Data Structures (CSC 212)
Fall Trimester 2022
Course Project
25 Marks

Due Date: Phase 1 (07 October 2022 11:59 pm).
Due Date: Phase 2 (29 October 2022 11:59 pm).

In this project, you will be using Java to develop a text analysis tool that will read, as an input, a text file (provided in .txt format), store it in the main memory, and then perform several word analytics tasks such as determining the number of occurrences and the locations of different words. Therefore, the main task of this project is to design a suitable ADT (call it ***WordAnalysis ADT***) to store the words in the text and enable the following operations to be performed ***as fast as*** possible:

(1) An operation to determine the total number of words in a text file (i.e., the length of the file).

(2) An operation to determine the total number of unique words in a text file.

(3) An operation to determine the total number of occurrences of a particular word.

(4) An operation to determine the total number of words with a particular length.

(5) An operation to display the unique words and their occurrences sorted by the total occurrences of each word (from the most frequent to the least).

(6) An operation to display the locations of the occurrences of a word starting from the top of the text file (i.e., as a list of line and word positions). Note that every new-line character '\n' indicates the end of a line.

(7) An operation to examine if two words are occurring adjacent to each other in the file (at least one occurrence of both words is needed to satisfy this operation).

Example:

Consider the following text:

"In computer science, a data structure is a collection of data values, the relationships among them, and the functions or operations that can be applied to the data."

The output of operation (1) would be 28.

The output of operation (2) would be 23.

The output of operation (3) for the word 'the' would be 3.

The output of operation (4) for word length 2 would be 6.

The output of operation (5) would be (the, 3), (data, 3), (a, 2), (in, 1), (computer, 1), (science, 1), (structure, 1) .... etc.

The output of operation (6) for the word 'data' would be (1, 5), (1, 11), (2, 14).

The output of operation (7) for the two words 'data' and 'the' would be True.

Remarks:

Assume that

- words are separated by at least one space.
- Single letter words (e.g., a, I) are counted as words.
- Punctuation (e.g., commas, periods, etc.) is to be ignored.
- Hyphenated words (e.g., decision-makers) or apostrophized words (e.g., customer's) are to be read as single words.

Phase 1 (10 Marks)

In the first phase of the project, you are asked to describe your suggested design of the ADT for the problem described above and perform the following tasks:

(a) Give a graphical representation of the ADT to show its structure. Make sure to label the diagram clearly.

(b) Write at least one paragraph describing your diagram from part (a). Make sure to clearly explain each component in your design. Also, discuss and justify the choices and the assumptions you make.

(c) Give a specification of the operations (1), (2), (3), (4), (5), (6), and (7) as well as any other supporting operations you may need to read the text from a text file and store the results in the ADT (e.g., insert).

(d) Provide the time complexity (worst case analysis) for all the operations discussed above using Big O notation. For operations (3) and (4), consider two cases: the first case, when the words in the text file have lengths that are evenly distributed among different lengths (i.e., the words should have different lengths starting from 1 to the longest with k characters), and the second case, when the lengths of words are not evenly distributed. For all operations, assume that the length of the text file is $\boldsymbol{n}$, the number of unique words is $\boldsymbol{m}$, and the longest word in the file has a length of $\boldsymbol{k}$ characters.

Phase 2 (15 Marks)

In the second phase of the project, you will implement the ADT and the operations specified in the first phase using Java programming language. Your program should contain a Main class that reads a specified user text file and perform each of the seven operations that are described in this assignment. **You will be asked to run a demo of your program and generate output using several test cases provided by the instructor. Failure to show for your demo will result in receiving ZERO for this phase.**

Note that All data structures used in this assignment must be implemented by the students and any use of Java collections or any other libraries is strictly forbidden (you may only use the ADTs you have studied in the course).

Submission Guidelines

- You are expected to work in teams, each team **must** contain two students.
- For **Phase 1**, submit a written report (in PDF format) answering the given questions to your instructor through LMS (use the instructor's section in LMS and DO NOT use the merged section). Your report should include the names and the IDs of your team's member.
- For **Phase 2**, submit your source code (compressed in ZIP format) to your instructor through LMS (use the instructor's section in LMS and DO NOT use the merged section). In addition, make sure to bring your LAPTOP, so you can present a demo of your work to your instructor.