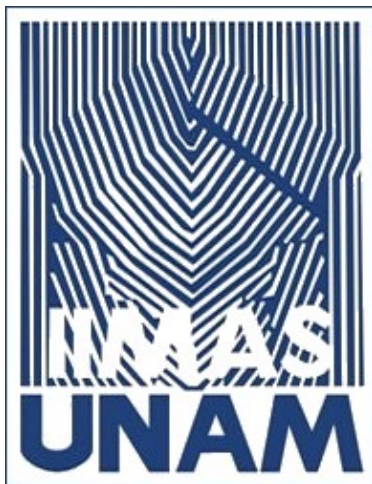


Procesamiento de Lenguaje Natural

Introducción



Dra. Helena Gómez Adorno

helena.gomez@iimas.unam.mx

Dra. Gemma Bel

gbele@iingen.unam.mx



Correo del curso:

pln.cienciadedatos@gmail.com

Asistente:

Luis Ramon Casillas

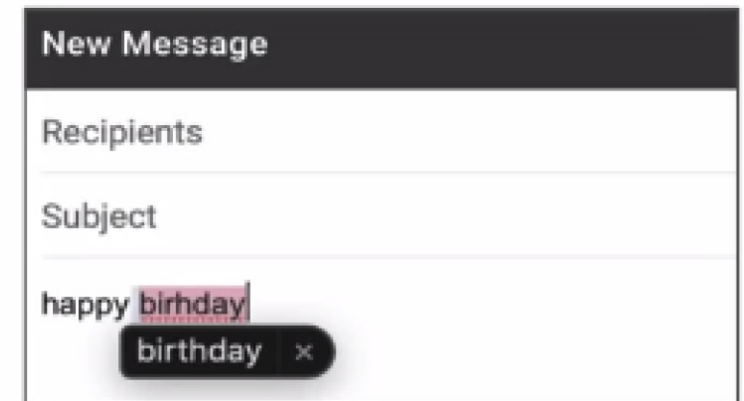
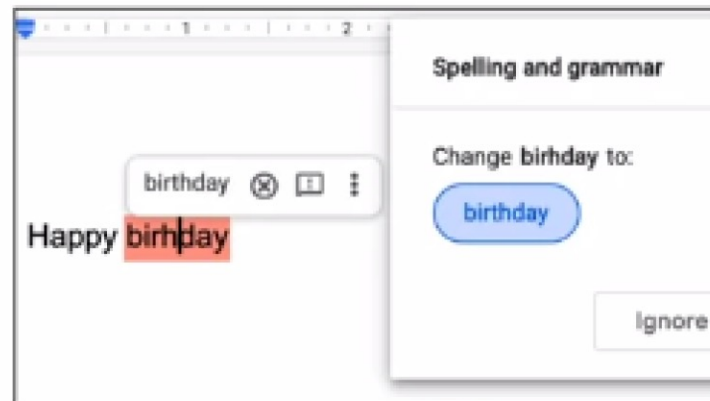
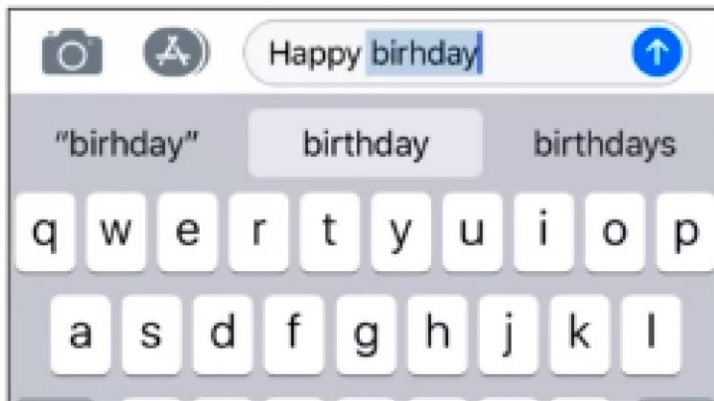
Contenido



- Qué es la autocorrección?
- Construcción del modelo
- Distancia mínima de edición
- Algoritmo de distancia mínima de edición

Qué es la autocorrección?

- Teléfonos
- Tabletas
- Computadoras



Qué es la autocorrección?



