

بسم الله الرحمن الرحيم

پردازش زبانهای طبیعی

جلسه ۷

احسان الدین عسگری

اسفند ۱۴۰۲



<http://language.ml/>

دانشکده مهندسی کامپیوتر

آزمایشگاه پردازش هوشمند متن و زبان و علوم انسانی محاسباتی

asgari@berkeley.edu

-اصلاح ایرادهای نگارشی



۱- تصحیح املاء

- کدام کلمه فرم صحیح «مسومیت» است؟
- مسئولیت
- مصونیت
- مسمومیت
- معصومیت

۲- بیوانفورماتیک: مقایسه رشته‌های زیستی

AGGCTATCACCTGACCTCCAGGCCGATGCCC
TAGCTATCACGACCGCGGTCGATTTGCCCGAC



—AGGCTATCACCTGACCTCCAGGCCGA—TGCCC—
TAG—CTATCAC—GACCGC—GGTCGATTTGCCCGAC

و کاربردهای دیگر مثل ترجمه ماشینی، استخراج اطلاعات، و تشخیص گفتار

فاصله ویرایشی

- فاصله کمینه ویرایشی بین دو رشته با کمترین تعداد عملگرهای برای تبدیل به رشته دیگر:
 - اضافه شدن
 - حذف
 - تغییر حرف

فاصله ویرایشی

تطبیق دو رشته

| | | | | | | |
|---|---|---|---|---|---|---|
| م | س | * | و | م | پ | ت |
| | | | | | | |
| م | س | ء | و | ل | پ | ت |
| | | ۱ | | ۲ | | |

کاربردهای دیگر در پردازش متن

○ ارزیابی ترجمه ماشینی یا تشخیص گفتار

R Spokesman confirms senior government adviser was appointed

H Spokesman said the senior adviser was appointed

S

I

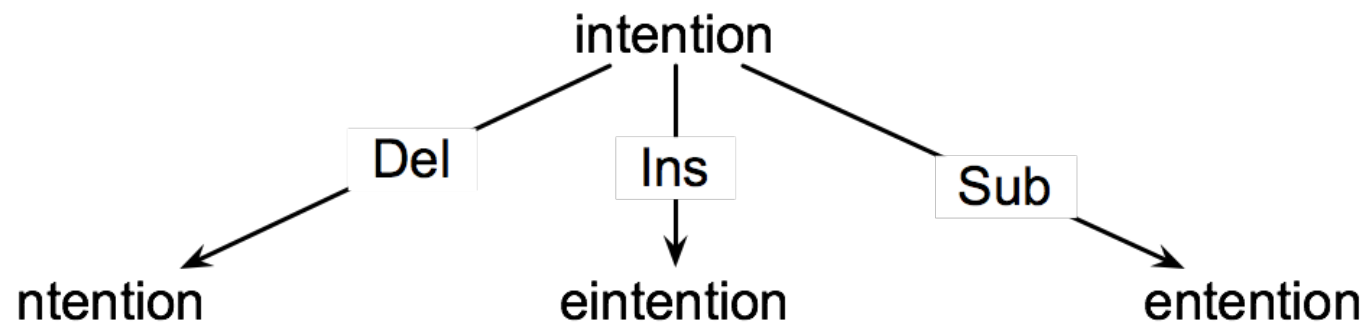
D

تجربیات بین دو رشته

پیدا کردن فاصله کمینه

تمام مسیرها عملی
و بهینه نیست

- شروع: رشته اولیه
- عملگرها: تولید، حذف، تبدیل
- هدف: رشته ثانویه
- تابع هدف: کمینه کردن هزینه‌ها



