Edit Distance

Given two strings, how many edits are needed to turn one string into another?

`SNOWY vs SUNNY`

Edit Distance

Need:

- Strings – Snowy, Sunny
- Cost function - character misalignment = +1

What are the costs of these two different alignments?

S – N O W Y	– S N O W – Y
S U N N – Y	S U N – – N Y
cost = ?	cost = ?

Edit Distance

Need:

- Strings – Snowy, Sunny
- Cost function - character misalignment = +1

What are the costs of these two different alignments?

S	–	N	O	W	Y			–	S	N	O	W	–	Y
S	U	N	N	–	Y			S	U	N	–	–	N	Y

cost = 3

cost = 5

Edit distance = cheapest possible alignment.

Edit Distance

Need:

- Strings – Snowy, Sunny
- Cost function - character misalignment = +1

What are the costs of these two different alignments?

S	–	N	O	W	Y			–	S	N	O	W	–	Y
S	U	N	N	–	Y			S	U	N	–	–	N	Y

cost = 3

cost = 5

Does a brute force solution sound like a good idea?

Edit Distance

We want to align two strings, $x = [x_1, \ldots, x_n]$ and $y = [y_1, \ldots, y_m]$.

Edit Distance

We want to align two strings, $x = [x_1, \ldots, x_n]$ and $y = [y_1, \ldots, y_m]$.

Dynamic Programming?

Edit Distance

We want to align two strings, $x = [x_1, \dots, x_n]$ and $y = [y_1, \dots, y_m]$.

$E(i, j)$ = optimal cost of aligning $[x_1, \dots, x_i]$ and $[y_1, \dots, y_j]$.

Edit Distance

We want to align two strings, $x = [x_1, \ldots, x_n]$ and $y = [y_1, \ldots, y_m]$.

$E(i, j)$ = optimal cost of aligning $[x_1, \ldots, x_i]$ and $[y_1, \ldots, y_j]$.

Can we say anything about optimal alignment of $[x_1, \ldots, x_i]$ and $[y_1, \ldots, y_j]$?

Edit Distance

We want to align two strings, $x = [x_1, \dots, x_n]$ and $y = [y_1, \dots, y_m]$.

$E(i, j)$ = optimal cost of aligning $[x_1, \dots, x_i]$ and $[y_1, \dots, y_j]$.

Can we say anything about optimal alignment of $[x_1, \dots, x_i]$ and $[y_1, \dots, y_j]$?

Specifically, how must the optimal alignments end?
(three possibilities).

Edit Distance

We want to align two strings, $x = [x_1, \ldots, x_n]$ and $y = [y_1, \ldots, y_m]$.

$E(i, j)$ = optimal cost of aligning $[x_1, \ldots, x_i]$ and $[y_1, \ldots, y_j]$.

Can we say anything about optimal alignment of $[x_1, \ldots, x_i]$ and $[y_1, \ldots, y_j]$?

Optimal alignments end in one of three ways:

	x_i	$-$	x_i
	$-$	y_j	y_j
Cost:	1	1	0,1

Edit Distance

We want to align two strings, $x = [x_1, \dots, x_n]$ and $y = [y_1, \dots, y_m]$.

$E(i, j)$ = optimal cost of aligning $[x_1, \dots, x_i]$ and $[y_1, \dots, y_j]$.

Can we say anything about optimal alignment of $[x_1, \dots, x_i]$ and $[y_1, \dots, y_j]$?

Optimal alignments end in one of three ways:

$$\begin{matrix} x_i & - & x_i \\ - & y_j & y_j \end{matrix}$$

Need to align $[x_1, \dots, x_{i-1}]$ with $[y_1, \dots, y_{j-1}]$:

$$\boldsymbol{E(i-1, j-1)}$$

Edit Distance

We want to align two strings, $x = [x_1, \dots, x_n]$ and $y = [y_1, \dots, y_m]$.

