Recap of the last lecture

Regular Expression (RE)

- A standard notation of characterizing a text sequence
- How can we search for any of the following:
  - woodchuck
  - woodchucks
  - Woodchuck
  - Woodchucks

- RE search requires a pattern and a corpus of texts to search through.

Morphology: Definition

The study of words, how they are formed, and their relationship to other words in the same language.

The Porter Stemmer (Porter, 1980)

- A simple rule-based algorithm for stemming
- An example of a HEURISTIC method
- Based on rules like:
  - ATIONAL -> ATE (e.g., *relational* -> *relate*)
- The algorithm consists of seven sets of rules, applied in order

Spelling Error: Minimum Edit Distance

How similar are two strings?

- Spell correction
  - The user typed “Appl”

    Which is closest?
    - App
    - Appeal
    - Apple

- Computational Biology
  - Align two sequences of nucleotides

    AGGCTATCACCTGACCTCCAGGCCGATGCCC
    TAGCTATCACGACCGCGGTCGATTTGCCCGAC
  - Resulting alignment:

    -AGG**CTATCAC**C**TGACCT**C**CA**G**G**C**CGA--TGCCC---**
    TAG-CTATCAC--GACCG**C--GG**T**CGA**TT**TGCCC**GAC

- Also for Machine Translation, Information Extraction, Speech Recognition

Edit Distance

- The minimum edit distance between two strings
- Is the minimum number of editing operations
  - Insertion **(I)**
  - Deletion **(D)**
  - Substitution **(S)**
- Need to transform one into the other

Minimum Edit Distance

- Two strings and their **alignment**: TRIAL vs ZEIL

```
T  R  I  A  L
|  |  |  |  |
Z  E  I  *  L
s  s     d
```

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

- If each operation has cost of 1
  - Distance between these is 3
- If substitutions cost 2 (Levenshtein)
  - Distance between them is 5

Other uses of Edit Distance in NLP

- Evaluating Machine Translation and speech recognition

Spokesman confirms senior government adviser was shot

Spokesman said the senior adviser was shot dead

S I D I

- **Named Entity Extraction** and **Entity Coreference**
  - IBM Inc. announced today
  - IBM profits
  - US President Donald Trump announced yesterday
  - for United States President Donald Trump

