Faculty of Computers and Information
Cairo University
Spring 2020

CS471: Parallel Processing
Assignment-2

Guidelines:

1. This assignment is groups of three (students given by the same TA can work together).
2. Deadline is 4 April 2020.
3. Total assignment grade is **out of 13**.
4. Upload it on Acadox with file named **"*A2_ID1_ID2_ID3_G#.c*" *else it will not be accepted.***
   eg.1 A2_20160555_20160343_2016000_CS_DS_1.c

Search Engine Helper:

Write a parallel program to search a given corpus and return the most relevant search results. You are given a corpus called Aristo Mini Corpus (https://www.kaggle.com/allenai/aristo-mini-corpus).

Aristo Mini Corpus:

The Aristo Mini corpus contains 1,197,377 science-relevant sentences drawn from public data. It provides simple science-relevant text that may be useful to help answer elementary science questions. You will work on 1500 sentence only divided across 50 File, each file is 30 lines.

Input: a given query in form of a sentence or a question.
Output: search results that contain all the words of the query.

Example:

Search query:
Capital of Egypt

If the corpus has the following sentences:

File1:
There is a capital for each country.
Capital of Egypt is Cairo.

File2:
The Capital of Egypt is Cairo.
You can visit the country you want.

Output should be:

Capital of Egypt is Cairo.

The Capital of Egypt is Cairo.

Pseudo code of search steps applied for each file:

```
For each Sentence in File:
    Match = true;
    For each word in the query:
        IF word not in CurrentSentence:
            MatchScore = false;
    IF MatchingScore is true:
        Store Sentence;
        ResultsFound += 1;
```

Parallel Scenario:

- ✓ You will use Master Slave Paradigm.
- ✓ Master will distribute the corpus files on slaves.
- ✓ Slaves will search the given part of a corpus.
- ✓ Each slave will return number of search results found and the corresponding relevant sentences.
- ✓ Master will collect the number of search results and write them to a file.

Expected input/output format:

Enter your query: sunlight energy nutrients

Output File:
Search Results Found = 2

Chlorophyll can make food the plant can use from carbon dioxide, water, nutrients, and energy from sunlight.
A process by which a plant produces its food using energy from sunlight, carbon dioxide from the air,and water and nutrients from the soil.

Requirements:

1- Study the MPI lab of the scatter and gather methods.
2- You have one week for questions about the assignment and the lab (22 Mar. to 28 Mar.).
3- Use all functions you learned so far in MPI library. (For Allreduce and Allgather it is not a must to use them).
4- You have to choose your functions carefully, which means if there is a value that should be sent to all slaves use MPI_Bcast, if there are values to be reduced using a specific operator use MPI_Reduce and so on.
5- Calculate the running time of the parallel program.
6- Run your code on the attached test cases, to ensure your result is right.

Grading Criteria:

Master workload distribution across slaves: • Using suitable MPI functions	50
Slave work: • Reading files and tokenizing queries. • Perform search and send back to master.	60
Master collection of results: • writing them to a file (# of Search Results, and the results itself)	50
Handling remaining workload	30
Running and valid output	30
Calculate the parallel running time	10
Total	**230**