Report on project - inverted_index

Subhrajyoti Pradhan

U79333962

Introduction -

For this lab you will use an alternate input, provided in the file invertedIndexInput.tgz. The aim of this project is to have a list of all the places where an individual word appears.

We format it such that the output contains the filename and the line number.

Description of Driver -

In order to format the date as a special input format, we use the code -

job.setInputFormatClass(KeyValueTextInputFormat.class;

We also to set all classes (eg -the output value, output key, mapper and reducer.)

Description of Mapper -

The mapper makes use of filename. Therefore, we are required to the get file names in order to include them as values. In order to do this we use the method getInputSplit and getPath. Besides the filename, the value also contains a separator (@) and the previous key which is the position of the word. The word is the new key.

Description of Reducer -

The reducer appends the values for every word using a semicolon. The word is the output key and the output is all the times that word appears expressed as an integer.

Data Flow -

The program first reads the entire play following which the reducer splits the lines by words. The words are used as output keys. The output is the name of the file.

Testing and Results -

We first check the file for errors in Eclipse and compile it to a jar file. We unzip the .tgz file and create a directory in hadoop. We execute everything using command -

Hadoop jar inv.jar stubs.InvertedIndex /user/training/invertedindex inv

Example output -

zeals timonofathens@746;

zed kinglear@1843;

zenith tempest@468;

zephyrs cymbeline@3615;

zir kinglear@4502;kinglear@4489;

zo kinglear@4495;

zodiac titusandronicus@802;

zodiacs measureforemeasure@446;

zone hamlet@5374;

zounds kingrichardiii@3583;1kinghenryiv@4370;1kinghenryiv@1020;

zwaggered kinglear@4494;