CS473HW4-Gelei Chen

2. Evaluation

Function	Mean Average Precision	NDCG	P@3
TF	0.024	0.056	0.080
TFIDF	0.107	0.220	0.267
LogTFIDF	0.192	0.333	0.467
Cosine	0.082	0.161	0.280
BM25	0.237	0.379	0.453
JM	0.174	0.290	0.360

3. Discuss

- TF is a very bad choice based on the score above, because TF has the lowest score among all score functions. Using only TF, we don't know if a word in the query is common or not, which results higher rank for some common words. TF doesn't have any advantages except a basic score function.
- TFIDF has much better MAP,NDCG and P@3 compared with TF, but is worse than both BM25's metrics and JM's metrics. TFIDF is better than TF, because TFIDF considers the frequency of a word in the whole corpus. TFIDF is also very good for stop word removal. The weakness of TFIDF is that its MAP score is low, which means that some documents ranked by TFIDF are not relevant. In addition, TFIDF doesn't consider the position of word in the document, and context.
- LogTFIDF has the biggest P@3 score among all score functions, so we can expect very good result on top 3 documents if we

use LogTFIDF, because LogTFIDF takes log of TF to represent TF in a linear model. Its MAP and NDCG are also high, which also suggest that it will provide good results. Like TFIDF, LogTFIDF doesn't consider the position of word in the document, and context, so the highest-ranked documents returned by LogTFIDF may not make sense with the topic of the document.

- Cosine provides reasonable top 3 documents, because of its P@3 score, but other two scores are low. In Geometry, cosine similarity only cares about the angel not the actual magnitude. Same idea here, in our cosine equation we normalized it by division. Failing to consider the actual value of freq(w,q)*freq(w,d) may explain the reason why other two scores are low.

- BM25 has the highest score of both MAP and NDCG, and good result on P@3, so it provides a very good result for a given query. BM25 not only provides a score for term, but also captures the relationship between documents and query. Also, Wikipedia says that BM25 is based on a probabilistic retrieval framework. I can't find any weakness of BM25.

- Jelinek-Mercer has reasonable P@3 score, but MAP score is relatively low.