

DISTANZA DI EDIT

UN ALLINEAMENTO RELATIVO ALLE OPERAZIONI

D = DELETE

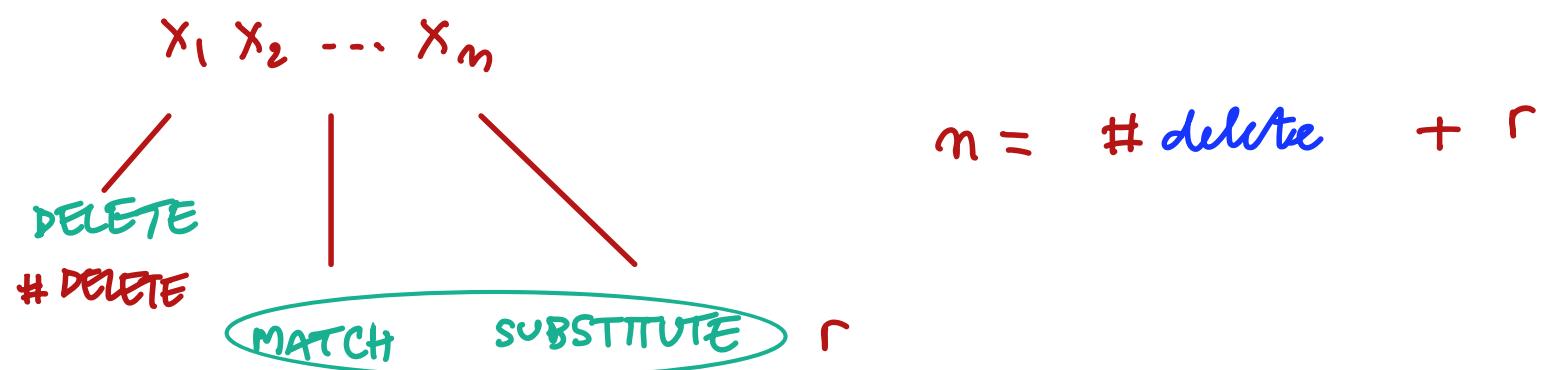
I = INSERT

M = MATCH

S = SUBSTITUTE

DI UNA STRINGA $X = x_1 x_2 \dots x_m$ IN UN'ALTRA
STRINGA $Y = y_1 y_2 \dots y_n$ E' UNA SEQUENZA a
DI LUNGHEZZA $m+n-r$ DI OPERAZIONI D, I, M, S
(DOVE r E' IL NUMERO COMPLESSIVO DI OPERAZIONI
DI TIPO M ED S IN a) CHE CONSENTE DI
TRASFORMARE X IN Y .

INFATTI



E QUINDI ...

$$\begin{aligned}\#\text{OPERAZIONI} &= \# \text{ delete} + \# \text{ insert} + r \\ &= \# \text{ delete} + \# \text{ insert} + r + r - r \\ &= (\# \text{ delete} + r) + (\# \text{ insert} + r) - r \\ &= n + m - r\end{aligned}$$

ESEMPI

M I - N E R V A -
M D I M S S S S I
M - A N T E L L O

M I N E R V A -
M S M S S S S I
M A N T E L L O

M I N - E R V A
M S M I M S S S
M A N T E L L O

IL COSTO DI UN ALLINEAMENTO a E' DATO DA

$$(\#D \text{ IN } a) \cdot \text{costo}(D) + (\#I \text{ IN } a) \cdot \text{costo}(I) \\ + (\#M \text{ IN } a) \cdot \text{costo}(M) + (\#S \text{ IN } a) \cdot \text{costo}(S).$$

LA SEQUENZA $(\text{costo}(D), \text{costo}(I), \text{costo}(M), \text{costo}(S))$
E' DETTA VETTORE DEI COSTI (DELL' ALLINEAMENTO).

QUANDO

$$\text{costo}(M) = 0$$

$$\text{costo}(D) = \text{costo}(I) = \text{costo}(S) = 1,$$

SI OTTIENE IL VETTORE DEI COSTI DI LEVENSHTEIN.

