Edit distance

- An alignment, or matched up, of two strings is simply a way of writing the strings one above the other.
Edit distance

- An **alignment**, or matched up, of two strings is simply a way of writing the strings one above the other.

  *example:* alignments of “SNOWY” and “SUNNY”:

  - snowy
    - snow - y
  - sunny
    - sun -- ny

  “−” indicates a “gap”
An **alignment**, or matched up, of two strings is simply a way of writing the strings one above the other.

*example: alignments of “SNOWY” and “SUNNY”:

```
  s - n o w y
  s u n n - y
```

“−” indicates a “gap”

The **cost** of an alignment is the number of columns in which the letters differ.
Edit distance

An alignment, or matched up, of two strings is simply a way of writing the strings one above the other.

example: alignments of “SNOWY” and “SUNNY”:

```
s - n o w y
 s u n n - y
```

“−” indicates a “gap”

The cost of an alignment is the number of columns in which the letters differ.

example: alignments of “SNOWY” and “SUNNY”:

```
s - n o w y
 s u n n - y
```

cost = 3
cost = 5
Edit distance

▶ An **alignment**, or matched up, of two strings is simply a way of writing the strings one above the other.

*example: alignments of “SNOWY” and “SUNNY”:

```
s - n o w y
sun n - y
```

“−” indicates a “gap”

▶ The **cost** of an alignment is the number of columns in which the letters differ.

*example: alignments of “SNOWY” and “SUNNY”:

```
s - n o w y
sun n - y
```

```
- s n o w - y
- s u n - - n y
```

cost = 3 cost = 5

▶ **Edit distance** between two strings is the **minimum cost** of their alignment, i.e., *the best possible alignment*
An alignment, or matched up, of two strings is simply a way of writing the strings one above the other.

**Example:** alignments of “SNOWY” and “SUNNY”:

```
s - n o w y
s u n n - y
```

“−” indicates a “gap”

The cost of an alignment is the number of columns in which the letters differ.

**Example:** alignments of “SNOWY” and “SUNNY”:

```
s - n o w y
s u n n - y
```

```
s - n o w y
s u n n - y
```

cost = 3 cost = 5

**Edit distance** between two strings is the minimum cost of their alignment, i.e., *the best possible alignment*

Edit distance is the minimum number of edits – insertions, deletions and substitutions of characters – need to transform the first string into the second.
Edit distance

- An **alignment**, or matched up, of two strings is simply a way of writing the strings one above the other.

  *example: alignments of “SNOWY” and “SUNNY”:

```
  s-n-o-w-y       -s-n-o-w-y
  s-u-n-n-y       s-u-n-n-y
```

  “-” indicates a “gap”

- The **cost** of an alignment is the number of columns in which the letters differ.

