# **TFIDF Using Lucene Comparison**

Task 4 :

Short query

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Evaluation metric | Your algorithm | page5image56832192Vector Space Model | BM25 | page5image56832768Language Model with Dirichlet Smoothing | page5image42187200Language Model with Jelinek Mercer Smoothing |
| P@5 | 0.4000 | 0.4000 | 0.6000 | 0.6000 | 0.4000 |
| P@10 | 0.4000 | 0.5000 | 0.5000 | 0.5000 | 0.5000 |
| P@20 | 0.3000 | 0.4000 | 0.3000 | 0.3500 | 0.2500 |
| P@100 | 0.1100 | 0.0900 | 0.1000 | 0.0900 | 0.1000 |
| Recall@5 | 0.0645 | 0.0645 | 0.0968 | 0.0968 | 0.0645 |
| Recall@10 | 0.1290 | 0.1613 | 0.1613 | 0.1613 | page5image568616960.1613 |
| Recall@20 | 0.1935 | 0.2581 | 0.1935 | 0.2258 | 0.1613 |
| Recall@100 | 0.3548 | 0.2903 | 0.3226 | 0.2903 | 0.3226 |
| MAP | 0.1291 | 0.1833 | 0.1894 | 0.1404 | 0.1462 |
| MRR | 0.3333 | 1.0000 | 1.0000 | 0.5000 | 1.0000 |
| NDCG@5 | 0.3008 | 0.5531 | 0.7227 | 0.4913 | 0.5531 |
| NDCG@10 | 0.3372 | 0.5801 | 0.6208 | 0.4666 | 0.5704 |
| NDCG@20 | 0.2920 | 0.4786 | 0.4341 | 0.3704 | 0.3681 |
| NDCG@100 | 0.3155 | 0.3804 | 0.4036 | 0.3180 | 0.3726 |

Long query

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Evaluation metric | Your algorithm | page5image56832192Vector Space Model | BM25 | page5image56832768page5image56833152Language Model with Dirichlet Smoothing | page5image42187200Language Model with Jelinek Mercer Smoothing |
| P@5 | 0.0000 | 0.6000 | 0.6000 | 0.4000 | 0.6000 |
| P@10 | 0.1000 | 0.5000 | 0.5000 | 0.3000 | 0.4000 |
| P@20 | 0.1500 | 0.3000 | 0.3000 | 0.3000 | 0.3000 |
| P@100 | 0.0500 | 0.0900 | 0.0900 | 0.1000 | 0.0900 |
| Recall@5 | 0.0000 | 0.0968 | 0.0968 | 0.0645 | 0.0968 |
| Recall@10 | 0.0323 | 0.1613 | 0.1613 | 0.0968 | 0.1290 |
| Recall@20 | 0.0968 | 0.1935 | 0.1935 | 0.1935 | 0.1935 |
| Recall@100 | 0.2258 | 0.2903 | 0.2903 | 0.3226 | 0.2903 |
| MAP | 0.0269 | 0.1486 | 0.1777 | 0.0919 | 0.1629 |
| MRR | 0.1429 | 1.0000 | 1.0000 | 0.2500 | 1.0000 |
| NDCG@5 | 0.0000 | 0.6399 | 0.7227 | 0.2773 | page5image568840320.7227 |
| NDCG@10 | 0.0734 | 0.5510 | 0.5989 | 0.2583 | 0.5424 |
| NDCG@20 | 0.1125 | 0.3903 | 0.4261 | 0.2690 | 0.4187 |
| NDCG@100 | 0.1531 | 0.3493 | 0.3792 | 0.2763 | 0.3712 |

## **Summary :**

1. My Algorithm for Short Query, can be easily compared equally with the other algorithms. The values of precision and recall are somewhat similar as others.
2. As the words in query increase the evaluation metrices drop their values i.e which happens in Long Query results for MyAlgorithm.
3. MyAlgorithm uses traditional TF\*IDF to score the documents. The performance of this TF\*IDF approach is poor because it is based on bags of words model and it doesn’t consider the position of text, its re-occurrence in documents etc. This affects the ranking and the scores calculation.
4. BM25 is the improved version of TF\*IDF which stands for ‘Best Match 25’ gives best results for both short as well as long queries which is clearly stated in above table. This model ranks a set of documents based on the query terms appearing in each document , regardless of inter-relationship between query terms within a document.[1]
5. Smoothing methods prevent zero probabilities. They also try to improve the accuracy of the model as a whole. This can improve the estimation greatly as observed above. LMD and LMJM models are also sufficient to give the exact results as smoothing technique is key step in those models.

Reference :

1) <https://en.wikipedia.org/wiki/Okapi_BM25>

Submitted by,

Suyash Poredi.

(sporedi@iu.edu)