Longest Common Subsequence

A subsequence of a given sequence is the given sequence with just some elements left out (order should be from left-to-right, not necessarily consecutive).. A common sequence of two sequences X and Y, is a subsequence of both X and Y.A longest common subsequence is the one with maximum length. For example, if X = {A,B,C,B,D,A,B} and Y = { B,D,C,A,B,A} then the longest common subsequence is of length 4 and they are {B,C,B,A} and {B,D,A,B}.

Finding the longest common subsequence has applications in areas like biology. The longest subsequence (LCS) problem has an optimal substructure property. Thus, dynamic programming method can be used to solve this problem.

Theorem used - Let X =< x1, x2, . . . , xm > and Y =< y1, y2, . . . , yn > be sequences, and let Z =< z1, z2, . . . , zk > be any LCS of X and Y .

1. If $x_m = y_n$, then $z_k = x_m = y_n$ and Z_{k-1} is an LCS of X_{m-1} and Y_{n-1}.
2. If $x_m = y_n$, then $z_k = x_m$ implies that Z is an LCS of X_{m-1} and Y.
3. If $x_m = y_n$, then $z_k = y_n$ implies that Z is an LCS of X and Y_{n-1}.

Algorithm:

S,T are two strings for which we have to find the longest common sub sequence. Input the two sequences. Now print the longest common subsequence using LongestCommonSubsequence function.

LongestCommonSubsequence function : This function takes the two sequences (S, T) as arguments and returns the longest common subsequence found.

Store the length of both the subsequences. Slength = strlen(S), Tlength = strlen(T).
We will Start with the index from 1 for our convenience (avoids handling special cases for negative indices).
Declare common[Slength][Tlength]. Where, common[i][j] represents length of the longest common sequence in S[1..i], T[1..j].
If there are no characters from string S, common[0][i]=0 for all i or if there are no characters from string T, common[i][0]=0 for all i.
Recurrence: for i=1 to Slength
for j=1 to Tlength
common[i][j] = common[i-1][j-1] + 1, if S[i]=T[j]. Else, common[i][j] = max(common[i-1][j],common[i][j-1]). Where max is a function which takes the two variables as arguments and returns the maximum of them.
Return, common[Slength][Tlength].

Property:

Time complexity is O(mn), where m and n are the length of two strings.

Example:

$S = \{A, B, C, B\}$, $T = \{B, D, C, A\}$

SLength = 4 and TLength = 4, common[0][0…4] = 0 and common[0…4][0] = 0

Recurrence

$i = 1$
$j = 1$
$A \neq B$, common[1][1] = max(common[1][0] , common[0][1]) = 0
$j = 2$
$A \neq D$, common[1][2] = max(common[1][1] , common[0][2]) = 0
$j = 3$
$A \neq C$, common[1][3] = max(common[1][2] , common[0][3]) = 0
$j = 4$
$A = A$, common[1][4] = common[0][2] + 1 = 1
$i = 2$
$j = 1$
$B = B$, common[2][1] = common[1][0] + 1 = 1
$j = 2$
$B \neq D$, common[2][2] = max(common[2][1] , common[1][2]) = 1
$j = 3$
$B \neq C$, common[2][3] = max(common[2][2] , common[1][3]) = 1
$j = 4$
$B \neq A$, common[2][4] = max(common[2][3] , common[1][4]) = 1
$i = 3$
$j = 1$
$C \neq B$, common[3][1] = max(common[3][0] , common[2][1]) = 1
$j = 2$
$C \neq D$, common[3][2] = max(common[3][1] , common[2][2]) = 1
$j = 3$
$C = C$, common[3][3] = common[2][2] + 1 = 2
$j = 4$
$C \neq A$, common[3][4] = max(common[3][3] , common[2][4]) = 2
$i = 4$
$j = 1$
$B \neq B$, common[4][1] = max(common[4][0] , common[3][1]) = 1
$j = 2$
$B \neq D$, common[4][2] = max(common[4][1] , common[3][2]) = 1
$j = 3$
$B \neq C$, common[4][3] = max(common[4][2] , common[3][3]) = 2

$j = 4$

B ≠ A, common[4][4] = max(common[4][3] , common[3][4]) = 2

common[4][4] = 2

Output: Longest common susequence is of length 2

An example: X_i = A B C B D A B and Y_i = B D C A B A

i/j	0	1	2	3	4	5	6		
	Y_j	B	D	C	A	B	A		
	0	0	0	0	0	0	0	0	X_i
	0	0	0	0	1	1	1	1	A
	0	1	1	1	1	2	2	2	B
	0	1	1	2	2	2	2	3	C
	0	1	1	2	2	3	3	4	B
	0	1	2	2	2	3	3	5	D
	0	1	2	2	3	3	4	6	A
	0	1	2	2	3	4	4	7	B

To reconstruct the elements of an LCS, follow the matrix arrows from the lower right-hand corner; the path is shaded. Each “↖ “on the path corresponds to an entry (highlighted) for which $X_i = Y_i$, is a member of LCS. The length of the common subsequence is 4.