ESEMPI (COSTI DI LEVENSHTEIN)

M	I	-	N	E	R	V	A	-
M	D	I	M	S	S	S	S	I
M	-	A	N	T	E	L	L	O

COSTO = 7

M	I	N	E	R	V	A	-
M	S	M	S	S	S	S	I
M	A	N	T	E	L	L	O

COSTO = 6

M	I	N	-	E	R	V	A
M	S	M	I	M	S	S	S
M	A	N	T	E	L	L	O

COSTO = 5

UN VETTORE DI COSTI E' SIMMETRICO SE

- PER OGNI COPPIA DI STRINGHE X, Y E
PER OGNI ALLINEAMENTO a DI X IN Y
ESISTE

UN ALLINEAMENTO a^T DI Y IN X TALE CHE
 $\text{costo}(a^T) = \text{costo}(a)$.

PROPRIETA'

IL VETTORE DEI COSTI DI LEVENSHTEIN E'
SIMMETRICO.

DIM.

DATE X E Y ED UN ALLINEAMENTO a
DI X IN Y , DETTO a^T L'ALLINEAMENTO
OTTENUTO SCAMBIANDO IN a LE D CON LE I
(E VICEVERSA),

- a^T E' UN ALLINEAMENTO DI Y IN X
- $\text{costo}(a^T) = \text{costo}(a)$.



ESEMPIO

M I - N E R V A -

M D I M S S S S S I

M - A N T E L L O

M - A N T E L L O

M I D M S S S S S D

M I - N E R V A -

$$- \text{M I D M S S S S S D} = (\text{M D I M S S S S I})^T$$

= CON I COSTI DI LEVENSHTEIN SI HA

costo(M I D M S S S S S D)

$$= \text{costo}(\text{M I D M S S S S S D}) = 7$$

LA DISTANZA DI EDIT DA X A Y E' IL
MINIMO COSTO DI UN ALLINEAMENTO DI X IN Y .

QUANDO IL VETTORE DEI COSTI E' SIMMETRICO, PARLEREMO
DI DISTANZA DI EDIT TRA X ED Y .

CON IL VETTORE DEI COSTI DI LEVENSHTEIN,
SI HA LA DISTANZA DI LEVENSHTEIN.

DISTANZA DI HAMMING

LA DISTANZA DI LEVENSHTEIN GENERALIZZA LA DISTANZA DI HAMMING.

QUEST'ULTIMA E' RISTRETTA A

- STRINGHE DELLA MEDESIMA LUNGHEZZA

- AMMETTE LE SOLE OPERAZIONI

M (MATCH) , S (SUBSTITUTION)

CON

$\text{COSTO}(M) = 0$ E $\text{COSTO}(S) = 1$,

A	L	B	E	R	O
S	S	M	S	M	M
L	A	B	B	R	O

CALCOLO DELLA DISTANZA DI LEVENSHTEIN TRA DUE STRINGHE MEDIANTE PROGRAMMAZIONE DINAMICA

SIANO DATE DUE STRINGHE

$$X = x_1 x_2 \dots x_m \quad E \quad Y = y_1 y_2 \dots y_n$$

- INDICHiamo CON $M[i,j]$ (CON $0 \leq i \leq m$ E $0 \leq j \leq n$)

LA DISTANZA DI EDIT TRA

$$X_i = x_1 x_2 \dots x_i \quad E \quad Y_j = y_1 y_2 \dots y_j.$$

- PER $1 \leq i \leq m$ E $1 \leq j \leq n$, PONIAMO

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

- LA MATRICE M PUÒ ESSERE DEFINITA RICORSIVAMENTE COME SEGUÉ:

$$M[i, j] = \begin{cases} j & \text{SE } i = 0 \\ i & \text{SE } j = 0 \\ \min \left\{ M[i, j-1] + 1, \right. & \text{-- INSERT } y_j \\ M[i-1, j] + 1, & \text{-- DELETE } x_i \\ \left. M[i-1, j-1] + p(i, j) \right\} & \text{SE } i \geq 1 \wedge j \geq 1 \\ & \text{-- MATCH } 0 \text{ SUBSTITUTE } x_i \leftrightarrow y_j \\ & (p(i, j) = 0) \quad (p(i, j) = 1) \end{cases}$$

INFATTI

- $DIST(\varepsilon, y_1, y_2, \dots, y_j) = j$ (j INSERIMENTI)
- $DIST(x_1, x_2, \dots, x_i, \varepsilon) = i$ (i CANCELLAZIONI)

