

:	:	2022-07-26
:	:	
:	:	0%
:	:	0.00
:	:	0.00

- # 1 - 2022-07-27 19:04 -
- () #6520 #6552() .
- # 2 - 2022-07-28 11:05 -
- 2022-07-28 105815.png (가) 가 .
- 2022-07-28 105834.png (가) 가 .
- # 3 - 2022-07-28 11:13 -
- clipboard-202207281106-snnkv.png (가) 가 .
- 2022-07-28 105815.png (가) 가 .
- clipboard-202207281107-codhi.png (가) 가 .
- clipboard-202207281107-gd0sl.png (가) 가 .
- clipboard-202207281109-o3oaf.png (가) 가 .
- clipboard-202207281110-sa7u3.png (가) 가 .
- clipboard-202207281111-malka.png (가) 가 .
- clipboard-202207281111-ncvoq.png (가) 가 .
- clipboard-202207281112-u1ilu.png (가) 가 .
- clipboard-202207281112-mb65i.png (가) 가 .

BM25 (Best Match 25)

- (1)
- 가 SOTA
 - BM25 IR TF-IDF
 - ElasticSearch 5.0 (default) BM25

(2) TF (Term Frequency)

$$\frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right)},$$

freq:
k1: 1.2(default)
b: 0.75(default)
dl: 가
avgdl:

- 가
- 가

(3) IDF (Inversed Document Frequency)

$$\text{IDF}(q_i) = \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5},$$

n:
N:

