Department of Computer Science and Engineering
Faculty of Engineering, University of Moratuwa

CS2022 – Data Structures and Algorithms
B. Sc. Engineering Semester 2

Programming Project

---

1 Introduction

1.1 The Idea

E-books are becoming more and more popular due to various reasons. In this project, you are required to create and maintain a keyword index for an e-book. The keywords are extracted from the book and are made available to you in the following format.

Keyword1	1
Keyword2	1
Keyword3	1
Keyword4	1
Keyword5	1
Keyword1	2
Keyword6	2
Keyword7	3
Keyword1	3
Keyword2	3

In the above list, each line contains a string and a number separated by a "tab" character. The string is the keyword and the number represents a page number the keyword appeared. You are expected to read the input file and store the keywords and the pages they appear. The system should provide the following functionality.

1. Given a keyword, output the first page the keyword appeared.
2. Given a keyword, output the sorted list of pages (in ascending order) a given keyword appeared.
3. Given a page number, output the sorted list of keywords (in ascending order) that appeared in the page.

The operations will be specified in the following format in the input file.

1. First<whitespace><Keyword>
    a. First Keyword7
2. List<whitespace><Keyword>
    a. List Keyword1
3. Keywords<whitespace><page_number>
    a. Keywords 2

The output of the operations should be displayed in the standard output in the following manner.

- The result of each query should be listed in a single line.
- For the second operation, the pages should be separated by a single whitespace.
- For the third operation, the keywords should be separated by a single whitespace.
- If the specified keyword is not present in the input, the output should be "Invalid Keyword".
- If the specified page number is not present in the input, the output should be "Invalid Page Number".

1.2 Additional Considerations:

- **The operations 1 and 2** (Searching the first page/pages the keyword appeared) **are the most frequent operations and needs to be implemented efficiently**. You have to consider that the number of keywords may be very high and all the keywords are specified at the start of the program.
- You are expected to implement all the data structures, and sorting and searching algorithms that you use in the program.
- Input to the program will be the name of the input file and all the outputs (of the queries) should be written to the standard output. Any output other than the results of queries (if there are any) should be written to a separate output file (not to the standard output).
- **For the implementation purposes**, you can assume that the maximum number of keywords to be 500 and the maximum number of pages to be 200.

2 Submissions Guidelines

You should submit the following to the assignment created in the Moodle.

1. Project Report

   A short (1 to 3 page) document which explains;

   a. The design of your solution.

   b. The data structures and algorithms you used and the rationale for selecting them.
   Please clearly indicate the data structures/algorithms you considered in your design and the reason you selected the one used in the final solution.

c. The assumptions you made.
d. The problems you faced (if any) and how you overcame them.
e. A short discussion on how your solution can be improved to improve the search performance for operations 1 and 2.

Please name your report as "<Your_Index_No>_CS2022PR.xxx" (e.g. 100112D_CS2022PR.doc)

2. A zip file which contains the following.
    a. A fully working program with the source code. All code should be commented, and all code files should start with a header comment that includes the name and the index number of the student.
    b. A Readme file which explains the steps required for compiling and executing the program. Please make sure you provide clear instructions on how to compile and execute your code.

    Please name your zip file as "<Your_Index_No>_CS2022SC.zip" (e.g. 100112D_CS2022SC.zip)

Important:

- You may use either "c" or "java" as the programming language for the implementation.
- You should submit a source code which can be compiled and executed in command line. The instructions you include in readme file should be how to compile and execute your program in command line. You can assume that a GNU C compiler/Java Development Kit (Java SE 7) is available.
- The programs may be graded using an automated grading tool. Therefore, please ensure that the output is exactly in the specified format.

3 Grading Policy

Following factors will be considered in evaluating your submission.

- The design of the solution.
- Use of most suitable data structures and algorithms in the solution.
    - Includes the explanation of why you selected them and the other options you selected.
- The correctness and quality of the submitted program.
- Adhering to the instructions.

All the submissions should be your own work. If plagiarism (Internet or peer) is detected, you will get **zero** marks for the complete project. Please make sure you acknowledge all the resources you referred.

4 Sample Input and Output

4.1 Input to the Program

Input.txt

4.2 Contents of the File "Input.txt"

```
Keyword List Start
Keyword1    3
Keyword2    1
Keyword3    1
Keyword4    1
Keyword5    1
Keyword1    2
Keyword3    2
Keyword6    2
Keyword7    3
Keyword1    1
Keyword2    3
Keyword List End
Queries Start
First Keyword1
First Keyword7
List Keyword1
List Keyword7
Keywords 1
Keywords 2
First Keyword9
List Keyword9
Keywords 5
Queries End
```

4.3 Output of the Program

```
1
3
1 2 3
3
Keyword1 Keyword2 Keyword3 Keyword4 Keyword5
Keyword1 Keyword3 Keyword6
Invalid Keyword
Invalid Keyword
Invalid Page Number
```