



ARTIFICIAL
INTELLIGENCE

Emojification for Thai Text



Kobkrit Viriyayudhakorn, Ph.D.

kobkrit@iapp.co.th

iApp Technology Limited

What is Emojification?

- ใส่ Emojicon ให้กับคำภาษาไทยได้อัตโนมัติ
- "รักเธอเหลือเกิน" -> "รักเธอเหลือเกิน ❤️"
- "ดีมากเลย" -> "ดีมากเลย 😄"
- "เศร้าแปป" -> "เศร้าแปป 😞"
- "กำลังหิวเลย" -> "กำลังหิวเลย 🍴"
- "มาเล่นเบสบอลกัน" -> "มาเล่นเบสบอลกัน ⚾"

Why?



- Great chatbot use **Emoji A LOT!!**
- Hand pick the list of emoji is super slow!!



สวัสดีค่ะ Gatedee ให้บริการทางเทคนิคการแพทย์ถึงประตูบ้าน
ท่าน

วันนี้สนใจทำอะไรดีคะ 🤗

บริการต่างๆ 📌

ดูตะกร้าสินค้า 🛒

ไม่มีไรละ ❌

หรือ โอนมาที่
นาย กอบกฤตย์ วัริยะยุทธกร
ธนาคารไทยพาณิชย์
383-253939-6
จำนวนเงิน 500.00 บาท 💰 ก็ได้นะคะ

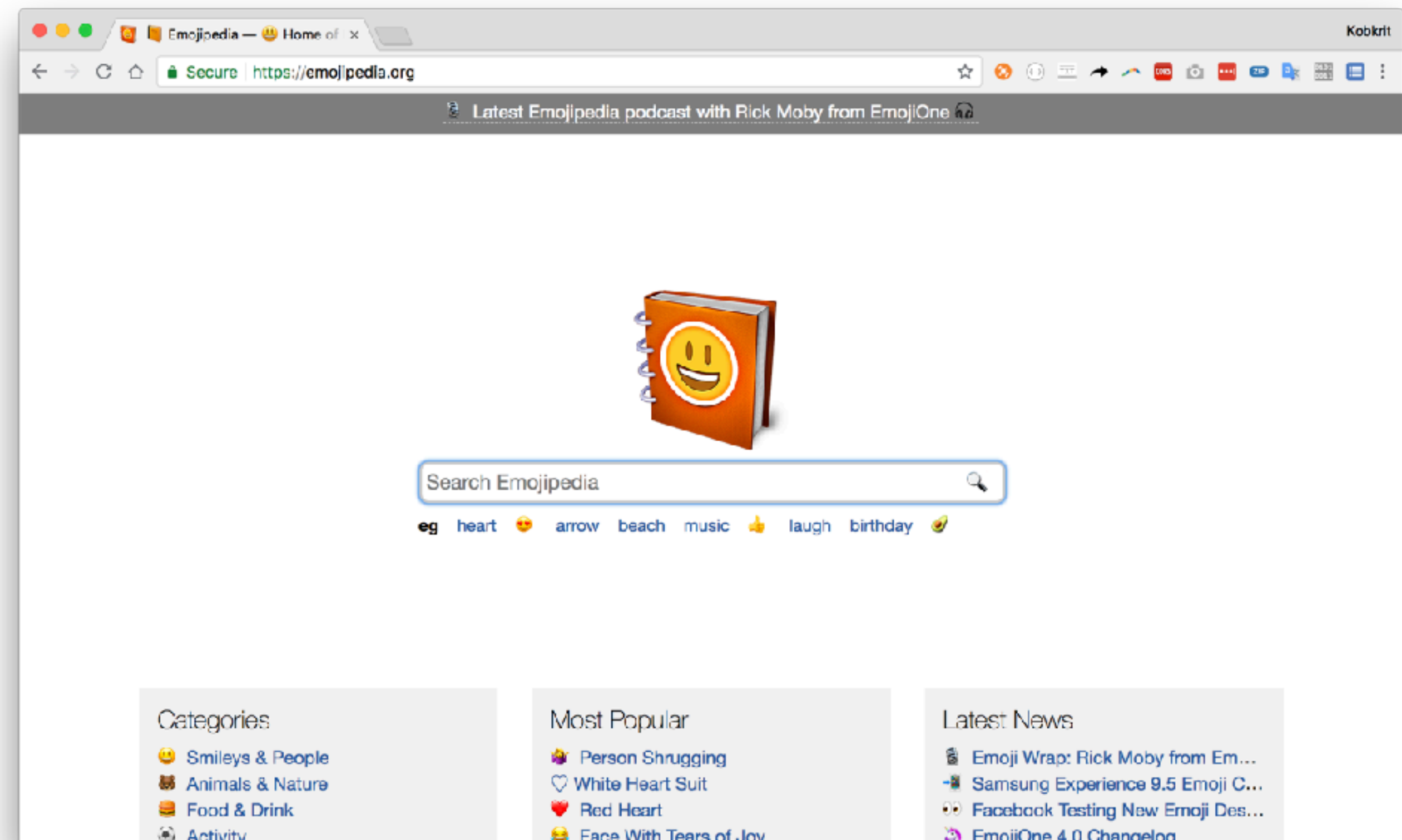
ทาง Call-center ของเราจะติดต่อกลับภายใน 15 นาที (ในเวลา
ทำการ) ที่เบอร์ 0863225858 หรือว่าตอนเช้าของวันทำการถัดไป

หรือหากท่านต้องการติดต่อเราด่วน สามารถติดต่อทางเราได้
โดยตรงเลยที่ 064-872-9266 (08.00น - 21.00น วันจันทร์-เสาร์)
ได้เลยนะคะ

โทรหาเราเลยตอนนี้ 📞

ขอบพระคุณมากๆค่ะ แล้วเจอกันนะคะ บริการ Gatedee บริการ
ดีถึงบ้านคุณ 🤗🤗🤗🤗🤗

เสร็จเรียบร้อยแล้ว



Dataset

You have a tiny dataset (X, Y) where:

- X contains 127 sentences (strings)
- Y contains a integer label between 0 and 4 corresponding to an emoji for each sentence

