Vidhixa Joshi
vidjoshi

Assignment 2- Report

Short Query

Evaluation metric	My algorithm	Vector space	BM25 Model	Dirichlet Model	JM Smoothing Model
map	0.1234	0.1976	0.2005	0.2059	0.1932
Rprec	0.1514	0.2197	0.2124	0.2203	0.2051
bpref	0.2546	0.2915	0.295	0.3003	0.2913
recip_rank	0.3818	0.4696	0.4772	0.4785	0.4637
P_5	0.256	0.296	0.308	0.348	0.288
P_10	0.21	0.302	0.3	0.326	0.28
P_15	0.2067	0.2773	0.2787	0.3053	0.2533
P_20	0.201	0.26	0.271	0.289	0.243
P_30	0.1913	0.2493	0.2487	0.266	0.238
P_100	0.128	0.1648	0.1678	0.169	0.1602
P_200	0.0919	0.1144	0.1164	0.1158	0.1078
P_500	0.055	0.0626	0.0642	0.064	0.0621
P_1000	0.0334	0.0367	0.0371	0.0381	0.0363

Long Query

Evaluation metric	My algorithm	Vector space	BM25 Model	Dirichlet Model	JM Smoothing Model
map	0.0696	0.1538	0.1676	0.1588	0.1518
Rprec	0.1081	0.1726	0.1898	0.1897	0.1813
bpref	0.2262	0.2712	0.2853	0.2791	0.2746
recip_rank	0.3055	0.449	0.4497	0.3483	0.3666
P_5	0.156	0.264	0.284	0.256	0.236
P_10	0.148	0.248	0.246	0.242	0.214
P_15	0.1267	0.228	0.2413	0.244	0.2173
P_20	0.119	0.221	0.234	0.234	0.213
P_30	0.1213	0.204	0.2233	0.2173	0.206
P_100	0.0922	0.141	0.1488	0.146	0.1382
P_200	0.0662	0.0962	0.1032	0.1022	0.0933
P_500	0.041	0.0534	0.056	0.0557	0.0526
P_1000	0.0268	0.0315	0.0331	0.033	0.0316

Vidhixa Joshi
vidjoshi

Summary:

Here are some of my observations based on above two table:

1. Ranking based on TF-IDF (My Algorithm) for short queries is comparable to other algorithms but gives low precision in case of larger queries. The reason being, it is a very trivial way of ranking without sophistications like performing smoothing or considering relation between terms.

2. The best amongst others is Dirichlet's model for shorter queries and BM25 for longer queries.
    - Dirichlet's search model uses the likelihood of terms in the as in Bayesian network. With smaller set of words, the likelihood probabilities have more meaning.
      Example from the topics file: Alternative/renewable energy. You expect such terms to appear together.
    - BM25 works on bag of words model without considering relation between other words. It basically ranks documents based on most matches it finds which is a good way to go for larger documents. Provided we have good analyzer to tokenize and prevent stopwords.

3. I believe, MAP alone is a good performance evaluation. As we can see from the results, MAP can represent the performance of a particular algorithm. The reason being, it estimates a value that takes into account the performance of system at all levels of recall and for all queries.

Note:

1. Please put the codes under one package because I have reused code.
2. Running searchTrec program takes longer than other two programs. But, it gives results for sure.
3. I read up online for TF-IDF formula and saw some implementations using log and some using log10. I decided on using log to base e as the ranks weren't making a lot of difference with either one.