  *example: alignments of “SNOWY” and “SUNNY”:

```
  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** between two strings is the *minimum cost* of their alignment, i.e., *the best possible alignment*

- Edit distance is the *minimum number of edits* – insertions, deletions and substitutions of characters – need to transform the first string into the second. *e.g. a spell checker.*
Edit distance

- Given strings $x[1 \cdots m]$ and $y[1 \cdots n]$. 
Edit distance

- Given strings $x[1 \cdots m]$ and $y[1 \cdots n]$. Define

\[ e(m, n) = \text{the edit distance between } x \text{ and } y \]
Edit distance

- Given strings $x[1 \cdots m]$ and $y[1 \cdots n]$. Define

$$e(m, n) = \text{the edit distance between } x \text{ and } y$$

Our objective is to compute $e(m, n)$ efficiently.
Edit distance

- Given strings \(x[1 \cdots m]\) and \(y[1 \cdots n]\). Define

\[ e(m, n) = \text{the edit distance between } x \text{ and } y \]

Our objective is to compute \(e(m, n)\) efficiently

- **Subproblem:**
Edit distance

- Given strings $x[1 \cdots m]$ and $y[1 \cdots n]$. Define

$$e(m, n) = \text{the edit distance between } x \text{ and } y$$

Our objective is to compute $e(m, n)$ efficiently

- Subproblem:

  edit distance $e(i, j)$ between $x[1 \cdots i]$ and $y[1 \cdots j]$
Edit distance

- Given strings $x[1 \cdots m]$ and $y[1 \cdots n]$. Define
  \[
e(m, n) = \text{the edit distance between } x \text{ and } y
  \]

  Our objective is to compute $e(m, n)$ efficiently

- **Subproblem**:
  
  edit distance $e(i, j)$ between $x[1 \cdots i]$ and $y[1 \cdots j]$  

- How to express $e(i, j)$ in terms of its subproblems, *recursively*?
Edit distance

- Given strings $x[1 \cdots m]$ and $y[1 \cdots n]$. Define

$$e(m, n) = \text{the edit distance between } x \text{ and } y$$

Our objective is to compute $e(m, n)$ efficiently

- **Subproblem:**

  edit distance $e(i, j)$ between $x[1 \cdots i]$ and $y[1 \cdots j]$

- How to express $e(i, j)$ in terms of its subproblems, *recursively*?

- **key observation:** the rightmost column of an alignment of $x[1 \cdots i]$ and $y[1 \cdots j]$ can only be one of the following three cases:

<table>
<thead>
<tr>
<th>Case 1</th>
<th>Case 2</th>
<th>Case 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x[i]$</td>
<td>$-$</td>
<td>$-$</td>
</tr>
<tr>
<td>$-$</td>
<td>$y[j]$</td>
<td>$y[j]$</td>
</tr>
</tbody>
</table>
Edit distance

- By the above key observation, then

\[
e(i, j) = \min \{ 1 + e(i - 1, j), \ 1 + e(i, j - 1), \ \text{diff}(i, j) + e(i - 1, j - 1) \}
\]

where

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

- Question: how to find the corresponding optimal alignment?
Edit distance

- The answers to all the subproblems $e(i, j)$ form a two-dimensional table, and the final answer (our objective) is at $e(m, n)$.

Example 1.

<table>
<thead>
<tr>
<th></th>
<th>s</th>
<th>u</th>
<th>n</th>
<th>n</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>s</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>n</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>o</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>w</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>y</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
</tr>
</tbody>
</table>

Therefore, the edit distance between $x$ and $y = e(5, 5) = 3$. 

4 / 5
Edit distance

- The answers to all the subproblems $e(i, j)$ form a two-dimensional table, and the final answer (our objective) is at $e(m, n)$.
- Initialization:

  \[
  e(0, 0) = 0;
  e(i, 0) = i \text{ for } i = 1, \ldots, m
  
  e(0, j) = j \text{ for } j = 1, \ldots, n
  \]
Edit distance

- The answers to all the subproblems $e(i, j)$ form a two-dimensional table, and the final answer (our objective) is at $e(m, n)$.

- Initialization:

  \[
  e(0, 0) = 0; \\
  e(i, 0) = i \text{ for } i = 1, \ldots, m \\
  e(0, j) = j \text{ for } j = 1, \ldots, n
  \]

- Pseudocode

Example 1.

\[
\begin{array}{cccccc}
  & s & u & n & n & y \\
  s & 0 & 1 & 2 & 3 & 4 \\
  n & 1 & 1 & 1 & 2 & 3 \\
  o & 2 & 2 & 2 & 2 & 3 \\
  w & 3 & 3 & 3 & 3 & 3 \\
  y & 4 & 4 & 4 & 4 & 3 \\
\end{array}
\]

Therefore, the edit distance between $x$ and $y = e(5, 5) = 3$. 

4 / 5
Edit distance

- The answers to all the subproblems $e(i, j)$ form a two-dimensional table, and the final answer (our objective) is at $e(m, n)$.
- Initialization:

\[
\begin{align*}
e(0, 0) &= 0; \\
e(i, 0) &= i \text{ for } i = 1, \ldots, m \\
e(0, j) &= j \text{ for } j = 1, \ldots, n
\end{align*}
\]

- Pseudocode
- Example 1. $x = 'snowy', y = 'sunny'$
