✔ Terminé

The data

In /user/fzanonboito/CISD/IEEEdata.csv you will find metadata on papers published by IEEE that contain "comput*" on the title field between 1962 and 2023. This data was obtained using a crawler created by Jonas M. Korndorfer (University of Basel) and converted to csv format with code written by Ahmed Eleliemy (University of Basel).

Data processing

1. You are asked to obtain the top 10 keywords for the whole period and also per decade. Each keyword must be accompanied by the corresponding number of papers.

2. Now imagine we want to periodically update this result by adding a new csv file containing papers published after the previous execution, but without reprocessing data unless strictly needed. Propose a solution for this, which may include modifying the code that you proposed for the first part.

Report

For the first report, which will make for 30% of your grade for this course, you are asked to:

- present and explain your solution in details.
- analyze your solution: what is the expected size of intermediate and output data, how many Map-Reduce jobs were used and why, how many reducers were used in each job and why, how its performance is expected to change if we increase/decrease the number of available machines and/or papers, etc.
- discuss intermediate data representation: text vs. integer values (for counters). You may, for example, add some results that compare alternatives to support your choice.
- if relevant, discuss the limitations of your solutions and how it could be improved.
- respect a page count limit of 3 and submit your work in .pdf format.
- work individually and submit your own report (i.e. one report per person).
- not copy anything from the internet or from colleagues, to not ask for solutions in online forums, and to cite all external sources you use.
- write in French or English (or Portuguese :).
- include the obtained results for the given file. That can be done as an appendix and does **not** count for the page count limitation.

Grades will be based on:

- quality of the report: readability, clarity, organization, etc;
- correctness of the provided arguments and explanations;
- correctness and performance of the provided solution.

Modifié le: Friday 28 October 2022, 10:34