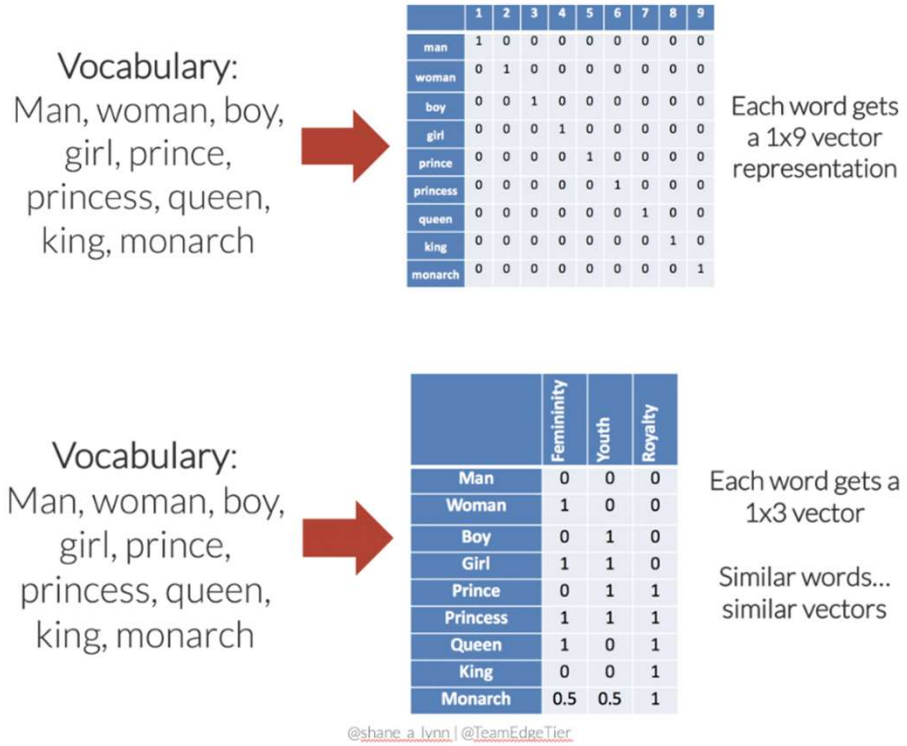
Week 3 – Word Embedding

EGCO467 Natural Language and Speech Processing

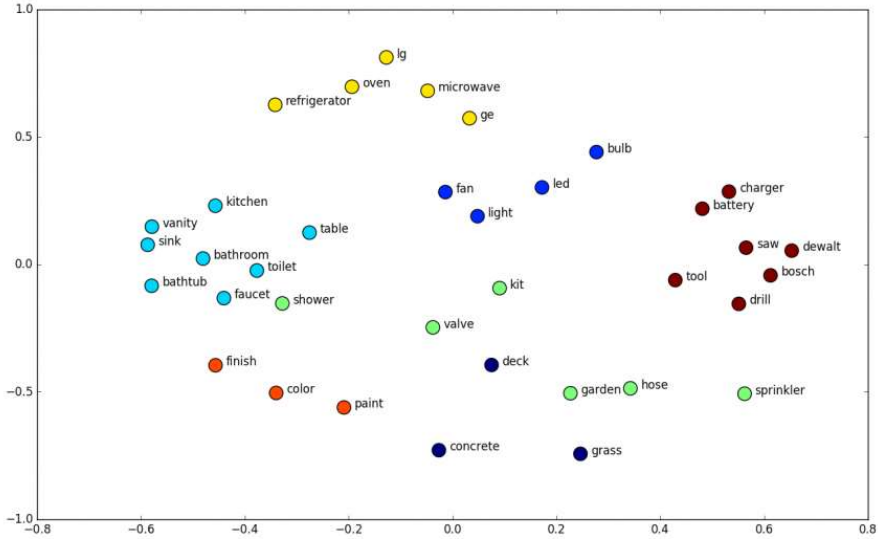
1/2564

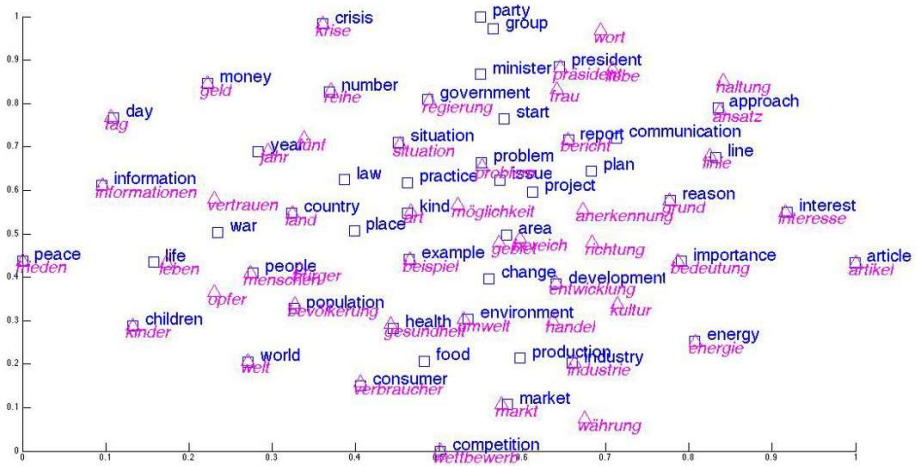
Word embedding



https://www.shanelynn.ie/get-busy-with-word-embeddings-introduction/

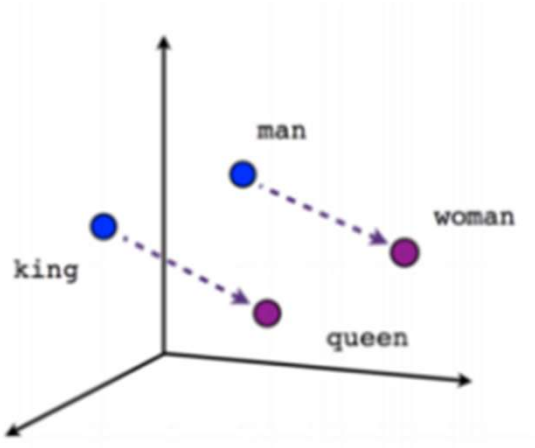
Word embedding





https://medium.com/deeper-learning/glossary-of-deep-learning-word-embedding-f90c3cec34ca

Word embedding



man is to woman
as king is to queen

walking is to walked
as swimming is to swam

cosine similarity

- used to the measure the "similarity" of embedded words

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

- value near 1 if similar
- near 0 if not related
- near -1 if opposite

Example

- find the cosine similarity
- $v_1 = [1.2, 3.2, -3.5]$
- $v_2 = [-1.5, 4.2, 3.3]$