پیدا کردن فاصله کمینه ویرایشی

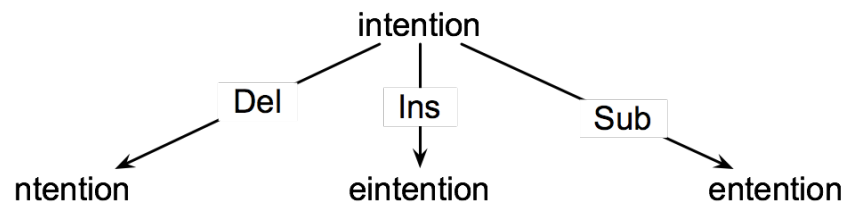
برای دو رشته

• X به طول n

• Y به طول m

را اینگونه تعریف می‌کنیم؟ $D(i, j)$

• فاصله کمینه ویرایشی $X[1..i]$ and $Y[1..j]$



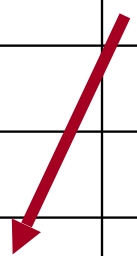
جدول فاصله کمینه

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| N | 9 | | | | | | | | | |
| O | 8 | | | | | | | | | |
| I | 7 | | | | | | | | | |
| T | 6 | | | X | | | | | | |
| N | 5 | | | Z | Y | | | | | |
| E | 4 | | | | | | | | | |
| T | 3 | | | | | | | | | |
| N | 2 | | | | | | | | | |
| 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 |

جدول فاصله کمینه

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 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 |

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



تعريف فاصله کمينه

- Initialization

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

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

- Recurrence Relation:

For each $i = 1 \dots M$

For each $j = 1 \dots N$

$$D(i, j) = \min \left\{ \begin{array}{l} D(i-1, j) + 1 \\ D(i, j-1) + 1 \\ D(i-1, j-1) + \left\{ \begin{array}{ll} 2; & \text{if } X(i) \neq Y(j) \\ 0; & \text{if } X(i) = Y(j) \end{array} \right. \end{array} \right.$$

- Termination:

$D(N, M)$ is distance

| | | | | | | | | | | |
|---|---|---|---|----|----|----|----|----|----|----|
| 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 |

فاصله کمینه ویرایشی و یافتن مسیر ویرایش

| | | | | | | | | | | | |
|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|--|
| 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:

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

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

Termination:

$$D(N, M) \text{ is distance}$$

Recurrence Relation:

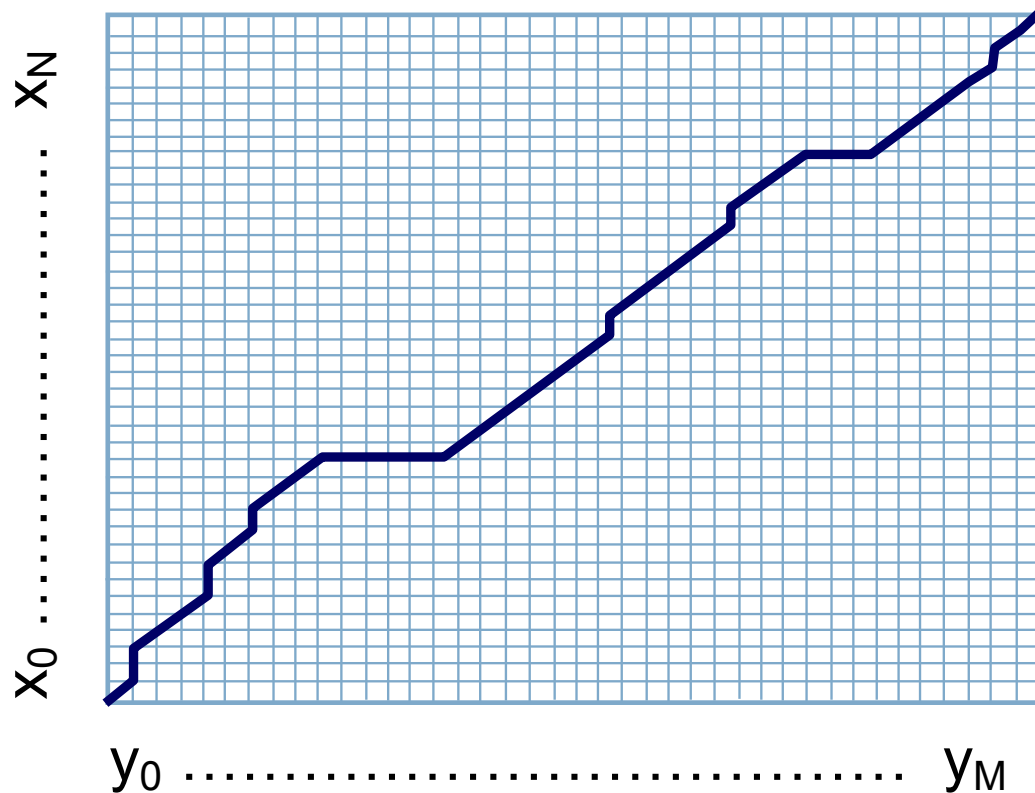
For each $i = 1 \dots M$

For each $j = 1 \dots N$

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

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

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

$O(n+m)$

وزن‌دهی الگوریتم شباهت رشته‌ها

Initialization:

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

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

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

Recurrence Relation:

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

| 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 | 2 | 0 | 8 | 0 | 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 | |



زبان‌های منظم



<http://language.ml/>

عبارات منظم برای تشخیص الگو

• زبانهای منظم

سلاااااام
سلاااااااااااام
سلااااااااااااام
سلااااااااااااام

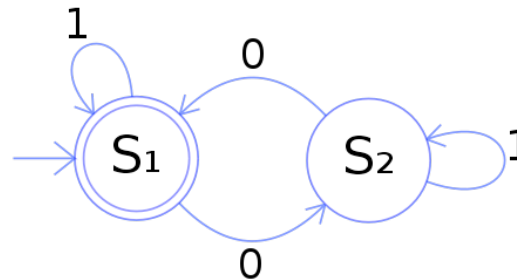
زبان منظم

• زبانهای منظم

- L is a *regular language* if and only if it is accepted by a DFA or NFA (or ϵ -NFA)

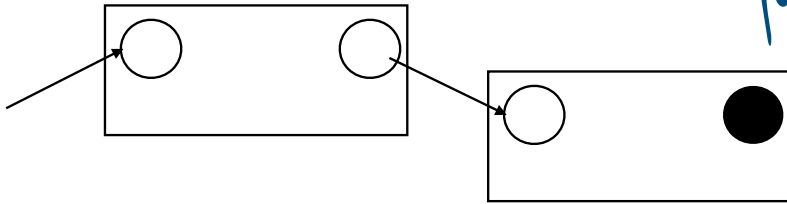
A deterministic finite automaton M is a 5-tuple, $(Q, \Sigma, \delta, q_0, F)$, consisting of

- a finite set of states Q
- a finite set of input symbols called the alphabet Σ
- a transition function $\delta : Q \times \Sigma \rightarrow Q$
- an initial or start state
- a set of accept states



- Regular languages can be specified without automata, but with *regular expressions*

عملگرهای یک زبان منظم



- ▶ Union: The union of two languages L and $M \rightarrow L \cup M$
- ▶ Dot: The concatenation of two languages L and $M \rightarrow L \cdot M$ (similar to Cartesian product $L \times M$)
- ▶ Star: The closure of a language, L^* , is defined as $L^0 \cup L^1 \cup L^2 \cup L^3 \cup \dots$, where $L^0 = \{\epsilon\}$ and $L^1 = L$

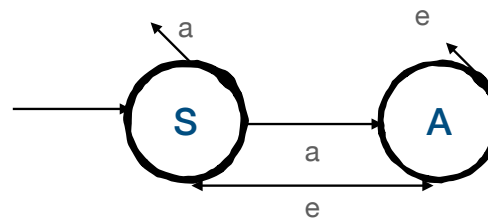
• زبانهای منظم

گرامر زبان منظم

• زبانهای منظم

گرامر خطی از راست برای زبان $L(N, \Sigma, P, S)$

1. $S \rightarrow a$
2. $S \rightarrow aA$
3. $S \rightarrow A$
4. $A \rightarrow \varepsilon$
5. $A \rightarrow S$



