UE INF2245
Word count on the Gutenberg dataset

Frédéric Raimbault

1. Test the WordCount program on (a small part of) the `hdfs:/data/Gutenberg` dataset (HDFS) ; the source code is given bellow.

2. Remplace the TextInputFormat with a CombineTextInputFormat, test it and explain the difference.

3. Add a combiner.

4. Modify the `WordCount` program to count the occurences of every words in the dataset, except whose which are given in a stop words file. This file will be transmitted by the driver to the mappers through the method job.addCacheFile(stopwords_path) and the contents will be read and stored in a hashset by the setup() method of the mappers.

   - Use the (HDFS) file `hdfs:/data/stop-words/stop-words-english4.txt` for your tests.
   - Add a counter REJECT_CNT to store the number of words filtered by the stoplist and add also a counter ACCEPTED_CNT to store the number of words retained. Test it and print the counter values in the driver at the end of the execution.

5. Write the Top100 MR program that prints the 100 most frequently used words in the Gutenberg books.

   - You will have to replace the default key comparator with job.setSortComparatorClass(LongWritable.DecreasingComparator.**class**) to ensure a descending order sort.
   - As your Top100 program will take as input the result of your preceeding `WordCount` program, its input format should be KeyValueTextInputFormat.**class**.