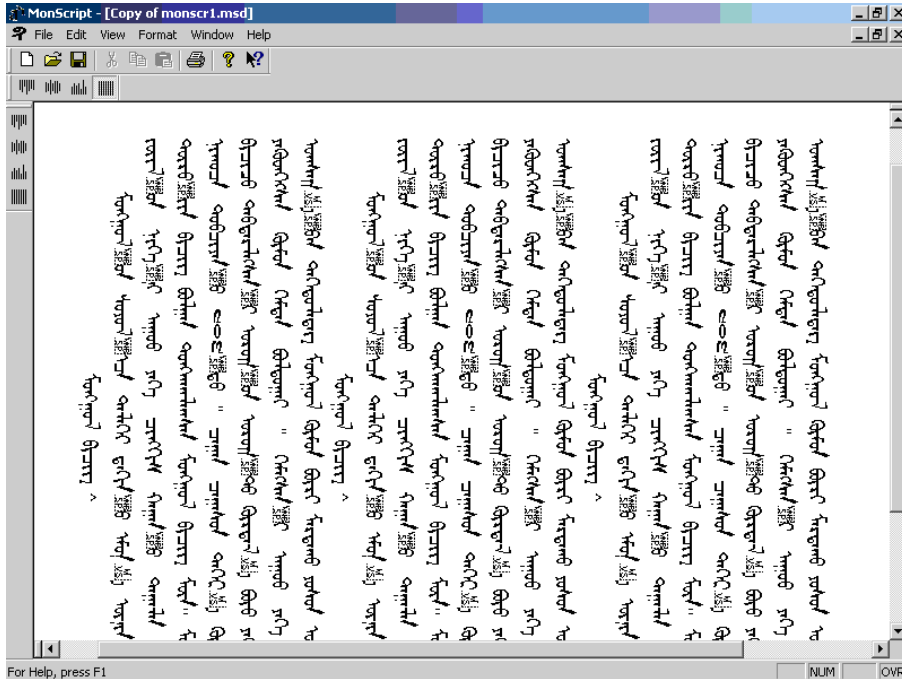


Хамгийн бага засварын хэмжээ

Лекц №6

Тодорхойлолт:
Хамгийн бага засварын
хэмжээ

2020 он



Хоёр тэмдэгт мөр хэр төстэй вэ?

- Зөв бичгийн алдаа зүгшрүүлэлт

- “graffe” гэж бичсэн бол

Аль нь хамгийн ойр вэ?

- graf
 - graft
 - grail
 - giraffe

- Тооцооллын биологи

- Нуклеотидуудын 2 дарааллын зэрэгцүүлэлт хий

```
AGGCTATCACCTGACCTCCAGGCCGATGCCC  
TAGCTATCACGACCGCGGTCGATTTGCCCGAC
```

- Зэрэгцүүлэлтийн үр дүн:

```
-AGGCTATCACCTGACCTCCAGGCCGA--TGCCC---  
TAG-CTATCAC--GACCGC--GGTCGATTTGCCCGAC
```

- Мөн машин орчуулга, мэдээлэл гаргах, яриа танихад ашиглагдана.

Засварын хэмжээ

- Хоёр тэмдэгт мөр хоорондын хамгийн бага засварын хэмжээ
- бол засах үйлдлийн хамгийн бага тоо юм.
 - Үсэг оруулах
 - Арилгах
 - Солих
- Нэг тэмдэгт мөр өөр тэмдэгт мөр болж хувирахад шаардлагатай

Хамгийн бага засварын хэмжээ

- Хоёр тэмдэгт мөр болон тэдгээрийн зэрэгцүүлэлт:

I	N	T	E	*	N	T	I	O	N
*	E	X	E	C	U	T	I	O	N

Хамгийн бага засварын хэмжээ

I N T E * N T I O N
| | | | | | | | | |
* E X E C U T I O N
d s s i s

- Хэрэв үйлдэл бүрийн зардлыг 1 гэвэл
 - Эдгээрийн хэмжээ 5
- Хэрэв солилтын зардлыг 2 гэвэл (Levenshtein)
 - Эдгээрийн хэмжээ 8