X (sentences)	Y (labels)
I love you	0
Congrats on the new job	2
I think I will end up alone	3
I want to have sushi for dinner!	4
It was funny lol	2
she did not answer my text	3
Happy new year	2
my algorithm performs poorly	3
he can pitch really well	1
you are failing this exercise	3
you did well on your exam.	2
What you did was awesome	2
I am frustrated	3





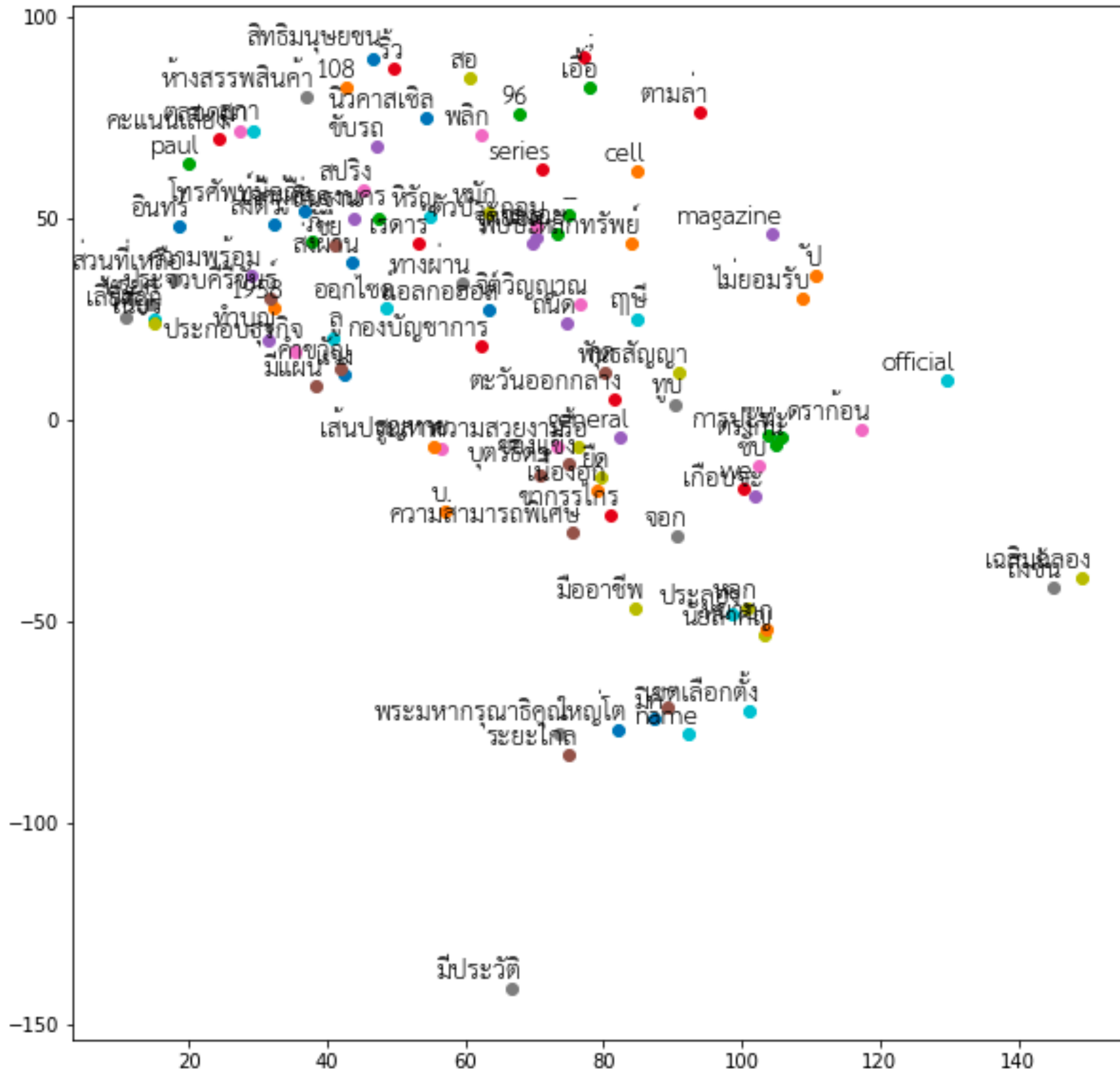
code	emoji	label
:heart:		0
:baseball:		1
:smile:		2
:disappointed:		3
:fork_and_knife:		4

Figure 1: EMOJISSET - a classification problem with 5 classes. A few examples of sentences are given here.



Thai2Vec



• คน =

[1.33543919e-01

1.07829292e+00

-3.72354955e-01

-1.46668282e+00

....{300 position}....

]

Data Preparation

- ฉันทักเธอ

- ตัดคำ: ฉันทักเธอ

- แปลงเป็น Vector:

[1.33543919e-01
1.07829292e+00
-3.72354955e-01
-1.46668282e+00
....{300 position}....
]

ฉันทัก

[2.33543919e-01
1.07829292e+00
-3.72354955e-01
-1.46668282e+00
....{300 position}....
]