$x_1 \dots x_i$	-	$x_1 \dots x_{i-1}$	x_i	$x_i \neq y_j$	$x_i = y_j$
.....	I	D	M
$y_1 \dots y_{j-1}$	y_j	$y_1 \dots y_j$	-	$y_1 \dots y_{j-1}$	y_j
$M[i, j-1]$	1	$M[i-1, j]$	1	$M[i-1, j-1]$	1
					$M[i-1, j-1]$

(CARATTERIZZAZIONE DI UNA SOLUZIONE OTTIMA)

CALCOLO DEL VALORE DI UNA SOLUZIONE OTTIMA

Distanza-edit (X,Y)

// Calcolo della distanza di edit tra due stringhe X e Y di dimensioni m ed n,
mediante la costruzione della matrice M[0:m,0:n]

for j := 0 **to** n **do** //inizializza la riga 0

M[0,j] := j;

for i := 1 **to** m **do** //inizializza la colonna 0

M[i,0] := i;

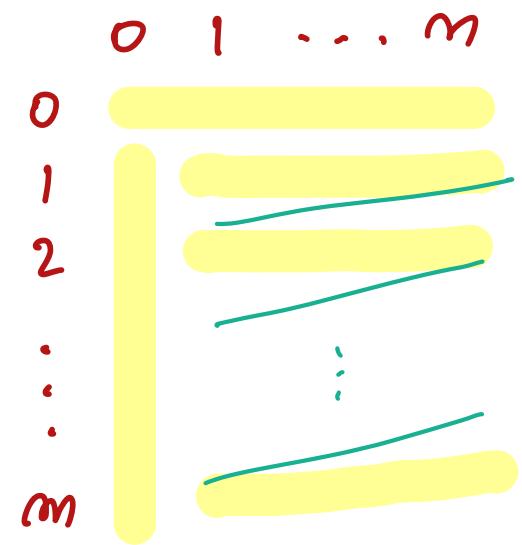
for i := 1 **to** m **do**

for j := 1 **to** n **do**

if X[i] = Y[j] **then** p := 0; **else** p := 1; **end_if**;

M[i,j] := min (M[i,j-1] + 1, M[i-1,j] + 1, M[i-1,j-1] + p);

return M[m,n];



COMPLESSITA`: $O(mn)$

ESEMPIO : DISTANZA DI EDIT TRA ALBERO E LABBRO

	0	1	2	3	4	5	6
0	∅	L	A	B	B	R	O
1	∅	∅	1	2	3	4	5
2	A	1	1	1	2	3	4
3	L	2	1	2	2	3	4
4	B	3	2	2	2	2	3
5	E	4	3	3	3	3	4
6	R	5	4	4	4	4	3
7	O	6	5	5	5	5	4

	∅	L	A	B	B	R	O
∅	∅	1	2	3	4	5	6
A	1	1	1	2	3	4	5
L	2	1	2	2	3	4	5
B	3	2	2	2	2	3	4
E	4	3	3	3	3	3	4
R	5	4	4	4	4	3	4
O	6	5	5	5	5	4	3

$k-i$	k	S
k	k	M
	↑	D
	←	I

COSTRUZIONE DI ALLINEAMENTI OTTIMI $(O(m+n))$

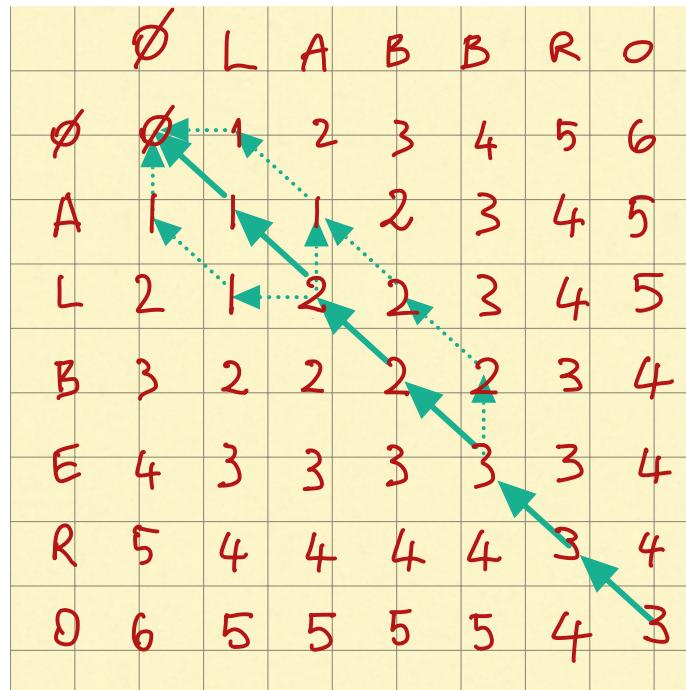
	\emptyset	L	A	B	B	R	O
\emptyset	\emptyset	1	2	3	4	5	6
A	1	1	1	2	3	4	5
L	2	1	2	2	3	4	5
B	3	2	2	2	2	3	4
E	4	3	3	3	3	3	4
R	5	4	4	4	4	3	4
O	6	5	5	5	5	4	3

