How to “Calculate” Meaning

DSI-TDA-03
THAI NLP

What is Meaning?

definition of ราชบัณฑิตยสภา (Royal Institute)

ความหมาย :
ใจความที่สื่อให้เข้าใจโดยหน่วยในภาษาหรือพฤติกรรมต่างๆ
ความหมายมีหลายประเภท แบ่งตามระดับของหน่วยภาษาเป็น ๓ ประเภท ได้แก่ ...

What is Meaning?

1. ความหมายประจำคำ ได้แก่ ความหมายของหน่วยคำหรือคำ
2. ความหมายเชิงโครงสร้างหรือความหมายทางไวยากรณ์
3. ความหมายที่ตีความจากเจตนาของผู้พูด

What is Meaning?

สนใจแค่นี้ก่อน!!

1. ความหมายประจำคำ ได้แก่ ความหมายของหน่วยคำหรือคำ **- word level meaning**
2. ความหมายเชิงโครงสร้างหรือความหมายทางไวยากรณ์ **- syntactic level meaning**
3. ความหมายที่ตีความจากเจตนาของผู้พูด **- pragmatic level meaning**

What is Meaning?

concept

sign of communication

reference

What is Meaning?

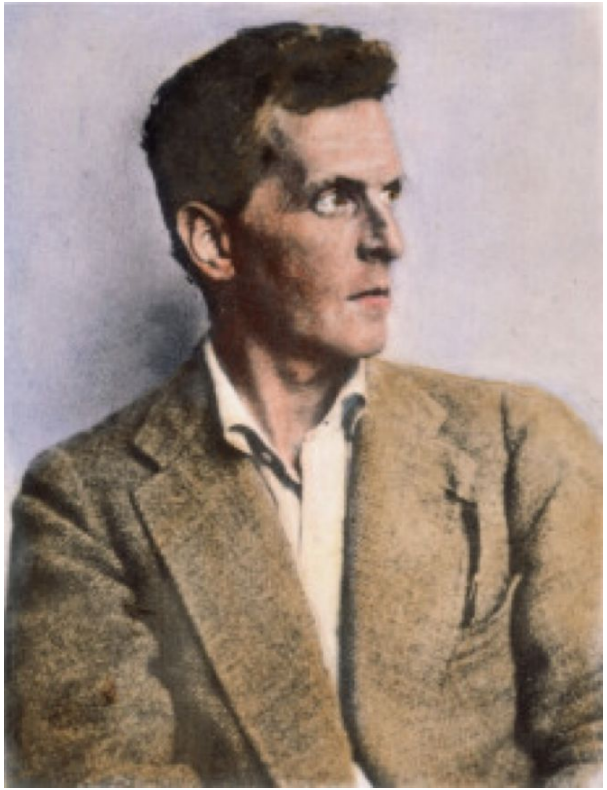
- “*love*” has no tangible reference, but we can understand the concept.
- Meaning is this abstraction process.
- Then, how we can do that? A baby who does not know language also can learn it.

What is Meaning?

Ludwig Wittgenstein

(1889 – 1951)

“In most cases, the meaning of a word is its use”



Context

ปิติ เดิน เข้า มา ใน ห้อง ท่าทาง ไม่ รื่นเริง เหมือน ทุก วัน ครู ไพลิน จึง ถาม ว่า “วันนี๊ ทำไม ปิติ จึง มา โรงเรียน สาย” ปิติ ตอบ ว่า “เจ้าแก่ ตาย แล้ว ครับ ผม จึง ต้อง ช่วย พ่อ ฝัง มัน ที โคน ต้นไม้ หลัง บ้าน” มานี ตกใจ ร้อง ว่า “อ้าว เจ้าแก่ ตาย แล้ว หรือ น่าสงสาร จริง เธอ คง เสียใจ มาก นะ เพราะ เธอ รัก มัน เหลือเกิน”

Context

before ว่า is verb

ปิติ เดิน เข้า มา ใน ห้อง ท่าทาง ไม่ รื่นเริง เหมือน ทุก วัน ครู ไพลิน จึง **ถาม** ว่า “วันนี๊ ทำไม ปิติ จึง มา โรงเรียน สาย” ปิติ **ตอบ** ว่า “เจ้าแก่ ตาย แล้ว ครับ ผม จึง ต้อง ช่วย พ่อ ฝัง มัน ที โคน ต้นไม้ หลัง บ้าน” มานี ตกใจ **ร้อง** ว่า “อ้าว เจ้าแก่ ตาย แล้ว หรือ น่าสงสาร จริง เธอ คง เสียใจ มาก นะ เพราะ เธอ รัก มัน เหลือเกิน”

