Computing at Scale in Machine Learning: Distributed computing and algorithmic approaches (14038)

B-TU Cottbus-Senftenberg · WiSe 2023
Prof. Dr. Alexander Schliep, Nathalie Gocht, M.Sc. and Aleksandra Khatova, M.Sc.

Problem set from January 10, 2024 · Due on January 19, 2024

For this assignment, you have to hand in AT LEAST ONE of the two problems. You can choose which one you like more. If you submit two problems, the pass/fail will be based on the task with the most points achieved. Example: you get 2p (out of 8p) for the problem 1, and 5p (out of 8p) for the problem 2. You pass the assignment because you've got more than 50% for the problem 2.

Problem 1 (Probability, 8 pt)**.** As we have seen, sometimes one is interested in the set of unique items present in a data set. If the unique items $1, ..., n$ occur with equal frequency we can make the assumption that the data generating process can be seen as uniform sampling with replacement from the set $1, ..., n$. What is the expected number of items one has to sample to collect every unique item at least once? Evaluate your result by simulating the experiment in a Python Script. Run at least 100 simulations and compare the results to the theoretical value from the formula you obtained.

Problem 2 (Tries and MapReduce, 8 pt)**.** To decide the amount of memory needed for storing all words (i.e. strings only containing the letters a-z) appearing in a large text in a trie, compute the average length of shared prefixes between consecutive words in a sorted list of unique words from the text. For example for the text "Barney's barn is burning", the sorted list of words consists of barn, barneys (removing any characters not in a-z), burning, is (note that case is ignored) and the shared prefixes are barn, b, "" with an average of (4+1+0) / 3. Implement a solution using MrJob returning the average length of the prefixes, the total number of unique words and the average word length.

Hint: Do not use one sort operation on all words at once in the text (or external sort programs). There is text data available in `06049.txt`.

Note: Please hand in either:

- A Python script named `problem1.py` and a PDF including the answer to Problem 1.
- Or a Python script named `problem2.py`

Python code must be handed in as '.py'. Code submission in PDF will be graded with 0 points. Please submit only the files specified above.

You are allowed to use the example solutions provided where they might be helpful for answering question on this problem set.

Please feel free to discuss the problems and approaches to solving them with other students. However *you* should understand results of discussions well enough to write up the solution by yourself and possibly explain the solution. When you do collaborate, please make other's contributions clear. Present your solutions concisely.