$E(i, j)$ = optimal cost of aligning $[x_1, \dots, x_i]$ and $[y_1, \dots, y_j]$.

Can we say anything about optimal alignment of $[x_1, \dots, x_i]$ and $[y_1, \dots, y_j]$?

Optimal alignments end in one of three ways:

$$\begin{array}{ccc} x_i & - & x_i \\ - & y_j & y_j \end{array}$$

Need to align $[x_1, \dots, x_i]$ with $[y_1, \dots, y_{j-1}]$: $E(i, j-1)$

Edit Distance

We want to align two strings, $x = [x_1, \dots, x_n]$ and $y = [y_1, \dots, y_m]$.

$E(i, j)$ = optimal cost of aligning $[x_1, \dots, x_i]$ and $[y_1, \dots, y_j]$.

Can we say anything about optimal alignment of $[x_1, \dots, x_i]$ and $[y_1, \dots, y_j]$?

Optimal alignments end in one of three ways:

x_i	$-$	x_i
$-$	y_j	y_j

Need to align $[x_1, \dots, x_{i-1}]$ with $[y_1, \dots, y_j]$: $E(i-1, j)$

Edit Distance

$$E(i,j) = \min \left\{ \quad ? \right.$$

Edit Distance

$$\begin{matrix} x_i & - & x_i \\ - & y_j & y_j \end{matrix}$$

$$E(i,j) = \min \begin{cases} ? \end{cases}$$

Need to align $[x_1, \ldots, x_{i-1}]$ with $[y_1, \ldots, y_j]$: $E(i-1, j)$

Edit Distance

$$\begin{matrix} x_i & - & x_i \\ - & y_j & y_j \end{matrix}$$

$$E(i,j) = \min \begin{cases} E(i-1,j)+1 \\ E(i,j-1)+1 \\ E(i-1,j-1)+\mathrm{diff}(i,j) \end{cases}$$

where $\mathrm{diff}(i,j) = \begin{cases} 0, & x_i = y_j \\ 1, & x_i \neq y_j \end{cases}$

Edit Distance

$$\begin{array}{ccc} x_i & - & x_i \\ - & y_j & y_j \end{array}$$

$$E(i,j) = \min \begin{cases} E(i-1,j) + 1 \\ E(i,j-1) + 1 \\ E(i-1,j-1) + \text{diff}(i,j) \end{cases}$$

$$\text{where diff}(i,j) = \begin{cases} 0, & x_i = y_j \\ 1, & x_i \neq y_j \end{cases}$$

Finding $E(n,m)$ requires finding all the other E's, which can be represented in a 2d table with the strings along the axes.