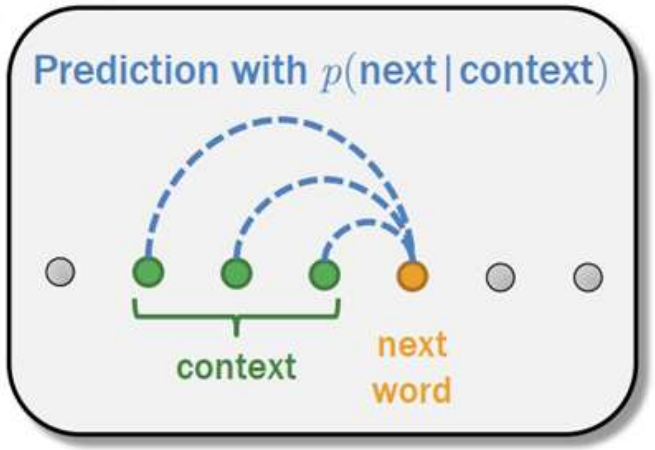
$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Word embeddings

- **word2vec**: Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781* (2013). (Google AI)
- **Glove**: Pennington, Jeffrey, Richard Socher, and Christopher Manning. "Glove: Global vectors for word representation." *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014. (Stanford)
- **Fasttext**: Bojanowski, Piotr, et al. "Enriching word vectors with subword information." Transactions of the Association for Computational Linguistics 5 (2017): 135-146. (Facebook AI)

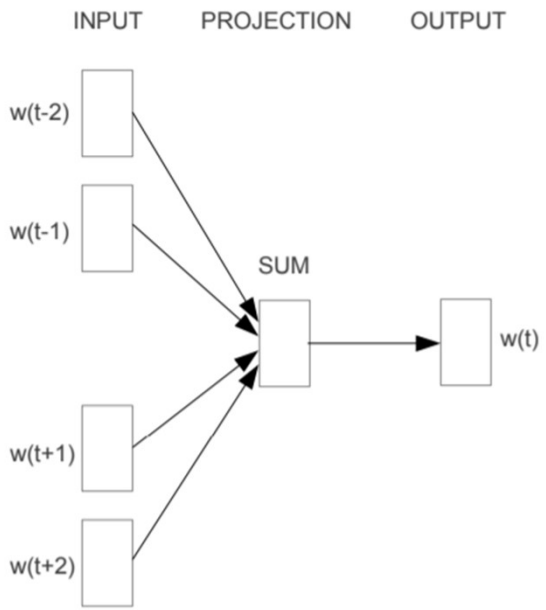
n-gram

- sequence of n consecutive n token
- E.g. "I love dogs they are so cute"
- trigrams: (I, love, dogs), (love, dog, they), (dogs, they, are), (they, are, so), (are, so, cute)