เธอ

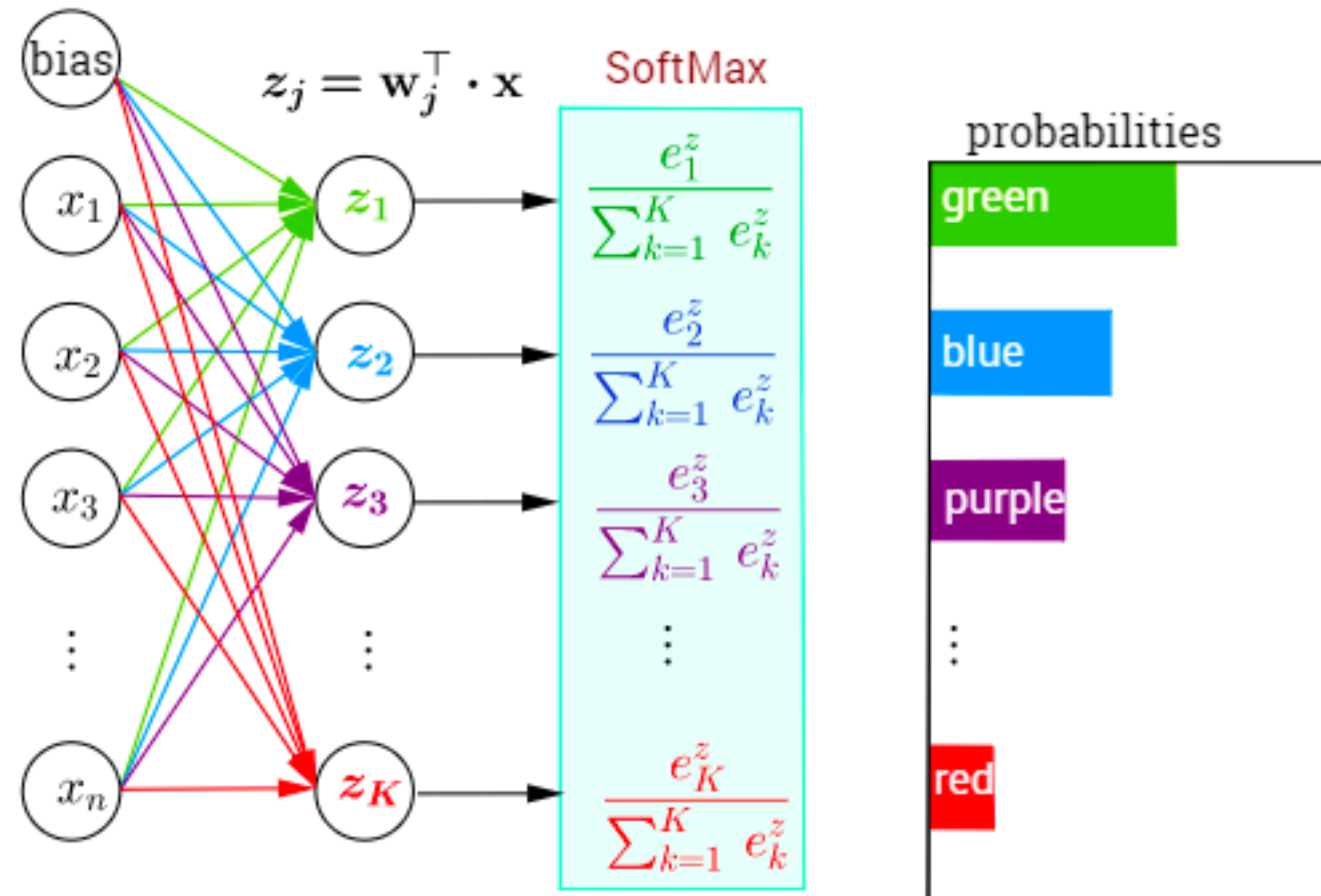
[0.33543919e-01
1.07829292e+00
-3.72354955e-01
-1.46668282e+00
....{300 position}....
]

ISO

Method 1: Compute Vector Average of Whole Sentence, and then, Softmax Classification

Multi-Class Classification with NN and SoftMax Function

$$\mathbf{z} = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ \vdots \\ z_K \end{bmatrix} = \begin{bmatrix} \mathbf{w}_1^\top \\ \mathbf{w}_2^\top \\ \mathbf{w}_3^\top \\ \vdots \\ \mathbf{w}_K^\top \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix}$$

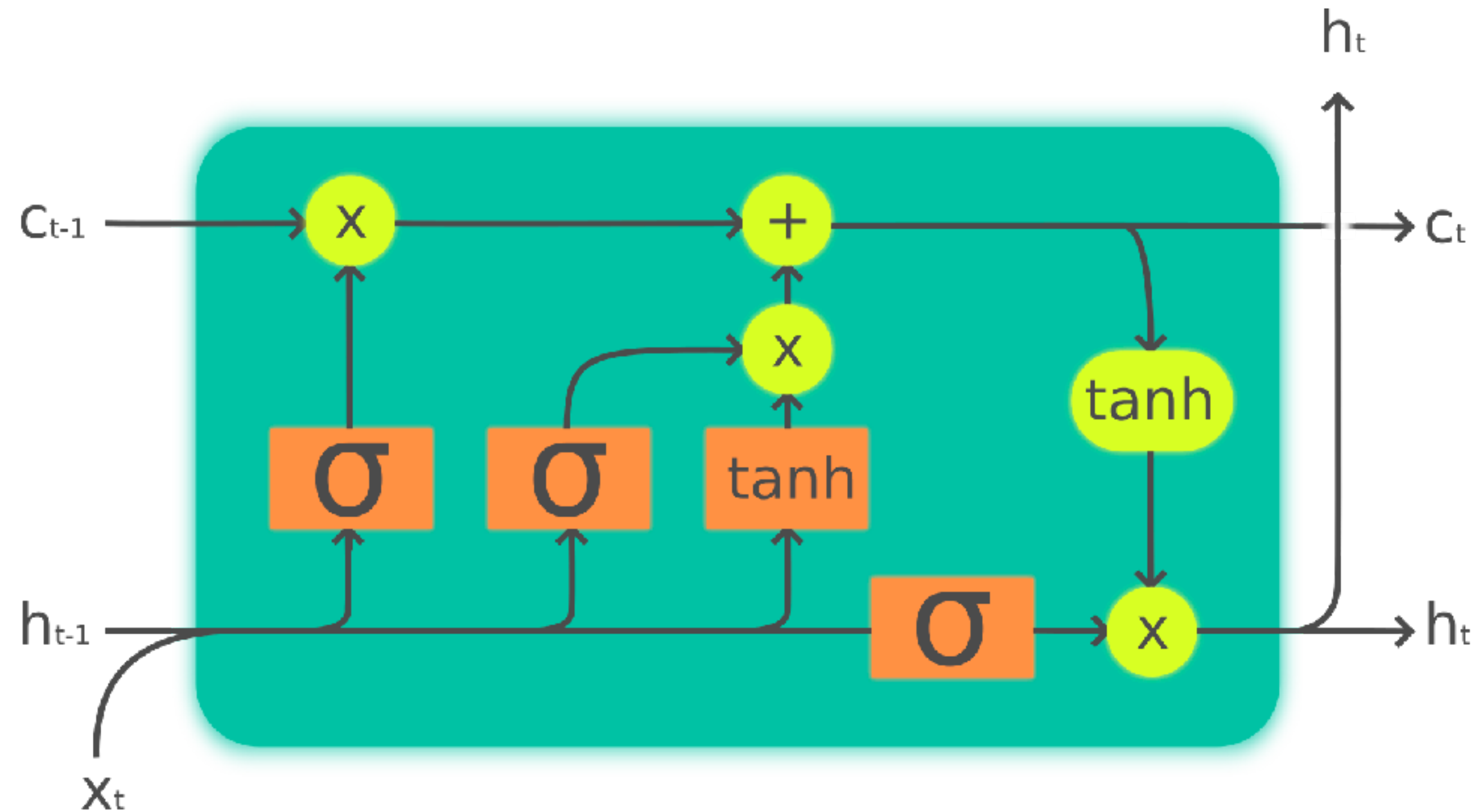


Method 1: Make Average of Whole Sentence, and then, Softmax Classification [Con]



