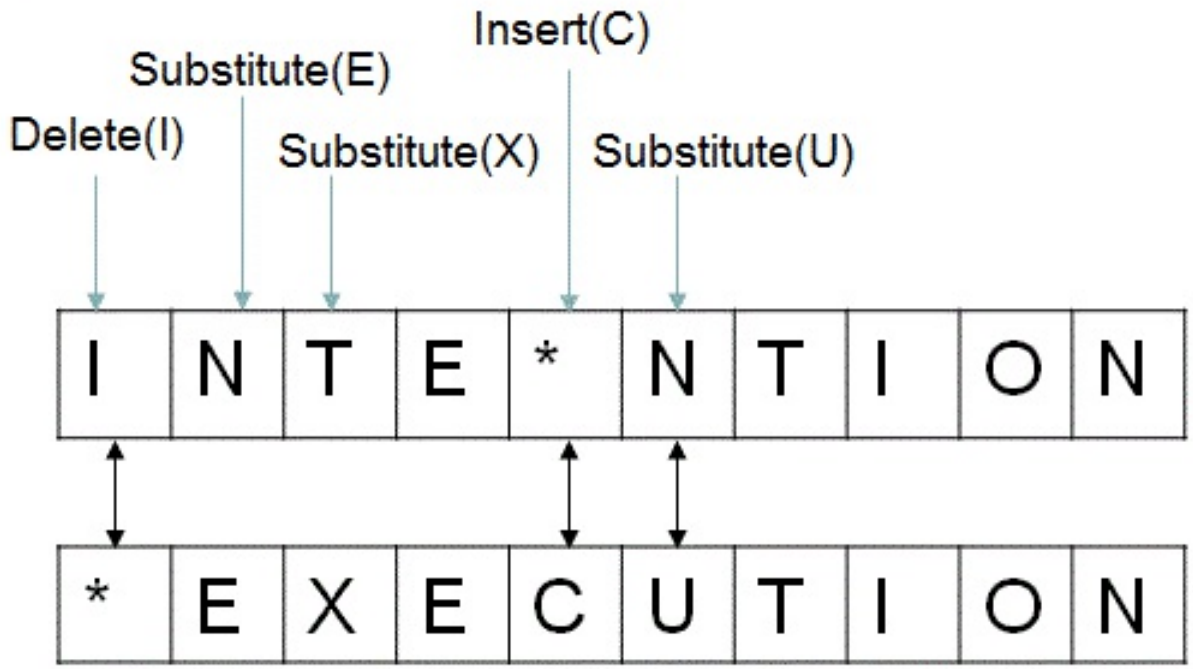

Given two strings, write a program to find out the minimum edit distance between them.

Applications

File Revision

The Unix command `diff f1 f2` finds the *difference* between files f1 and f2, producing an *edit script* to convert f1 into f2. If two (or more) computers share copies of a large file F, and someone on machine-1 edits `F=F.bak`, making a few changes, to give `F.new`, it might be very expensive and/or slow to transmit the whole revised file `F.new` to machine-2. However, `diff F.bak F.new` will give a *small* edit script which can be transmitted quickly to machine-2 where the local copy of the file can be updated to equal `F.new`.

`diff` treats a whole line as a "character" and uses a special edit-distance algorithm that is fast when the "alphabet" is large and there are few chance matches between elements of the two strings (files). In contrast, there are many chance character-matches in DNA where the alphabet size is just 4, {A,C,G,T}.

Try `` `man diff' `` to see the manual entry for diff.

Remote Screen Update Problem

If a computer program on machine-1 is being used by someone from a screen on (distant) machine-2, e.g. via `rlogin` etc., then machine-1 may need to update the screen on machine-2 as the computation proceeds. One approach is for the program (on machine-1) to keep a "picture" of what the screen currently is (on machine-2) and another picture of what it should become. The differences can be found (by an algorithm related to edit-distance) and the differences transmitted... saving on transmission band-width.

Spelling Correction

Algorithms related to the edit distance may be used in spelling correctors. If a text contains a word, w, that is not in the dictionary, a `close' word, i.e. one with a small edit distance to w, may be suggested as a correction.

Transposition errors are common in written text. A transposition can be treated as a deletion plus an insertion, but a simple variation on the algorithm can treat a transposition as a single point mutation.

Plagiarism Detection

The edit distance provides an indication of similarity that might be too close in some situations ... think about it.

Molecular Biology

The edit distance gives an indication of how `close' two strings are. Similar measures are used to compute a distance between DNA sequences (strings over {A,C,G,T}, or protein sequences (over an alphabet of 20 amino acids), for various purposes, e.g.:

1. to find genes or proteins that may have shared functions or properties
2. to infer family relationships and evolutionary trees over different organisms

Example

An example of a DNA sequence from `Genebank' can be found [here]. The simple edit distance algorithm would normally be run on sequences of *at most* a few thousand bases.

Speech Recognition

Algorithms similar to those for the edit-distance problem are used in some speech recognition systems: find a close match between a new utterance and one in a library of classified utterances.