- Ejemplo:

“Que yo vaia al cumpleaños no cambia nada”

Qué es la autocorrección?



- Ejemplo:

“Que yo valla al cumpleaños no cambia nada”



??



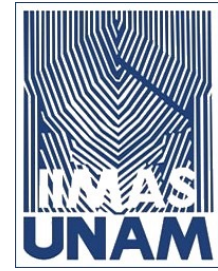
??

Qué es la autocorrección?



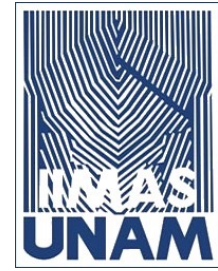
- Ejemplo:

“Que yo  vaya al cumpleaños no cambia nada”



Cómo funciona?

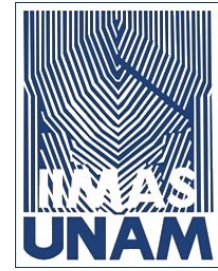
1. Identificar palabras mal escritas
2. Encontrar cadenas a una distancia de edición N
3. Filtrar candidatos
4. Calcular probabilidades de palabras



Cómo funciona?

1. Identificar palabras mal escritas
2. Encontrar cadenas a una distancia de edición N
3. Filtrar candidatos
4. Calcular probabilidades de palabras

vaya



Cómo funciona?

1. Identificar palabras mal escritas
2. Encontrar cadenas a una distancia de edición N
3. Filtrar candidatos
4. Calcular probabilidades de palabras

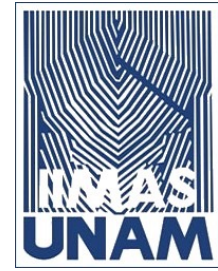
vaia

v_ia

va_a

vai_

etc..



Cómo funciona?

1. Identificar palabras mal escritas
2. Encontrar cadenas a una distancia de edición N
3. Filtrar candidatos
4. Calcular probabilidades de palabras

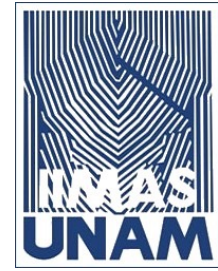
vaia

veia

vaya

vais

etc..



Cómo funciona?

1. Identificar palabras mal escritas
2. Encontrar cadenas a una distancia de edición N
3. Filtrar candidatos
4. Calcular probabilidades de palabras

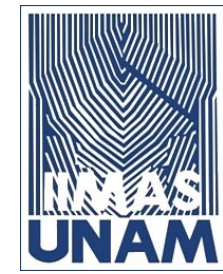
vaia

veia

vaya

vais

etc..



Cómo funciona?

1. Identificar palabras mal escritas
2. Encontrar cadenas a una distancia de edición N
3. Filtrar candidatos
4. Calcular probabilidades de palabras

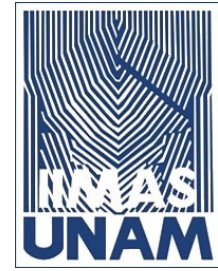
vaia → vaya ✓

veia

vaya

vais

etc..



Construcción del modelo

1. Identificar palabras mal escritas

Vaia?? 🤔

Construcción del modelo

1. Identificar palabras mal escritas

```
if word not in vocab:  
    misspelled = True
```

Vaia?? 🤔

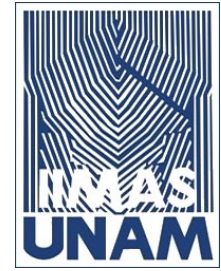
Construcción del modelo

1. Identificar palabras mal escritas

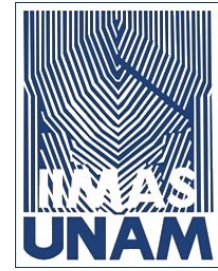
```
if word not in vocab:  
    misspelled = True
```

Vaia??**X**

Construcción del modelo



2. Encontrar cadenas a una distancia de edición N



Construcción del modelo

2. Encontrar cadenas a una distancia de edición N

Edición: Una operación realizada a la cadena para cambiarla

Construcción del modelo

2. Encontrar cadenas a una distancia de edición N

Edición: Una operación realizada a la cadena para cambiarla

Insert (agregar una letra)

