Information Extraction

Data ข้อมูล —> Information ความรู้

lingling (˘˘)❁ リン

16 127 559

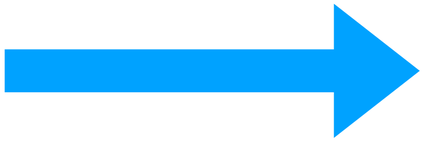
22 ก.ค. 2018

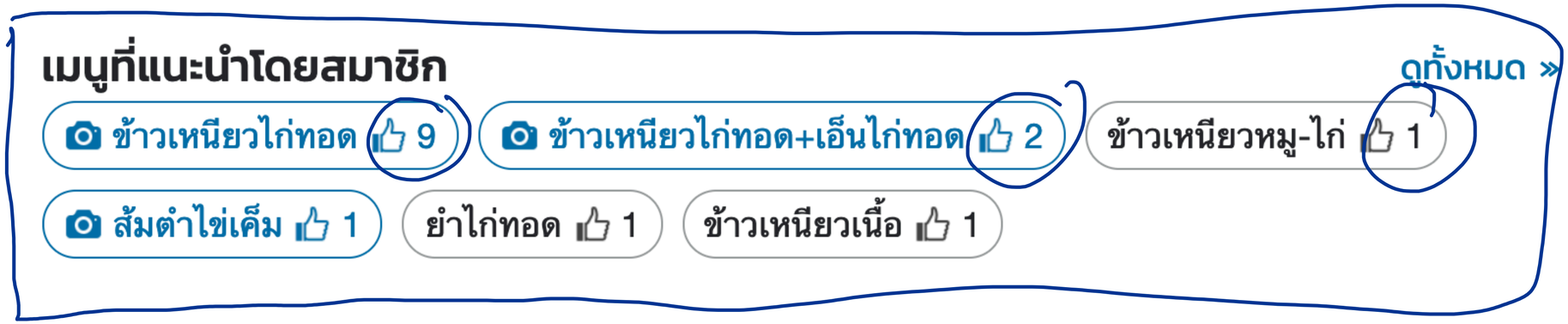
★★★★★ Quality Review

ร้านในตำนานอักษรศาสตร์ จุฬาฯ

ขอยกให้ร้านนี้เป็นร้านในตำนานอักษรศาสตร์ จุฬาฯ ^^ เราทานมาตั้งแต่สมัยเรียนปริญญาตรี เหมือนนานมากมาแล้ว และห่างหายไปนาน จนล่าสุดมาเจอในแอป NOW Food Delivery โดยบังเอิญ เลยลองสั่งมาทานค่ะ

เซตที่สั่งมาเป็นข้าวเหนียวไก่ทอด+ข้าวเหนียว ราคาตามในแอปคือ 33 บาท ใส่ในกล่องกระดาษวัสดุอย่างดี แพ็กมาสวยงามน่าทานตามรูปเลยค่ะ โดยรสชาติความอร่อยยังเหมือนเดิม ทำให้หายคิดถึงไปได้มากเลยค่ะ





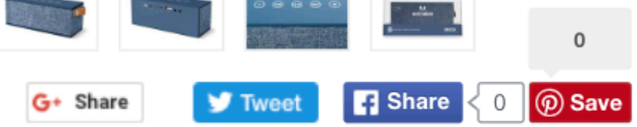
ภาษาจัดเป็นข้อมูลแบบไม่มีโครงสร้าง

- เปิดเพลงอะไรก็ได้ของปาล์มมีตอนหกโมงเย็น

Artist ID	Artist Name
1	Atom ชนกันต์
2	Palmy
3	Stamp
4	แจนจัง

18:00 น.

Unstructured Data



Rockbox Brick เป็นลำโพงไร้สายที่มีเบสขนาดใหญ่ที่มีรูปร่างอิฐแบบคลาสสิค
ใช้บลูทูธเพื่อเชื่อมต่อแบบไร้สายกับอุปกรณ์
สามารถเชื่อมต่อเข้ากับ โทรศัพท์ แท็บเล็ตหรือโน้ตบุ๊ค
สามารถเป็น Powerbank แบตเตอรี่ที่มีกำลังไฟ 4000 mAh
ฟังเพลงต่อเนื่อง 20 ชั่วโมงจากการชาร์จไฟครั้งเดียวไฟ
ขนาด 15.5 x 5.9 x 5.9 ซม.
ในชุดประกอบด้วย
Rockbox Brick
Micro-USB charging cable
3.5 mm audio cable พบสินค้าเพิ่มเติมจาก FRESHN REBEL

