PD2 2023

HW5

CSIE@NCKU 2023
Deadline: 2023/5/15 11:59pm

Homework Description:
In this homework, we exercise a simple search engine. Given a set of strings (in English) as training corpus, each string has a string ID (S_ID in the positive number format), and the context of each string will be contained in double quotation marks “”. A line contains S_ID and its corresponding string.
Example input corpus file, named corpus1.txt:

```
1, "A survey of user opinion of computer system response time"
2, "Relation of user perceived response time to error measurement"
3, "The generation of random binary unordered trees"
4, "The intersection graph of paths in trees"
5, "Graph minors IV Widths of trees and well quasi ordering"
```

Another input file will contain different queries, and each query may contain 1~3 keywords (separated by space). You should return the set of S_ID whose context contains all keywords.
Example query file, named query1.txt:

```
Survey time
Trees
Trees ordering
Ordered trees
```

Same as previous homeworks, use ‘cout’ to output in the console. We will use shell operator “>” to copy output to a created file named “result”. For the example input, the output should be:

```
1
3 4 5
5
-1
```

The output only considers “exactly contain” cases. “Partial contain” is not considered. If you don’t find any string contains all query keywords, return “-1” as the “Ordered trees” case in the example.

You don’t need to remove stop words. In addition, no word stem needs to be considered. Please let “capital letter” be equal to its small letter. For each query, the result should be ordered according to S_ID in the ascending order.

Deadline:
2023/5/15 11:59pm (Monday).
HW5 should be submitted before the deadline. No excuse to submit your code after the deadline. TA will copy your code at 2023/5/16 00:01. If your code is not in the hw5 Folder, you will get score ‘0’.

Environment:
1. uname -a
   Linux version 5.15.0-67-generic (buildd@lcy02-amd64-116) (gcc (Ubuntu 11.3.0-1ubuntu1~22.04) 11.3.0, GNU ld (GNU Binutils for Ubuntu) 2.38) #74-Ubuntu SMP Wed Feb 22 14:14:39 UTC 2023
2. IP: 140.116.246.230
3. Please remember you should connect to the server with a NCKU IP. Make sure you use NCKU VPN or connect to the server in our school.

Spec:

Note:
1. It is encouraged that you can “use chatGPT”.
2. A string may contain more than 200 words.
3. Any punctuation mark can be removed.
4. The corpus may be larger than 100 MB.
5. More than 10,000 queries could be used.
6. STL map or hash_map is a good implementation for indexing.
7. Check the TRIE structure if you have time.
8. S_ID is not always ordered, but it is a positive integer value.

9. Remember the strict output ordering policy. We will examine your correctness by our shell script without any excuse.
10. Please list the result to console. We will use the shell operator “>” to copy all your homework output to a created file named “result”.
11. You need to declare the executable file named “**hw5**” which will be generated by using your makefile with “make all”.
12. The csv file name will be given as the input argument without any exception of file handle error.
13. For the convenience of checking your homework, **output** must not contain any exceptional character, otherwise your score will be deducted.
14. We count the result correction as the base of scoring (100%).
15. The execution efficiency will also be counting. Top 10% submissions will get the bonus of 20% score.
16. The memory usage will also be counting. Top 10% small usages will get the bonus of 20% score.
17. If you know how to use scp (you may use Windows-based PowerShell), you could try scp in powershell like:

```
scp hw5.cpp ktchuang@140.116.246.230:~/hw5
```

How to Submit:

Please pay attention to the following instructions when submitting homework:

1. Under your account folder, create the folder named “**hw5**”.
   (*Please note that you must pay attention to the correct capitalization. If we cannot correctly copy the foler hw5, you will get score ‘0’.)

2. Put every necessary files under the folder, including :
   i. Your **main program**, such as main.cpp, hw5.cpp, main.h, program.h, etc.
   ii. **makefile** (*Name your executable file as “hw5”)
3. Make sure it works normally under the folder and all files are in the correct path.

Examples of files in your folder:

```
netdb@2023pd2:/home/vs6112030/hw5$ pwd
/home/vs6112030/hw5
netdb@2023pd2:/home/vs6112030/hw5$ ls
hw5  main.cpp  makefile
```

How we execute:

```
netdb@2023pd2:/home/vs6112030/hw5$ make all
netdb@2023pd2:/home/vs6112030/hw5$./hw5 corpus.txt query.txt >
result
```