Edit Distance

i \ j		0	1	2	3	4	5
			S	U	N	N	Y
0		0					
1	S						
2	N						
3	O					●	
4	W						
5	Y						

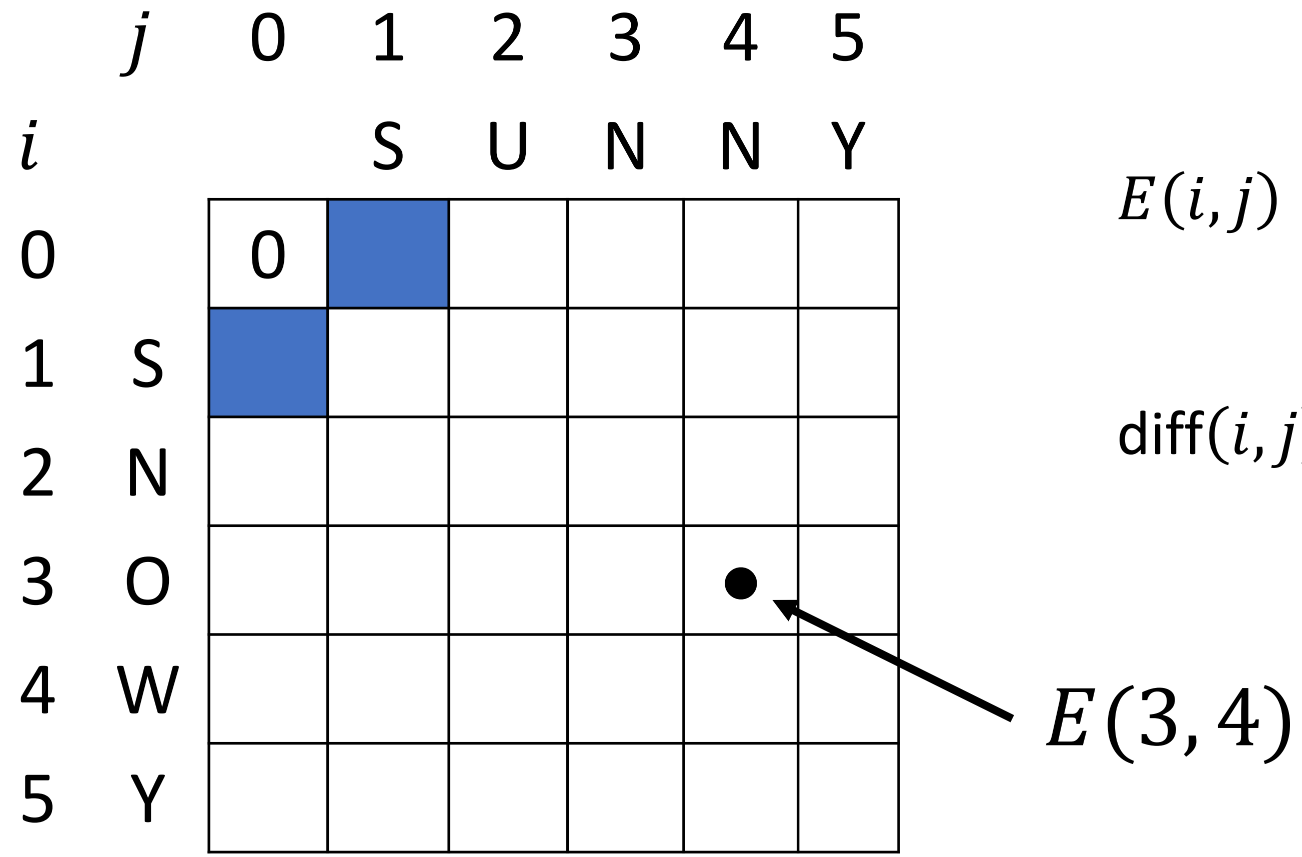
$E(3,4)$

$$E(i,j) = \min \begin{cases} E(i-1,j)+1 \\ E(i,j-1)+1 \\ E(i-1,j-1)+\text{diff}(i,j) \end{cases}$$

$$\text{diff}(i,j) = \begin{cases} 0, & x[i] = y[j] \\ 1, & x[i] \neq y[j] \end{cases}$$

Where can we start?

Edit Distance



$$E(i,j) = \min \begin{cases} E(i-1,j)+1 \\ E(i,j-1)+1 \\ E(i-1,j-1)+\text{diff}(i,j) \end{cases}$$

$$\text{diff}(i,j) = \begin{cases} 0, & x[i] = y[j] \\ 1, & x[i] \neq y[j] \end{cases}$$

Where can we start?

$E(0,1)$ or $E(1,0)$

Edit Distance

i \ j		0	1	2	3	4	5
			S	U	N	N	Y
0		0					
1	S						
2	N						
3	O						
4	W						
5	Y						

$$E(i,j) = \min \begin{cases} E(i-1,j)+1 \\ E(i,j-1)+1 \\ E(i-1,j-1)+\text{diff}(i,j) \end{cases}$$

$$\text{diff}(i,j) = \begin{cases} 0, & x[i] = y[j] \\ 1, & \text{otherwise} \end{cases}$$

$$E(0,1) = \min \begin{cases} E(-1,1)+1 \\ E(0,0)+1 \\ E(-1,0)+1 \end{cases} = ?$$