- สี : Indigo
- กำลังไฟฟ้า (วัตต์) : 4000 mAh

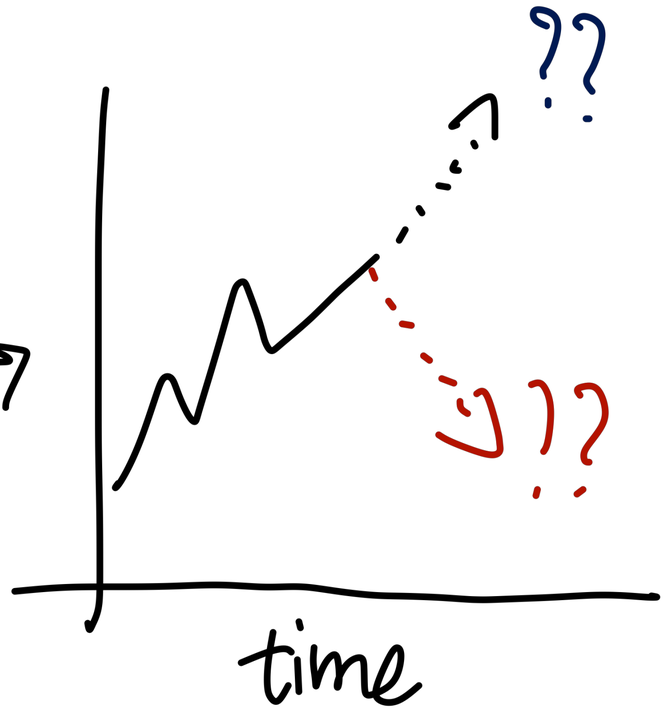
Structured Data

```
Brand = FRESHN REBEL
Color = Indigo
Type  = Portable Speaker
        Bluetooth Speaker
        Home Decor
```

Finances

financial report

company	event
APL	increase
UTD	unite
SET	fall
⋮	



Marketing + Brand Monitoring

web
review
facebook
pantip

product	sentiment
ours 1	positive
ours 1	negative
theirs 1	positive
theirs 1	positive
⋮	⋮

ours
+ 72%
− 28%

insights!

theirs
+ 63%
− 37%

Election Forecast

newspaper
pantip
facebook

where	candidate	sentiment
BKK	8	neutral
BKK	8	positive
BKK	6	positive
CNX	7	neutral
CNX	1	negative
⋮	⋮	

→ campaign at CNX

Drug Administration

Clinical notes

patient	time	event	diseases
1	0	x-ray	-
1	1	drug1	pneumonia
2	0	MRI	-
2	1	x-ray	-
2	2	drug2	infection

สกัดอะไรได้บ้าง

- ชื่อคน สถานที่ทางภูมิศาสตร์ ร้านค้า ชื่อองค์กร
- ชื่อเพลง ชื่อศิลปิน ชื่ออัลบั้ม
- วัน เวลา วันที่ เหตุการณ์
- ยีนส์ โปรตีน ชื่อยา อาการทางแพทย์ เครื่องมือการวินิจฉัย ชื่อเชื้อโรค ชื่อโรค

การสกัดความรู้ (Information Extraction)

