Comment comparison[1]

Cartesian product pattern (Challenge 6_d)

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 without filter

Problem: Given a groomed data set of StackOverflow comments, find pairs of comments that are similar based on the number of like words between each pair.

1.2 Implementation

Look at page 132 of the book "Map Reduce design patterns" and see the proposed `Driver, Record reader, Map` code. 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.

- Describe the general principle of the solution proposed by the book.
- What is the role of the `Input format code`? Explain the use of `getInputSplits`.
- What does de `driver code` do? And the `record reader code`?
- How is the Cartesian product similar to the self join?
- 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. 132.