عبارات منظم – Disjunctions

• زبانهای منظم

| Pattern | Matches |
|---------------------------|----------------------|
| <code>[wW]oodchuck</code> | Woodchuck, woodchuck |
| <code>[1234567890]</code> | Any digit |

| Pattern | Matches | |
|--------------------|----------------------|---|
| <code>[A-Z]</code> | An upper case letter | <u>D</u> renched Blossoms |
| <code>[a-z]</code> | A lower case letter | <u>m</u> y beans were impatient |
| <code>[0-9]</code> | A single digit | Chapter <u>1</u> : Down the Rabbit Hole |

عبارات منظم - $.*?+$

| Pattern | Matches | |
|---------------|----------------------------|---|
| $colou?r$ | Optional previous char | <u>color</u> <u>colour</u> |
| $oo*h!$ | 0 or more of previous char | <u>oh!</u> <u>ooh!</u> <u>oooh!</u> <u>ooooh!</u> |
| $o+h!$ | 1 or more of previous char | <u>oh!</u> <u>ooh!</u> <u>oooh!</u> <u>ooooh!</u> |
| $o\{5,10\}h!$ | 5-10 of prev. char | |
| $baa+$ | | <u>baa</u> <u>baaa</u> <u>baaaa</u> <u>baaaaa</u> |
| $beg.n$ | | <u>begin</u> <u>begun</u> <u>begun</u> <u>beg3n</u> |

عبارات منظم - مرزها

- زبانهای منظم

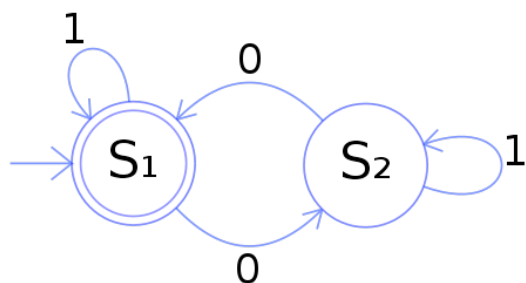
| RE | Match |
|----|-------------------|
| ^ | start of line |
| \$ | end of line |
| \b | word boundary |
| \B | non-word boundary |

عبارات منظم – اعداد و حروف

• زبانهای منظم

| RE | Expansion | Match | First Matches |
|----|--------------|-----------------------------|---------------|
| \d | [0-9] | any digit | Party_of_5 |
| \D | [^0-9] | any non-digit | Blue_moon |
| \w | [a-zA-Z0-9_] | any alphanumeric/underscore | Daiyu |
| \W | [^\w] | a non-alphanumeric | !!!! |
| \s | [\r\t\n\f] | whitespace (space, tab) | |
| \S | [^\s] | Non-whitespace | in_Concord |

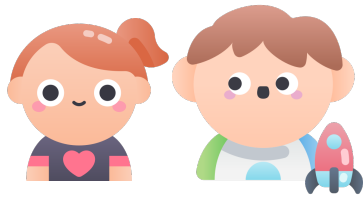
زبانهای منظم — نمایشهای معادل



$$(1|(01^*0))^*$$

یادگیری واژگان و زبان

- در دو سالگی به طور میانگین هر دو ساعت یک کلمه
- گاهی از یک کاربرد یاد می گیرند.
- تعمیم بیش از حد: سوزیدم



- یک بزرگسال ۶۰۰۰۰ کلمه

– یک بزرگسال تحصیل کرده – ۱۲۰ هزار



حیوانات هم می توانند تطابق کلمه را یاد بگیرند.



یادگیری زبان



THIS IS A WUG.



NOW THERE IS ANOTHER ONE.

THERE ARE TWO OF THEM.

THERE ARE TWO _____?