- A L B E R O
 I M S M D M M
 L A B B - R O

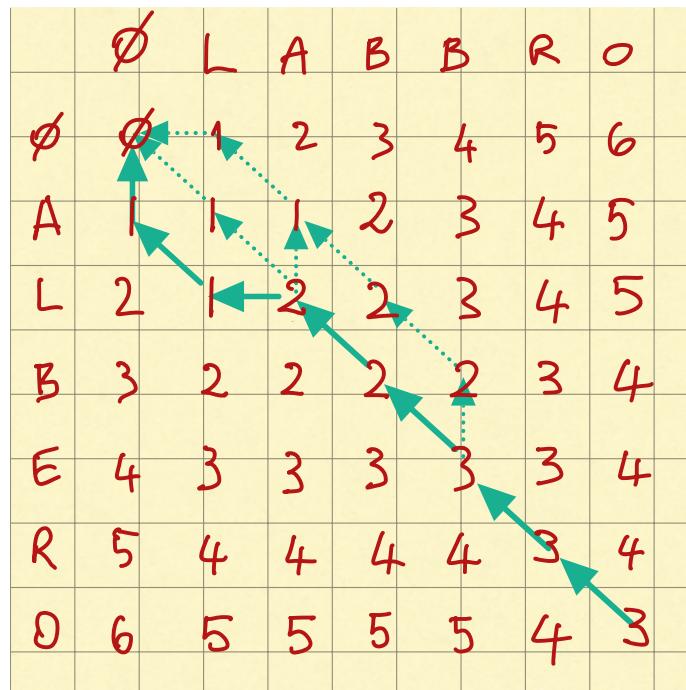
	\emptyset	L	A	B	B	R	O
\emptyset	\emptyset	1	2	3	4	5	6
A	1	1	1	2	3	4	5
L	2	1	2	2	3	4	5
B	3	2	2	2	2	3	4
E	4	3	3	3	3	3	4
R	5	4	4	4	4	3	4
O	6	5	5	5	5	4	3

- A L B E R O
 I M D M S M M
 L A B B R O

$k-1$	k	S
k	k	M
\uparrow		D
\leftarrow		I



ALBERO
 SSMSSMM
 LABBRO



AL-BERO
 D M I M S M M
 -LABBRO

ALLINEAMENTI OTTIMI PER SOTTO PROBLEMI

	\emptyset	L	A	B	B	R	O
\emptyset	\emptyset	1	2	3	4	5	6
A	1	1	1	2	3	4	5
L	2	1	2	2	3	4	5
B	3	2	2	2	2	3	4
E	4	3	3	3	3	3	4
R	5	4	4	4	4	3	4
O	6	5	5	5	5	4	3

ALBERO
-L-----

ALBERO
-LA-B-

ALBERO
D M D D D
-L-----

	\emptyset	L	A	B	B	R	O
\emptyset	\emptyset	1	2	3	4	5	6
A	1	1	1	2	3	4	5
L	2	1	2	2	3	4	5
B	3	2	2	2	2	3	4
E	4	3	3	3	3	3	4
R	5	4	4	4	4	3	4
O	6	5	5	5	5	4	3

ALBERO
D M S D S D
-L A -B -

RICERCA DI OCCORRENZE APPROXIMATE

DI UN DATO PATTERN (BREVE) X IN UN TESTO (LUNGO) Y

SI ANO DATE DUE STRINGHE

$$X = x_1 x_2 \dots x_m \quad (\text{PATTERN})$$

$$Y = y_1 y_2 \dots y_n \quad (\text{TESTO})$$

INDICHIAMO CON $N[i,j]$ (DOVE $0 \leq i \leq m$ E $0 \leq j \leq n$)

