ILS Z534 (Search) – Assignment 2

Yash Sumant Ketkar (yketkar@indiana.edu)

Findings:

Long Query:

Evaluation metric	Your algorithm	Vector Space Model	BM25	Language Model with Dirichlet Smoothing	Language Model with Jelinek Mercer Smoothing
P@5	0.0	0.2	0.6	0.0	0.2
P@10	0.0	0.3	0.3	0.2	0.3
P@20	0.15	0.25	0.3	0.25	0.25
P@100	0.04	0.1	0.1	0.1	0.09
Recall@5	0.0	0.0323	0.0968	0.0	0.0323
Recall@10	0	0.0968	0.0968	0.0645	0.0968
Recall@20	0.968	0.1613	0.1935	0.1613	0.1613
Recall@100	0.1290	0.3226	0.3226	0.3226	0.2903
MAP	0.0210	0.0966	0.1333	0.0693	0.0853
MRR	0.0833	0.5	1.0	0.125	0.2
NDCG@5	0	0.214	0.6399	0.0	0.1312
NDCG@10	0	0.2785	0.4153	0.1357	0.2369
NDCG@20	0.1060	0.2545	0.3732	0.1924	0.2259
NDCG@100	0.0982	0.2881	0.3519	0.2399	0.2478

ILS Z534 (Search) – Assignment 2

Yash Sumant Ketkar (yketkar@indiana.edu)

Short Query:

Evaluation metric	Your algorithm	Vector Space Model	BM25	Language Model with Dirichlet Smoothing	Language Model with Jelinek Mercer Smoothing
P@5	0.2	0.4	0.6	0.6	0.4
P@10	0.1	0.5	0.5	0.5	0.5
P@20	0.1	0.4	0.3	0.35	0.25
P@100	0.06	0.09	0.10	0.09	0.1
Recall@5	0.0323	0.06	0.0968	0.0968	0.0645
Recall@10	0.0323	0.16	0.1613	0.1613	0.1613
Recall@20	0	0.25	0.1935	0.2258	0.1613
Recall@100	0.1935	0.29	0.3226	0.2903	0.3226
MAP	0.0631	0.1833	0.1894	0.1404	0.1462
MRR	1.0	1.00	1.0	0.5	1.0
NDCG@5	0.3392	0.5531	0.7227	0.4913	0.5531
NDCG@10	0.2201	0.5801	0.6208	0.4666	0.5704
NDCG@20	0.1804	0.4786	0.4341	0.3704	0.3681
NDCG@100	0.2112	0.3804	0.4036	0.3180	0.3726

Summary of Findings:

In our algorithm we are not using the inbuilt function like topDocs and have used Classic Similarity in implementation. Default scoring implementation which encodes norm values as a single byte before being stored. Compression of the norm values to a byte saves memory at search time, because once a field is referenced at search time, its norms - for all documents - are maintained in memory.The rationale supporting such lossy compression of norm values is that given the difficulty and inaccuracy of users to express their true information need by a query, only big differences matter. This is clearly reflected in the table above.

The default BM25 algorithm values are used here with values of k1 = 1.2 and b = 0.75. The BM25 algorithm takes into account scaling by term frequency and also by document length as observed in the table above.

ILS Z534 (Search) – Assignment 2

Yash Sumant Ketkar (yketkar@indiana.edu)

Smoothing has techniques for adjusting the maximum likelihood estimate of probabilities to produce more accurate probabilities. 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.