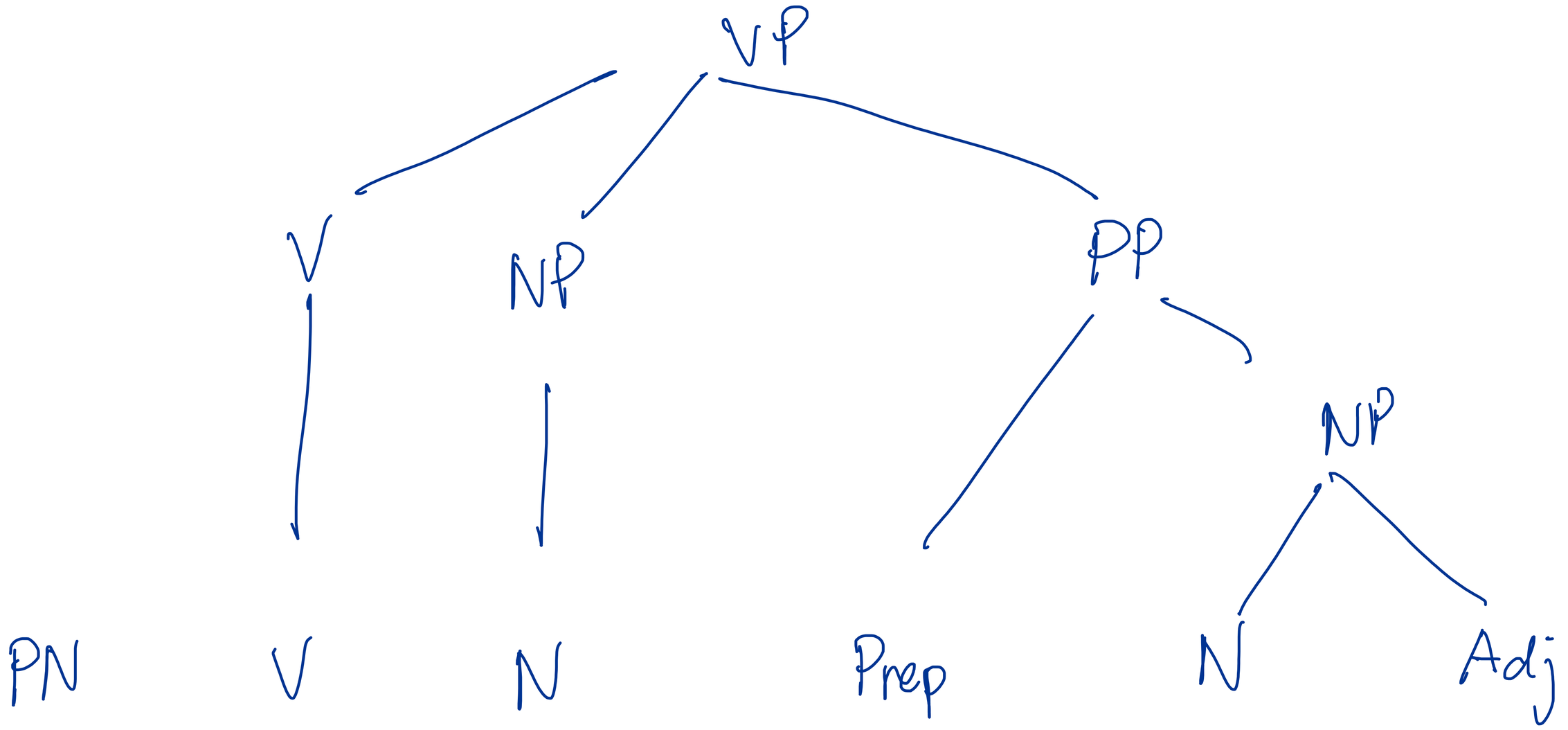
- การเปลี่ยน unstructured data (data ที่เป็น text นำไปใช้ได้ยาก) เป็น structured data (data ที่เป็นตารางสามารถนำไปใช้ได้ง่าย)

