Inf-2202 Mandatory Assignment 3

Out: Thursday 24.10.2013
Due: Tuesday 05.11.2013
Individual assignment.

In the previous assignment you implemented a system for deduplication of data sent over a network. In this assignment you will measure the compression ratio that can be achieved using deduplication when removing non-essential fields in the records in the UniProt dataset.

You should do the measurements by implementing a MapReduce program that:

1. Splits the input files into records.
2. Computes a SHA-1 fingerprint for each record.
3. Uses the SHA-1 fingerprints and the size of records to find the compression ratio.

You must also measure the execution time of the MapReduce program and discuss how it compares to the execution time of your program in Mandatory Assignment 2.

You may also extend the program such that records fields can be set to a default value before calculating the SHA-1 fingerprint.

Further you can compare the compression ratio achieved when modifying different fields in the UniProt records.

You can implement the program in either Java or Python (using Pydoop).

The hand in comprises the code for your MapReduce program, a short report describing the design choices and assumptions made in your program, and the results. The report should be maximum 2 pages.

The code and report are to be sent by e-mail to Ibrahim.

Lars Ailo Bongo (larsab@cs.uit.no), Fall 2013