COMP6791 Project 1

Current Version: 0.1
Date: Oct. 21, 2012
Author: Yuan Tao
Instructor: Dr. Sabine Bergler

Concordia University

Table of Contents

1. Design

1.1. Assumption

Because the major memory consumer is the dictionary and postings, the memory size used by other parts of the program is ignored, such as the buffer used to load the file, the memory used to do the sorting, the memory used by function calling, etc.

1.2. Term convention

- String
  The string consists of letters, following are valid strings:
  a word

- Number
  The number consists of digits, following are valid numbers:
  0 1 001 123

1.3. Statistics of lossy compression

	Terms			Non-positional postings		
	number	△%	T%	number	△%	T%
Unfiltered	70,878	-	-	2,360,096	-	-
No numbers	63,408	10.54	10.54	1,899,121	19.53	19.53
Case folding	45,535	28.19	35.76	1,739,020	8.43	26.32
25 stop words	45,510	0	35.79	1,516,927	12.77	35.73
Stemming	34,066	25.15	51.94	1,451,940	4.28	38.48

1.4. Data structure of Inverted Index

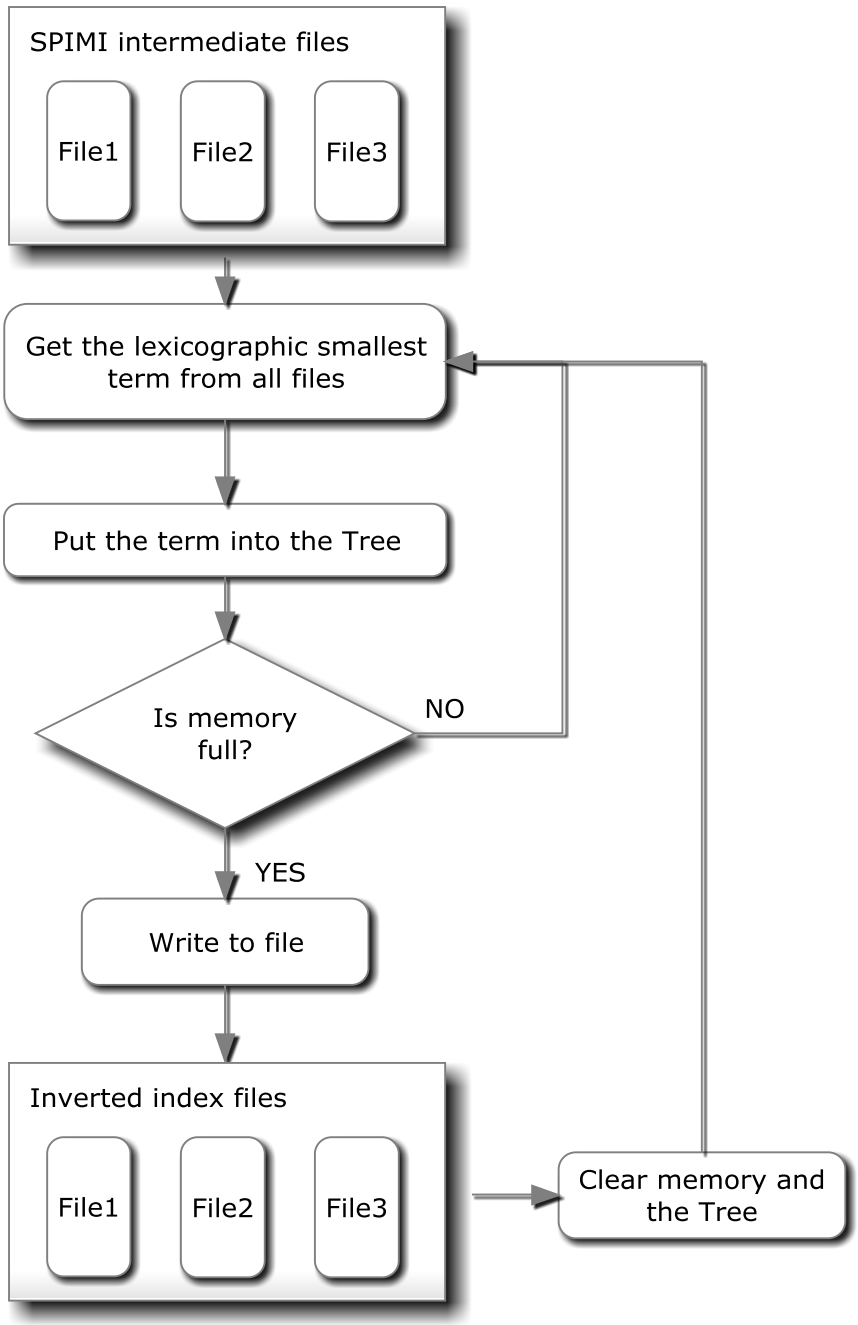
```
private HashMap<ByteArrayWrapper, InvertedIndex> mapUnsortedIndex = new
HashMap<ByteArrayWrapper, InvertedIndex>();

public class InvertedIndex {
    public static final int POSTINGSLIST_INIT_SIZE = 30;
```

```
    public static final int SIZE_OF_POINTER = 4;  // to calculate the memsize
    public static final int SIZE_OF_LONG = 4;     // to calculate the memsize

    //public byte[] term;  // the pointer of the term is stored in the key of hashmap
    public int docFreq;    // the actual length of postings
    public long[] postings = new long[POSTINGSLIST_INIT_SIZE];
}
```

1.5. The way of merging SPIMI intermediate files



1.6. Searching commands

- Blank (Boolean AND. It means intersection of the terms).
  For example:

term1
term1 term2 term3

- OR (Capital letters, otherwise it means a term.)
  Term1 OR term2
  Term1 OR (term2 term3)
  (Term1 term2) OR term3
  (Term1 term2 term3) OR (term4 term5 term6)

- AND NOT (Capital letters, otherwise it means two terms.)
  Term1 AND NOT term2
  Term1 AND NOT (term2 term3)
  (Term1 term2) AND NOT term3
  (Term1 term2 term3) AND NOT (term4 term5 term6)

2. Sample test cases

1. Ronald Reagan
2. black Monday
3. Wall Street crisis
4. (Black Monday) OR (Wall Street crisis)
5. (Black Monday) AND NOT (Wall Street crisis)

All the results are stored to the directory:
../ SampleQueries/

3. Optimization

- Read file

The way to read raw files for this project is got from following URL:
http://nadeausoftware.com/articles/2008/02/java_tip_how_read_files_quickly

The project chooses the way of "BufferedInputStream with byte array reads" with array size of 16K bytes.

- Write to file

When writing the sorted inverted index to file, the program buffers the data to be written. Once the buffer reaches a size, or it gets the last element of the index, it starts to write the buffer to the file.

- Sort terms

Java does not provide any sorting algorithms for byte array (byte[]). If the program converts the byte[] to String, it might greatly increase the size of dictionary. As a result, this program implements the Mergesort algorithm for byte[]. And the following statistics shows that it is faster than using TreeSet:

1. Mergesort(byte[][])

Around 55 milliseconds for 52980 terms

2. Convert to TreeSet<String>

Around 88 milliseconds for 52980 terms

Appendix A: Revision History

Version	Date	Author	Remark
V0.1	Oct. 21, 2012	Yuan Tao	Draft