University of Texas at Dallas

CS 6322.501 :

Information Retrieval Fall 2015

Homework # 2

Sumit K Sikhwal

sks130730

Run the following commands to run the application : -

javac -cp Externaljars\\*; IndexBuilding.java

java -cp Externaljars\\*; IndexBuilding "/people/cs/s/sanda/cs6322/Cranfield/ " "/people/cs/s/sanda/cs6322/resourcesIR/stopwords "

The jars needed for running this code are present in the folder Externaljars.

Output is in the following documents

1)v1\_compress\_lemmas :- compressed dictionary and posting files for lemmas

2)v1\_uncompressed\_lemmas:-uncompressed dictionary and posting files for lemmas

3)v2\_compress\_stems :- compressed dictionary and posting files for stems

4)v2\_uncompressed\_stems :- uncompressed dictionary and posting files for stems

5)max\_tf : the document name ,the stem with highest frequency and its count.

Format :- Documentname:stem:count

6)doc\_length :- the document and the number of terms in the document.

Format :- Documentname:count

The program reads in the cranfield documents and removes special characters from the words and creates lemmas and stems out of the words. Then it creates the dictionary and posting files for these words and then compresses the dictionary and posting files in another file.

Statistics (for cs1.utdallas.edu cluster)

Time elapsed to build uncompressed Version 2 is 48015 milli Secs

Time elapsed to build uncompressed Version 1 is 62580 milli Secs

Time elapsed to build compressed Version 1 is 97479 milli Secs

Time elapsed to build compressed Version 2 is 84352 milli Secs

The size of the Index Version 1 uncompressed : 1273235.0 Bytes

The size of the Index Version 2 uncompressed : 1202209.0 Bytes

The size of the Index Version 1 compressed : 817683.0 Bytes

The size of the Index Version 2 compressed : 746419.0 Bytes

The number of inverted list in Index Version 1 uncompressed: 7224

The number of inverted list in Index Version 2 uncompressed: 6081

The number of inverted list in Index Version 1 compressed: 7224

The number of inverted list in Index Version 2 compressed: 6081

the df, tf, and inverted list for the terms: "Reynolds", "NASA", "Prandtl", "flow", "pressure", "boundary", "shock" can be found in their respective output files.

Naming convention for files :- word\_stem or word\_lemma.

Output format for the word\_stem or word\_lemma files :- word-df tf

For eg the df, tf, and inverted list for the term Reynolds(lemma):

reynolds-3 0487=1 1215=1 1395=1

Inverted list length for the term reynold (stem) : 1000 Bytes

Inverted list length for the term nasa (stem) : 1725 Bytes

Inverted list length for the term prandtl (stem) : 2040 Bytes

Inverted list length for the term flow (stem) : 5690 Bytes

Inverted list length for the term pressur (stem) : 8440 Bytes

Inverted list length for the term boundari (stem) : 10775 Bytes

Inverted list length for the term shock (stem) : 11970 Bytes

Inverted list length for the term reynolds lemma : 15 Bytes

Inverted list length for the term nasa lemma : 740 Bytes

Inverted list length for the term prandtl lemma : 1055 Bytes

Inverted list length for the term flow lemma : 4705 Bytes

Inverted list length for the term pressure lemma : 7410 Bytes

Inverted list length for the term boundary lemma : 9745 Bytes

Inverted list length for the term shock lemma : 10935 Bytes