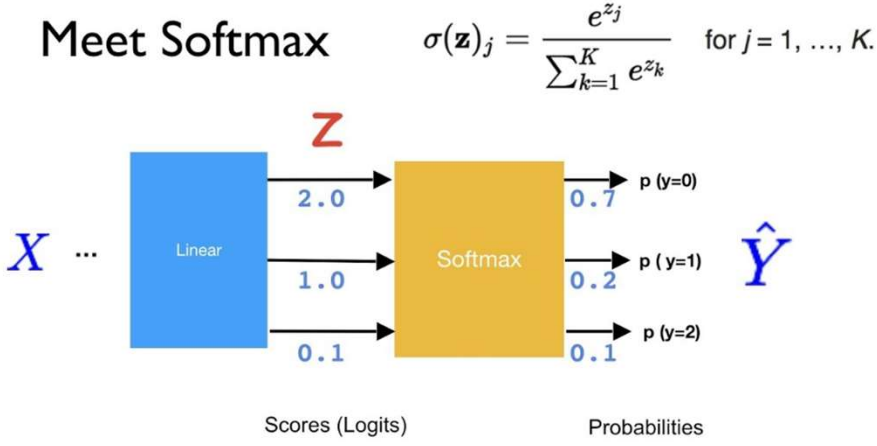
Continuous bag of words (cbow)

- Context C = N words before and N words after current word $w(t)$
- Input = one-hot embedding of each word



Softmax

- Map N real numbers (-inf, inf) -> (0,1)
- all the outputs sum to 1
- for classification



credit: https://www.youtube.com/watch?v=lvNdl7yg4Pg

word2vec – single word C

given word k, predict what is word j

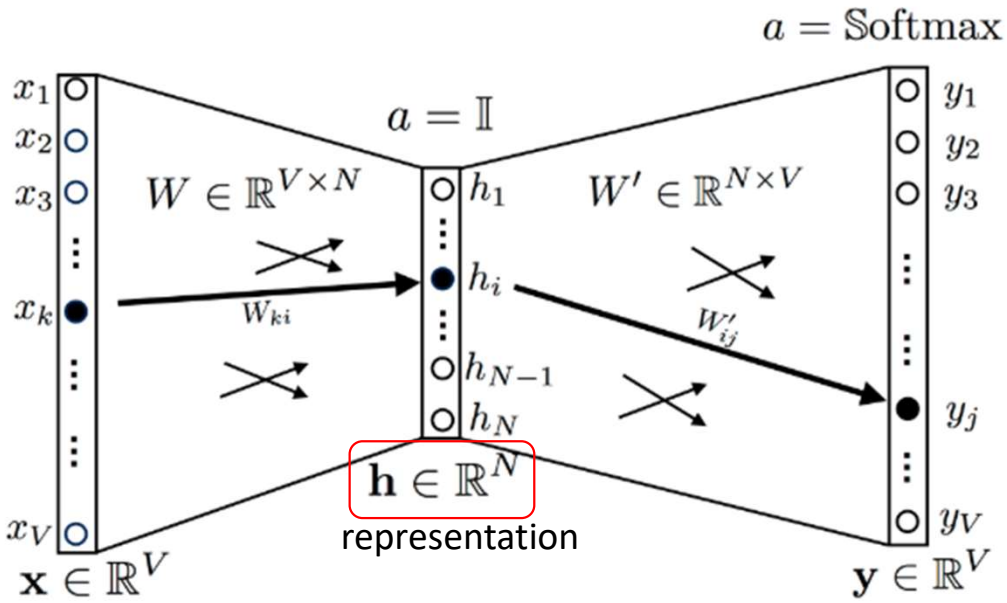


Figure 2. Topology of the one-word Continuous Bag-of-Words model.

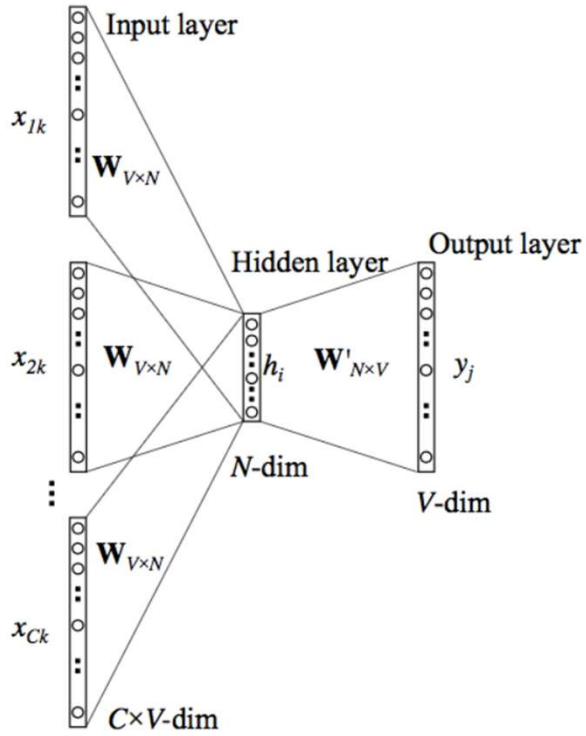
Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781* (2013).

