Project 2 Report

Junrong Yan

April 8, 2014

1 Expectations

The scoring functions peform as:
Okapi TF < Okapi TF-IDF < Okapi BM25
Okapi Laplace Smoothing < Okapi Jelinek-Mercer Smoothing

I expect that Okapi TF-IDF will get higher score than Okapi TF, because the Okapi TF estimates the importance of a term by the frequence of the term occuring in each document, that is, tf(d,i). However, the Okapi TF-IDF estimates the importance of a term by the frequence of the term occuring in all documents, that is $\log(D/df(i))$.

Based on Wikipedia, BM25 is a bag-of-words retrieval function that ranks a set of documents based on the query terms appearing in each document, regardless of the inter-relationship between the query terms within a document. I think it be de different based ont then lenght of document. Thus, I could't say which one is better, though BM25 may be better thant Okapi TF-IDF, as BM25 is a method for scoring documents with relation to a query.

The Jelinek-Mercer Smoothing will be better than Laplace Smoothing. It uses formual $p_i = \lambda$ p + (1- λ)Q by training model on training set, which is based on the experience. And it will be better than Okapi TF, Okapi TF-IDF

Laplace Smoothing is also called add-one smoothing for solving the problem with 0 probability. Compare with all above scoring functions, it will be less effective because it gives too much probability mass to term which does not exist.