The answers to all the subproblems $e(i, j)$ form a two-dimensional table, and the final answer (our objective) is at $e(m, n)$.

Initialization:

$$e(0, 0) = 0;$$
$$e(i, 0) = i \text{ for } i = 1, \ldots, m$$
$$e(0, j) = j \text{ for } j = 1, \ldots, n$$

Pseudocode

Example 1. $x = \text{'snowy'}, \ y = \text{'sunny'}$

<table>
<thead>
<tr>
<th></th>
<th>s</th>
<th>u</th>
<th>n</th>
<th>n</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>s</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>n</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>o</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>w</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>y</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

Therefore, the edit distance between $x$ and $y = e(5, 5) = 3$. 

4 / 5
Edit distance

- The answers to all the subproblems \( e(i, j) \) form a two-dimensional table, and the final answer (our objective) is at \( e(m, n) \).
- Initialization:

\[
\begin{align*}
e(0,0) &= 0; \\
e(i,0) &= i \quad \text{for } i = 1, \ldots, m \\
e(0,j) &= j \quad \text{for } j = 1, \ldots, n
\end{align*}
\]

- Pseudocode
- Example 1. \( x = \text{'snowy'}, y = \text{'sunny'} \)

<table>
<thead>
<tr>
<th></th>
<th>s</th>
<th>u</th>
<th>n</th>
<th>n</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>s</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>s</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>n</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>o</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>w</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>y</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

Therefore, the edit distance between \( x \) and \( y = e(5, 5) = 3 \).
Edit distance

Example 2. $x = \text{'heroically'}, y = \text{'scholarly'}$
## Edit distance

**Example 2.** $x = 'heroically'$, $y = 'scholarly'$

<table>
<thead>
<tr>
<th></th>
<th>s</th>
<th>c</th>
<th>h</th>
<th>o</th>
<th>l</th>
<th>a</th>
<th>r</th>
<th>l</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>h</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>e</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>r</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>o</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>i</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>c</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>a</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>l</td>
<td>7</td>
<td>7</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>y</td>
<td>8</td>
<td>8</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>7</td>
</tr>
</tbody>
</table>

Therefore, the edit distance between $x$ and $y = e(10, 9) = 6$

Note: $LCS(x, y) = 5$
Edit distance

Example 2. $x = 'heroically', y = 'scholarly'$

<table>
<thead>
<tr>
<th></th>
<th>s</th>
<th>c</th>
<th>h</th>
<th>o</th>
<th>l</th>
<th>a</th>
<th>r</th>
<th>l</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>h</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>e</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>r</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>o</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>i</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>c</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>a</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>6</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>l</td>
<td>7</td>
<td>7</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>l</td>
<td>8</td>
<td>8</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>y</td>
<td>9</td>
<td>9</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>y</td>
<td>10</td>
<td>10</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>7</td>
</tr>
</tbody>
</table>

Therefore, the edit distance between $x$ and $y = e(10, 9) = 6$
Example 2. $x = \text{'heroically'}, \ y = \text{'scholarly'}$

<table>
<thead>
<tr>
<th></th>
<th>s c h o l a r l y</th>
</tr>
</thead>
<tbody>
<tr>
<td>h</td>
<td>0 1 2 3 4 5 6 7 8 9</td>
</tr>
<tr>
<td>e</td>
<td>1 1 2 2 3 4 5 6 7 8</td>
</tr>
<tr>
<td>r</td>
<td>2 2 2 3 3 4 5 6 7 8</td>
</tr>
<tr>
<td>o</td>
<td>3 3 3 3 4 4 5 5 6 7</td>
</tr>
<tr>
<td>i</td>
<td>4 4 4 4 3 4 5 6 6 7</td>
</tr>
<tr>
<td>c</td>
<td>5 5 5 5 4 4 5 6 7 7</td>
</tr>
<tr>
<td>a</td>
<td>6 6 6 6 5 5 5 6 7 8</td>
</tr>
<tr>
<td>l</td>
<td>7 7 7 6 6 6 6 5 6 7 8</td>
</tr>
<tr>
<td>l</td>
<td>8 8 7 7 7 6 6 6 6 7</td>
</tr>
<tr>
<td>y</td>
<td>9 9 8 8 8 7 7 7 6 7</td>
</tr>
<tr>
<td>y</td>
<td>10 10 9 9 9 8 8 8 7 6</td>
</tr>
</tbody>
</table>

Therefore, the edit distance between $x$ and $y = e(10, 9) = 6$

Note: $LCS(x, y) = 5$