- Does not handle the sequence well.
- **ฉันไม่เศร้าและ ลู! == ฉันไม่ลูและ เศร้า!**
- = Vector ของ ฉัน + Vector ของ ไม่ +
Vector ของ เศร้า + Vector ของ และ +
Vector ของ ลู + Vector ของ !
- มีค่า Vector Average เท่ากับ
[1.33543919e-01
1.07829292e+00
-3.72354955e-01
-1.46668282e+00
....{300 position}....
]

Method 2: LSTM (Long Short Term Memory)



Legend:

Layer

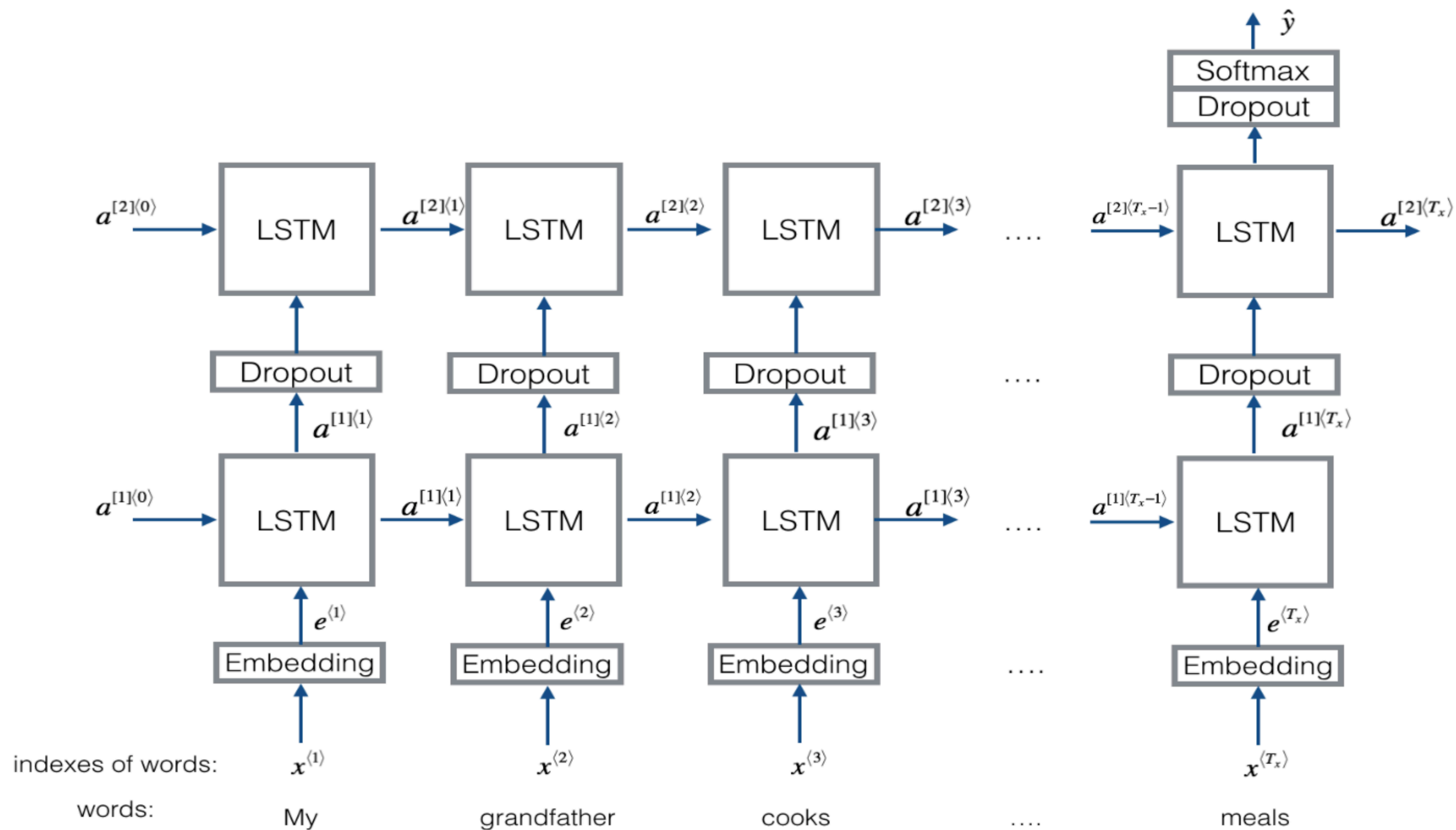


Pointwise op



Copy







```
def Emojify_V2(input_shape, word_to_vec_map, word_to_index):
    """
    Function creating the Emojify model's graph.

    Arguments:
    input_shape -- shape of the input, usually (max_len,)
    word_to_vec_map -- dictionary mapping every word in a vocabulary into its 50-dimensional vector representation
    word_to_index -- dictionary mapping from words to their indices in the vocabulary (400,001 words)

    Returns:
    model -- a model instance in Keras
    """

    # Define sentence_indices as the input of the graph, it should be of shape input_shape and dtype 'int32' (as it contains
    sentence_indices = Input(input_shape, dtype='int32')

    # Create the embedding layer pretrained with GloVe Vectors (~1 line)
    embedding_layer = pretrained_embedding_layer(word_to_vec_map, word_to_index)

    # Propagate sentence_indices through your embedding layer, you get back the embeddings
    embeddings = embedding_layer(sentence_indices)

    # Propagate the embeddings through an LSTM layer with 128-dimensional hidden state
    # Be careful, the returned output should be a batch of sequences.
    X = LSTM(128, return_sequences=True)(embeddings)
    # Add dropout with a probability of 0.5
    X = Dropout(0.5)(X)
    # Propagate X through another LSTM layer with 128-dimensional hidden state
