Assignment – 2

Search ILS Z – 534

By, Supreeth Keragodu Suryaprakash

Task 1:

Kindly run easySearch.java. The program first calculates the TF-IDF for each term in the query and then run it for the whole term as given in the assignment.

Task 2:

Kindly run searchTRECtopics.java. The Program will output two files, one for short Query and one for long Query. The output will be in .trec format for easy evaluation.

Task3:

Please run compareAlgorithms.java. The Program will compare each and every algorithm given in the assignment. The outputs can be seen in the txt files that are produced.

Task4:

Long Query:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *Evaluation*  *Metric* | *Your*  *Algorithm* | *Vector*  *Space model* | *Bm25* | *Language*  *Model with*  *Dirichlet*  *Smoothing* | *Language*  *Model with*  *Jelinek*  *Mercer*  *Smoothing* |
| P@5 | 0.05 | 0.24 | 0.25 | 0.24 | 0.20 |
| P@10 | 0.04 | 0.21 | 0.22 | 0.22 | 0.19 |
| P@20 | 0.15 | 0.189 | 0.20 | 0.20 | 0.18 |
| P@100 | 0.24 | 0.127 | 0.14 | 0.14 | 0.12 |
| Recall@5 | 0.001 | 0.02 | 0.02 | 0.04 | 0.03 |
| Recall@10 | 0.003 | 0.043 | 0.05 | 0.05 | 0.02 |
| Recall@20 | 0.09 | 0.076 | 0.1 | 0.10 | 0.12 |
| Recall@100 | 0.12 | 0.12 | 0.3 | 0.3 | 0.3 |
| MAP | 0.31 | 0.35 | 0.14 | 0.13 | 0.1 |
| MRR | 0.08 | 0.24 | 1.36 | 0.26 | 0.22 |
| NDCG@5 | 0.00 | 0.22 | 0.25 | 0.2 | 0.17 |
| NDCG@10 | 0.00 | 0.21 | 0.23 | 0.22 | 0.30 |
| NDCG@20 | 0.10 | 0.22 | 0.22 | 0.23 | 0.23 |
| NDCG@100 | 0.09 | 0.29 | 0.24 | 0.29 | 0.27 |

Short Query:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *Evaluation*  *Metric* | *Your*  *Algorithm* | *Vector*  *Space model* | *Bm25* | *Language*  *Model with*  *Dirichlet*  *Smoothing* | *Language*  *Model with*  *Jelinek*  *Mercer*  *Smoothing* |
| P@5 | 0.19 | 0.26 | 0.26 | 0.00 | 0.20 |
| P@10 | 0.15 | 0.264 | 0.3 | 0.29 | 0.40 |
| P@20 | 0.14 | 0.235 | 0.25 | 0.30 | 0.25 |
| P@100 | 0.06 | 0.151 | 0.15 | 0.11 | 0.10 |
| Recall@5 | 0.03 | 0.047 | 0.04 | 0.00 | 0.03 |
| Recall@10 | 0.03 | 0.109 | 0.09 | 0.08 | 0.12 |
| Recall@20 | 0.06 | 0.132 | 0.13 | 0.17 | 0.16 |
| Recall@100 | 0.20 | 0.313 | 0.32 | 0.34 | 0.32 |
| MAP | 0.07 | 0.175 | 0.14 | 0.08 | 0.09 |
| MRR | 0.99 | 0.417 | 0.85 | 0.12 | 0.20 |
| NDCG@5 | 0.33 | 0.269 | 0.63 | 0.00 | 0.13 |
| NDCG@10 | 0.22 | 0.2759 | 0.41 | 0.20 | 0.30 |
| NDCG@20 | 0.17 | 0.271 | 0.37 | 0.23 | 0.23 |
| NDCG@100 | 0.21 | 0.2879 | 0.35 | 0.27 | 0.27 |

BM25. In general, it is known to perform just as good or even better than TF-IDF, especially on collections with short documents. Vector Space Models Rank documents with respect to the query by score Returning the top K documents. As the sample space grows, Vector space model won’t perform better than TF-IDF. Both the Smoothing algorithms perform way better than TF-IDF.