Delete (eliminar una letra)

Switch (intercambiar 2 letras adyacentes)

Replace (cambiar una letra por otra)

Construcción del modelo

2. Encontrar cadenas a una distancia de edición N

Edición: Una operación realizada a la cadena para cambiarla

Insert (agregar una letra)

Delete (eliminar una letra)

Switch (intercambiar 2 letras adyacentes)

Replace (cambiar una letra por otra)

ir: 'irá', 'iré'

unos: 'uno', 'nos'

peor: 'pero', 'epor'

vasta: 'basta', 'vasto'



Construcción del modelo

2. Encontrar cadenas a una distancia de edición N

- Dada una cadena, encontrar todas las posibles cadenas que están a una distancia de edición N , usando:
 - Insert
 - Delete
 - Switch
 - Replace

Construcción del modelo

2. Encontrar cadenas a una distancia de edición N

- Dada una cadena, encontrar todas las posibles cadenas que están a una distancia de edición N , usando:
 - Insert
 - Delete
 - Switch
 - Replace

vaia
V_ia
va_a
vai_
etc..

Construcción del modelo

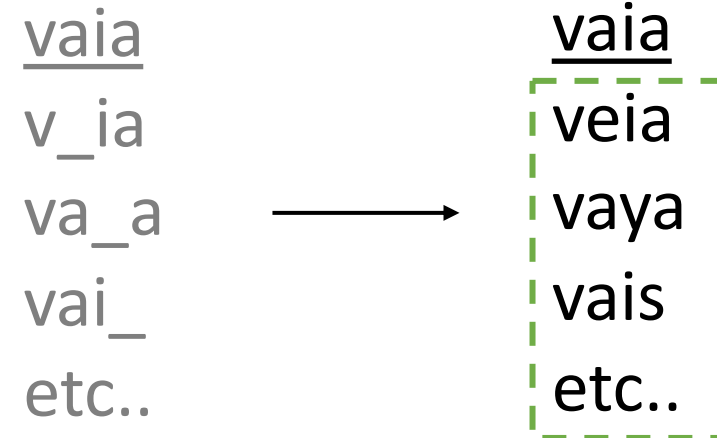


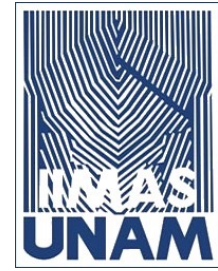
3. Filtrar candidatos

vaia
v_ia
va_a
vai_
etc..

Construcción del modelo

3. Filtrar candidatos





Construcción del modelo

4. Calcular probabilidades

Ejemplo: “I am happy because I am learning”

Construcción del modelo

4. Calcular probabilidades

Ejemplo: “I am happy because I am learning”

Word	Count
I	2
am	2
happy	1
because	1
learning	1

Total: 7

Construcción del modelo

4. Calcular probabilidades

Ejemplo: “I am happy because I am learning”

$$P(w) = \frac{C(w)}{V}$$

$$P(\text{am}) = \frac{C(\text{am})}{V} = \frac{2}{7}$$

$P(w)$ Probability of a word

$C(w)$ Number of times the word appears

V Total size of the corpus

Word	Count
I	2
am	2
happy	1
because	1
learning	1

Total: 7

Construcción del modelo



4. Calcular probabilidades

vaia

veia

vaya

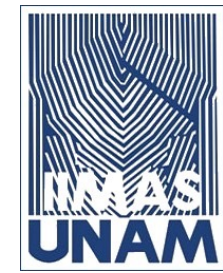
vais

etc..

Construcción del modelo

4. Calcular probabilidades

vaia → vaya ✓
veia
vaya
vais
etc..



Resumen

1. Identificar palabras mal escritas
2. Encontrar cadenas a una distancia de edición N

Insert
Delete
Switch
Replace

1. Filtrar candidatos
2. Calcular probabilidades de palabras

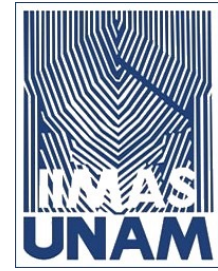
$$P(w) = \frac{C(w)}{V}$$

vaia → vaya ✓
veia
vaya
vais
etc..