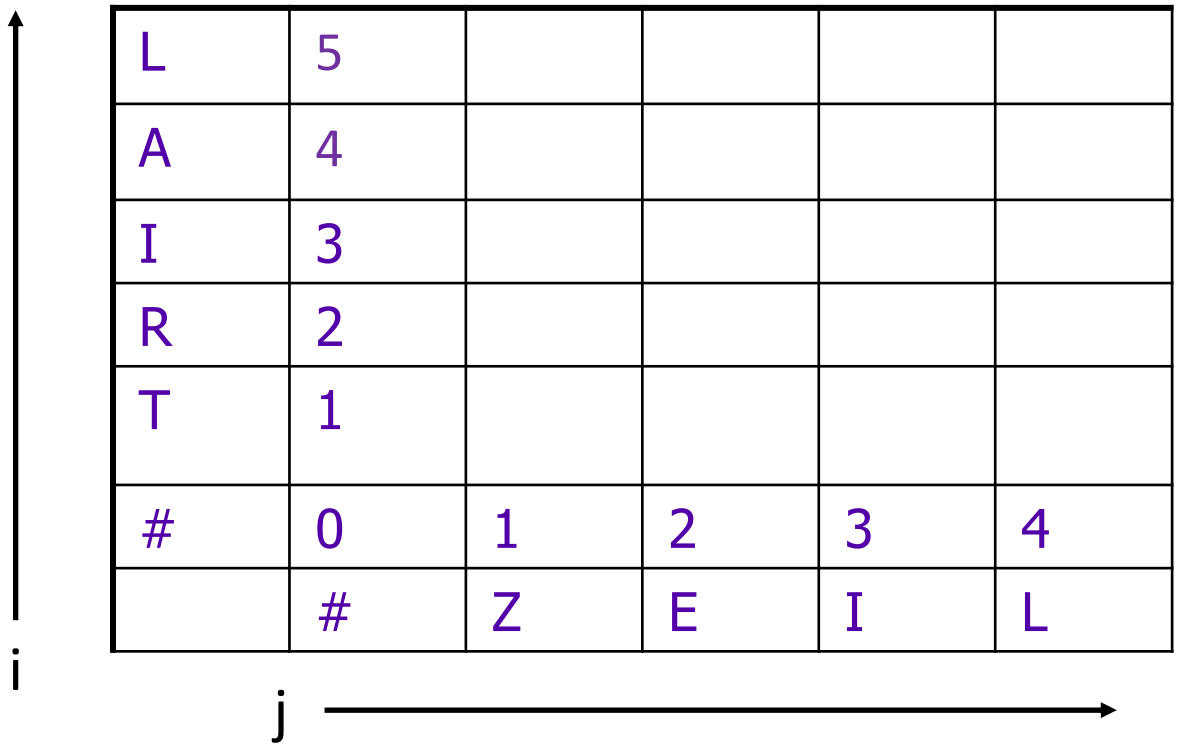
Minimum Edit as Search

- But the space of all edit sequences is huge!
  - We can't afford to navigate naïvely
  - Lots of distinct paths wind up at the same state.
    - We don't have to keep track of all of them

Defining Min Edit Distance

- For two strings
  - X of length n
  - Y of length m
- We define D(i,j)
  - the edit distance between X[1..i] and Y[1..j]
    - i.e., the first i characters of X and the first j characters of Y
- The edit distance between X and Y is thus D(n,m)

The Edit Distance Table



L	5				
A	4				
I	3				
R	2				
T	1				
#	0	1	2	3	4
	#	Z	E	I	L

The Edit Distance Table

i		j=0	1	2	3	4
5	L	5	6	7	6	**5**
4	A	4	5	6	5	6
3	I	3	4	5	4	5
2	R	2	3	4	5	6
1	T	1	2	3	4	5
0	#	0	1	2	3	4
		#	Z	E	I	L

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

The Edit Distance Table

(INTENSION vs EXECUTION)

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

Computing Alignments

- Edit distance isn't sufficient
  - We often need to **align** each character of the two strings to each other
- We do this by keeping a "backtrace"
- Every time we enter a cell, remember where we came from
- When we reach the end,
  - Trace back the path from the upper right corner to read off the alignment

Edit Distance with Backtrace

L	5	6	7	6	**5**
A	4	5	6	5	6
I	3	4	5	4	5
R	2	3	4	5	6
T	1	2	3	4	5
#	0	1	2	3	4
	#	Z	E	I	L

Substitute

No Change

Delete

T R I A L

| | | | | |

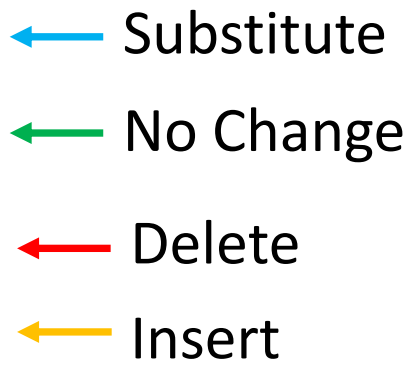
Z E I * L

s s d

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

Edit Distance with Backtrace (Another Path)

L	5	6	7	6	**5**
A	4	5	6	5	6
I	3	4	5	4	5
R	2	3	4	5	6
T	1	2	3	4	5
#	0	1	2	3	4
	#	Z	E	I	L



Cost is same, i.e., 5

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

MinEdit with Backtrace

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**	

Adding Backtrace to Minimum Edit Distance

- Base conditions:

  $D(i,0) = i$ $\qquad$ $D(0,j) = j$

- Recurrence Relation:

