UE INF2245
Hadoop Spark: first steps on the Gutenberg dataset

Frédéric Raimbault

1. Write the SparkWordCountAWS, equivalent to the MRWordCountAWS of the first lab on MapReduce programming. Help you with several fragments exposed during the course.

2. Test it on a small part of the `s3:///ubs-datasets/gutenberg` dataset and compare it against the results obtained with the MapReduce version.

3. Write the SparkTop100 program that prints the 100 most frequently used words in the Gutenberg books with their frequency.