Distancia mínima de edición

- Cómo evaluar la similitud entre 2 cadenas?
- Número de ediciones mínimas necesarias para transformar una cadena en otra
- Aplicaciones:
 - Corrección de ortografía, similitud de documento, traducción automática, secuencia de DNA, y más



Distancia mínima de edición

Ediciones:

Insert (agregar una letra)

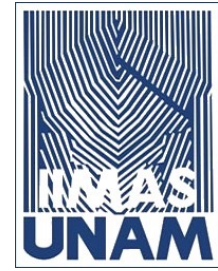
Delete (eliminar una letra)

Switch (intercambiar 2 letras adyacentes)

ir: 'irá', 'iré'

unos: 'uno', 'nos'

peor: 'pero', 'epor'



Distancia mínima de edición

Ediciones:

Insert (agregar una letra)

Delete (eliminar una letra)

Switch (intercambiar 2 letras adyacentes)

ir: 'irá', 'iré'

unos: 'uno', 'nos'

peor: 'pero', 'epor'

Distancia mínima de edición

Ejemplo:

Fuente:

p	l	a	y
---	---	---	---



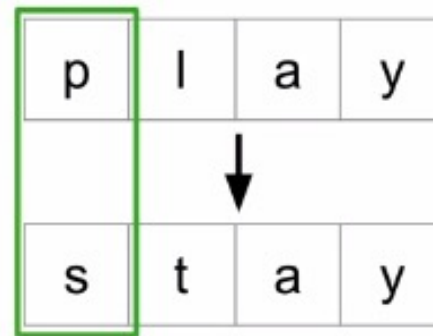
Objetivo:

s	t	a	y
---	---	---	---

Distancia mínima de edición

Ejemplo:

Fuente:



Objetivo:

$p \rightarrow s$: replace

Distancia mínima de edición

Ejemplo:

Fuente:

p	l	a	y
---	---	---	---

Objetivo:

s	t	a	y
---	---	---	---

p → s : replace
l → t : replace

Distancia mínima de edición

Ejemplo:

Fuente:

p	l	a	y
---	---	---	---



Objetivo:

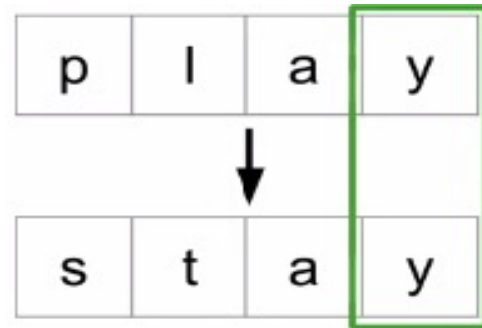
s	t	a	y
---	---	---	---

$p \rightarrow s$: replace
 $l \rightarrow t$: replace

Distancia mínima de edición

Ejemplo:

Fuente:



Objetivo:

$p \rightarrow s$: replace
 $l \rightarrow t$: replace

Distancia mínima de edición

Ejemplo:

Fuente:

p	l	a	y
---	---	---	---



Objetivo:

s	t	a	y
---	---	---	---

p → s : replace
l → t : replace

 } edits = 2

Distancia mínima de edición

Ejemplo:

Fuente:

p	l	a	y
---	---	---	---



Objetivo:

s	t	a	y
---	---	---	---

p → s : replace
l → t : replace

 } edits = 2

Edit cost:

Insert 1

Delete 1

Replace 2

Distancia mínima de edición

Ejemplo:

Fuente:

p	l	a	y
---	---	---	---



s	t	a	y
---	---	---	---

Objetivo:

p → s : replace
l → t : replace

 } edits = 2

Edit cost:

Insert 1

Delete 1

Replace 2

edit distance = $2 * 2 = 4$

Distancia mínima de edición

Fuente: stay → Objetivo: play

		0	1	2	3	4
		#	s	t	a	y
0	#					
1	p					
2	l					
3	a					
4	y					

Distancia mínima de edición

Fuente: stay → Objetivo: play

D[]

	0	1	2	3	4
	#	s	t	a	y
0	#				
1	p				
2	l				
3	a				
4	y				

Distancia mínima de edición

Fuente: stay \rightarrow Objetivo: play

D[]

