String Editing

String Editing

given two strings $X = x_1, x_2, \ldots, x_n$ and $Y = y_1, y_2, \ldots, y_m$, where x_i, $1 \leq i \leq n$, and y_j, $1 \leq j \leq m$, are members of a finite set of symbols known as the *alphabet*.

Example : X = REAP and Y = CREAM ; n= 4 and m = 5

To transform X into Y using a sequence of *edit operations* on X The permissible edit operations are insert, delete, and change (a symbol of X into another)

there is a cost associated with performing each.

The cost of a sequence of operations is the sum of the costs of the individual operations in the sequence. The problem of string editing is to identify a minimum-cost sequence of edit operations that will transform X into Y.

Let $D(x_i)$ be the cost of deleting the symbol x_i from X, $I(y_j)$ be the cost of inserting the symbol y_j into X, and $C(x_i, y_j)$ be the cost of changing the symbol x_i of X into y_j.

Solution

Define $cost(i, j)$ to be
the minimum cost of any edit sequence for transforming $x_1, x_2, \ldots, x_i$ into $y_1, y_2, \ldots, y_j$ (for $0 \leq i \leq n$ and $0 \leq j \leq m$). Compute $cost(i, j)$ for each i and j. Then $cost(n, m)$ is the cost of an optimal edit sequence.

For i = 0 and j = 0 (Both are empty String and hence identical) cost(i, j) = 0

For i > 0 and j = 0 (Transforming a string to empty string – transform X into Y by a sequence of deletes)

$$cost(i, 0) = cost(i-1, 0) + D(x_i)$$

For i = 0 and j > 0 (Transforming an empty string to another string – transform X into Y by a sequence of insertions)

$$cost(0, j) = cost(0, j-1) + I(y_j)$$

If (i ≠ 0 and j ≠ 0) and $x_i = y_j$ (characters are same – no need for edit operation)

cost(i j) = cost(i -1, j-1)

If $i \neq 0$ and $j \neq 0$, $x_1, x_2, \ldots, x_i$ can be transformed into $y_1, y_2, \ldots, y_j$ in one of three ways:

1. Transform $x_1, x_2, \ldots, x_{i-1}$ into $y_1, y_2, \ldots, y_j$ using a minimum-cost edit sequence and then delete x_i. The corresponding cost is
$$cost(i-1, j) + D(x_i)$$

2. Transform $x_1, x_2, \ldots, x_{i-1}$ into $y_1, y_2, \ldots, y_{j-1}$ using a minimum-cost edit sequence and then change the symbol x_i to y_j. The associated cost is $cost(i-1, j-1) + C(x_i, y_j)$.

3. Transform $x_1, x_2, \ldots, x_i$ into $y_1, y_2, \ldots, y_{j-1}$ using a minimum-cost edit sequence and then insert y_j. This corresponds to a cost of
$$cost(i, j-1) + I(y_j)$$

Recurrence Relation

$$cost(i,j) = \begin{cases} 0 & i = j = 0 \\ cost(i-1,0) + D(x_i) & j = 0,\ i > 0 \\ cost(0,j-1) + I(y_j) & i = 0,\ j > 0 \\ cost(i-1,\ j-1) & i > 0,\ j > 0 \quad \text{and } x_i = y_j \end{cases}$$

$$\min \{ \quad cost(i-1,j) + D(x_i), \quad cost(i-1,j-1) + C(x_i,y_j), \quad cost(i,j-1) + I(y_j) \ \} \qquad i > 0,\ j > 0 \quad \text{and } x_i \neq y_j$$

Example: X= REAP Y= CREAM

$$cost(i,j) = \begin{cases} 0 & i=j=0 \\ cost(i-1,0)+D(x_i) & j=0,\ i>0 \\ cost(0,j-1)+I(y_j) & i=0,\ j>0 \\ cost(i-1,j-1) & i>0,\ j>0 \text{ and } x_i = y_j \\ \min\{ cost(i-1,j)+D(x_i), \\ \quad cost(i-1,j-1)+C(x_i,y_j), & i>0,\ j>0 \text{ and } x_i \neq y_j \\ \quad cost(i,j-1)+I(y_j) \} \end{cases}$$

Let $D(x_i) = 1$
$I(y_j) = 1$
$C(x_i, y_j) = 2$

	0	C 1	R 2	E 3	A 4	M 5
0	0	1	2	3	4	5
R 1	1	2	1	2	3	4
E 2	2	3	2	1	2	3
A 3	3	4	3	2	1	2
P 4	4	5	4	3	2	**3**

Left : Insert
Up : Delete
Diag : Change

Edit Operations

	0	C 1	R 2	E 3	A 4	M 5
0	0	1	2	3	4	5
R 1	1	2	1	2	3	4
E 2	2	3	2	1	2	3
A 3	3	4	3	2	1	2
P 4	4	5	4	3	2	**3**

X : R E A P

R E A P M (Left : insert M)

R E A M (Up : Delete P)

C R E A M (Left : Insert C) => Y

Exercise

- Transform String “BASKET” to “BARK” using minimum edit operations.