Context

similar meaning/function words tend to be located in the same position in the sentence

- ผมแดกอาหาร
- ผมกินอาหาร
- ผมทานอาหาร
- ผมรับประทานอาหาร

Context : table of co-occurrence (bigram)

	ข้าว	อาหาร	ผม	มาก	ที่	งาม
กิน	132	210	5	25	82	0
ทาน	190	341	1	12	190	0
สวย	0	0	3	498	201	170
จุฬา	15	23	0	0	12	0
ลุง	0	0	102	31	45	2

Context

similar word has similar distribution

	ข้าว	อาหาร	ผม	มาก	ที่	งาม
กิน	132	210	5	25	82	0
ทาน	190	341	1	12	190	0
สวย	0	0	3	498	201	170
จุฬา	15	23	0	0	12	0
ลุง	0	0	102	31	45	2

Vector Semantics

$$\overrightarrow{kin} = (132, 210, 5, 25, 82, 0)$$

$$\overrightarrow{thaan} = (191, 341, 1, 12, 190, 0)$$

we can define the vector according to the count of co-occurrence - **Embedding**

Vector Semantics

$$\overrightarrow{kin} = (132, 210, 5, 25, 82, 0)$$

$$\overrightarrow{thaan} = (191, 341, 1, 12, 190, 0)$$

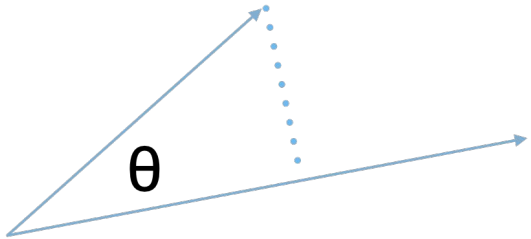
these 2 vectors are different in **size**,

but similar in **direction** → measure the angle

Vector Semantics

$$\vec{a} = (x_1, y_1, z_1, \dots) \qquad \vec{b} = (x_2, y_2, z_2, \dots)$$

$$\cos\theta = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}||\vec{b}|} = \frac{x_1x_2 + y_1y_2 + z_1z_2 + \cdots}{\sqrt{x_1^2 + y_1^2 + z_1^2 \cdots}\sqrt{x_2^2 + y_2^2 + z_2^2 \cdots}}$$



cosθ shows the similarity of the two vectors, it is called **cosine similarity**. (This can be used to compare other vectors, such as document similarity.)

Vector Semantics

$$\overrightarrow{kin} = (132, 210, 5, 25, 82, 0)$$

$$\overrightarrow{thaan} = (191, 341, 1, 12, 190, 0)$$

$$\overrightarrow{kin} \cdot \overrightarrow{thaan} = 132 \cdot 191 + 210 \cdot 341 + 5 \cdot 1 + 25 \cdot 12 + 82 \cdot 190 + 0 \cdot 0 = 112707$$

$$|\overrightarrow{kin}| = \sqrt{132^2 + 210^2 + 5^2 + 25^2 + 82^2 + 0^2} = 262.5$$

$$|\overrightarrow{thaan}| = \sqrt{191^2 + 341^2 + 1^2 + 12^2 + 190^2 + 0^2} = 434.7$$

$$\cos\theta = \frac{112707}{262.5 \cdot 434.7} = 0.9877$$

(cosθ = 1 means the same direction)

