Karthik Vegi| kvegi@iu.edu Melita Dsouza | dsouzam@iu.edu

HBASE FREQINDEXBUILDER

Project Report – Hbase FreqIndexBuilder

Indiana University

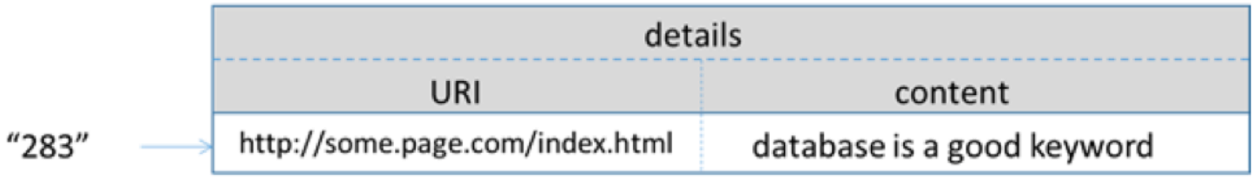
Revision	Date	Description	Author
1.0	15-Mar-2017	Coding	Karthik Vegi
1.1	18-Mar-2017	Testing and report	Melita Dsouza

HBASE FREQINDEXBUILDER

PROJECT DESCRIPTION

In this project, we build an inverted index table which has the unique term's occurrences in all documents from the clueWeb09 dataset.

The table schema for the Clueweb09 data set is shown below:



	details	
	URI	content
"283"	http://some.page.com/index.html	database is a good keyword

Data Flow and Logic

Hbase FreqIndexBuilder is ***Map Only*** parallel program which only has the Mapper code

Mapper Code:

- The Map function receives as input the ***<key, value>*** pair in the form of ***<ImmutableBytesWritable rowKey, Result result >*** where row is the row key of the HBase record related to a specified URI and result is the stored text of that URI.
- The following function is used to access the content in the Hbase table

```
byte[] contentBytes =
result.getValue(Constants.CF_DETAILS_BYTES,Constants.QUAL_CONTENT_BYTES);
```