• کودکان ۴-۷ ساله

- مدل زبانی
- انواع مدل زبانی
- مدل‌های زبانی n-gram
- ارزیابی
- هموار کردن
- مدل‌های زبانی شبکه عصبی

یادگیری زبان

- حفظ کردن واژه‌هاست یا قاعده؟
- رفتار
- بدرفتار – خوش رفتار
- بدرفتاری کردن / خوش رفتاری کردن



مدل زبانی

- کلمات محدود Σ
- جملات نامحدود Σ^*
- آیا تمام جملات ممکن با این الفبا و کلمات تولید می شوند؟

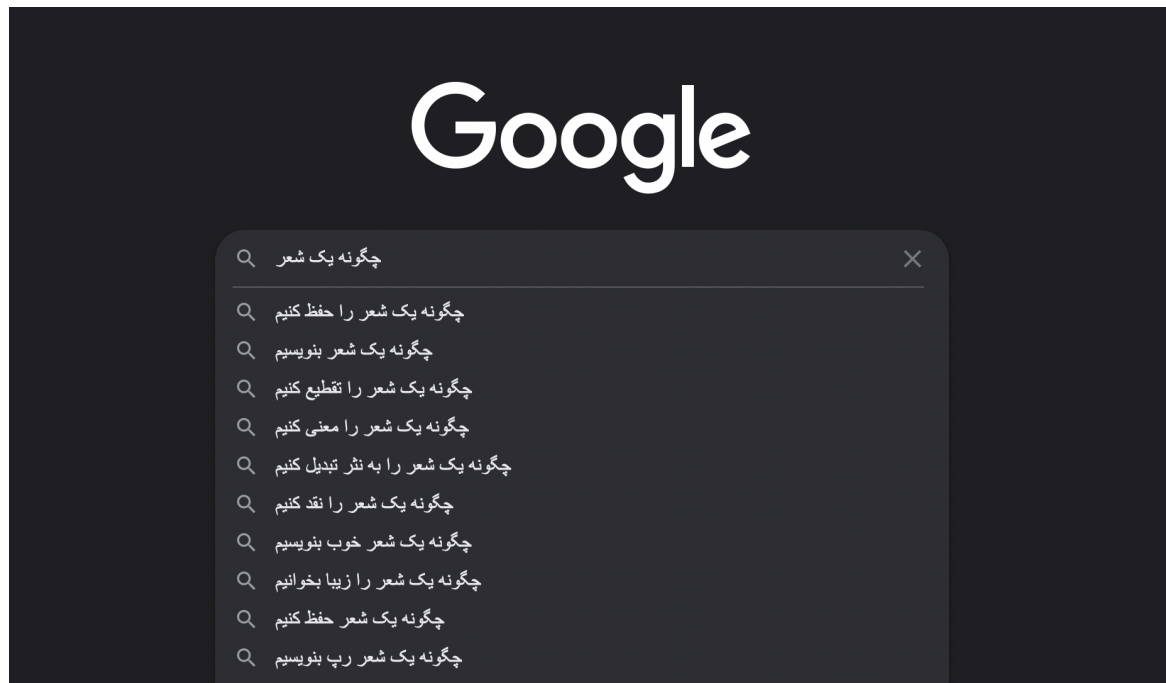
$$p(\text{start}, w_1, w_2, \dots, w_n, \text{stop})$$

- مدل زبانی
- انواع مدل زبانی
- مدل‌های زبانی **n-gram**
- ارزیابی
- هموار کردن
- مدل‌های زبانی شبکه عصبی

$p(\text{start}, w_1, w_2, \dots, w_n, \text{stop})$

کاربردها

- تکمیل خودکار



- مدل زبانی
- انواع مدل زبانی
- مدل‌های زبانی n-gram
- ارزیابی
- هموار کردن
- مدل‌های زبانی شبکه عصبی

کاربردها

$p(\text{start}, w_1, w_2, \dots, w_n, \text{stop})$

• OCR (تبدیل متن تصویری به متن دیجیتایز شده)



