CS426
Spring 2019
Project 2
Due 31/03/2019

1. Problem Definition

In this homework , you will implement a parallel document search system. The algorithm that you will implement is similar to Supervised Search. However, it is a very simple version of it. Since this is a simplified, more abstract version tailored for this class project, do not overthink things. Instead carefully read everything about problem definition and what is expected of you.

In our system, each document i is represented with a weight vector w_i. A weight vector w_i is composed of D number of elements such that each of the weight values $w_{i,j}$ corresponds to the relationship between the file i and word j in our dictionary, and D is the dictionary size. D is only an integer here, we do not have an actual dictionary.

Set of documents W which contains F documents can be represented as follows:

$$W = \{\, w_i \mid w_i \; is\ a\ vector\ and\ w_{i,j} \geq 0\ and\ w_{i,j}\ is\ an\ integer \,\}\ where\ 0 \leq i \leq F\ and\ 0 \leq j \leq D$$

Then, we will insert some queries to the system. A query Q has same vector definition as a document vector w_i.

$$Q = \{q_j \mid q_j \geq 0\ and\ q_j\ is\ an\ integer\}\ where\ 0 \leq j \leq D$$

Finally, your program will take $W, Q, D\ and\ K$ as input and will give least K related documents as output according to given *similarity function*.

You will give a document file as an input file to your system. Each line of the input file contains id of document, consecutive D integers which are seperated by a single space. Format of a line is as follows:

Doc_id: weight$_0$ weight$_1$ ... weight$_d$

Example line for document with id 5 and a system with dictionary size 3

5: 0 5 3

And query will also be given with another input file. However, it does not contain an id field.

E.g. for a query with a dictionary size 3

2 0 7

Now, we will define our *similarity function*. Similarity of a document i to a query Q, s_i , is defined as follows:

$$s_i = \sum_{j=0}^{D} w_{i,j}^{q_j}$$

E.g. we can calculate similarity value for previous weight vector and query pair as follows:

$$0^2 + 5^0 + 3^7 = 2188$$

To summarize, you will read documents file and query file. You will find similarity values for each document and output least related K documents' ids, ie. With least similarity values.

2. Implementation Details

First you will implement a *kreduce* function with the following function prototype.

*void kreduce(int * leastk, int * myids, int * myvals, int k, int world_size, int my_rank)*

kreduce function is a collective communication function like MPI_Bcast or MPI_Reduce. All processes will call it, however only the master process will have k ids that correspond to least k values in ascending order according to their corresponding values. Meaning that *leastk* is only used by master process. Each process pass two integer arrays to this function. *Myids* array keeps ids of documents of a process and *myvals* array keeps similarity values for these documents. *k* tells that how many least elements will be selected. Last two parameters of this function is optional and their names are self explanatory. Sizes of myids and myvals arrays are equal to k. And the values in myvals array is sorted in ascending order.

Let say, we have 2 processes running and their inputs to this function as follows:

k = 3

Process 0 (Master):

My_ids = [5, 6, 7]

My_vals = [4, 8, 14]

Process 1 (Slave):

My_ids = [1, 2, 11]

My_vals = [7, 10, 22]

And as an output of this function, process 0 should have the following values in leastk array:

leastk = [5, 1, 6]

You are not allowed to use any collective communication operations outside of this kreduce function. The only collective communication operation that you can use in main part of your program is kreduce

function and you have to use kreduce function in your main program. Better implementation of kreduce function will get better grades.

3. Output

You will give outputs for three things. First you will print out the computation time for sequential part of your program e.g. time for reading files. Secondly, you will print out the computation time for parallel part of your program e.g. calculating least k documents for the query. Lastly, you will print out least k ids. Output format should be as follows:

```
Sequential Part: 1.50 ms

Parallel Part: 5.22 ms

Total Time: 7.12 ms

Least k = 3 ids:

5

1

6
```

4. Submission

Send a single zip file (**yourname_lastname_p2.zip**) that includes:

- Your code:
  - utils.h, utils.c files which are implemented for reading input files.
  - main.c implementation of the program.
  - Your code should be compiled with the following command:
    - mpicc –o documentSearch main.c utils.h utils.c –lm
  - Your executable should be able run with the following command:
    - mpirun –n X ./documentSearch dictionarySize kValue documents.txt query.txt
  - Your code should include explanatory comments.
  - It is important that your program will work with these commands. For example, documents.txt will have a fomat like

    ```
    ...
    ...
    5: 24 56 3 6
    8: 2 12 23 8
    ...
    81: 52 34 6 13
    ...
    ```

From this you will parse out file contents and calculate similarities using the query.txt etc.
And you will create necessary input to your kreduce function using these files. Then kreduce will do the actual parallel processing.

- Put several example outputs of your program. Place all output examples in directory named as **outputs**. Each output example should include some information about input parameters of your program.
- Run your code with various parameters. Decide which parameters affects the performance of your program. Disccuss your reasoning for choosing these parameters in your report.
- Your report:
    - **Very detailed explanation of implementation of kreduce function.**
    - Explanation of your main program.
    - **Plot several graphs** for choosen parameters. (**Hint:** There are at least three different parameters that will affect your programs performance. You have already discovered two of them in your first homework e.g. number of processes.)
    - Discussion of selected parameters and and performance results according to these parameters. Remember that you have also taken sequential part execution and parallel part execution times. How can you interpret those values?

Email: kaan.akyol@bilkent.edu.tr

Email subject: CS426_Project2 (Without this subject, your project will not be evaluated).

Zip File name: yourname_lastname_p2.zip (Without this name, will not be evaluated).

No Late Submission Allowed!