Programming Assignment 5

1 Assignments

1. Generate the inverted index given a set of input files.
   - The input contains a set of text files. Each file contains words without punctuation marks in multiple lines. Assume that the size of a file is less than 128 MB.
   - The files are named as "file0", "file1", "file2", ...
     - Please download 4 input files from blackboard.
   - The generated inverted index is distributively saved in multiple files. An output file contains multiple lines. Each line consists of a term (i.e., a word) and the posting list.
   - In each output file, those lines are listed in an alphabetic order.
   - Each posting list is in such a format as "file_name:# of occurrence;file_name:# of occurrence...". The posting list needs to be in the order of file names. Example: "file0:18;file1:20;file2:3".
     - Please see one example output file that can be downloaded from blackboard
   - Specify 3 reducers.
2. You are not allowed to perform the sorting on the reducer side. Instead, you need to design your program such that the "shuffle and sort phase" will sort for you.
3. More details:
   - Use pairs approach to generate complex keys, i.e., (term, filename), in the mapper stage. The value is the total number of occurrences the term appears in the file.
   - Apply in-mapper combining to figure out the total number of occurrences a term appears in a file.
   - Refer to the example on the following link to obtain the file name in mapper stage.
     - `https://acadgild.com/blog/building-inverted-index-mapreduce`
     - However, do NOT use the approach in the above link for producing inverted index.

2 Submission

- Due date: March 8, 2019 @ 11:59 AM.
- Submission
  - Name two files as wordpair.java and invertedindex.java.
  - tar -cvf pa5_<your last name>.tar wordpair.java invertedindex.java
  - Upload the tar file to blackboard before deadline.