Part-of-Speech Tagging

PN V Noun

ดีแทค ลด ค่าโทร

Subject Object



ดีแทค ลด ค่าโทร สำหรับ ลูกค้า ใหม่

subject V object modifier

Universal POS Tag

Open-class words

- ADJ
- ADV
- INTJ
- NOUN
- PROPN
- VERB

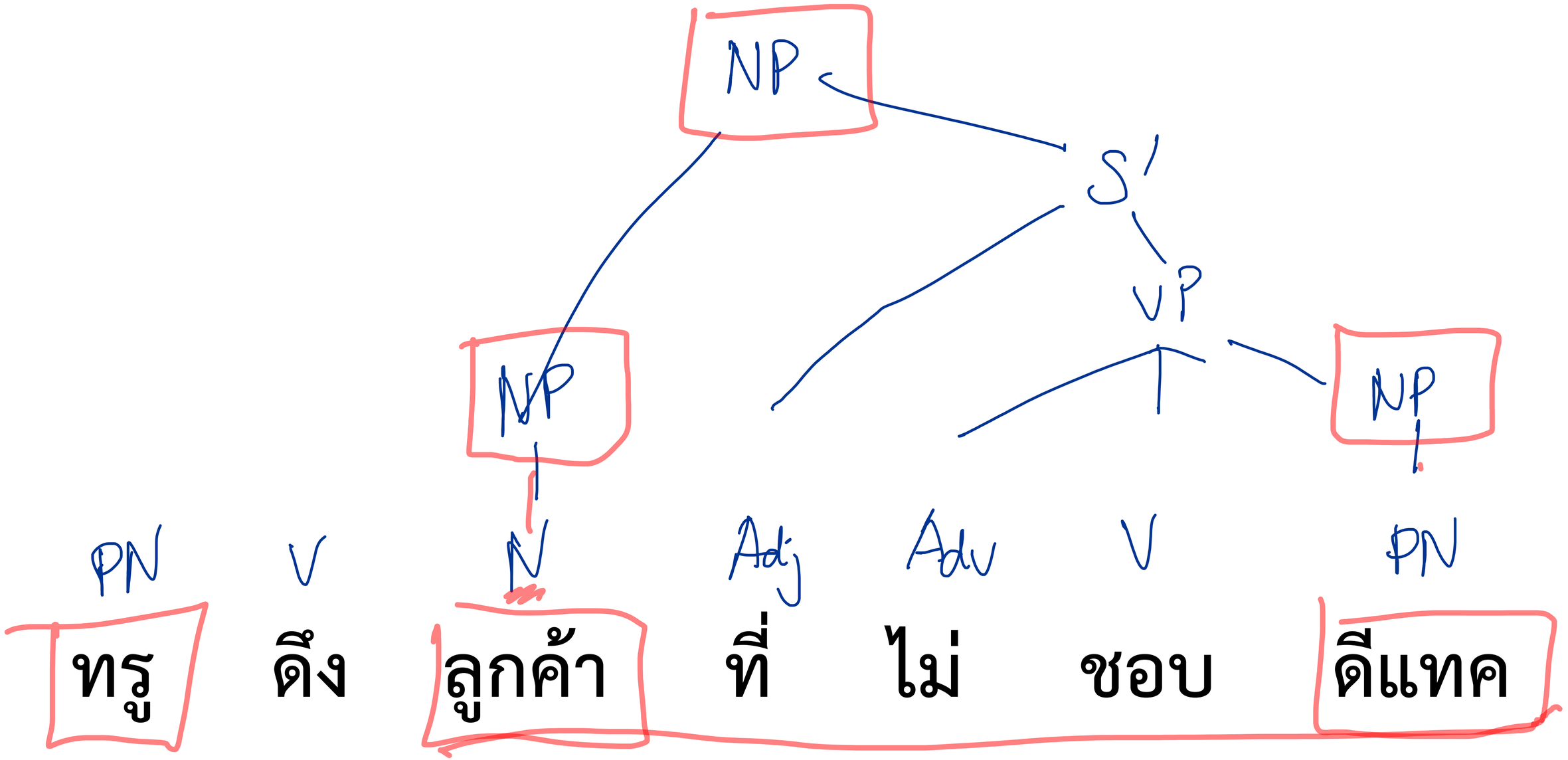
content word

Closed-class words

function word

- ADP — preposition, grammer
- AUX
- CCONJ
- DET — a, an, the, my, which
- NUM
- PART
- PRON — he, she, they
- SCONJ

Base NP chunk



1) PN* 2) N Adj 3) N

PN V N Adj Adj Adv V PN

ใหม่

ทรู ดึง ลูกค้า ที่ ไม่ ชอบ ดีแทค

V

Sequence Labeling

Part-of-Speech Tagging + Base NP

- Sequence labeling task
- การแปะชนิดของคำทำให้เราเข้าถึงความหมายได้ระดับนึง
- Base NP Chunking ช่วยสกัดความรู้เกี่ยวกับ คน สัตว์ สิ่งของ สถานที่ และสิ่งนามธรรมอื่นๆ

Named-Entity Recognition (NER)

displaCy Named Entity Visualizer

When Sebastian Thrun PERSON started working on self-driving cars at Google ORG in 2007 DATE, few people outside of the company took him seriously. “I can tell you very senior CEOs of major American NORP car companies would shake my hand and turn away because I wasn’t worth talking to,” said Thrun PERSON, now the co-founder and CEO of online higher education startup Udacity, in an interview with Recode ORG earlier this week DATE.