    # Be careful, the returned output should be a single hidden state, not a batch of sequences.
    X = LSTM(128, return_sequences=False)(X)
    # Add dropout with a probability of 0.5
    X = Dropout(0.5)(X)
    # Propagate X through a Dense layer with softmax activation to get back a batch of 5-dimensional vectors.
    X = Dense(5)(X)
    # Add a softmax activation
    X = Activation('softmax')(X)

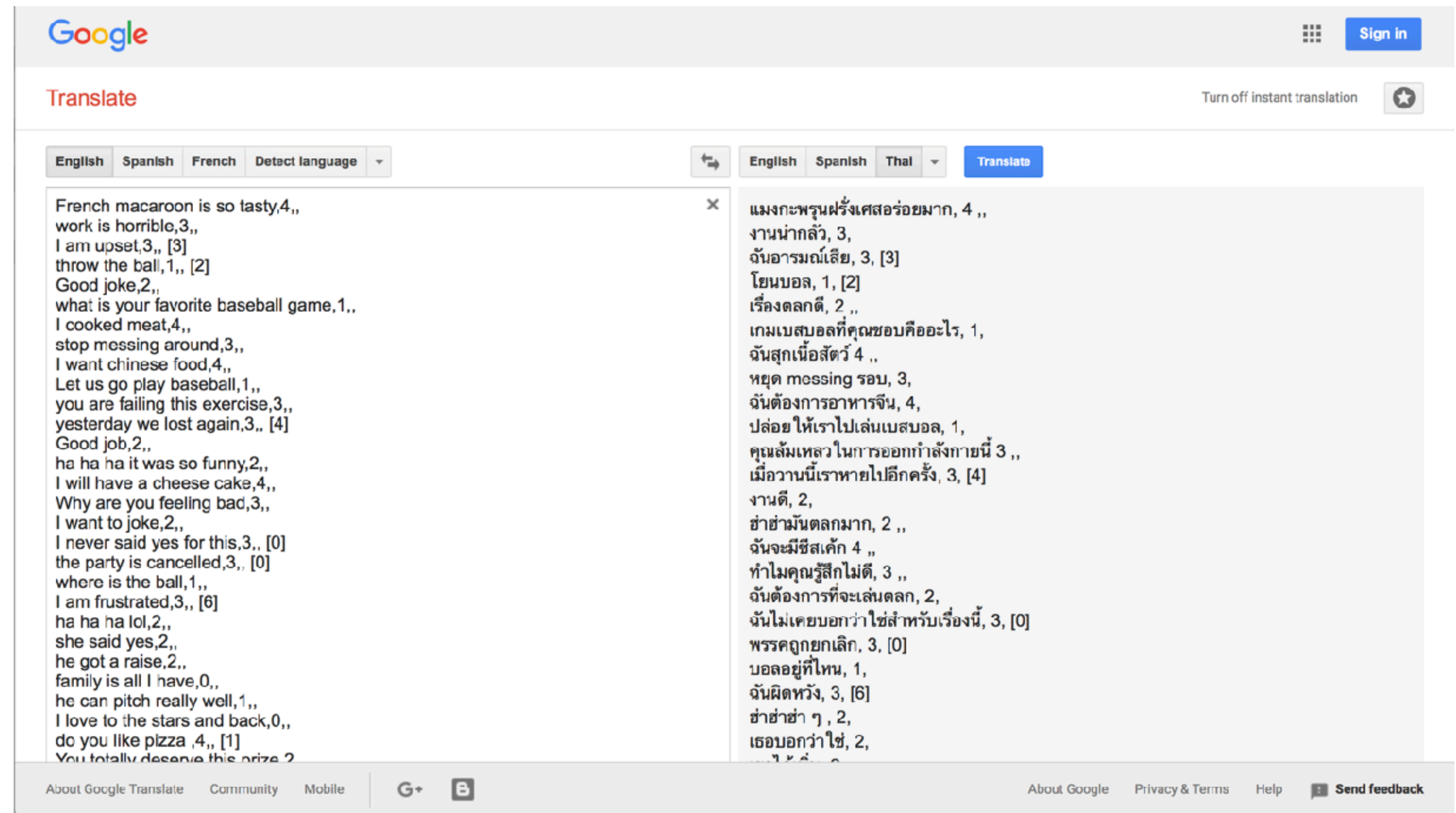
    # Create Model instance which converts sentence_indices into X.
    model = Model(inputs=sentence_indices, outputs=X)

    return model
```


How can I make Thai Dataset?



- Google Translate
- Training Data = 128 records
- Testing Data = 55 records



Google Translate vs GT+Manually Cleaned



jupyter training_th.csv ✓ a minute ago

File Edit View Language

```
1 แมงกะพรุนฝรั่งเศสอร่อยมาก,4
2 งานน่ากลัว,3
3 ฉันอารมณ์เสีย,3
4 โยนลูกบอล,1
5 เรื่องตลกดี,2
6 เกมเบสบอลที่คุณชอบคืออะไร,1
7 ฉันสนุก,4
8 หยุดยุ่งอยู่รอบ ๆ,3
9 ฉันต้องการอาหารจีน,4
10 ปลอຍให้เราไปเล่นเบสบอล,1
11 คุณล้มเหลวในการออกกำลังกายนี้,3
12 เมื่อวานนี้เราหายไปอีกครั้ง,3
13 งานที่ดี,2
14 ซ้ำซ้ำมันตลกมาก,2
15 ฉันจะมีชีวิตรัก,4
16 ทำไมคุณรู้สึกไม่ดี,3
17 ฉันต้องการที่จะเล่นตลก,2
18 ฉันไม่เคยพูดว่าใช่สำหรับเรื่องนี้,3
19 ฝ่ายถูกยกเลิก,3
20 ลูกบอลอยู่ที่ไหน,1
21 ฉันผิดหวัง,3
22 ซ้ำซ้ำซ้ำ ๆ,2
23 เธอตอบตกลง,2
24 เขาได้รับเงินเพิ่ม,2
25 ครอบครัวคือสิ่งที่ฉันมี,0
26 เขาสามารถเล่นได้จริงๆ,1
27 ฉันรักดวงดาวและด้านหลัง,0
28 คุณชอบพืชชาไหม,4
29 คุณสมควรได้รับรางวัลนี้ทั้งหมด,2
30 ฉันคิดถึงคุณมากเหลือเกิน,0
31 ฉันชอบเสื้อของคุณ,2
32 เธอทำให้ฉันเป็นของขวัญ,0
33 คุณจะ เป็นวาเลนไทน์ของฉันไหม,0
34 คุณล้มเหลวในระยะกลาง,3
35 ใครลงไปร้านอาหาร,4
36 วันวาเลนไทน์อยู่ใกล้,0
```

