COMP6791 Project 2

Current Version: 0.1
Date: Nov. 17, 2012
Author: Yuan Tao
Instructor: Dr. Sabine Bergler

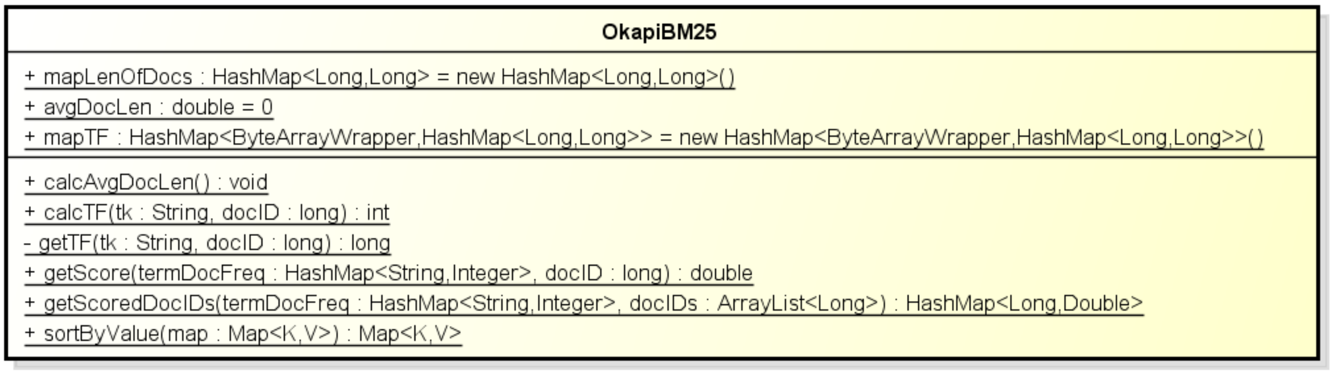
Concordia University

Table of Contents

1. Design

1.1. OkapiBM25 class

Compared with Project 1, a new class is added:



```
OkapiBM25
+ mapLenOfDocs : HashMap<Long,Long> = new HashMap<Long,Long>()
+ avgDocLen : double = 0
+ mapTF : HashMap<ByteArrayWrapper,HashMap<Long,Long>> = new HashMap<ByteArrayWrapper,HashMap<Long,Long>>()
+ calcAvgDocLen() : void
+ calcTF(tk : String, docID : long) : int
- getTF(tk : String, docID : long) : long
+ getScore(termDocFreq : HashMap<String,Integer>, docID : long) : double
+ getScoredDocIDs(termDocFreq : HashMap<String,Integer>, docIDs : ArrayList<Long>) : HashMap<Long,Double>
+ sortByValue(map : Map<K,V>) : Map<K,V>
```

- void calcAvgDocLen()
  It used to calculate the average length of all documents in the collection.

- int calcTF(String tk, long docID)
  It used to calculate the term frequency for all term in all documents and store the results into memory. It is done during the tokenizing step.

- long getTF(String tk, long docID)
  It used to get the term frequency for the term in the document with document ID equals to docID.

- double getScore(HashMap<String, Integer> termDocFreq, long docID)
  It used to calculate the BM25 sore for the query and the document.

- HashMap<Long, Double> getScoredDocIDs(HashMap<String, Integer> termDocFreq, ArrayList<Long> docIDs)
  It returns a sorted hash map whose key is the document ID and value is the BM25 score. The map is sorted by its value from high to low.

2. Tuning parameters of k1 and b in BM25

$$\text{score}(D,Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i,D)\cdot(k_1+1)}{f(q_i,D)+k_1\cdot\left(1-b+b\cdot\frac{|D|}{\text{avgdl}}\right)},$$

$$\text{IDF}(q_i) = \log\frac{N-n(q_i)+0.5}{n(q_i)+0.5},$$

This formula is got from http://en.wikipedia.org/wiki/Okapi_BM25.

In BM25 formula, k1 (k1 >= 0) is tuning parameter controlling the term frequency scaling. b (0 <= b <= 1) is the tuning parameter controlling the scaling by the document length.

In this project, following experiments are done to show the different results for different values of k1 and b.

Query: food company china

The result of the query has following documents:

Document ID	**Document Length**
6215	264
8143	144
12878	413

Term Frequency	6215	8143	12878
Food	2	2	2
Company	2	2	8
China	1	7	1

Term	**Document Frequency**
Food	443
Company	6186
China	313

Here are the different BM25 scores for different values of k1 and b:

k1	b	Document ID (BM25 Score)
0.1	0.1	8143 (9.5751) 6215 (9.1640) 12878 (9.1404)
1.0	0.1	8143 (13.7351) 12878 (10.2550) 6215 (10.2149)
2.0	0.1	8143 (16.9801) 12878 (11.1024) 6215 (10.8295)

100.0	0.1	8143 (37.1814) 12878 (15.6943) 6215 (12.5555)
0.1	1.0	8143 (9.5643) 6215 (8.6920) 12878 (8.2387)
1.0	1.0	8143 (13.6273) 6215 (7.7921) 12878 (6.5482)
2.0	0.75	8143 (16.8329) 6215 (8.2484) 12878 (7.1289)
100.0	0.75	8143 (36.2967) 6215 (8.1961) 12878 (7.6894)

3. Sample queries

Details of query results can be found from $Program\SampleQueries\.