Named-Entity Recognition

Person

ลุงตู่ต่อว่าผู้สื่อข่าวที่ตึกไทยคู่ฟ้าเมื่อเช้า

Place

Time

IOB tagging NER Task Formulation

BI*

ลุงตู่ต่อว่าผู้สื่อข่าวที่ตึกไทยคู่ฟ้าเมื่อเช้า

O I-Time I-PER

ลุง	ตู่,	ต่อว่า	ผู้สื่อข่าว	ที่	ตึกไทยคู่ฟ้า	เมื่อ	เช้า
Noun	PNoun	Verb	Noun	Adj	PNoun	ADP	Noun
B-PER	I-PER	O	O	O	B-PLACE	B-TIME	I-TIME

begin inside outside begin place inside time

sequence labeling

ชื่อคนไทย

- สนธยา คุณปลื้ม
- กุ้งอบวุ้นเส้น
- ปรัชญาเศรษฐกิจพอเพียง
- วัฒนากระโปรงแดง

ชื่อฝรั่ง

- Are you Rich?
- Play a song by Ke$ha
- A new mission to Mars
- Dean's new single

ชื่อยีนส์ ชื่อโปรตีน

- rpsL = ribosomal protein, small S12
- polA = DNA polymerase I
- gal = galactose
- cat = chloramphenicol resistance
- amp
- azi

ชื่อยา

- quertiapine = Seroquel XR
- PN = penicillin != pneumonia
- IUPAC = 7-{4-[4-(2,3-dichlorophenyl) piperazin-1-yl] butoxy}-3,4-dihydroquinolin-2(1H)-one
- loop, potassium-sparing and thiazide diuretics (Dai et al, 2017)

การรู้จำเอ็นทิตี้

- NER มักถูกแก้ด้วย sequence labeling model โดยใช้ IOB label
- ยังจำเป็นต้องนึกถึงธรรมชาติของข้อมูลว่าประหลาดอย่างไร

Sequence Labeling Model

Sequence Labeling vs Classification

ลุง	ตู่	ต่อว่า	ผู้สื่อข่าว	ที่	ตึกไทยคู่ฟ้า	เมื่อ	เช้า
Noun	PNoun	Verb	Noun	Adj	PNoun	ADP	Noun

Sequence ของหน่วยทางภาษาต่าง ๆ

ลุง	ตู่	ต่อว่า	ผู้สื่อข่าว	ที่	ตึกไทยคู่ฟ้า	เมื่อ	เช้า
B-PER	I-PER	O	O	O	B-PLACE	B-TIME	I-TIME

ล	ุ	ง	ต	ู	่	ต	่	อ	ว	่	า	ผ	ู	้	ส	ื	่	อ	ข	่	า	ว
BP	IP	IP	IP	IP	IP	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O

Sequence ของหน่วยทางภาษาต่าง ๆ

ล ุ ง ต ู ่ ต ่ อ ว ่ า ผ ู ้ ส ื ่ อ ข ่ า ว

B O O O O O B O O O O O B O O O O O O O O O

Sequence ของหน่วยทางภาษาต่าง ๆ

เรื่องนี้คนแสดงนำหล่อ แต่เนื้อเรื่องน่าเบื่อ จบได้จืดมาก

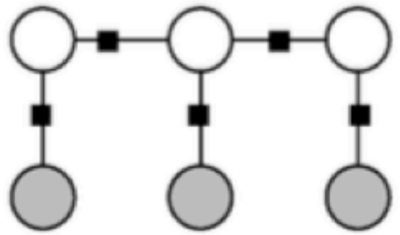
positive negative negative

Sequence Model ที่ฮิตอยู่ขณะนี้

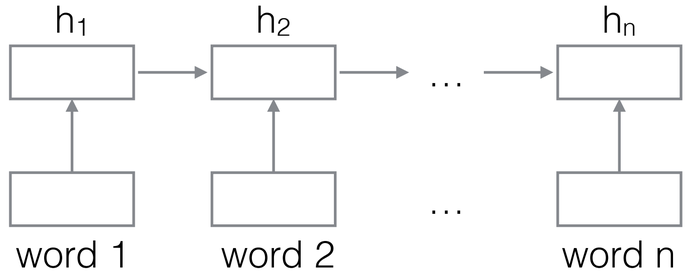
MaxEnt

Linear-chain

- Conditional Random Fields (CRF)