jupyter training_th_cleaned.csv ✓ 7 minutes ago

File Edit View Language

```
1 มาการูนฝรั่งเศสอร่อยมาก,4
2 งานแย่มาก,3
3 ฉันอารมณ์เสีย,3
4 โยนบอล,1
5 เรื่องตลกดี,2
6 เกมเบสบอลที่คุณชอบคืออะไร,1
7 ฉันกำลังปรุงเนื้อสัตว์,4
8 หยุดกวนแฉนี้ได้แล้ว,3
9 ฉันต้องการอาหารจีน,4
10 ปลอຍให้เราไปเล่นเบสบอล,1
11 คุณล้มเหลวในการออกกำลังกายนี้,3
12 เมื่อวานนี้เราแพ้อีกครั้ง,3
13 งานดี,2
14 ซ้ำซ้ำมันตลกมาก,2
15 ฉันจะมีชีวิตรัก,4
16 ทำไมคุณรู้สึกไม่ดี,3
17 ฉันต้องการที่จะเล่นตลก,2
18 ฉันไม่เคยบอกว่าใช่สำหรับเรื่องนี้,3
19 งานเลี้ยงถูกยกเลิก,3
20 บอลอยู่ที่ไหน,1
21 ฉันผิดหวัง,3
22 ซ้ำซ้ำซ้ำ ,2
23 เธอบอกว่าใช่,2
24 เขาได้เงินเดือนเพิ่ม,2
25 ครอบครัวเป็นสิ่งสำคัญที่สุด,0
26 เขาสามารถโยนลูกได้ดีจริงๆ,1
27 ฉันรักดวงดาวและกลับ,0
28 คุณชอบพืชชาไหม,4
29 คุณสมควรได้รับรางวัลนี้ทั้งหมด,2
30 ฉันคิดถึงคุณมาก,0
31 ฉันชอบเสื้อของคุณ,2
32 เธอมอบของขวัญให้ฉัน,0
33 คุณจะ เป็นคนรักของฉัน,0
34 ทำข้อสอบไม่ได้,3
35 ใครอยากไปร้านอาหาร,4
36 วันวาเลนไทน์อยู่ใกล้แล้ว,0
```

Test accuracy =
0.6545454523780129

Test accuracy =
0.6363636461171237





Misprediction from Google Translate

emoji: ❤️	ฉันกำลังมองหาวันที่😞
emoji: 🍴	คำแนะนำสำหรับอาหารค่ำ😄
emoji: 😊	เธอมีความสุข😞
emoji: 😊	เธอยิ้มให้มาก❤️
emoji: 🍴	คุณก็เป็นสิ่งที่ดี😄
emoji: 😞	ฉันเกลียดเขา❤️
emoji: 😞	ฉันผิดหวังมาก 🍴
emoji: 😞	ฉันทำงานในวันเกิดของฉัน😄
emoji: 😞	ฉันไม่ได้ทานอาหารเช้า 🍴
emoji: ❤️	สุนัขของฉันมีลูกสุนัขเพียงไม่กี่คน😞
emoji: ⚾	เขาต้องวิ่งกลับบ้าน😞
emoji: 😊	คุณร้ายแรงไหม 🍴
emoji: ❤️	คุณทั้งสองน่ารัก😞
emoji: 😞	เครื่องคิดเลข ใงนี้ไม่ทำงาน😄
emoji: 😊	สิ่งที่คุณทำก็น่ากลัว😞
emoji: 😞	ชีวิตฉันน่าเบื่อ❤️
emoji: ⚾	ช่วยให้ออกกำลังกาย😄
emoji: 😊	คุณสดใสในวันของฉัน❤️
emoji: 🍴	ช่วยให้รันชวัน😄

Misprediction from Google Translate + Human

emoji: ❤️	อยากจะเดทกับคุณ😄
emoji: 🍴	คำแนะนำสำหรับอาหารเย็น😄
emoji: 😊	เธอมีความสุข⚾
emoji: 😊	เธอยิ้มให้มาก❤️
emoji: 😊	ฉันได้รับการอนุมัติ 🍴
emoji: 🍴	คุณก็เป็นสิ่งที่ดี😄
emoji: 😞	ฉันเกลียดเขา❤️
emoji: ⚾	ฉันจะไปที่สนามกีฬา 🍴
emoji: 😞	ฉันผิดหวังมาก😄
emoji: 😞	เธอคนนี้กำลังยุ่งอยู่กับชีวิตฉัน❤️
emoji: 😞	ไปไกลๆเลย😄
emoji: 🍴	ฉันหิว😄
emoji: ❤️	ฉันขอบคุณมาก😄
emoji: 😊	หยุดทำเรื่องตลกนี้ซ้ำๆ😞
emoji: 😞	ชีวิตฉันน่าเบื่อ❤️
emoji: ⚾	ช่วยออกกำลังกาย😄
emoji: 😊	คุณสดใสในวันของฉัน❤️
emoji: 😊	ฉันจะไปเต้น 🍴
emoji: 😊	เต้นรำกับฉัน❤️
emoji: 😞	เธอเป็นคนพาล❤️