Edit Distance

i \ j		0	1	2	3	4	5
			S	U	N	N	Y
0		0					
1	S						
2	N						
3	O						
4	W						
5	Y						

$$E(i,j) = \min \begin{cases} E(i-1,j)+1 \\ E(i,j-1)+1 \\ E(i-1,j-1)+\text{diff}(i,j) \end{cases}$$

$$\text{diff}(i,j) = \begin{cases} 0, & x[i] = y[j] \\ 1, & \text{otherwise} \end{cases}$$

$$E(0,1) = \min \begin{cases} \cancel{E(-1,1)+1} \\ E(0,0)+1 \\ \cancel{E(-1,0)+1} \end{cases} = ?$$

Edit Distance

	j	0	1	2	3	4	5
i			S	U	N	N	Y
0		0	1				
1	S						
2	N						
3	O						
4	W						
5	Y						

$$E(i,j) = \min \begin{cases} E(i-1,j)+1 \\ E(i,j-1)+1 \\ E(i-1,j-1)+\text{diff}(i,j) \end{cases}$$

$$\text{diff}(i,j) = \begin{cases} 0, & x[i]=y[j] \\ 1, & \text{otherwise} \end{cases}$$

$$E(0,1) = \min \begin{cases} \cancel{E(-1,1)+1} \\ E(0,0)+1 \\ \cancel{E(-1,0)+1} \end{cases} = 1$$

Edit Distance

i \ j		0	1	2	3	4	5
			S	U	N	N	Y
0		0	1				
1	S						
2	N						
3	O						
4	W						
5	Y						

$$E(i,j) = \min \begin{cases} E(i-1,j)+1 \\ E(i,j-1)+1 \\ E(i-1,j-1)+\text{diff}(i,j) \end{cases}$$

$$\text{diff}(i,j) = \begin{cases} 0, & x[i]=y[j] \\ 1, & \text{otherwise} \end{cases}$$

$$E(1{,}1) = \min \begin{cases} E(0,1)+1 \\ E(1,0)+1 =? \\ E(0,0)+0 \end{cases}$$

Edit Distance

i \ j		0	1	2	3	4	5
			S	U	N	N	Y
0		0	1				
1	S						
2	N						
3	O						
4	W						
5	Y						

$$E(i,j) = \min \begin{cases} E(i-1,j)+1 \\ E(i,j-1)+1 \\ E(i-1,j-1)+\text{diff}(i,j) \end{cases}$$

$$\text{diff}(i,j) = \begin{cases} 0, & x[i]=y[j] \\ 1, & \text{otherwise} \end{cases}$$

$$E(1,1) = \min \begin{cases} E(0,1)+1 \\ \boldsymbol{E(1,0)}+1 =? \\ E(0,0)+0 \end{cases}$$

Not calculated yet!

Edit Distance

i \ j		0	1 S	2 U	3 N	4 N	5 Y
0		0	1				
1	S						
2	N						
3	O						
4	W						
5	Y						

$$E(i,j) = \min \begin{cases} E(i-1,j)+1 \\ E(i,j-1)+1 \\ E(i-1,j-1)+\text{diff}(i,j) \end{cases}$$

$$\text{diff}(i,j) = \begin{cases} 0, & x[i]=y[j] \\ 1, & \text{otherwise} \end{cases}$$

Need upper left hand corner filled out before we can progress.