- Recurrent Neural Network (RNN)



- RNN + CRF

Feedforward

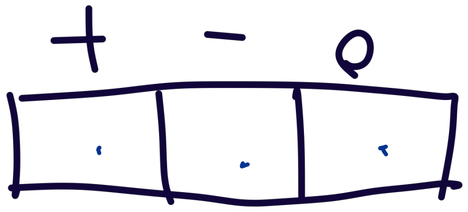
Sequence Labeling Model

- ภาษามีการเรียงตัวกันเป็นลำดับ
- ผลจะแม่นยำขึ้นถ้า Label มีความเกี่ยวเนื่องกันใน sequence และจำนวน Label = จำนวนหน่วย

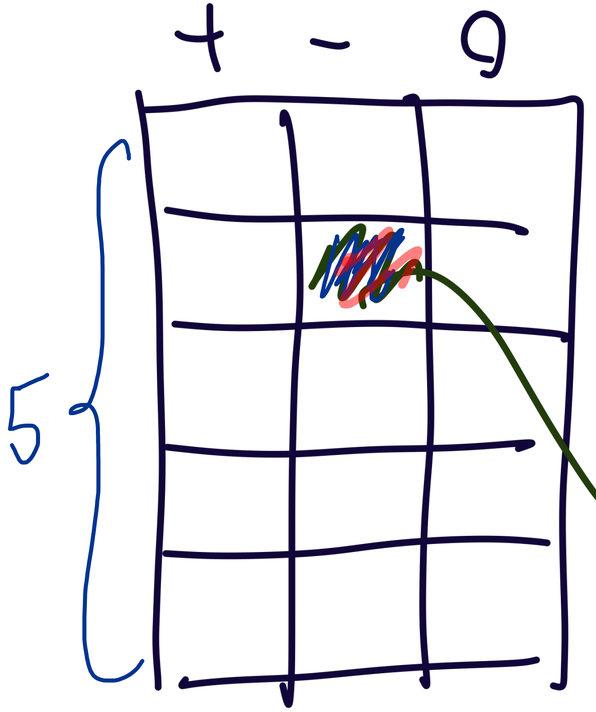
Conditional Random Fields (CRF)

