Information Retrieval

SS2014

Exercise 2 - Group 19

Our task in this exercise builds upon our experiments and experiences of the previous exercise. We were asked to integrate in our project Lucene functionality and to compare it with our results from the previous exercise.

Lucene (https://lucene.apache.org/) is an open source information retrieval software library developed (and supported and released) by the Apache Software Foundation. It is originally developed in Java, so the integration in our project was not a problem.

For our experiments we used the same 20 Newsgroups Subset document collection provided to us for the first exercise, and also the same 20 search topics for testing. We used the default Lucene implementation to index the documents and search them. For comparison reasons, we used our indexing from the first exercise, the Lucene default Similarity, and the Lucene BM25. Moreover, we implemented a BM25L Similarity score calculator, as described in the paper “When Documents Are Very Long, BM25 Fails!” published by Yuanhua Lv and ChengXiang Zhai (<http://sifaka.cs.uiuc.edu/~ylv2/pub/sigir11-bm25l.pdf>) and used this one also for our experiments.

The BM25 retrieval function has been the state of the art for nearly decades. BM25 scores a Document *D* with respect to Query *Q* as follows:

where *c(q,D)* is the count of *q* in *Q*, *N* is the total number of documents, *df(q)* is the document frequency of *q*, and is a parameter. The term frequency is calculated by the following formula (TF normalization formula):

where represents document length, stands for average document length, *c(q,D)* is the raw Term Frequency of *q* in *D*, and *b* and are two parameters. is the normalized Term Frequency by document length using pivoted length normalization:

As is shown in the above article, these formulas perform very well in the general case, but tend to penalize very long documents. So it introduces an improved TF normalization formula:

This equation still satisfies the constraints for BM25, but performs better with very long documents by adding a shift parameter. By substituting into the first equation we obtain a new retrieval function BM25L.

For each of the 20 given topics, the result appears in a ranked list, with each line having the following format:

*topic Q0 document-id rank score run-name*

where

**topic** is ‘’topic# with # the number of the given topic

**document-id** is an identifier for the document given by us and containing the

parent folder in the documents subset and the file name.

**rank** is an integer holding the rank of the object in the list

**score** is the similarity score result calculated by Lucene

**run-name** a name given from us to our experiment