Misprediction from Google Translate

emoji: ❤️	ฉันกำลังมองหาวันที่😞	
emoji: 🍴	คำแนะนำสำหรับอาหารค่ำ😄	
emoji: 😊	เธอมีความสุข😞	
emoji: 😊	เธอยิ้มให้มาก❤️	OK
emoji: 🍴	คุณก็เป็นสิ่งที่ดี😄	
emoji: 😞	ฉันเกลียดเขา❤️	
emoji: 😞	ฉันผิดหวังมาก 🍴	
emoji: 😞	ฉันทำงานในวันเกิดของฉัน😄	
emoji: 😞	ฉันไม่ได้ทานอาหารเช้า 🍴	
emoji: ❤️	สุนัขของฉันมีลูกสุนัขเพียงไม่กี่คน😞	
emoji: ⚾	เขาต้องวิ่งกลับบ้าน😞	
emoji: 😊	คุณร้ายแรงไหม 🍴	
emoji: ❤️	คุณทั้งสองน่ารัก😞	
emoji: 😞	เครื่องคิดเลข ใงนี้ไม่ทำงาน😄	
emoji: 😊	สิ่งที่คุณทำก็น่ากลัว😞	
emoji: 😞	ชีวิตฉันน่าเบื่อ❤️	
emoji: ⚾	ช่วยให้ออกกำลังกาย😄	
emoji: 😊	คุณสดใสในวันของฉัน❤️	OK
emoji: 🍴	ช่วยให้บรันชวัน😄	

Misprediction from Google Translate + Human

emoji: ❤️	อยากจะเดทกับคุณ😄	
emoji: 🍴	คำแนะนำสำหรับอาหารเย็น😄	
emoji: 😊	เธอมีความสุข⚾	
emoji: 😊	เธอยิ้มให้มาก❤️	OK
emoji: 😊	ฉันได้รับการอนุมัติ 🍴	
emoji: 🍴	คุณก็เป็นสิ่งที่ดี😄	
emoji: 😞	ฉันเกลียดเขา❤️	
emoji: ⚾	ฉันจะไปที่สนามกีฬา 🍴	
emoji: 😞	ฉันผิดหวังมาก😄	
emoji: 😞	เธอคนนี้กำลังยุ่งอยู่กับชีวิตฉัน❤️	
emoji: 😞	ไปไกลๆเลย😄	
emoji: 🍴	ฉันหิว😄	
emoji: ❤️	ฉันขอบคุณมาก😄	
emoji: 😊	หยุดทำเรื่องตลกนี้ซ้ำซ้ำ😞	
emoji: 😞	ชีวิตฉันน่าเบื่อ❤️	
emoji: ⚾	ช่วยออกกำลังกาย😄	
emoji: 😊	คุณสดใสในวันของฉัน❤️	OK
emoji: 😊	ฉันจะไปเต้น 🍴	
emoji: 😊	เต้นรำกับฉัน❤️	OK
emoji: 😞	เธอเป็นคนพาล❤️	

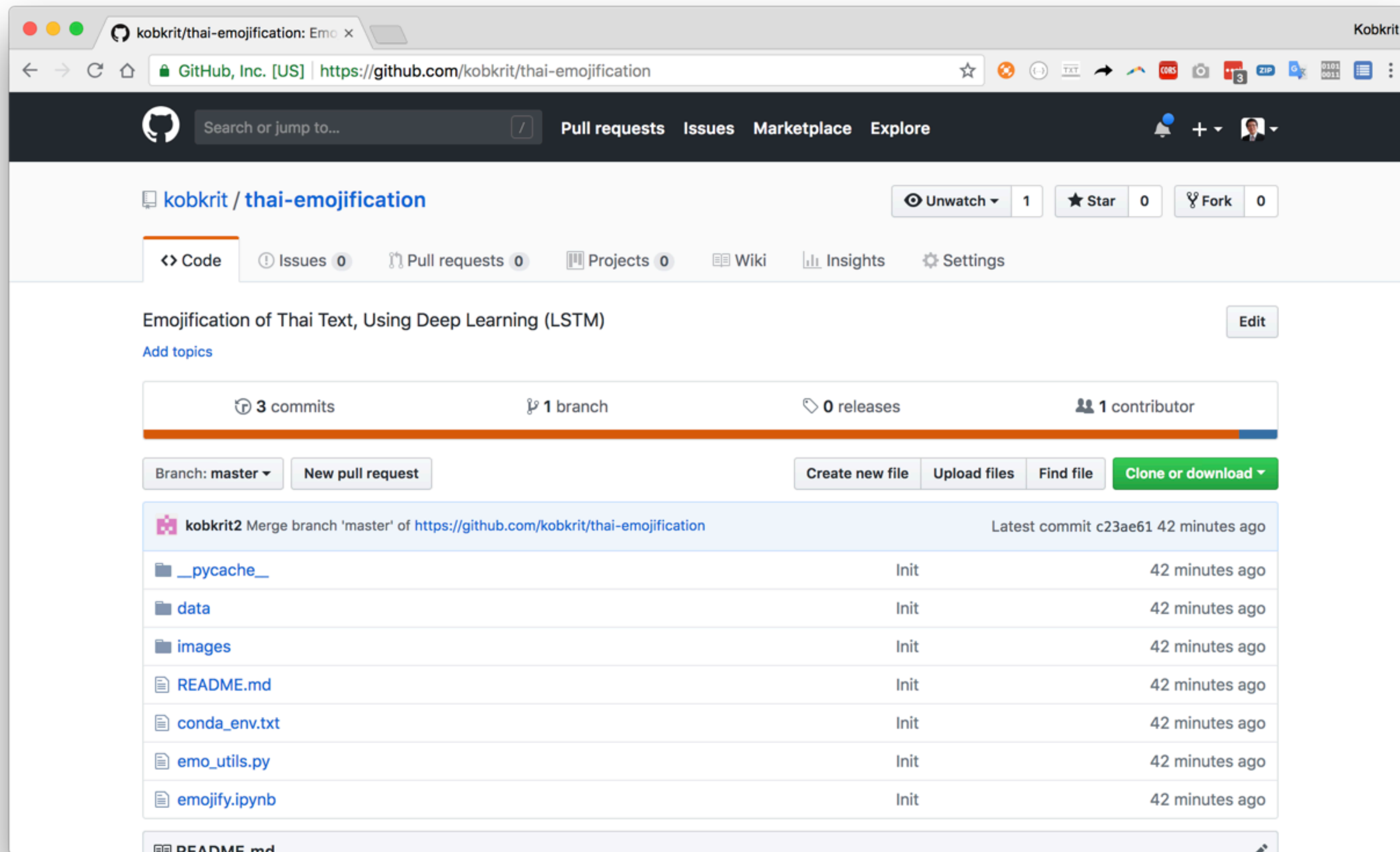
Configuration

- Thai2Vec 0.2 (300d)
- Word Segmentation: PyThaiNLP NewMM (Followed Thai2Vec)
- Tensorflow + Keras
- Pretrained embedding layer using Thai2Vec
- LSTM = 128 hidden layers
- Dropout = 0.5
- maxLen = 12 words