Logistic Regression

MaxEnt



bias



parameter matrix

feature นี้กับ - นี่ดีมั้ย

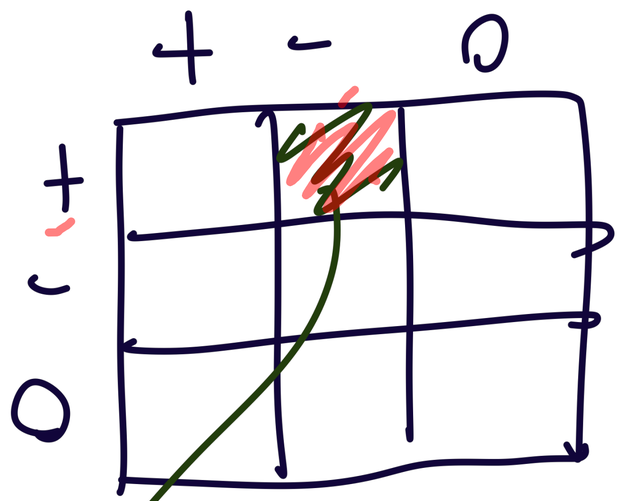
5 x 3

+

CRF

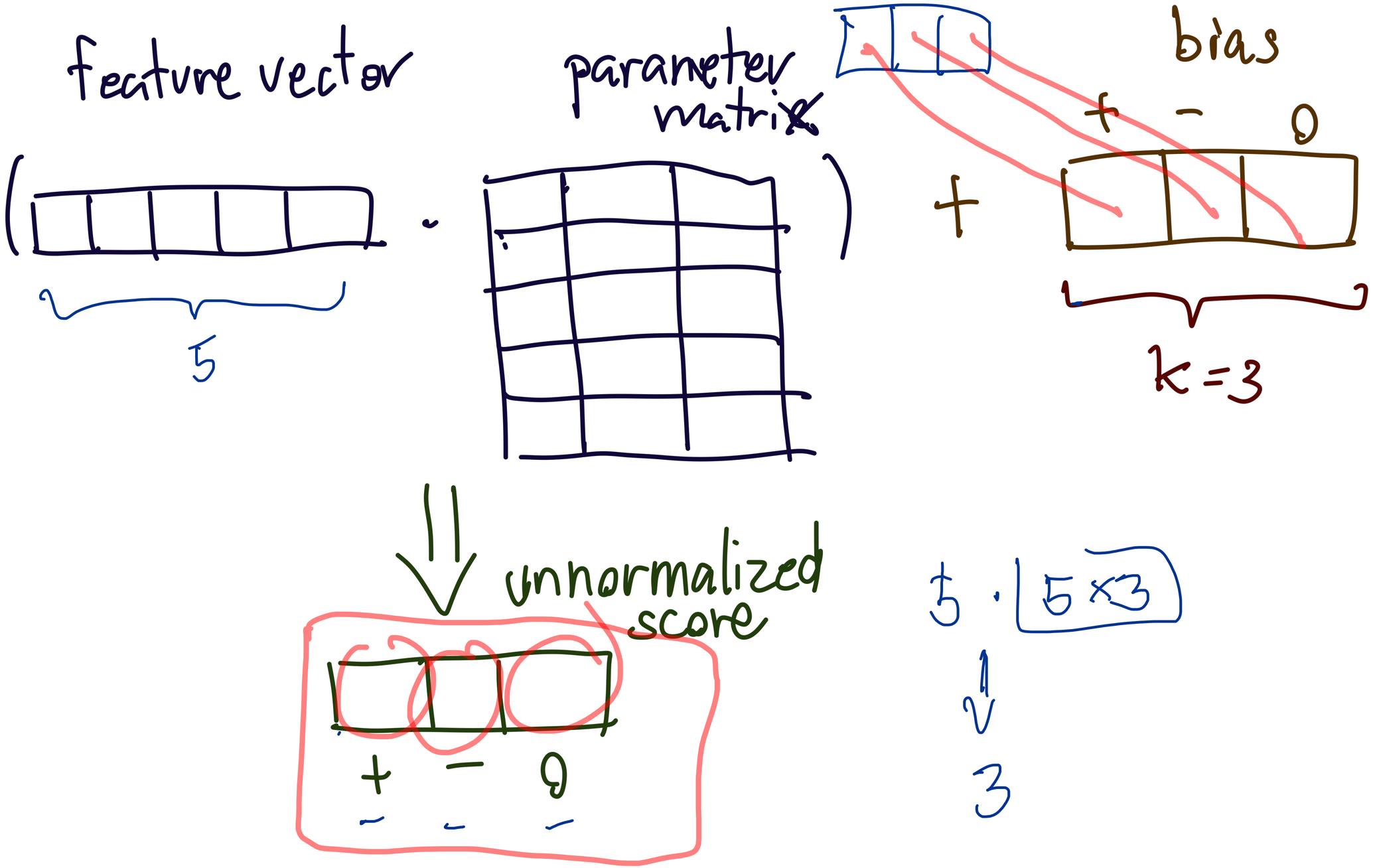
transition

label compatibility



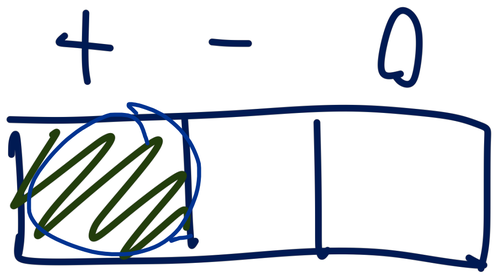
+ ตามด้วย - นี่ดีมั้ย

ค่าบวก +, -


feature vector
parameter matrix
bias
+
−
0
5
k=3
unnormalized score
+
−
0
5 · 5×3
3

