Anonymizing comments[1]

Shuffle pattern (Challenge 5_e)

For this exercise you have to prepare a report including:

- Your data collections comment describing what they represent.
- Your source codes: a file precising in a comment the compiling and execution commands.
- Propose a test battery and report execution measures.
- Compress everything and sent a .zip or a .tar to genoveva.vargas@gmail.com

1.1 Problem statement

Problem: Given a large data set of StackOverflow comments, anonymize each comment by removing IDs, removing the time from the record, and then randomly shuffling the records within the data set.

1.2 Implementation

Look at page 101 of the book "Map Reduce design patterns" and see the proposed `Map`, and `Reduce` codes. Prepare a data collection of your choice and implement the solution. As in previous challenges, you can use StackOverflow or any other source for generating your collection. Program the first part of the example.

- Explain the principle of the solution. Use examples for illustrating your answer.
- Prepare collections of different sizes to run your tests trying to get to the limits of your solution.
- Make comparisons. Do not hesitate to prepare graphics.

[1] This challenge is an example proposed in the book MapReduce design patterns, pp. 101.