UE INF2245
Hadoop MapReduce: first steps on the Gutenberg dataset

Frédéric Raimbault

1. Test the MRWordCountAWS program on (a small part of) the `s3:///ubs-datasets/gutenberg` dataset ; the source code presented during the course is given on the ENT.

2. Is it possible to obtain the result of MRWordCountAWS in a single file and in descending order of the number of occurrences of the words?

3. Write the MRTop100AWS 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 MRTop100AWS should take as input the result of the preceeding MapReduce program, its input format should be KeyValueTextInputFormat, with a Text as key's type and a LongWritable as value's type.
   - The resulting list should be print on the screen as such:

```
// affichage du contenu du fichier resultat
System.out.println("Top100: ");
FSDataInputStream inputStream = hdfs.open(new Path(output_folder+"/part-r-00000"));
IOUtils.copyBytes(inputStream, System.out, 4096, false);
inputStream.close();
```