A2 - Neighborhood Score: MapReduce

Write an additional implementation of A0 using Hadoop & MapReduce, then re-evaluate them by analogy to A1. Use a pseudo-distributed cluster to evaluate the MapReduce implementation.

Collect and analyze execution-time information of all implementations. Use the following big corpus (http://violet.ele.fit.cvut.cz/~kondziu/pdpmr/big-corpus.tar) (SHA512 (http://violet.ele.fit.cvut.cz/~kondziu/pdpmr/big-corpus.sha512)).

Prepare a report using R Markdown highlighting the differences in the execution profile of the variants. In particular, compare the difference in execution profiles of the Hadoop/MR vs other variants and comment on the reasons for performance increases or decreases.

Deliverables

Deliver the code of the implementation and the performance report via a private repository on CCS GitHub (https://github.ccs.neu.edu) or Github (https://github.com) and share the repository with the instructors and TAs:

- if you are creating private repository on `github.ccs.neu.edu`, share it with `aviralgoel`, `samarthshetty1990`, and `ksiek`.
- if you are creating private repository on `github.com`, share it with `aviralg`, `kondziu`, `janvitek`, and `shettysamarth`.

Repository name: `pdpmr-f17-your-name-a2` (replace `your-name` with your name in lowercase letters with dashes between words).

Required files:

- `src/` (directory containing the sources of the implentation)
- `README.md` (Markdown file containing the description of the implementations)
- `Makefile` (Configuration file for the make command with the following rules: `build` — builds all the implementations, `run` — runs the variants and generate the report, `all` – build and run; be sure that `build` works on somebody else's machine)
- `report.Rmd` (The report as described in the previous section)
- `report.html` (An HTML rendering of `report.Rmd`)
- `input/books/` (A directory containing input files used in the report for A0/A1; please `gzip` all text files

to save space)

- `input/big-corpus/` (An empty directory. Keep your copy of the big corpus there, but do not push it to git. A code reviewer will put their own copy of the corpus there too.)

Submissions not adhering to the prescribed structure will be penalized.