3.1. Drug company bankruptcies

Document ID list (In total: 2):

```
  3091 (14.2894)    2127 (11.8888)
```

3.2. George Bush

Document ID list (In total: 13):

```
20891 (15.9528)    8593 (13.1173)    4008 (12.3950)  16780 (11.4868)  20719 (10.4883)    3560 (10.2646)
  7525 ( 9.1157) 20860 ( 8.9045)    8500 ( 8.8002)    2796 ( 7.7240)     965 ( 6.0912)    5405 ( 6.0723)
   854 ( 6.0556)
```

3.3. Democrats' welfare and healthcare reform policies

The program cannot get any document for the query. But it gets some documents for the following queries:

- Democrats welfare

  Document ID list (In total: 3):

  ```
    8072 (10.2716) 10125 ( 9.5712)    1895 ( 4.7047)
  ```

- Healthcare

  Document ID list (In total: 30):

  ```
  17560 (13.1158)    1595 (12.8980) 17697 (12.8685) 19173 (12.8276)    5093 (12.6468) 15791 (12.6063)
  20423 (12.4764) 15630 (12.3566)    2916 (12.0944)    3266 (11.8654) 20686 (11.6776) 19929 (11.3966)
  17740 (11.2363)    6276 (11.1741)     535 (10.6721) 19829 (10.2391)    6323 ( 9.5377)    1188 ( 9.4264)
  19133 ( 8.7681) 20449 ( 8.4031) 19453 ( 8.4031)    1760 ( 8.1652)    7433 ( 7.9404) 18722 ( 7.8478)
    5659 ( 7.7574) 10251 ( 7.5826)    6329 ( 7.4981)    8289 ( 6.7918)    3899 ( 6.4656)    3619 ( 6.3838)
  ```

- reform policies

  Document ID list (In total: 109):

```
  6940 (11.9501)   8310 (11.7374) 11204 (11.6582) 18161 (11.5532)   4006 (10.8353) 19845
(10.7119)
  2417 (10.4929) 12750 (10.1200) 17934 ( 9.6468) 20091 ( 9.5641) 12552 ( 9.5229) 16778
( 9.3590)
19591 ( 9.1934) 20061 ( 9.1543) 17940 ( 9.1092) 16092 ( 9.0600)   2480 ( 9.0126)   2068
( 8.9427)
12879 ( 8.8947)   3366 ( 8.8068)   2242 ( 8.7970) 12806 ( 8.7179) 16226 ( 8.6692) 14874
( 8.5368)
15140 ( 8.5050)   1088 ( 8.4618) 16060 ( 8.4081)   7219 ( 8.3322) 16196 ( 8.3009)   7375
( 8.1952)
  8627 ( 8.0842)   8128 ( 8.0761) 16540 ( 7.9616) 17197 ( 7.9017)   7651 ( 7.7587) 17106
( 7.6478)
17420 ( 7.3789)   5944 ( 7.2688)   6606 ( 7.2512)   9694 ( 7.2119)   8635 ( 7.1932)   8667
( 7.1783)
18598 ( 7.1655)   8835 ( 7.1655) 12598 ( 7.1368)   8069 ( 7.0357) 15058 ( 7.0079)   6600
( 6.9793)
18507 ( 6.9665) 10230 ( 6.9550)   5943 ( 6.9022) 17185 ( 6.8543)   3030 ( 6.7422)    862
( 6.6938)
11770 ( 6.6664) 16051 ( 6.5822) 20829 ( 6.5165)   3537 ( 6.4866) 17381 ( 6.4465)   5453
( 6.4227)
16292 ( 6.3998)   8764 ( 6.3136) 19543 ( 6.2695)   7493 ( 6.2096) 10355 ( 6.0991) 12438
( 6.0895)
18403 ( 6.0445) 16168 ( 6.0174) 15369 ( 5.9810)    672 ( 5.8500) 16948 ( 5.8435) 17075
( 5.8343)
  6052 ( 5.7988) 16198 ( 5.7593) 10059 ( 5.4765) 18378 ( 5.4281)   8390 ( 5.4101) 16544
( 5.3840)
15382 ( 5.2089) 18428 ( 5.1958)   4727 ( 5.1917)   4625 ( 5.1803)    742 ( 5.1008)   2052
( 5.0863)
15957 ( 5.0799)   2078 ( 5.0544) 11768 ( 5.0270) 20955 ( 4.9867)   9795 ( 4.9425) 13323
( 4.9099)
  9055 ( 4.9099) 13637 ( 4.8777)   2132 ( 4.8470) 13046 ( 4.4629)   8077 ( 4.4584)   8202
( 4.4495)
   238 ( 4.4495)   5759 ( 4.4402)   5169 ( 4.4056) 17366 ( 4.3796) 18420 ( 4.2221) 20462
( 4.0024)
17070 ( 3.7915) 17402 ( 3.7659)   5318 ( 3.6517)   5761 ( 3.6428) 11214 ( 3.5276) 20614
( 3.2453)
   335 ( 2.7291)
```

- Democrats welfare policies
  Document ID list (In total: 1):
  1895 (8.0831)

- Democrats welfare reform
  Document ID list (In total: 1):
  8072 (15.0529)

4. How to run

1. Extract xxx.zip to dir $DIR.
2. cd $DIR

(NOTE: Before running the program, please first read the documents under $DIR\doc.)

3. copy raw reuters21578 files into $DIR\input
4. java -jar COMP6791Prj2.jar

The result will be stored in the file $DIR\output\search-results.txt

The latest source code can be gotten from here:
https://all-projects-concordia.googlecode.com/svn/comp6791/Project1

Appendix A: Revision History

Version	Date	Author	Remark
V0.1	Nov. 17, 2012	Yuan Tao	Draft