- 가
- 가 가 가
- 가 , 가

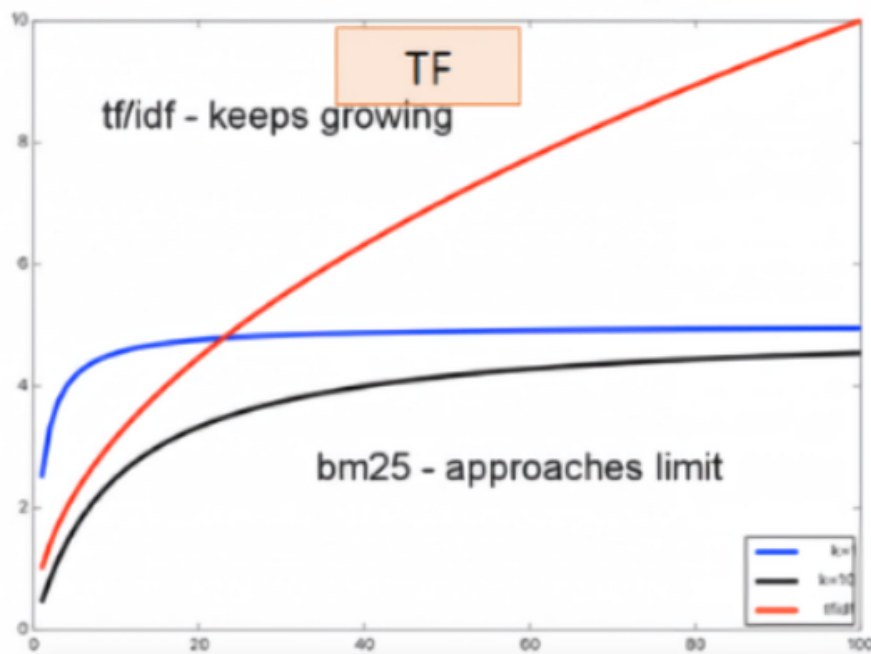
(4) Score

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

$$\text{IDF}(q_i) = \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5},$$

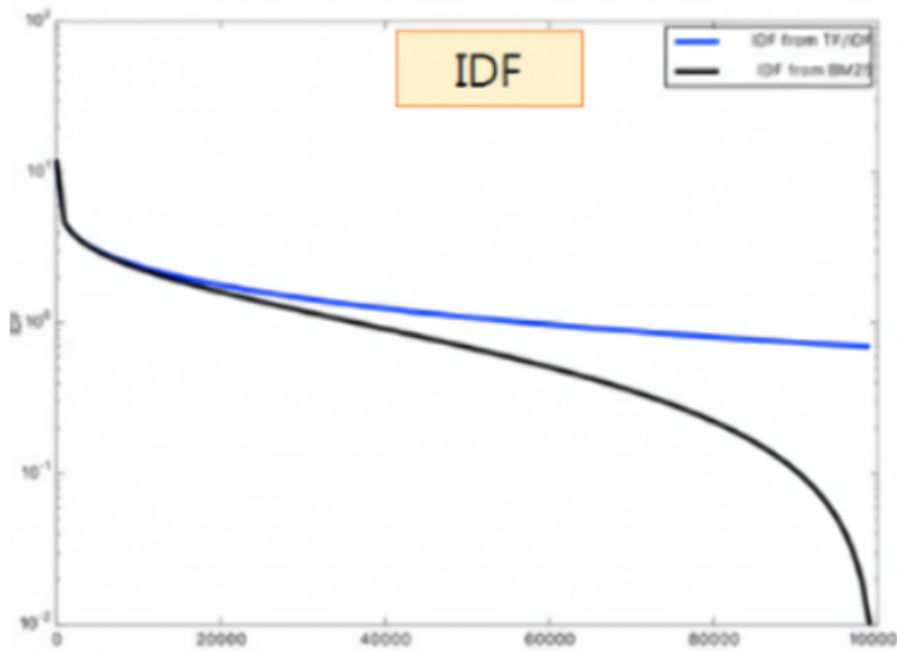
(5) TF-IDF vs BM25

1) TF



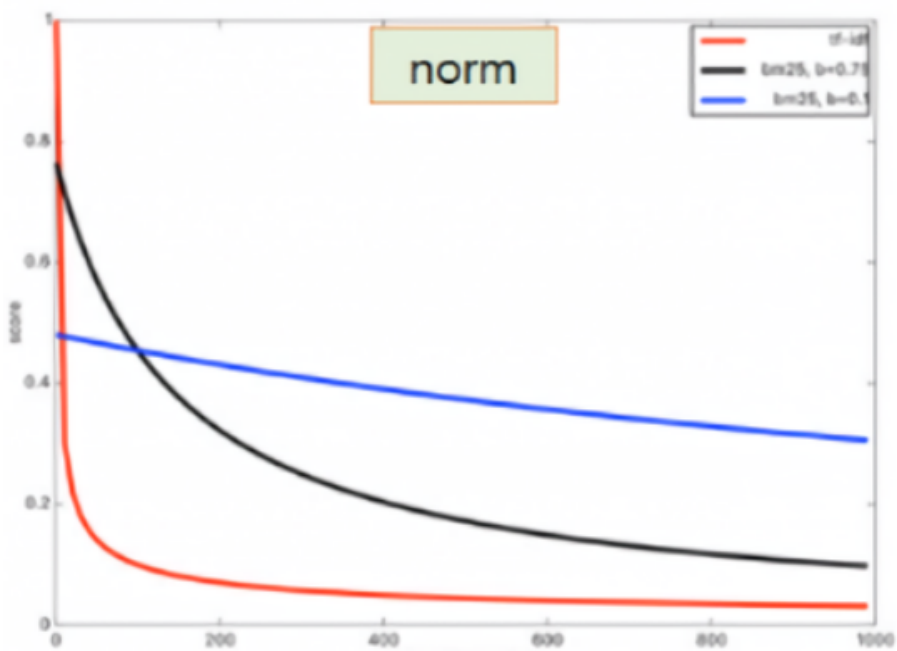
- TF 가 , BM25

2) IDF



• BM25 DF가 가 0 , (stop words)가

3) norm



• BM25 (avgdl) , 가

2022-07-28 105815.png	10.8 KB	2022-07-28
2022-07-28 105834.png	8.89 KB	2022-07-28
clipboard-202207281106-snnkv.png	10.8 KB	2022-07-28
2022-07-28 105815.png	10.8 KB	2022-07-28
clipboard-202207281107-codhi.png	10.8 KB	2022-07-28
clipboard-202207281107-gd0sl.png	8.89 KB	2022-07-28
clipboard-202207281109-o3oaf.png	8.89 KB	2022-07-28
clipboard-202207281110-sa7u3.png	27.2 KB	2022-07-28
clipboard-202207281111-malka.png	27.2 KB	2022-07-28
clipboard-202207281111-ncvoq.png	107 KB	2022-07-28

clipboard-202207281112-u1ilu.png	127 KB	2022-07-28
clipboard-202207281112-mb65i.png	123 KB	2022-07-28