Cairo University

Faculty of Engineering

Computer Engineering Dept.

Assignment 1

Distributed Word Count using Go lang

Problem Definition:

Input: a file containing English text, as in "ExampleIn.txt"

Output: a file containing each unique word and its associated count as appeared in the input text. The output file format should follow the "ExampleOut.txt" provided. (Note that the final output will be checked using a test script).

Method:

The input file should be divided evenly among **5** go routines, each routine computes the word counts for the portion of the file it is responsible for. After each routine finishes, it writes the output to a shared map that is handled by another routine, call it "reducer". Note that only one routine should access the map at a time. The output correctness should also be guaranteed e.g. if one routine has the word "assignment" 3 times and another has the same word 7 times then the map entry for the word assignment should now be 10. When all routines write the output to the shared map, the "reducer" should write the output in the file sorted by the frequency (sorting is not distributed, "reducer" can be responsible for it). If two or more words have the same frequency, then sort them alphabetically.

Language and Tools:

Implement the requirement using Go. Make use of the concurrency techniques in the language such as mutex and channels. "Concurrency Control is a MUST".

Notes:

Start working using any input example and before the delivery by 1 day, you will be provided with an input file that will be used for testing. This means that your code should be generic (work on any given English file). For simplicity, no data cleaning is required, you only get the words by splitting on the space character. However, it is necessary to convert all words to lower case.

Deliverables:

- Work in teams of 2 members. Due date is Monday 26 April 2021 at 11:59 p.m.
- Each team should deliver:
  - All go code file(s) and one output file generated for the input file "test.txt" (which will be provided later), name this file "WordCountOutput.txt"
- Only 1 member of the team should submit the required files in one zipped folder named by the names of both team members.
- Submissions should be done via blackboard "Assignment-1".