• چو در دهند ندای شفاعت کبری

• چو دردمند ندای شفاعت کبری

• مدل زبانی

• انواع مدل زبانی

• مدل‌های زبانی n-gram

• ارزیابی

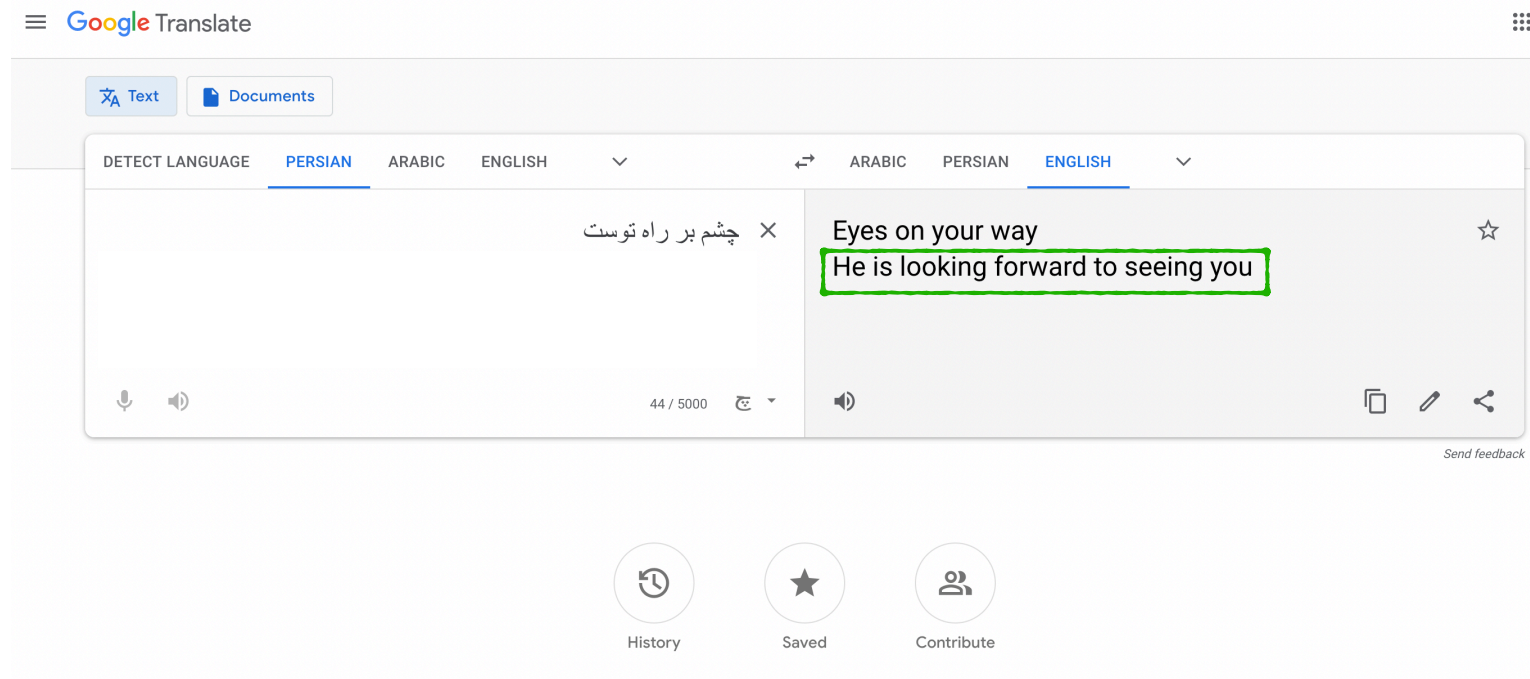
• هموار کردن

• مدل‌های زبانی شبکه عصبی

$p(\text{start}, w_1, w_2, \dots, w_n, \text{stop})$

کاربردها

• ترجمه ماشینی



- مدل زبانی
- انواع مدل زبانی
- مدل‌های زبانی n-gram
- ارزیابی
- هموار کردن
- مدل‌های زبانی شبکه عصبی