CS69003: Computing Systems Lab I
Autumn 2008

Assignment 2

Implementation of text search using hash tables

Due: August 13, 2008 (Wednesday)

In this exercise, you are required to write a C program to make an indexed search in several text files. The behavior of your program would simulate the working of modern search engines for the Internet.

Part I **(50)**

You are given a set of text files. Isolate the words in the files and store the frequencies of the words in different files in a hash table. Each entry in the hash table must store a word along with its frequency of occurrence in each of the database files. A linked list may be used to store the list of frequencies of a word. A file not containing a word must not be present in the list associated with the word.

Let the user enter a word. Assume that the word occurs in one or more of the database files. List all the files in which the query word appears. Your output listing should be sorted in the decreasing order of the frequency of the query word in the database files. That is, files with larger numbers of occurrences of the query word should be listed earlier than those with smaller numbers of occurrences.

For simplicity, make your search case-insensitive. This means that the words 'computer', 'Computer' and 'COMPUTER' are treated the same so long as your search is concerned. Assume also that each word consists of alphabetic letters (a–z and A–Z) only.

Part II **(50)**

Now, suppose that the query word entered by the user exists in none of the database files. In this case, your program should notify the user that no database file matches the query. In addition, your program should supply a list of words that are present in the database and have the edit distance one (defined below) from the query word.

A basic operation on a word is the following: changing a single letter in the word, inserting a letter at some position in the word and deleting a single letter from the word. For example, each of the words 'conputer', 'compuiter' and 'cmputer' is obtained by a single basic operation on the word 'computer'. The edit distance between two words w_1 and w_2 is the minimum number of basic operations that need to be applied on w_1 in order to obtain w_2. Thus, two words w_1 and w_2 are at edit distance one, if a single basic operation transforms w_1 to w_2. Note that the notion of edit distance is symmetric with respect to its two arguments, that is, the edit distance from w_1 to w_2 is the same as the edit distance from w_2 to w_1.

You are required to submit a file with the name `<your_roll_no>-assgn2.c` solving both the parts of this assignment.