  For each $i = 1 \ldots M$

  For each $j = 1 \ldots N$

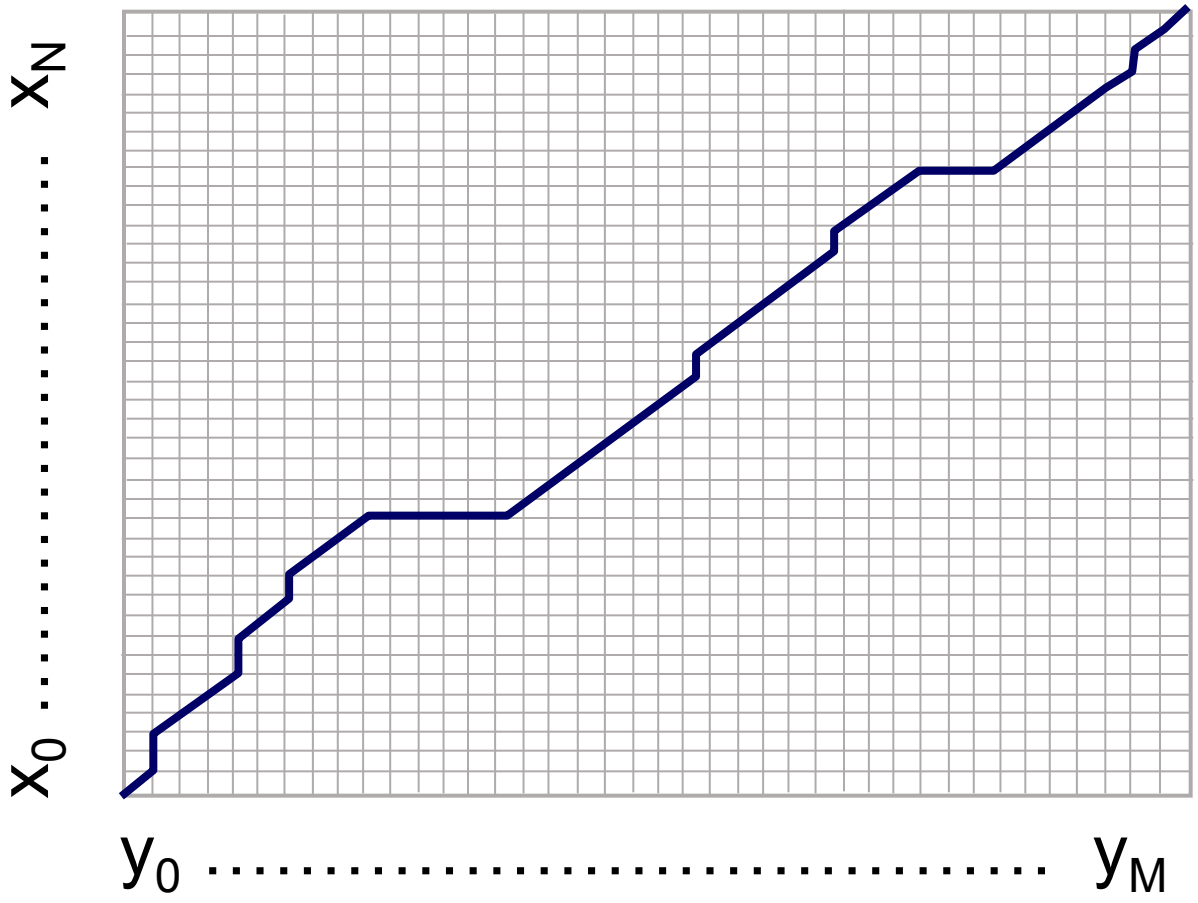
$$D(i,j) = \min \begin{cases} D(i-1,j) + 1 & \text{deletion} \\ D(i,j-1) + 1 & \text{insertion} \\ D(i-1,j-1) + \begin{cases} 2; & \text{if } X(i) \neq Y(j) \\ 0; & \text{if } X(i) = Y(j) \end{cases} & \text{substitution} \end{cases}$$

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

Termination:

$D(N,M)$ is distance

The Distance Matrix



Every non-decreasing path

from (0,0) to (M, N)

corresponds to
an alignment
of the two sequences

An optimal alignment is composed of optimal subalignments

Performance

- Time: $O(nm)$
- Space: $O(nm)$
- Backtrace: $O(n+m)$