LA MINIMA DISTANZA DI EDIT TRA $X_i = x_1 x_2 \dots x_i$
E TUTTI I $(j+i)$ SUFFISSI DI $Y_j = y_1 y_2 \dots y_j$:

$$y_1 \dots y_j, \quad y_2 \dots y_j, \quad \dots, \quad y_{j-1} y_j, \quad y_j, \quad \varepsilon.$$

PONENDO COME PRIMA

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

PER $0 \leq i \leq m$ E $0 \leq j \leq n$, SI HA:

$$N[i,j] = \begin{cases} 0 & \text{SE } i=0 \\ i & \text{SE } j=0 \\ \min \left\{ N[i,j-1] + 1, \right. \\ \quad N[i-1,j] + 1, \\ \quad \left. N[i-1,j-1] + p(i,j) \right\} & \text{SE } i \geq 1 \text{ E } j \geq 1 \end{cases}$$

PER $0 \leq i \leq m$ E $0 \leq j \leq n$.

E

Ricerca-occorrenze-approssimate(X,Y)

// Ricerca delle occorrenze approssimate di X in Y, rispettivamente di dimensioni m ed n, mediante la costruzione della matrice N[0:m,0:n]

for j := 0 **to** n **do** //inizializza la riga 0

N[0,j] := 0;

for i := 1 **to** m **do** //inizializza la colonna 0

N[i,0] := i;

for i := 1 **to** m **do**

for j := 1 **to** n **do**

if X[i] = Y[j] **then** p := 0; **else** p := 1; **end_if**;

N[i,j] := min (N[i,j-1] + 1, N[i-1,j] + 1, N[i-1,j-1] + p);

return N;

ESEMPIO: RICERCA DELLE OCCORRENZE APPROXIMATE
DI $X = RAT$ IN $Y = SERRATURA$

0	1	2	3	4	5	6	7	8	9
\emptyset	S	E	R	R	A	T	U	R	A
\emptyset									
R	1	1	1	\emptyset	\emptyset	1	1	1	\emptyset
A	2	2	2	1	1	\emptyset	1	2	1
T	3	3	3	2	2	1	\emptyset	1	2

0	1	2	3	4	5	6	7	8	9
S	E	R	R	A	T	U	R	A	
3	3	3	2	2	1	\emptyset	1	2	1

- C'È UN'OCCHIATA DI RAT (TERMINANTE) IN POSIZIONE 6
- CI SONO OCCHIATE A DISTANZA 1 NELLE POSIZIONI 5, 7, 9
- CI SONO OCCHIATE A DISTANZA 2 NELLE POSIZIONI 3, 4, 8

0	1	2	3	4	5	6	7	8	9
Ø	S	E	R	R	A	T	U	R	A
Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø
R	I	I	I	Ø	Ø	I	I	I	I
A	2	2	2	1	1	Ø	1	2	Ø
T	3	3	3	2	2	1	Ø	1	1

0 1 2 3 4 5 6 7 8 9
S E R R A T U R A
R A T

DISTANZA 0 POS. 6

0 1 2 3 4 5 6 7 8 9
S E R R A - T U R A
R A T

DISTANZA 1 POS. 5

0 1 2 3 4 5 6 7 8 9
S E R - - R A T U R A
R A T

DISTANZA 2 POS. 3

0 1 2 3 4 5 6 7 8 9
S E R R A T U R A
R A T -

DISTANZA 1 POS. 7

0 1 2 3 4 5 6 7 8 9
S E R R A T U R A
R A T -

DISTANZA 2 POS. 4

0 1 2 3 4 5 6 7 8 9
S E R R A T U R A -
R A T

DISTANZA 1 POS. 9

0 1 2 3 4 5 6 7 8 9
S E R R A T U R A
R A T - -

DISTANZA 2 POS. 8

ECC.

	T	A	L	B	E	R
\emptyset	1	2	3	4	5	6
A	1	1	2	3	4	5
L	2	2	2	1	2	3
B	3	3	3	2	1	2
E	4	4	4	3	2	1
R	5	5	5	4	3	2
D	6	6	6	5	4	3

DISTANZA DI HAMMING

A L B E R O
 S S S S S S
 T A L B E R

6

DISTANZA DI LEVENSHTEIN

- A L B E R O
 I M M M M M D
 T A L B E R -

2