Тооцооллын биологи дахь зэрэгцүүлэлт

- Сууриудын дараалал өгөгдөнө

AGGCTATCACCTGACCTCCAGGCCGATGCCC
TAGCTATCACGACCGCGGGTCGATTGCCCCGAC

- Зэрэгцүүлэлт:

–AGGCTATCACCTGACCTCCAGGCCGA–TGCCC––
TAG–CTATCAC–GACCGC–GGTCGATTGCCCCGAC

- Өгөгдсөн хоёр дарааллаас, үсэг эсвэл зөрүү байгааг үсэг бүрийг зэрэгцүүлж олно.

ЭХБ дахь засварын хэмжээний бусад хэрэглээ

- машин орчуулга болон яриа танилтыг үнэлэх

R Spokesman confirms senior government adviser was shot

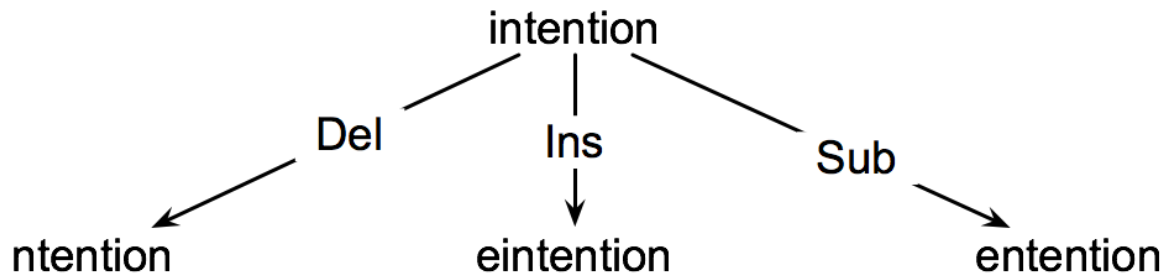
H Spokesman said the senior adviser was shot dead

S I D I

- Оноосон нэрийг олборлолт болон төлөө үгийг олоход
 - IBM Inc. announced today
 - IBM profits
 - Stanford President John Hennessy announced yesterday
 - for Stanford University President John Hennessy

Хам.бага.засврын хэмжээг хэрхэн олох вэ?

- Эхний тэмдэг мөрөөс төгсгөлийн тэмдэгт мөр хүртэлх нэг (засварын дарааллыг) замыг хайна:
 - **Эхний төлөв:** хувирах үг
 - **үйлдлүүд:** оруулах, арилгах, солих
 - **Зорилгын төлөв:** болох үг
 - **Замын зардал:** хамгийн бага байлгах: засварын тоо



Хамгийн бага засварыг хайх

- Гэхдээ бүх боломжит засварын дарааллын тоо асар их!
 - Энгийн аргаар олох боломжгүй
 - Олон алгаатай замууд ижил төлөвт хүрч болно.
 - Тэдгээрийн бүх алхамуудыг хадгалах шаардлагагүй
 - Эдгээр шинэчилсэн төлөв бүрээс зөвхөн хамгийн богино замыг

Хамгийн бага засварын хэмжээг тодорхойлох нь

- Хоёр тэмдэгт мөрийн хувьд
 - X –н урт n
 - Y –н урт m
- $D(i,j)$ –ийг тодорхойлно
 - $X[1..i]$ болон $Y[1..j]$ хоорондын засварын хэмжээ
 - ж.нь., X –н эхний i тэмдэгт ба Y –н эхний j тэмдэгт
 - X болон Y хоорондын засварын хэмжээ нь $D(n,m)$ байна.

