Part 4

Explanation of My Ranking Algorithm

My ranking algorithm makes the following two improvements on the original BM25 formula:
1) Add weight to TF of documentation title and TF of documentation description.
2) Add collection frequency to the original document frequency.

Part 1 Weighted TF of document title and description

Suppose we have a document A, whose number of certain term is 0 in title and all in description. For the purpose of experiment, we suppose the frequency to be [0,2,4,6,8,10,12,14]. Another document B, whose number of certain term is half in title and half in description. We suppose the frequency to be [0,1,2,3,4,5,6,7] in title and description respectively. Notice that document A and document B have same total term frequency.

Intuitively, we know that B might have a greater correlation with the query containing that term than A since

- Title is more important than description.
- If no term occurs in any of the part (title of description), the probability that the document have relationship with the query is low.

However, in the original BM25 method, they are considered to be the same in ranking score. Hence, we propose the **linear combination of tf of title and description**: *tf_new = (1 + weight) * title_tf + (1 - weight) * desc_tf*

It should be noticed that the new formula maintain the property that if title_tf=desc_tf, the score is the same as the original BM25 score, which means that we do not to re-tune the pre-optimized k1 parameter. Also, when weight=0, the tf_new have no bias towards title_tf and desc_tf, which reduces to the orignal BM25 formula.

As shown in the following figure, the new structured weighted ranking algorithm performs well on identifying document B as a more correlated document with a higher score.

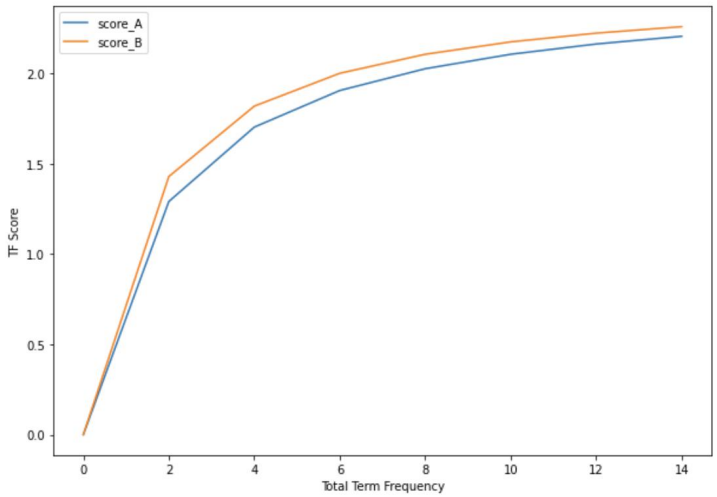

Figure 1. Biased TF Score of Document A and Document B

Another interesting thing we can get from the above formula is that *(1 + w) * x + (1 - w) * y > x + y* if and only if x > y when w > 0.

That is to say, if the title frequency is larger than the description frequency and we assume that title is more important than description (w > 0), the resulting score will be higher than baseline (w = 0). On the contrary, if title frequency is smaller than description frequency and we assume that title is more important than description (w>0), the resulting score will be lower than baseline (w = 0).

To illustrate, suppose we have a document C, whose number of certain term is [0,1,2,3,4,5,6] in title and [0,2,4,6,8,10,12] in description. For another document D, whose number of certain term is [0,2,4,6,8,10,12] in title and [0,1,2,3,4,5,6] in description. Notice that document C and document D have same total term frequency.

As shown in the figure, by introduce the weights, we successfully identify document D as a more correlated document and the baseline lies between score of document C and document D.

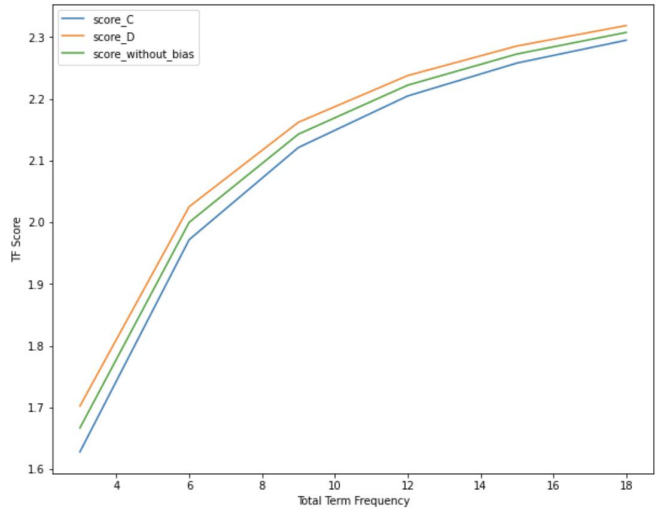

Figure 2. Biased and Unbiased TF Score of Document C and D

Part 2 Collection frequency added to document frequency.

Consider the following case:

Doc1: ‘who’: 3, “usb”: 2

Doc2: ‘who’: 10, “usb”: 1

In the original BM25 formula, it is likely that the score of Doc1 of query ‘who use usb’ will be lower than Doc2 since the document frequency for ‘usb’ and ‘who’ are the same and Doc2 has higher term frequency. However, intuitively, we know ‘usb’ is a more informative word, hence higher term frequency of ‘usb’ should result in a higher relevance score.

By taking collection frequency into consideration, although the document frequency for ‘usb’ and ‘who’ are the same, the collection frequency of ‘who’ is much higher than ‘usb’, resulting in a lower score. In this way, the new formula reveals the informativeness of ‘usb’ versus ‘who’.

Reference

[1] Stephen Robertson, Hugo Zaragoza, and Michael Taylor. 2004. Simple BM25 extension to multiple weighted fields. In Proceedings of the thirteenth ACM international conference on Information and knowledge management (CIKM '04). Association for Computing Machinery, New York, NY, USA, 42–49. DOI:https://doi.org/10.1145/1031171.1031181

[2] Jimenez, Sergio et al. ‘BM25-CTF: Improving TF and IDF Factors in BM25 by Using Collection Term Frequencies’. 1 Jan. 2018 : 2887 – 2899.