PyThaiNLP


Open Source



<https://github.com/kobkrit/thai-emojification>



Demo

ARTIFICIAL
INTELLIGENCE

HOME

Thai Text Emojification


Write down some Thai text, e.g., "รักเธอเหลือเกิน" -> "รักเธอเหลือเกิน ❤️", "ดีมากเลย" -> "ดีมากเลย 😊", "เศร้าแปป" -> "เศร้าแปป 😞", "อาหารอร่อยมาก" -> "อาหารอร่อยมาก 🍴" เป็นต้น

Now support only 5 Emoji ❤️ 😊 😞 🍴 🏠

Type a Thai Text:

Result

It's a มาเล่นเบสบอลกัน 🏈 with condifence score 0.9994849.

ARTIFICIAL
INTELLIGENCE

HOME

Analysed successfully. It's a หายเศร้าและ 😊 with confidence score 0.99947184

Thai Text Emojification

Write down some Thai text, e.g., "รักเธอเหลือเกิน" -> "รักเธอเหลือเกิน ❤️", "ดีมากเลย" -> "ดีมากเลย 😊", "เศร้าแปป" -> "เศร้าแปป 😞", "อาหารอร่อยมาก" -> "อาหารอร่อยมาก 🍴" เป็นต้น

Now support only 5 Emoji ❤️ 😊 😞 🍴 🏠

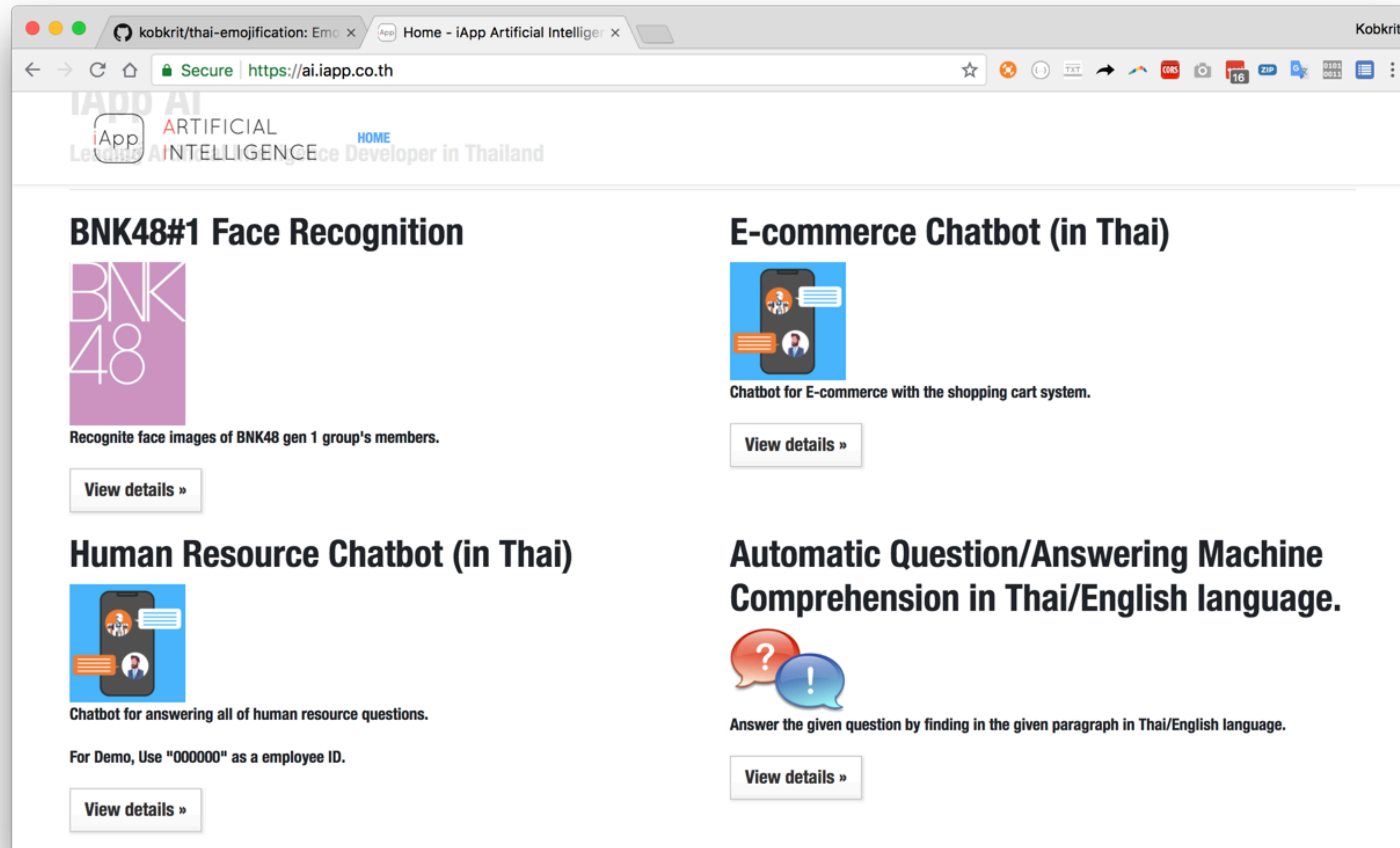
Type a Thai Text:

Result

It's a หายเศร้าและ 😊 with condifence score 0.99947184.

<http://ai.iapp.co.th>

You might want to try other AI as well.



<http://ai.iapp.co.th>



ARTIFICIAL
INTELLIGENCE

Q/A

(or kobkrit@iapp.co.th)

All Slides and Source code is at
<https://github.com/kobkrit/thai-emojification>