Хамгийн бага засварын хэмжээний динамик програмчлал

- **Динамик програмчлал:** $D(n, m)$ –н хүснэгтэн тооцоолол
- Дэд асуудлуудын шийдлийг нэгтгэх замаар асуудлыг шийдэх.
- Доороос-дээшээ
 - жижиг i, j -н хувьд $D(i, j)$ –г тооцоолно
 - Мөн жижиг утгуудын өмнөх тооцоолол дээр суурилж арай том $D(i, j)$ –г тооцоолно
 - ж.нь., бүх i ($0 < i < n$) болон j ($0 < j < m$) –н хувьд $D(i, j)$ –г тооцоолно

Хамгийн бага засварын хэмжээг (Levenshtein) тодорхойлох нь

- Эхлэл

$$D(i, 0) = i$$

$$D(0, j) = j$$

- Рекурент хамаарал:

For each $i = 1 \dots M$

For each $j = 1 \dots N$

$$D(i, j) = \min \begin{cases} D(i-1, j) + 1 \\ D(i, j-1) + 1 \\ D(i-1, j-1) + \begin{cases} 2; & \text{if } X(i) \neq Y(j) \\ 0; & \text{if } X(i) = Y(j) \end{cases} \end{cases}$$

- Дуусгавар:

$D(N, M)$ бол хамгийн бага засварын хэмжээ


Засварын хэмжээний хүснэгт

N	9									
O	8									
I	7									
T	6									
N	5									
E	4									
T	3									
N	2									
I	1									
#	0	1	2	3	4	5	6	7	8	9
	#	E	X	E	C	U	T	I	O	N

Засварын хэмжээний хүснэгт

N	9									
O	8									
I	7									
T	6									
N	5									
E	4									
T	3									
N	2									
I	1									
#	0	1	2	3	4	5	6	7	8	9
	#	E	X	E	C	U	T	I	O	N

$$D(i,j) = \min \begin{cases} D(i-1,j) + 1 \\ D(i,j-1) + 1 \\ D(i-1,j-1) + \begin{cases} 2; & \text{if } S_1(i) \neq S_2(j) \\ 0; & \text{if } S_1(i) = S_2(j) \end{cases} \end{cases}$$



Засварын хэмжээ

$$D(i,j) = \min \begin{cases} D(i-1,j) + 1 \\ D(i,j-1) + 1 \\ D(i-1,j-1) + \begin{cases} 2; & \text{if } S_1(i) \neq S_2(j) \\ 0; & \text{if } S_1(i) = S_2(j) \end{cases} \end{cases}$$

N	9									
O	8									
I	7									
T	6									
N	5									
E	4									
T	3									
N	2									
I	1	2,2,2								
#	0	1	2	3	4	5	6	7	8	9
	#	E	X	E	C	U	T	I	O	N

Засварын хэмжээ

$$D(i,j) = \min \begin{cases} D(i-1,j) + 1 \\ D(i,j-1) + 1 \\ D(i-1,j-1) + \begin{cases} 2; & \text{if } S_1(i) \neq S_2(j) \\ 0; & \text{if } S_1(i) = S_2(j) \end{cases} \end{cases}$$

N	9									
O	8									
I	7									
T	6									
N	5									
E	4									
T	3									
N	2	3,3,3								
I	1	2								
#	0	1	2	3	4	5	6	7	8	9
	#	E	X	E	C	U	T	I	O	N

Засварын хэмжээ

$$D(i,j) = \min \begin{cases} D(i-1,j) + 1 \\ D(i,j-1) + 1 \\ D(i-1,j-1) + \begin{cases} 2; & \text{if } S_1(i) \neq S_2(j) \\ 0; & \text{if } S_1(i) = S_2(j) \end{cases} \end{cases}$$

N	9									
O	8									
I	7									
T	6									
N	5									
E	4									
T	3									
N	2	3								
I	1	2								
#	0	1	2	3	4	5	6	7	8	9
	#	E	X	E	C	U	T	I	O	N