D[2,4] = pl \rightarrow sta

		0	1	2	3	4
		#	s	t	a	y
0	#					
1	p					
2	l					
3	a					
4	y					

Distancia mínima de edición

fuente: stay → objetivo: play

D[]

D[2,3] = pl --> sta

D[2,3] = fuente[:2] --> objetivo[:3]

		0	1	2	3	4
		#	s	t	a	y
0	#					
1	p					
2	l					
3	a					
4	y					

Distancia mínima de edición

fuente: stay → objetivo: play

$D[i]$

$D[2,3] = \text{pl} \rightarrow \text{sta}$

$D[2,3] = \text{fuente}[:2] \rightarrow \text{objetivo}[:3]$

$D[i,i] = \text{fuente}[:i] \rightarrow \text{objetivo}[:j]$

		0	1	2	3	4
		#	s	t	a	y
0	#					
1	p					
2	l					
3	a					
4	y					

Distancia mínima de edición

Fuente: stay → Objetivo: play

$D[i, j]$

$D[i, i] = \text{fuente}[i] \rightarrow \text{objetivo}[i]$

	0	1	2	3	4
	#	s	t	a	y
0	#				
1	p				
2	l				
3	a				
4	y				

Distancia mínima de edición

Fuente: stay → Objetivo: play

$D[i]$

$D[i,i] = \text{fuente}[:i] \rightarrow \text{objetivo}[:i]$

$D[m,n] = \text{fuente} \rightarrow \text{objetivo}$

		0	1	2	3	4
		#	s	t	a	y
0	#					
1	p					
2	l					
3	a					
4	y					

Distancia mínima de edición

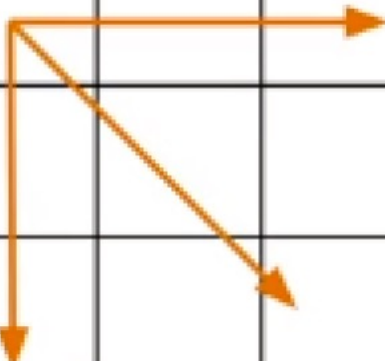
Fuente: stay → Objetivo: play

$D[i]$

$D[i,i] = \text{fuente}[:i] \rightarrow \text{objetivo}[:i]$

$D[m,n] = \text{fuente} \rightarrow \text{objetivo}$

		0	1	2	3	4
		#	s	t	a	y
0	#					
1	p					
2	l					
3	a					
4	y					



Distancia mínima de edición

Fuente: stay → Objetivo: play

Costo:

Insert:1, delet:1, replace:2

		0	1	2	3	4
		#	s	t	a	y
0	#					
1	p					
2	l					
3	a					
4	y					

Distancia mínima de edición

Fuente: stay \rightarrow Objetivo: play

Costo:

Insert:1, delet:1, replace:2

\rightarrow

	0	1	2	3	4
	#				
0	#				
1					
2					
3					
4					

Distancia mínima de edición

Fuente: stay \rightarrow Objetivo: play

Costo:

Insert:1, delet:1, replace:2

p \rightarrow #

		0	1	2	3	4
		#				
0	#	0				
1	p					
2						
3						
4						

Distancia mínima de edición

Fuente: stay \rightarrow Objetivo: play

Costo:

Insert:1, delet:1, replace:2

p \rightarrow #

delete

		0	1	2	3	4
		#				
0	#	0				
1	p	1				
2						
3						
4						

Distancia mínima de edición

Fuente: stay \rightarrow Objetivo: play

Costo:

Insert:1, delet:1, replace:2

\rightarrow s

		0	1	2	3	4
		#	s			
0	#	0				
1	p	1				
2						
3						
4						

Distancia mínima de edición

Fuente: stay \rightarrow Objetivo: play

Costo:

Insert:1, delet:1, replace:2

\rightarrow s

insert

	0	1	2	3	4
	#	s			
0	#	0	1		
1	p	1			
2					
3					
4					

Distancia mínima de edición

Fuente: stay \rightarrow Objetivo: play

Costo:

