much effect field-length normalization should have, a value of 0.0 disables normalization completely, and a value of 1.0 normalizes fully.

The second part, probabilistic IDF, serves for the same purpose as it does on normal TF-IDF equation (penalizing the insignificant common terms (e.g. “and”, “or”, “the”) which overwhelmingly occur in every document and mainly are not important in the retrieval. However, it responds more to the changes in DF having steeper curves as the log equation is modified subject to Robertson-Spärck Jones weight (0.5).

We can see that BM25 is an extension to the baseline methods, where BM25 $\propto$ TF-IDF where *k1=0*, *b=0* or *k1*=∞, *b=0*. We can conclude that, although BM25 performs better than baseline methods, it still doesn’t promise significant improvements as it is only a refined version of the existing baseline model. Nevertheless, when training a corpus, a much better result can be achieved by fine-tuning the BM25 parameters.

2.2 Why are the results with the top 50 sentences worse than documents?

In table 1. we can see that the results of BM25 on the top 50 sentences are worse than those of the retrieved documents. But why does this happen? In order to understand why, we need to take a closer look at the term frequency and the BM25 equation. This very problem is caused by the fact that BM25 performs better with short queries than the very long query. This is mainly because as the query gets longer and complicated (multi-term queries), the smaller the term frequency gets. Moreover, this is also because the BM25 model is more sensitive to multi-term

queries. We know that due to term saturation modification, terms will contribute to the rank score more when all of them are found in the document than when only some of terms are found. So with larger multi-terms queries, the score equation is not only likely to have small term frequency, but also to receive small contributions by the terms as not all of them are found in the document. Therefore, rank scores become smaller, which in its turn results in poor assessment of documents. Similarly to the query being very long, if our documents are too small, then even a relatively smaller query becomes large compared to the documents. This leaves us with the same problem of long query kind mentioned above. That is why splitting documents to sentences challenges BM25 ranking and consequently leads to poorer results.

3. Conclusion

Consequently, we found that BM25 outperforms the baseline model. Test results also concluded that BM24 performs better on a corpus with the n most relevant documents than the corpus which consists of top n sentences within the relevant documents. Moreover conducting fine-tuning, we found out that the BM25 model performs better on the corpora we tested with small term-frequency saturation and field-length normalization values.