Засварын хэмжээ

$$D(i,j) = \min \begin{cases} D(i-1,j) + 1 \\ D(i,j-1) + 1 \\ D(i-1,j-1) + \begin{cases} 2; & \text{if } S_1(i) \neq S_2(j) \\ 0; & \text{if } S_1(i) = S_2(j) \end{cases} \end{cases}$$

N	9									
O	8									
I	7									
T	6									
N	5									
E	4									
T	3									
N	2	3								
I	1	2	3,3,3							
#	0	1	2	3	4	5	6	7	8	9
	#	E	X	E	C	U	T	I	O	N

Засварын хэмжээ

$$D(i,j) = \min \begin{cases} D(i-1,j) + 1 \\ D(i,j-1) + 1 \\ D(i-1,j-1) + \begin{cases} 2; & \text{if } S_1(i) \neq S_2(j) \\ 0; & \text{if } S_1(i) = S_2(j) \end{cases} \end{cases}$$

N	9									
O	8									
I	7									
T	6									
N	5									
E	4									
T	3									
N	2	3								
I	1	2	3							
#	0	1	2	3	4	5	6	7	8	9
	#	E	X	E	C	U	T	I	O	N

Засварын хэмжээний хүснэгт

N	9	8	9	10	11	12	11	10	9	8
O	8	7	8	9	10	11	10	9	8	9
I	7	6	7	8	9	10	9	8	9	10
T	6	5	6	7	8	9	8	9	10	11
N	5	4	5	6	7	8	9	10	11	10
E	4	3	4	5	6	7	8	9	10	9
T	3	4	5	6	7	8	7	8	9	8
N	2	3	4	5	6	7	8	7	8	7
I	1	2	3	4	5	6	7	6	7	8
#	0	1	2	3	4	5	6	7	8	9
	#	E	X	E	C	U	T	I	O	N

Зэрэгцүүлэлтийн тооцоолол

- Засварын хэмжээ нь хангалтгүй
 - Хоёр тэмдэгт мөрийн тэмдэгт бүрийг нөгөөгийн тэмдэгт мөртэй харгалзаж байгааг **зэрэгцүүлэх** шаардлага байнга гардаг.
- “буцах мөр”-ийг хадгалах замаар үүнийг олдог
- Үргэлж нэг нүд рүү очдог, аль нүднээс ирсэнээ санана
- Төгсгөлд хүрсэн үедээ,
 - Баруун дээд өнцгөөс эхлэн шилжилтийн дагуу буцаж мөрдөнө

Засварын хэмжээ

$$D(i,j) = \min \begin{cases} D(i-1,j) + 1 \\ D(i,j-1) + 1 \\ D(i-1,j-1) + \begin{cases} 2; & \text{if } S_1(i) \neq S_2(j) \\ 0; & \text{if } S_1(i) = S_2(j) \end{cases} \end{cases}$$

N	9	8	9	10	11	12	11	10	9	8
O	8	7	8	9	10	11	10	9	8	9
I	7	6	7	8	9	10	9	8	9	10
T	6	5	6	7	8	9	8	9	10	11
N	5	4	5	6	7	8	9	10	11	10
E	4	3	4	5	6	7	8	9	10	9
T	3	4	5	6	7	8	7	8	9	8
N	2	3	4	5	6	7	8	7	8	7
I	1	2	3	4	5	6	7	6	7	8
#	0	1	2	3	4	5	6	7	8	9
	#	E	X	E	C	U	T	I	O	N