Edit Distance

j		0	1	2	3	4	5
i			S	U	N	N	Y
0		0	1	2			
1	S						
2	N						
3	O						
4	W						
5	Y						

$$E(i,j) = \min \begin{cases} E(i-1,j)+1 \\ E(i,j-1)+1 \\ E(i-1,j-1)+\text{diff}(i,j) \end{cases}$$

$$\text{diff}(i,j) = \begin{cases} 0, & x[i]=y[j] \\ 1, & \text{otherwise} \end{cases}$$

$$E(0,2) = \min \begin{cases} E(-1,2)+1 \\ E(0,1)+1 \\ E(-1,1)+1 \end{cases} = 2$$

Edit Distance

i \ j		0	1	2	3	4	5
			S	U	N	N	Y
0		0	1	2			
1	S	1					
2	N						
3	O						
4	W						
5	Y						

$$E(i,j) = \min \begin{cases} E(i-1,j)+1 \\ E(i,j-1)+1 \\ E(i-1,j-1)+\text{diff}(i,j) \end{cases}$$

$$\text{diff}(i,j) = \begin{cases} 0, & x[i]=y[j] \\ 1, & \text{otherwise} \end{cases}$$

$$E(1,0) = \min \begin{cases} E(0,0)+1 \\ E(1,-1)+1 = 1 \\ E(0,-1)+1 \end{cases}$$

Edit Distance

	j	0	1	2	3	4	5
i			S	U	N	N	Y
0		0	1	2			
1	S	1	0				
2	N						
3	O						
4	W						
5	Y						

$$E(i,j) = \min \begin{cases} E(i-1,j)+1 \\ E(i,j-1)+1 \\ E(i-1,j-1)+\text{diff}(i,j) \end{cases}$$

$$\text{diff}(i,j) = \begin{cases} 0, & x[i]=y[j] \\ 1, & \text{otherwise} \end{cases}$$

$$E(1,1) = \min \begin{cases} E(0,1)+1 \\ E(1,0)+1 = 0 \\ E(0,0)+0 \end{cases}$$

Edit Distance

	j	0	1	2	3	4	5
i			S	U	N	N	Y
0		0	1	2	3	4	5
1	S	1	0	1	2	3	4
2	N	2	1	1	1	2	3
3	O	3	2	2	2	2	3
4	W	4	3	3	3	3	3
5	Y	5	4	4	4	4	3

$$E(i,j) = \min \begin{cases} E(i-1,j)+1 \\ E(i,j-1)+1 \\ E(i-1,j-1)+\text{diff}(i,j) \end{cases}$$

$$\text{diff}(i,j) = \begin{cases} 0, & x[i] = y[j] \\ 1, & \text{otherwise} \end{cases}$$

Running Time?

Edit Distance

	j	0	1	2	3	4	5
i			S	U	N	N	Y
0		0	1	2	3	4	5
1	S	1	0	1	2	3	4
2	N	2	1	1	1	2	3
3	O	3	2	2	2	2	3
4	W	4	3	3	3	3	3
5	Y	5	4	4	4	4	3

$$E(i,j) = \min \begin{cases} E(i-1,j)+1 \\ E(i,j-1)+1 \\ E(i-1,j-1)+\text{diff}(i,j) \end{cases}$$

$$\text{diff}(i,j) = \begin{cases} 0, & x[i]=y[j] \\ 1, & \text{otherwise} \end{cases}$$

Running Time?

Fill out $n \times m$ table with constant operations: $O(nm)$

Edit Distance

	j	0	1	2	3	4	5
i			S	U	N	N	Y
0		0	1	2	3	4	5
1	S	1	0	1	2	3	4
2	N	2	1	1	1	2	3
3	O	3	2	2	2	2	3
4	W	4	3	3	3	3	3
5	Y	5	4	4	4	4	**3**

Edit distance = **3**.

How can we recreate the actual alignments?

Backtracking.

Ask the question: “How did we get here?”

Edit Distance

i	j	0	1	2	3	4	5
			S	U	N	N	Y
0		0	1	2	3	4	5
1	S	1	0	1	2	3	4
2	N	2	1	1	1	2	3
3	O	3	2	2	2	2	3
4	W	4	3	3	3	3	3
5	Y	5	4	4	4	4	**3**

How did we get to $E(5,5)$?