Insert:1, delet:1, replace:2

p \rightarrow s

	0	1	2	3	4
	#	s			
0	#	0	1		
1	p	1			
2					
3					
4					

Distancia mínima de edición

Fuente: stay \rightarrow Objetivo: play

Costo:

Insert:1, delet:1, replace:2

p \rightarrow s

insert+delete: p \rightarrow ps \rightarrow s: 2

	0	1	2	3	4
	#	s			
0	#	0	1		
1	p	1			
2					
3					
4					

Distancia mínima de edición

Fuente: stay \rightarrow Objetivo: play

Costo:

Insert:1, delet:1, replace:2

$p \rightarrow s$

insert+delete: $p \rightarrow ps \rightarrow s$: 2

delete+insert: $p \rightarrow \# \rightarrow s$: 2

		0	1	2	3	4
		#	s			
0	#	0	1			
1	p	1				
2						
3						
4						

Distancia mínima de edición

Fuente: stay \rightarrow Objetivo: play

Costo:

Insert:1, delet:1, replace:2

$p \rightarrow s$

insert+delete: $p \rightarrow ps \rightarrow s$: 2

delete+insert: $p \rightarrow \# \rightarrow s$: 2

replace: $p \rightarrow s$: 2

		0	1	2	3	4
		#	s			
0	#	0	1			
1	p	1	2			
2						
3						
4						

Distancia mínima de edición

Fuente: stay → Objetivo: play

Costo:

Insert:1, delet:1, replace:2

		0	1	2	3	4
		#	s	t	a	y
0	#	0	1			
1	p	1	2			
2	l					
3	a					
4	y					

Distancia mínima de edición

Fuente: stay \rightarrow Objetivo: play

Costo:

Insert:1, delet:1, replace:2

play \rightarrow #

		0	1	2	3	4
		#	s	t	a	y
0	#	0	1			
1	p	1	2			
2	l					
3	a					
4	y					

Distancia mínima de edición

Fuente: stay → Objetivo: play

Costo:

Insert:1, delet:1, replace:2

play → #

$D[i,j] = D[i-1,j] + \text{del_cost}$

		0	1	2	3	4
		#	s	t	a	y
0	#	0	1			
1	p	1	2			
2	l					
3	a					
4	y					

Distancia mínima de edición

Fuente: stay \rightarrow Objetivo: play

Costo:

Insert:1, delet:1, replace:2

play \rightarrow #

$D[i,j] = D[i-1,j] + \text{del_cost}$

		0	1	2	3	4
		#	s	t	a	y
0	#	0	1			
1	p	1	2			
2	l	2				
3	a	3				
4	y	4				

Distancia mínima de edición

Fuente: stay → Objetivo: play

Costo:

Insert:1, delet:1, replace:2

play → #

$D[i,j] = D[i-1,j] + \text{del_cost}$

$D[4,0] = \text{play} \rightarrow \#$

=fuente[:4] → objetivo[0]

		0	1	2	3	4
		#	s	t	a	y
0	#	0	1			
1	p	1	2			
2	l	2				
3	a	3				
4	y	4				

Distancia mínima de edición

Fuente: stay \rightarrow Objetivo: play

Costo:

Insert:1, delet:1, replace:2

\rightarrow play

		0	1	2	3	4
		#	s	t	a	y
0	#	0	1			
1	p	1	2			
2	l	2				
3	a	3				
4	y	4				

Distancia mínima de edición

Fuente: stay → Objetivo: play

Costo:

Insert:1, delet:1, replace:2

→ stay

$D[i,j] = D[i-1,j] + \text{ins_cost}$

		0	1	2	3	4
		#	s	t	a	y
0	#	0	1			
1	p	1	2			
2	l	2				
3	a	3				
4	y	4				

Distancia mínima de edición

Fuente: stay → Objetivo: play

Costo:

Insert:1, delet:1, replace:2

p → s

$$D[i, j] = \min \begin{cases} D[i-1, j] + \text{del_cost} \\ D[i, j-1] + \text{ins_cost} \\ D[i-1, j-1] + \begin{cases} \text{rep_cost}; & \text{if } \text{src}[i] \neq \text{tar}[j] \\ 0; & \text{if } \text{src}[i] = \text{tar}[j] \end{cases} \end{cases}$$