Буцах мөртэй ХаБаЗаХэмжээ

n	9	↓ 8	↙←↓ 9	↙←↓ 10	↙←↓ 11	↙←↓ 12	↓ 11	↓ 10	↓ 9	↙ 8	
o	8	↓ 7	↙←↓ 8	↙←↓ 9	↙←↓ 10	↙←↓ 11	↓ 10	↓ 9	↙ 8	← 9	
i	7	↓ 6	↙←↓ 7	↙←↓ 8	↙←↓ 9	↙←↓ 10	↓ 9	↙ 8	← 9	← 10	
t	6	↓ 5	↙←↓ 6	↙←↓ 7	↙←↓ 8	↙←↓ 9	↙ 8	← 9	← 10	←↓ 11	
n	5	↓ 4	↙←↓ 5	↙←↓ 6	↙←↓ 7	↙←↓ 8	↙←↓ 9	↙←↓ 10	↙←↓ 11	↙↓ 10	
e	4	↙ 3	← 4	↙← 5	← 6	← 7	←↓ 8	↙←↓ 9	↙←↓ 10	↓ 9	
t	3	↙←↓ 4	↙←↓ 5	↙←↓ 6	↙←↓ 7	↙←↓ 8	↙ 7	←↓ 8	↙←↓ 9	↓ 8	
n	2	↙←↓ 3	↙←↓ 4	↙←↓ 5	↙←↓ 6	↙←↓ 7	↙←↓ 8	↓ 7	↙←↓ 8	↙ 7	
i	1	↙←↓ 2	↙←↓ 3	↙←↓ 4	↙←↓ 5	↙←↓ 6	↙←↓ 7	↙ 6	← 7	← 8	
#	0	1	2	3	4	5	6	7	8	9	
	#	e	x	e	c	u	t	i	o	n	

Хамгийн бага засварын хэмжээнд буцах мөрийг НЭМЭХ НЬ

- Үндсэн нөхцөлүүд:

$$D(i, 0) = i$$

$$D(0, j) = j$$

Дуусгавар:

$D(N, M)$ is distance

- Рекурент хамаарал:

For each $i = 1 \dots M$

For each $j = 1 \dots N$

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

арилгах

оруулах

$$\begin{cases} 2; & \text{if } X(i) \neq Y(j) \\ 0; & \text{if } X(i) = Y(j) \end{cases}$$

орлуулах

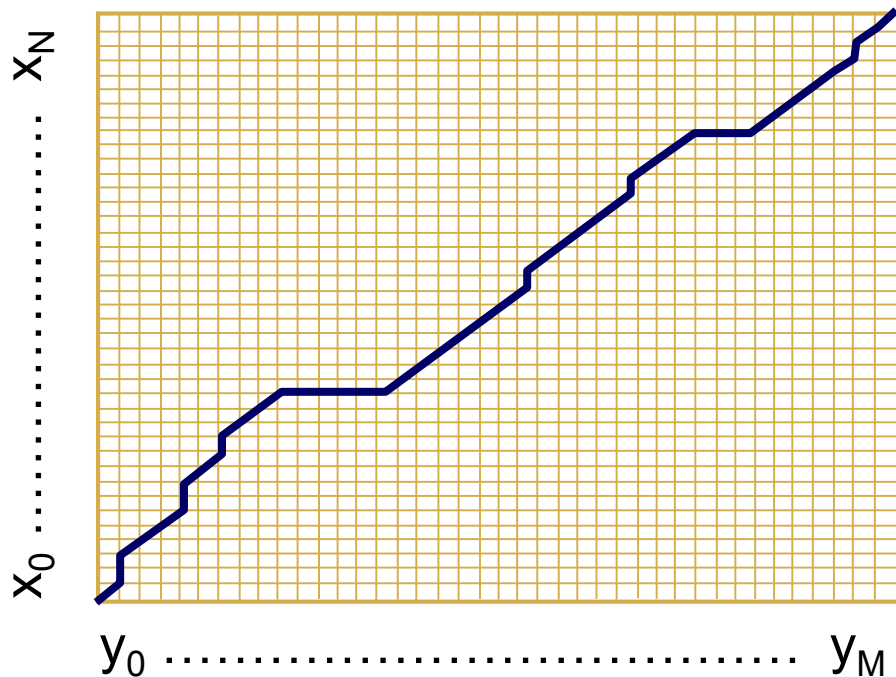
$$\text{ptr}(i, j) = \begin{cases} \text{LEFT} \\ \text{DOWN} \\ \text{DIAG} \end{cases}$$