Edit Distance

	j	0	1	2	3	4	5
i			S	U	N	N	Y
0		0	1	2	3	4	5
1	S	1	0	1	2	3	4
2	N	2	1	1	1	2	3
3	O	3	2	2	2	2	3
4	W	4	3	3	3	3	3
5	Y	5	4	4	4	4	← **3**

How did we get to $E(5,5)$?

From $E(5,4)$?

Edit Distance

j		0	1	2	3	4	5
i			S	U	N	N	Y
0		0	1	2	3	4	5
1	S	1	0	1	2	3	4
2	N	2	1	1	1	2	3
3	O	3	2	2	2	2	3
4	W	4	3	3	3	3	3
5	Y	5	4	4	4	4	←**3**

How did we get to $E(5,5)$?

From $E(5,4)$? – No. Can never go down in cost.

Edit Distance

i	j	0	1	2	3	4	5
			S	U	N	N	Y
0		0	1	2	3	4	5
1	S	1	0	1	2	3	4
2	N	2	1	1	1	2	3
3	O	3	2	2	2	2	3
4	W	4	3	3	3	3	3
5	Y	5	4	4	4	4	↑ **3**

How did we get to $E(5,5)$?

From $E(5,4)$? – No. Can never go down in cost.

From $E(4,5)$?

Edit Distance

i \ j		0	1	2	3	4	5
			S	U	N	N	Y
0		0	1	2	3	4	5
1	S	1	0	1	2	3	4
2	N	2	1	1	1	2	3
3	O	3	2	2	2	2	3
4	W	4	3	3	3	3	3
5	Y	5	4	4	4	4	**3**

How did we get to $E(5,5)$?

From $E(5,4)$? – No. Can never go down in cost.

From $E(4,5)$? – No. Need +1 to move that direction.

$$E(i,j) = \min \begin{cases} E(i-1,j)+1 \\ E(i,j-1)+1 \\ E(i-1,j-1)+\mathrm{diff}(i,j) \end{cases}$$

Edit Distance

	j	0	1	2	3	4	5
i			S	U	N	N	Y
0		0	1	2	3	4	5
1	S	1	0	1	2	3	4
2	N	2	1	1	1	2	3
3	O	3	2	2	2	2	3
4	W	4	3	3	3	3	3
5	Y	5	4	4	4	4	**3**

How did we get to $E(5,5)$?

From $E(5,4)$? – No. Can never go down in cost.

From $E(4,5)$? – No. Need +1 to move that direction.

From $E(4,4)$?

$$E(i,j) = \min \begin{cases} E(i-1,j)+1 \\ E(i,j-1)+1 \\ E(i-1,j-1)+\text{diff}(i,j) \end{cases}$$

Edit Distance

i \ j		0	1 S	2 U	3 N	4 N	5 Y
0		0	1	2	3	4	5
1	S	1	0	1	2	3	4
2	N	2	1	1	1	2	3
3	O	3	2	2	2	2	3
4	W	4	3	3	3	3	3
5	Y	5	4	4	4	4	**3**

How did we get to $E(5,5)$?

- From $E(5,4)$? – No. Can never go down in cost.
- From $E(4,5)$? – No. Need +1 to move that direction.
- From $E(4,4)$? – Yes. Match Y's.

$$E(i,j) = \min \begin{cases} E(i-1,j)+1 \\ E(i,j-1)+1 \\ E(i-1,j-1)+\text{diff}(i,j) \end{cases}$$

Edit Distance

Continuing the process yields all of the optimal solutions.

i \ j		0	1	2	3	4	5
			S	U	N	N	Y
0		0	1	2	3	4	5
1	S	1	0	1	2	3	4
2	N	2	1	1	1	2	3
3	O	3	2	2	2	2	3
4	W	4	3	3	3	3	3
5	Y	5	4	4	4	4	3

Diagonal move indicates ?

Vertical move indicates ?

