Top ten users by reputation[1]

Filtering pattern (Challenge 4_c)

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 list of user information, output the information of the top ten users based on reputation.

Determining the top ten records of a data set is an interesting use of MapReduce. Each mapper determines the top ten records of its input split and outputs them to the reduce phase. The mappers are essentially filtering their input split to the top ten records, and the reducer is responsible for the final ten.

Attention: *Remember to configure your job to only use one reducer! Multiple reducers would shard the data and would result in multiple "top ten" lists.*

1.2 Implementation

Look at page 63 of the book "Map Reduce design patterns" and see the proposed `Map` and `Reduce` codes.

- Explain the principle and role of the `TreeMap` object. Use examples for illustrating your explanation.
- Explain the role of the `cleanup` method of the mapper code proposed in the book.
- Explain why it would be equivalent to directly use a `reduce` method than using the `cleanup` method for obtaining one top ten list.
- Test your implementation with different data collection sizes and discuss results.

[1] This challenge is an example proposed in the book MapReduce design patterns, pp. 63.