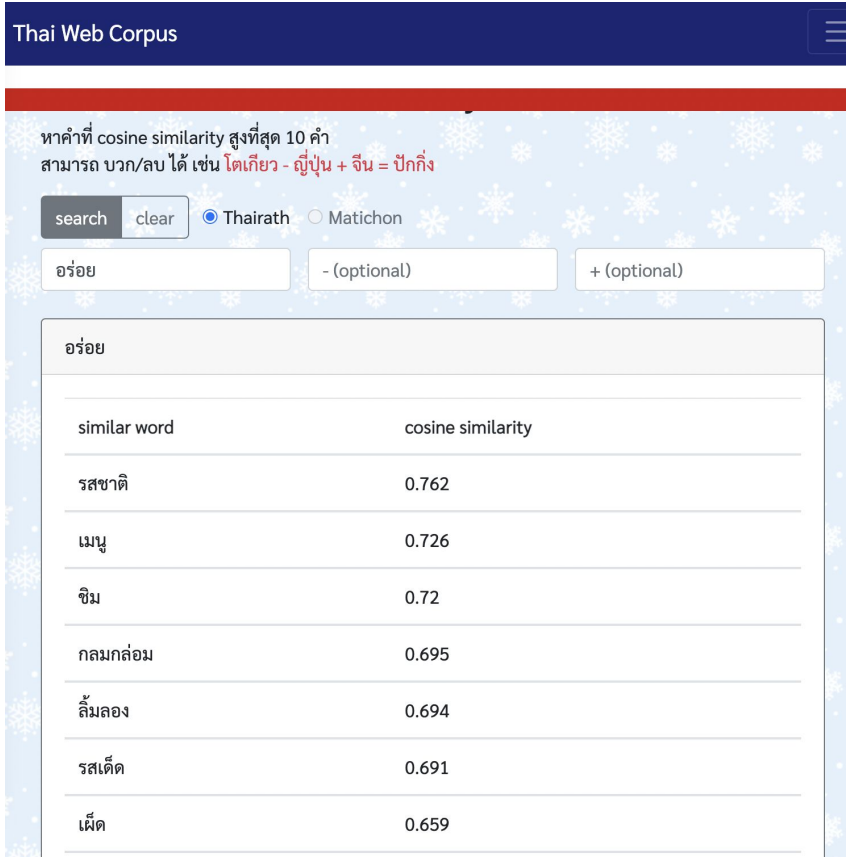
Vector Semantics

- there are many varieties of embedding methods
- actually, just to count co-occurrence is not used, because it has too many dimensions
- SVD and Deep Learning are primary methods
- Whatever method you take to get vectors, you can calculate the cos sim by the same formula

Vector Semantics

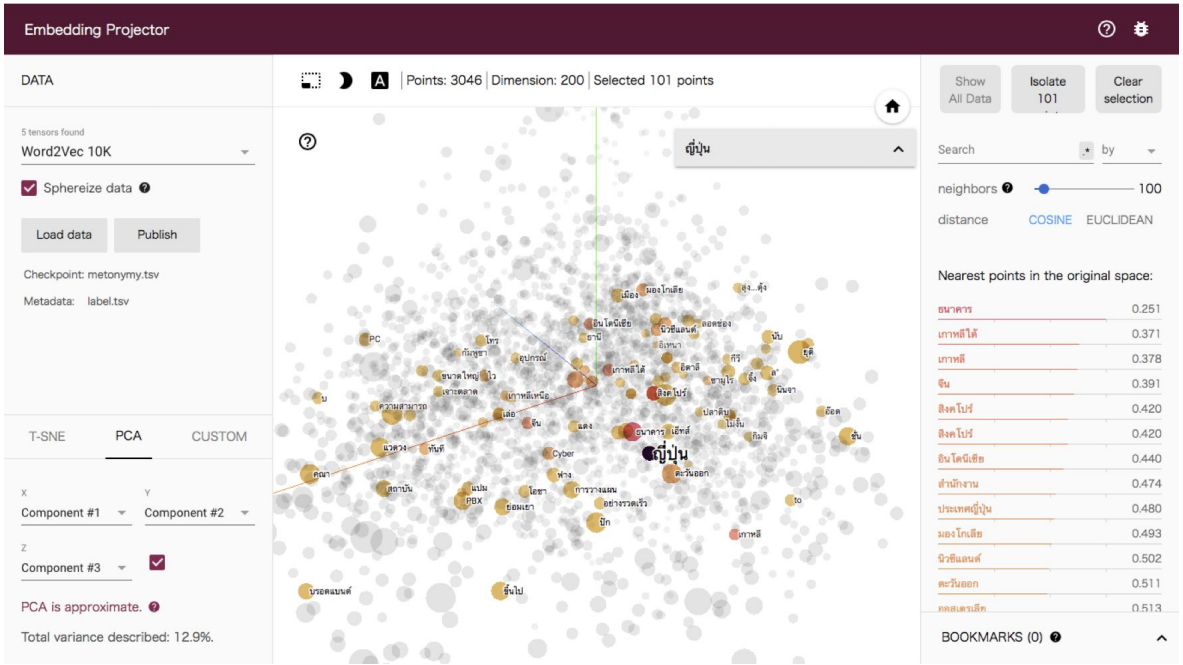
high cosine similarity of “อร่อย”