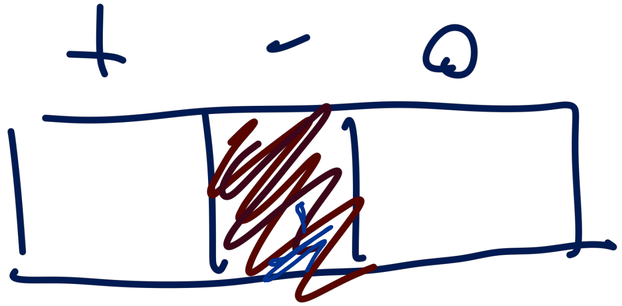
$t=0$

sentence 1



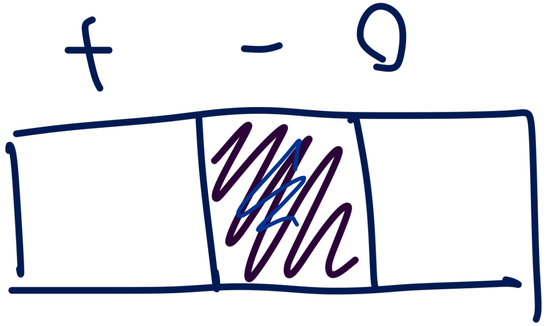
$t=1$

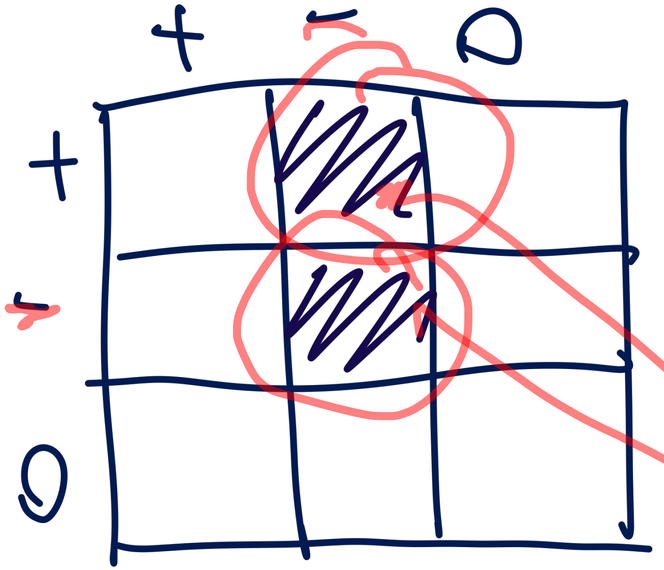
sentence 2



$t=2$

sentence 3





$$score(+,-) = score_0(+) + score_1(-) + score_2(-) + tscore(+,-) + tscore(-,-)$$

$q = 3 + 3 + 3 = 27$

$3 \times 3 \times 3$ combinations = too slow

label sequence	unnormalized	probability
+ + +	~	—
+ + −		softmax(· ·)
+ + 0		
+ − +		
+ − −		
+ − 0		
⋮		
⋮		

sum = 1

Conditional Random Fields

- Training ต้องใช้ algorithm ที่หา probability ได้เร็วๆ
- Decoding ต้องใช้ algorithm ที่หา label sequence ที่ดีสุดได้เร็วๆ

หา sequence of labels ที่ดีที่สุด

Decoding

ประโยค	เรื่องนี้คนแสดงนำหล่อ	แต่เนื้อเรื่องน่าเบื่อ	จบได้จืดมาก
Sentiment	?	?	?

คำ	They	can	fish
POS tag	? Pron	? V	? V

Training CRF

Simple Classification

label	sentence	
+	S_1	A_1
−	S_2	
−	S_3	
0	S_4	A_2
−	S_5	
+	S_6	A_3
+	S_7	
+	S_8	

Sequence Labeling

label sequence	sentence sequence
$[+, -, -]$	$[S_1\ S_2\ S_3]$
$[0, -]$	$[S_4\ S_5]$
$[+, +, +]$	$[S_6\ S_7\ S_8]$

$3 \times 3 \times 3$ combinations = too slow

label sequence	unnormalized	probability
+ + +		
+ + −		
+ + 0		
+ − +		
+ − −		
+ − 0		
⋮		
⋮		

sum = 1

Objective Function

sequence of feature vectors

$$L(\theta) = -\sum_{\text{sequences} \in D} \log P(\bar{Y} \mid \bar{X})$$

sequence of labels

gradient

[+ + 0, - -]

Training Algorithm

- Forward-Backward algorithm
- Averaged Structured Perceptron

Forward-Backward Algorithm

- คำนวณ log-likelihood ของ data (forward)
- คำนวณ gradient ของแต่ละ parameter (forward-backward)

Averaged Structured Perceptron

mini batch

30 sequences

update

- Viterbi algorithm ในการ decode แล้วปรับแก้ parameter โดยไม่ต้องใช้ gradient