

問題 12 - 編輯距離(Hamming Distance)

(15 分)

前言

在人工智慧的領域中, 模糊的快速比對是極常見的運算. 因此常需要計算兩個字串間的相似或相異度.

編輯距離是用來表示字串間相異度的一種方法(編輯距離指的是經過多少操作後可以把兩個字串變成一樣).

問題描述

請依下列定義製作一個程式來計算將第一個字串轉換成第二個字串兩字串的編輯距離:

若兩個字串 A 與 B, 最少需要 X 的編輯代價, 才能把 A 完全變成 B, 則稱 A 字串到 B 字串的編輯距離為 X.

可進行的編輯動作有下列三種, 每做一次都要付出相對的編輯代價:

編輯動作	編輯代價	說明
刪除	3	刪除一個位置上的一個字元
插入	4	在某一位置插入任一個字元
取代	5	將某一位置的字元取代成任一個字元

輸入格式

第一個字串

第二個字串

輸出格式

X

X = 正整數, 編輯距離

資料範圍

輸入資料可為任何可由鍵盤輸入的字串, 含英文, 數字, 以及常用符號。

資料範例

輸入範例 1

ABCDEFK

BCGDHK

輸出範例 1

15

輸入範例 2

There is no good player.
Who say the player is not good?

輸出範例 2

92

範例解釋

範例 1 的輸出 15, 表示"ABCDEFK" 到 "BCGDHK" 的編輯距離為 15, 因為最少需要編輯代價 15(刪除 x2, 插入 x1, 取代 x1).

將"ABCDEFK"編輯成"BCGDHK")最佳編輯過程有好幾種, 下面是其中一個例子:

原始字串 : "ABCDEFK"

編輯動作 1 : 將位置 0(此時為'A')的字元刪除 ==> "BCDEFK" 3

編輯動作 2 : 在位置 1(此時為'C')之後插入'G' ==> "BCGDEFK" $4 + 3 = 7$

編輯動作 3 : 將位置 4(此時為'E')取代成'H' ==> "BCGDHFK" $5 + 7 = 12$

編輯動作 4 : 將位置 5(此時為'F')的字元刪除 ==> "BCGDHK" $3 + 12 = 15$

範例 2 則可以用 3 次取代, 7 次刪除, 14 次插入將 "There is no good player." 轉換成 "Who say the player is not good?",

因此編輯距離是 $3 \times 5 + 7 \times 3 + 14 \times 4 = 92$. 沒有更小的, 所以輸出 92