- รสชาติ
- เมนู
- ชิม
- กลมกล่อม



Vector Semantics

vector projector (PCA 3 dimension)



- ญี่ปุ่น
- เกาหลีใต้
- สิงคโปร์

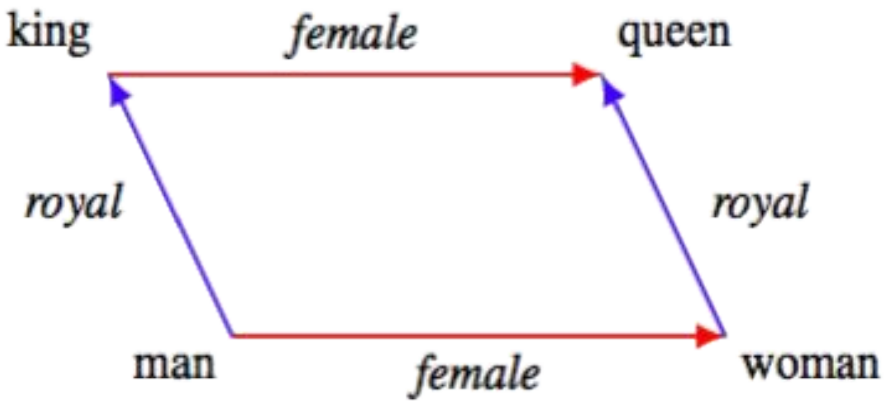
countries are near position

Word Arithmetic

not only find similar word,

but can add/subtract vectors

to calculate meaning

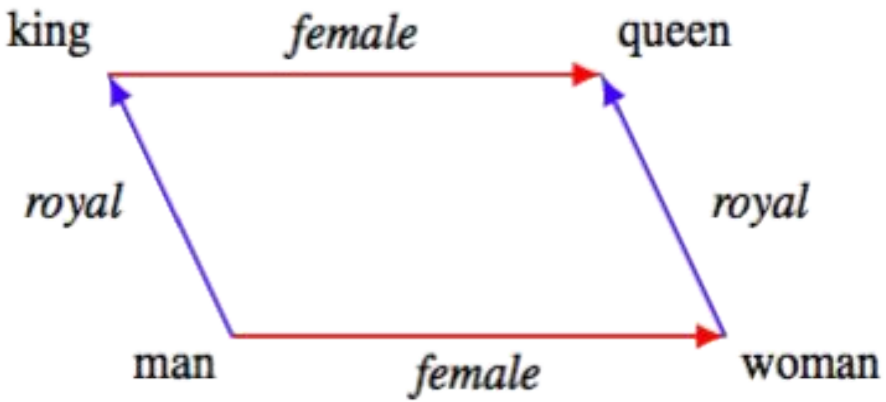


Word Arithmetic

for example,

“king” - “man” + “woman”

= “queen”



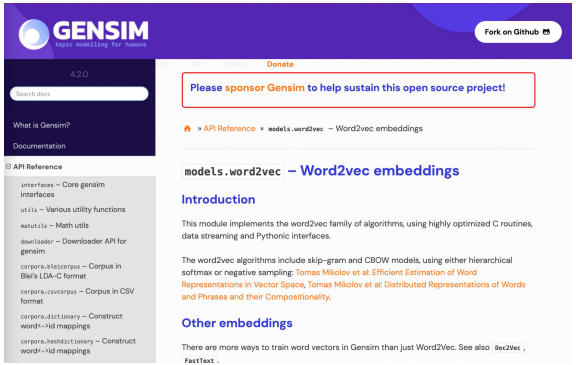
What for ?

- to find similar words in documents
- to find topic of the documents (like Text Classification)
- **to use as pre-trained word vector for Deep Learning**

not only “word embedding”, there are also other embeddings

e.g. document embedding, character embedding

Python Package



`gensim` gives a lot of functions for word embedding. You can train model by using your own data.

Since word embedding is **Unsupervised Learning**, you can easily try without annotating data.

Websites

- https://aiforthai.in.th/service_bn.php
  “Word Similarity”
- https://www.thaicorpus.net/w2v
  (my website)