# SDL Research Programming Test

**Problem 1**

Write a program that reads in a text file and prints out the 10 most frequent words from that file together with their counts.

A word is defined as a space-separated token.

If several words have the same frequency, they are considered to occupy a single “slot” from 1 to 10; and the program should print them all out.

The program should take as input argument a file, and print to STDOUT the list of words and their counts.

**Problem 2**

Write a program that computes the edit distance between two strings.

The edit distance is defined by the series of operations that transforms the second string into the first one. The allowed operations are:

* Insertion of a character
* Deletion of a character
* Substitution of a character for another

For example, the string “kitten” can be transformed into “splitting” as follows: substitute “k” with “s”, insert “p” at position 1, insert “l” at position 2, substitute “e” with “I”, and insert “g” at the end. The sequence of operations is not unique; another possibility is to delete the “k” (one operation), then insert “s” “p” and “l”. Each operation incurs a cost.

The edit distance is the cost of the minimum-cost series of operations that accomplishes the desired transformation.

For this problem, the strings will be composed only of letters a-zA-Z. The costs of the operations are as follows:

* insertion of a symbol, cost 3
* deletion of a symbol, cost 2
* substitution of a vowel with another vowel, cost 0.5 (vowels are aeiouAEIOU)
* substitution of a symbol with another (where not both are vowels), cost 1

The program should take as arguments the two strings, and write out the edit distance (the cost of the minimum-cost sequence of operations), as well as the operations themselves.

Here is the output that should be generated foir the example above:

Edit distance: 10.5

Operations:

Replace character from position 0 with character s

Insert character p at position 1

Insert character l at position 2

Replace character from position 6 with e

Insert character g at position 8

If there are several sequences with the same cost, print them all out.