		0	1	2	3	4
		#	s	t	a	y
0	#	0	1	2	3	4
1	p	1	2			
2	l	2				
3	a	3				
4	y	4				

Distancia mínima de edición

Fuente: stay → Objetivo: play

Costo:

Insert:1, delet:1, replace:2

p → s

$$D[i, j] = \min \begin{cases} D[i-1, j] + \text{del_cost} \\ D[i, j-1] + \text{ins_cost} \\ D[i-1, j-1] + \begin{cases} \text{rep_cost}; & \text{if } \text{src}[i] \neq \text{tar}[j] \\ 0; & \text{if } \text{src}[i] = \text{tar}[j] \end{cases} \end{cases}$$

		0	1	2	3	4
		#	s	t	a	y
0	#	0	1	2	3	4
1	p	1	2			
2	l	2				
3	a	3				
4	y	4				

Distancia mínima de edición

Fuente: stay → Objetivo: play

Costo:

Insert:1, delet:1, replace:2

p → s

$$D[i, j] = \min \begin{cases} D[i-1, j] + \text{del_cost} \\ D[i, j-1] + \text{ins_cost} \\ D[i-1, j-1] + \begin{cases} \text{rep_cost}; & \text{if } \text{src}[i] \neq \text{tar}[j] \\ 0; & \text{if } \text{src}[i] = \text{tar}[j] \end{cases} \end{cases}$$

		0	1	2	3	4
		#	s	t	a	y
0	#	0	1	2	3	4
1	p	1	2			
2	l	2				
3	a	3				
4	y	4				

Distancia mínima de edición

Fuente: stay → Objetivo: play

Costo:

Insert:1, delet:1, replace:2

p → s

$$D[i, j] = \min \begin{cases} D[i-1, j] + \text{del_cost} \\ D[i, j-1] + \text{ins_cost} \\ D[i-1, j-1] + \begin{cases} \text{rep_cost}; & \text{if } \text{src}[i] \neq \text{tar}[j] \\ 0; & \text{if } \text{src}[i] = \text{tar}[j] \end{cases} \end{cases}$$

		0	1	2	3	4
		#	s	t	a	y
0	#	0	1	2	3	4
1	p	1	2			
2	l	2				
3	a	3				
4	y	4				

Distancia mínima de edición

Fuente: stay → Objetivo: play

Costo:

Insert:1, delet:1, replace:2

p → s

$$D[i, j] = \min \begin{cases} D[i-1, j] + \text{del_cost} \\ D[i, j-1] + \text{ins_cost} \\ D[i-1, j-1] + \begin{cases} \text{rep_cost}; & \text{if } \text{src}[i] \neq \text{tar}[j] \\ 0; & \text{if } \text{src}[i] = \text{tar}[j] \end{cases} \end{cases}$$

		0	1	2	3	4
		#	s	t	a	y
0	#	0	1	2	3	4
1	p	1	2			
2	l	2				
3	a	3				
4	y	4				

Distancia mínima de edición

Fuente: stay \rightarrow Objetivo: play

Costo:

Insert:1, delet:1, replace:2

p \rightarrow s

$$D[i-1, j] + 1 = 2$$

$$D[i, j-1] + 1 = 2$$

$$D[i-1, j-1] + 2 = 2$$

		0	1	2	3	4
		#	s	t	a	y
0	#	0	1	2	3	4
1	p	1	2			
2	l	2				
3	a	3				
4	y	4				

Distancia mínima de edición

Fuente: stay → Objetivo: play

Costo:

Insert:1, delet:1, replace:2

play → stay

$D[m,n] = 4$

		0	1	2	3	4
		#	s	t	a	y
0	#	0	1	2	3	4
1	p	1	2	3	4	5
2	l	2	3	4	5	6
3	a	3	4	5	4	5
4	y	4	5	6	5	4

Distancia mínima de edición

Fuente: stay → Objetivo: play

Costo:

Insert:1, delet:1, replace:2

- Distancia de Levenshtein
- Programación dinámica

		0	1	2	3	4
		#	s	t	a	y
0	#	0	1	2	3	4
1	p	1	2	3	4	5
2	l	2	3	4	5	6
3	a	3	4	5	4	5
4	y	4	5	6	5	4