Data Structures and Algorithms

LEXICON AND TEXT ANALYSIS

REQUIREMENTS:

You are tasked with writing a program that takes as input a possibly large text document and creates as output an alphabetized index showing the frequency that each word appears.

As a brief example:

Input: "How much wood could a woodchuck chuck if a woodchuck could chuck wood?"

Output:

A	2
Chuck	2
Could	2
How	1
If	1
Much	1
Wood	2
Woodchuck	2

Your program should additionally calculate a statistical report for the file processed containing the following information:

- total number of words
- number of unique words
- number of unique words of more than three letters
- average word length
- average sentence length

SPECIFICATIONS:

You must use the binary search tree ADT to solve this problem.

Input

From the keyboard: The name of the file containing the text to be analyzed.

Output

Write the following information to a file *and* print to the screen:

1. The name of the file
2. A listing of the file
3. The lexicon with word frequencies
4. The statistical summary

Definitions

Word:

A sequence of letters ending in a blank, a period, an exclamation point, a question mark, a colon, a comma, a single quote, a double quote, or a semicolon. Numbers do not appear in the words; they may be ignored.

Unique word:

Words that are spelled the same, ignoring uppercase and lowercase distinctions.

Sentence:

All words between any two periods, exclamation points, or question marks.

Deliverables

1. Listing of the source code of your program.
2. Sample Run.
3. Test Plan.