Sort users by last visit[1]

Total order sorting pattern (Challenge 5_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

Problem: The user data in our StackOverflow data set is in the order of the account's creation. Instead, we'd like to have the data ordered by the last time they have visited the site.

1.2 Implementation

Look at page 96 of the book "Map Reduce design patterns" and see the proposed `Map`, `Driver`, 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 `Driver` and its role in the whole task. Why is it divided into to sections? Use examples for illustrating your answer.
- Comment on the configuration of the number of reducers.
- Explain the role of the `InputSampler` utility.
- The book states "*If your data distribution is unlikely to change, it would be worthwhile to keep this partition file around. It can then be used over and over again for this job in the future as new data arrives on the system.*" Explain why.
- Explain the role of the `analyse mapper`, `order mapper` and `order reducer` codes.
- 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. 95.