Taylor Peer 0725922
Michael Wagner 0827376
Group 3
Information Retrieval 2014

Exercise 2 - Lucene, BM25, Evaluation

Prototype

The prototype application uses the Apache Lucene library to index a collection of documents and perform similarity searches upon it. Depending on the parameters used to start it, the application runs in different modes.

- Indexer (usage: java -jar task2.jar: -i)
  If the -i parameter is present when running the application, an index will be created on the document collection. This is done using the Standard Analyzer of Lucene, which tokenizes the collection while also removing stopwords, normalizing tokens to lower case and filtering out tokens above a given length. All documents are indexed using fields for their path and content.
- Search (usage: java -jar task2.jar: -s [-l | -d delta -k1 k -b b])
  By using the -s parameter, the application performs a search on a previously-generated index. With only the -s parameter, the search uses the BM25 similarity method. Optional parameters k1 and b are tuning parameters for the BM25 similarity and are set to k1 = 1.2 and b = 0.75 if not otherwise specified. Additional parameters modify the similarity method to be used during the search: the -l parameter enables the default Lucene similarity method and the -d parameter causes the modified BM25L similarity method to be used, which makes use of the delta argument given with the -d parameter.

BM25LSimilarity modification

For this exercise we modified the BM25Similarity class from Lucene to add another parameter, *delta*, to the scoring, as in this formula:

$$f'(q,D) = \begin{cases} \frac{(k_1+1)\cdot[c'(q,D)+\delta]}{k_1+[c'(q,D)+\delta]} & \text{if } c'(q,D) > 0 \\ 0 & \text{otherwise} \end{cases}$$

Where c'(q,D) is the normalized term frequency by document length.

$$c'(q,D) = \frac{c(q,D)}{1 - b + b\frac{|D|}{avdl}}$$

With this change f'(q,D) favors small c'(q,D) values more, which should boost the score of long documents. Since adding the delta was the only change in the implemented formula, we copied the Lucene base BM25Similarity class and changed the scoring and explainScore methods.

Explain

We used the explain function of the searcher with the query "camera sale" and used the first and second documents of the results. The explanation shows how the score is calculated according to the BM25L formula.

Q0 misc.forsale\76359 1 13.996054 BM25L-run

13.996053 = (MATCH) sum of:
  8.481009 = (MATCH) weight(content:camera in 2650) [BM25LSimilarity], result of:
        8.481009 = score(doc=2650,freq=2.0 = termFreq=2.0
), product of:
        4.8682847 = idf(docFreq=61, maxDocs=8000)
        1.7420938 = tfNorm, computed from:
        2.0 = termFreq=2.0
        1.2 = parameter k1
        0.5 = parameter delta
        0.75 = parameter b
        259.1086 = avgFieldLength
        83.591835 = fieldLength
  5.515044 = (MATCH) weight(content:sale in 2650) [BM25LSimilarity], result of:
        5.515044 = score(doc=2650,freq=2.0 = termFreq=2.0
), product of:
        3.1657562 = idf(docFreq=337, maxDocs=8000)
        1.7420938 = tfNorm, computed from:
        2.0 = termFreq=2.0
        1.2 = parameter k1
        0.5 = parameter delta
        0.75 = parameter b
        259.1086 = avgFieldLength
        83.591835 = fieldLength

Q0 misc.forsale\76357 2 12.782103 BM25L-run

12.782101 = (MATCH) sum of:
  7.267057 = (MATCH) weight(content:camera in 2648) [BM25LSimilarity], result of:
        7.267057 = score(doc=2648,freq=1.0 = termFreq=1.0
), product of:
        4.8682847 = idf(docFreq=61, maxDocs=8000)
        1.4927346 = tfNorm, computed from:
        1.0 = termFreq=1.0
        1.2 = parameter k1
        0.5 = parameter delta
        0.75 = parameter b
        259.1086 = avgFieldLength
        83.591835 = fieldLength
  5.515044 = (MATCH) weight(content:sale in 2648) [BM25LSimilarity], result of:
        5.515044 = score(doc=2648,freq=2.0 = termFreq=2.0
), product of:
        3.1657562 = idf(docFreq=337, maxDocs=8000)
        1.7420938 = tfNorm, computed from:
        2.0 = termFreq=2.0
        1.2 = parameter k1
        0.5 = parameter delta
        0.75 = parameter b
        259.1086 = avgFieldLength
        83.591835 = fieldLength

Evaluation Results

	our index	**Lucene Default Similarity**	**BM25**	**BM25L**
topic1	0.1433	0.1736	0.1523	0.0973
topic2	0.2498	0.3775	0.5373	0.3375
topic3	0.4684	0.4687	0.604	0.4282
topic4	0.1667	0.5333	0.527	0.5118
topic5	0.4029	0.5759	0.7399	0.4814
topic6	0.006	0.5	0.5	0.5
topic7	0.5787	0.57	0.6632	0.5565
topic9	0.0135	0.6429	0.75	0.5909

topic10	0.5	1	1	1
topic11	0.2023	0.2112	0.191	0.2052
topic13	0.3692	0.445	0.4638	0.4068
topic14	0.3967	0.5954	0.5638	0.5753
topic15	0.5062	0.676	0.7181	0.6274
topic16	0.2812	0.4854	0.4912	0.4819
topic17	0.033	0.3011	0.439	0.264
topic18	0.3981	0.5174	0.5131	0.5139
topic19	0.3333	0.6667	0.6667	0.6667
average	0.297	0.5141	0.56	0.485