оруулалт

арилгалт

орлуулалт

Хэмжээний матрици



Бүх зам доошилдоггүй

$(0,0)$ -ээс (M, N) лүү

хоёр дарааллын
зэрэгцүүлэлттэй холбоотой

Оновчтой зэрэгцүүлэлт нь
оновчтой дэд зэрэгцүүлэлтүүдээс
бүрдэнэ

Буцах мөрийн үр дүн

- Хоёр тэмдэгт мөр ба тэдгээрийн **зэрэгцүүлэлт**:

I	N	T	E	*	N	T	I	O	N
*	E	X	E	C	U	T	I	O	N

Гүйцэтгэл

- Хугацаа:

$O(nm)$

- Санах ой:

$O(nm)$

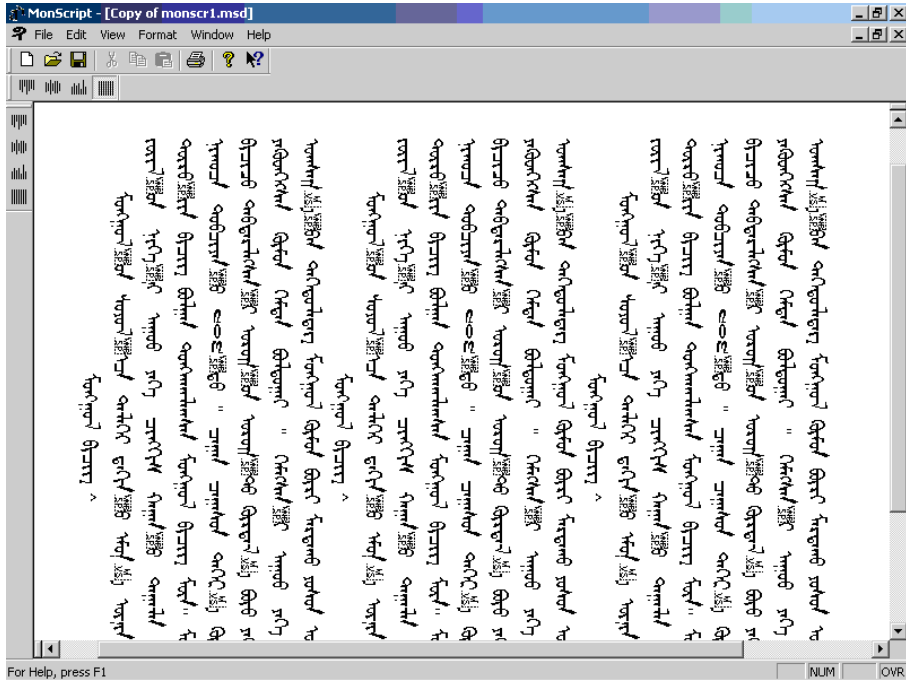
- Буцах мөр

$O(n+m)$

Хамгийн бага

засварын хэмжээ

Жигнэсэн хамгийн бага
засварын хэмжээ



Жигнэсэн засварын хэмжээ

- Яагаад тооцоололд жинг нэмсэн вэ?
 - Зөв бичгийн алдаа зүгшрүүлэлт: зарим үсгүүд нь бусдыгаа бодвол буруу бичигдсэн байх магадлал илүү байдаг
 - Биологи: арилгах эсвэл оруулах зарим төрөл бусдыгаа бодвол илүү магадлалтай байдаг

Зөв бичгийн алдааны андуурлын матрици

sub[X, Y] = Substitution of X (incorrect) for Y (correct)

