CSCI 36200: Data Structures

Programming Assignment 2

Instructor: Dr. Snehasis Mukhopadhyay **Due date: March 21, 2017**

The objective of this programming project is to construct a *cross-reference index* for a given text file. Such index structures have applications in the design of compilers and databases.

Our task is to write a program that while reading a text file collects all words of the text and retains the numbers of the lines in which each word occurred. When this scan is terminated, a table is printed showing all collected words in alphabetical order with lists of line numbers where they occurred. There would be only one line for each word.

Represent the words encountered in the text by a binary search tree (also called a lexicographic tree). For example, if there were three words 'abracadabra', 'hocuspocus', and 'watchamaycalli', 'hocuspocus' would be the root with 'abracadabra' its left child and 'watchamaycalli' its right child. Each node not only contains a word as key value, but is also the head of a list of line numbers.

Make the following assumptions regarding the text file:

1. Only lower-case letters, digits, punctuation marks (*e.g.*, '.', ',' *etc.*), blanks, and the newline control character '\n' are present in the input text file. The end of the file is signified by the character '#'.

2. A word is considered as any sequence of letters and digits starting with a letter. The end of a word is marked by a blank, a punctuation mark, or a newline character.

3. Punctuation marks are not part of any word.

4. It is desirable that, for a particular word, the line numbers are printed in ascending order in the cross-reference index.

5. Each word is at most ten characters long. All characters beyond that in a word are ignored. Two words are the same if their first ten characters match.

A sample run of the program on a text file

civilization of science.
science is knowledge.
knowledge is our destiny.
#

should produce

civilizati	*1*	
destiny	*3*	
is	*2*	*3*
knowledge	*2*	*3*
means	*2*	
of	*1*	
our	*3*	
science	*1*	*2*

Check the correctness of your program on the following text, which represents a code written in an as yet undiscovered programming language called Almost-English:

set an integer variable x1 to be 2.
set y2 to be 2.
yes, y2 is a variable.
yes, y2 is an integer.
add y2 to x1.
print the result.
convert the result to an integer.
thank you.
#

You should submit the source code and the output file. You are required also to submit a project report consisting of: (a) the project description (in your words), and (b) any other comments and conclusions that you may prefer to include. The report should be roughly about two printed pages.

As usual, you will be graded for the project on the following points: (i) correctness of the program, (ii) organization of the program, (iii) readability of the program, and (iv) your report. Your code will be tested to verify correctness.