Horizontal move indicates ?

$$E(i,j) = \min \begin{cases} E(i-1,j)+1 \\ E(i,j-1)+1 \\ E(i-1,j-1)+\text{diff}(i,j) \end{cases}$$

Edit Distance

Continuing the process yields all of the optimal solutions.

i \ j		0	1	2	3	4	5
			S	U	N	N	Y
0		0	1	2	3	4	5
1	S	1	0	1	2	3	4
2	N	2	1	1	1	2	3
3	O	3	2	2	2	2	3
4	W	4	3	3	3	3	3
5	Y	5	4	4	4	4	3

Diagonal move indicates match.

Vertical move indicates ?

Horizontal move indicates ?

$$E(i,j) = \min \begin{cases} E(i-1,j)+1 \\ E(i,j-1)+1 \\ E(i-1,j-1)+\text{diff}(i,j) \end{cases}$$

Edit Distance

i \ j		0	1 S	2 U	3 N	4 N	5 Y
0		0	1	2	3	4	5
1	S	1	0	1	2	3	4
2	N	2	1	1	1	2	3
3	O	3	2	2	2	2	3
4	W	4	3	3	3	3	3
5	Y	5	4	4	4	4	3

Continuing the process yields all of the optimal solutions.

Diagonal move indicates match.

Vertical move indicates space inserted in j.

Horizontal move indicates ?

$$E(i,j) = \min \begin{cases} E(i-1,j)+1 \\ E(i,j-1)+1 \\ E(i-1,j-1)+\text{diff}(i,j) \end{cases}$$

Edit Distance

i \ j		0	1	2	3	4	5
			S	U	N	N	Y
0		0	1	2	3	4	5
1	S	1	0	1	2	3	4
2	N	2	1	1	1	2	3
3	O	3	2	2	2	2	3
4	W	4	3	3	3	3	3
5	Y	5	4	4	4	4	3

Continuing the process yields all of the optimal solutions.

Diagonal move indicates match.

Vertical move indicates space inserted in j.

Horizontal move indicates space inserted in i.

$$E(i,j) = \min \begin{cases} E(i-1,j)+1 \\ E(i,j-1)+1 \\ E(i-1,j-1)+\text{diff}(i,j) \end{cases}$$

Edit Distance

i \ j		0	1	2	3	4	5
			S	U	N	N	Y
0		0	1	2	3	4	5
1	S	1	0	1	2	3	4
2	N	2	1	1	1	2	3
3	O	3	2	2	2	2	3
4	W	4	3	3	3	3	3
5	Y	5	4	4	4	4	3

Diagonal move indicates match.

Vertical move indicates space inserted in j.

Horizontal move indicates space inserted in i.

S – N O W Y

S U N N – Y

Edit Distance

	j	0	1	2	3	4	5
i			S	U	N	N	Y
0		0	1	2	3	4	5
1	S	1	0	1	2	3	4
2	N	2	1	1	1	2	3
3	O	3	2	2	2	2	3
4	W	4	3	3	3	3	3
5	Y	5	4	4	4	4	3

Diagonal move indicates match.

Vertical move indicates space inserted in j.

Horizontal move indicates space inserted in i.

Alignment?

Edit Distance

	j	0	1	2	3	4	5
i			S	U	N	N	Y
0		0	1	2	3	4	5
1	S	1	0	1	2	3	4
2	N	2	1	1	1	2	3
3	O	3	2	2	2	2	3
4	W	4	3	3	3	3	3
5	Y	5	4	4	4	4	3

Diagonal move indicates match.

Vertical move indicates space inserted in j.

Horizontal move indicates space inserted in i.

S – N O W Y

S U N – N Y

Edit Distance

j		0	1	2	3	4	5
i			S	U	N	N	Y
0		0	1	2	3	4	5
1	S	1	0	1	2	3	4
2	N	2	1	1	1	2	3
3	O	3	2	2	2	2	3
4	W	4	3	3	3	3	3
5	Y	5	4	4	4	4	3

Diagonal move indicates match.

Vertical move indicates space inserted in j.

Horizontal move indicates space inserted in i.

S N O W Y

S U N N Y