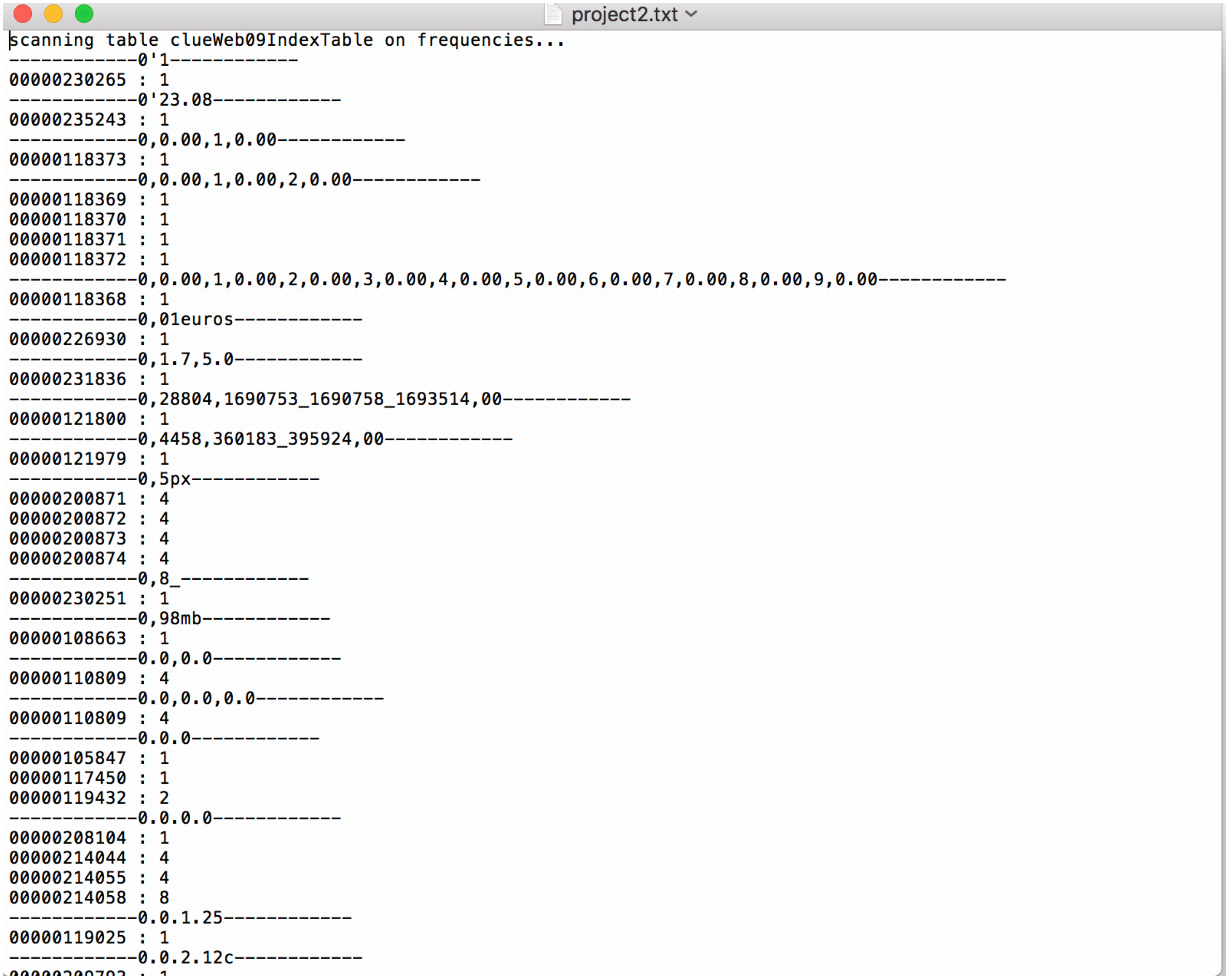
- ***Constants.CF_DETAILS_BYTES*** is the column family "details" and ***Constants.QUAL_CONTENT_BYTES*** is the column content of the table shown in the table schema
- We create a HashMap object named ***freqs*** which returns the frequency of each word returned by the function ***getWordFreq()***
- Once we get the frequencies for each distinct word, a Put object is created which will add a row in the ***FreqIndexTable*** in Hbase using the following command:

```
FreqIndexTable.add(Bytes.toBytes("frequencies"), docIdBytes, Bytes.toBytes(freq));
```

- where ***docIdBytes*** is the rowkey of an HBase record. The output ***<key, value>*** pair of this function is ***<Text word, <docId, frequency>>***
- Write the put object to context which inserts it into the ***clueWeb09IndexTable***

Output

Partial output is shown below. The complete output is attached.



```
scanning table clueWeb09IndexTable on frequencies...
------------0'1------------
00000230265 : 1
------------0'23.08------------
00000235243 : 1
------------0,0.00,1,0.00------------
00000118373 : 1
------------0,0.00,1,0.00,2,0.00------------
00000118369 : 1
00000118370 : 1
00000118371 : 1
00000118372 : 1
------------0,0.00,1,0.00,2,0.00,3,0.00,4,0.00,5,0.00,6,0.00,7,0.00,8,0.00,9,0.00------------
00000118368 : 1
------------0,01euros------------
00000226930 : 1
------------0,1.7,5.0------------
00000231836 : 1
------------0,28804,1690753_1690758_1693514,00------------
00000121800 : 1
------------0,4458,360183_395924,00------------
00000121979 : 1
------------0,5px------------
00000200871 : 4
00000200872 : 4
00000200873 : 4
00000200874 : 4
------------0,8_------------
00000230251 : 1
------------0,98mb------------
00000108663 : 1
------------0.0,0.0------------
00000110809 : 4
------------0.0,0.0,0.0------------
00000110809 : 4
------------0.0.0------------
00000105847 : 1
00000117450 : 1
00000119432 : 2
------------0.0.0.0------------
00000208104 : 1
00000214044 : 4
00000214055 : 4
00000214058 : 8
------------0.0.1.25------------
00000119025 : 1
------------0.0.2.12c------------
```