X	Y (correct)																									
	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z
a	0	0	7	1	342	0	0	2	118	0	1	0	0	3	76	0	0	1	35	9	9	0	1	0	5	0
b	0	0	9	9	2	2	3	1	0	0	0	5	11	5	0	10	0	0	2	1	0	0	8	0	0	0
c	6	5	0	16	0	9	5	0	0	0	1	0	7	9	1	10	2	5	39	40	1	3	7	1	1	0
d	1	10	13	0	12	0	5	5	0	0	2	3	7	3	0	1	0	43	30	22	0	0	4	0	2	0
e	388	0	3	11	0	2	2	0	89	0	0	3	0	5	93	0	0	14	12	6	15	0	1	0	18	0
f	0	15	0	3	1	0	5	2	0	0	0	3	4	1	0	0	0	6	4	12	0	0	2	0	0	0
g	4	1	11	11	9	2	0	0	0	1	1	3	0	0	2	1	3	5	13	21	0	0	1	0	3	0
h	1	8	0	3	0	0	0	0	0	0	2	0	12	14	2	3	0	3	1	11	0	0	2	0	0	0
i	103	0	0	0	146	0	1	0	0	0	0	6	0	0	49	0	0	0	2	1	47	0	2	1	15	0
j	0	1	1	9	0	0	1	0	0	0	0	2	1	0	0	0	0	0	5	0	0	0	0	0	0	0
k	1	2	8	4	1	1	2	5	0	0	0	0	5	0	2	0	0	0	6	0	0	0	4	0	0	3
l	2	10	1	4	0	4	5	6	13	0	1	0	0	14	2	5	0	11	10	2	0	0	0	0	0	0
m	1	3	7	8	0	2	0	6	0	0	4	4	0	180	0	6	0	0	9	15	13	3	2	2	3	0
n	2	7	6	5	3	0	1	19	1	0	4	35	78	0	0	7	0	28	5	7	0	0	1	2	0	2
o	91	1	1	3	116	0	0	0	25	0	2	0	0	0	0	14	0	2	4	14	39	0	0	0	18	0
p	0	11	1	2	0	6	5	0	2	9	0	2	7	6	15	0	0	1	3	6	0	4	1	0	0	0
q	0	0	1	0	0	0	27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
r	0	14	0	30	12	2	2	8	2	0	5	8	4	20	1	14	0	0	12	22	4	0	0	1	0	0
s	11	8	27	33	35	4	0	1	0	1	0	27	0	6	1	7	0	14	0	15	0	0	5	3	20	1
t	3	4	9	42	7	5	19	5	0	1	0	14	9	5	5	6	0	11	37	0	0	2	19	0	7	6
u	20	0	0	0	44	0	0	0	64	0	0	0	0	2	43	0	0	4	0	0	0	0	2	0	8	0
v	0	0	7	0	0	3	0	0	0	0	0	1	0	0	1	0	0	0	8	3	0	0	0	0	0	0
w	2	2	1	0	1	0	0	2	0	0	1	0	0	0	0	7	0	6	3	3	1	0	0	0	0	0
x	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0
y	0	0	2	0	15	0	1	7	15	0	0	0	2	0	6	1	0	7	36	8	5	0	0	1	0	0
z	0	0	0	7	0	0	0	0	0	0	0	7	5	0	0	0	0	2	21	3	0	0	0	0	3	0



Жигнэсэн хамгийн бага засварын хэмжээ

- Эхлэл:

$$D(0, 0) = 0$$

$$D(i, 0) = D(i-1, 0) + \text{del}[x(i)]; \quad 1 < i \leq N$$

$$D(0, j) = D(0, j-1) + \text{ins}[y(j)]; \quad 1 < j \leq M$$

- Рекурэнт хамаарал:

$$D(i, j) = \min \begin{cases} D(i-1, j) & + \text{del}[x(i)] \\ D(i, j-1) & + \text{ins}[y(j)] \\ D(i-1, j-1) & + \text{sub}[x(i), y(j)] \end{cases}$$

- Дуусгавар:

$D(N, M)$ бол хэмжээ