Embedding of k^{th} word

Decoding an embedding

word2vec – multiple words C

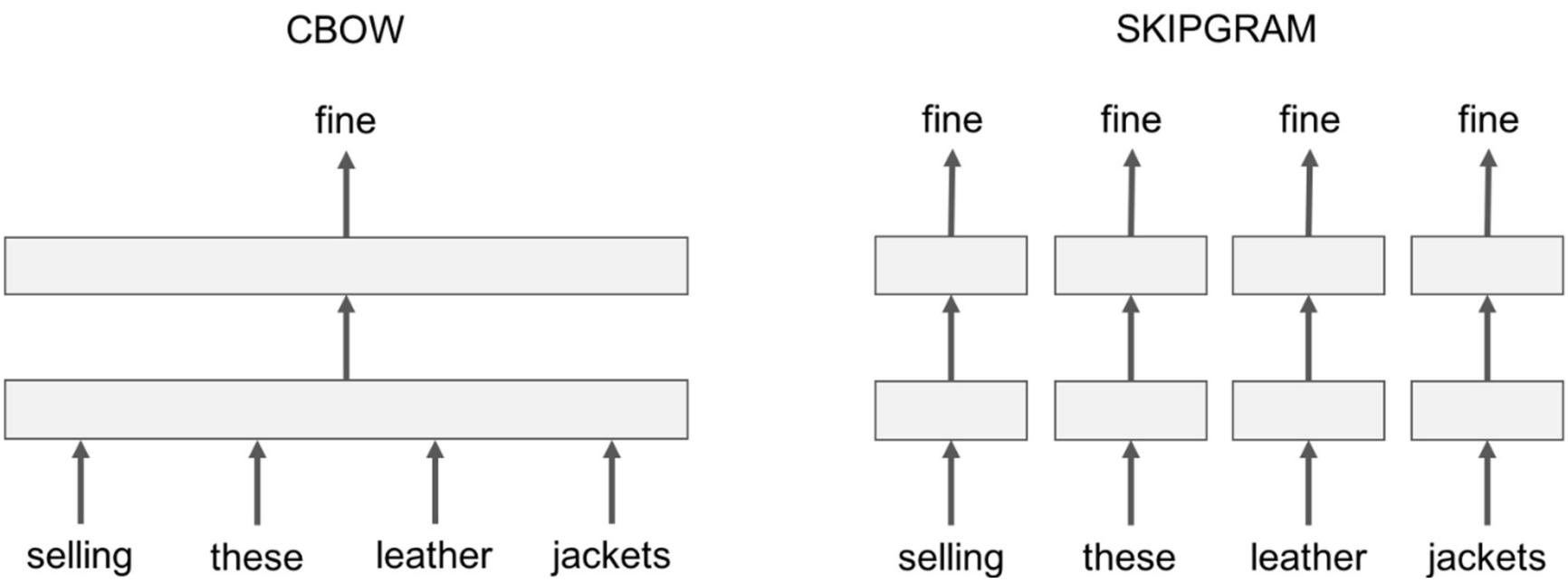
- Given words j-1, j-2, j+1, j+2 (context C)
- Predict word j



Skip-gram (Fasttext's version)

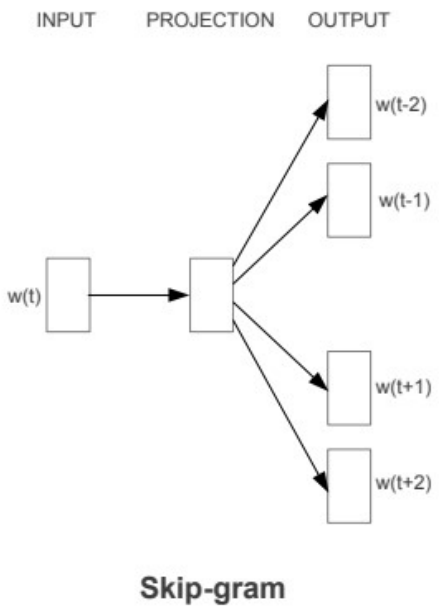
use sum of all nearby words as context

choose a random nearby word to use as context



*selling these **fine** leather jackets*

Skip-gram (in general)



Skip-gram

Vector representation for sentence

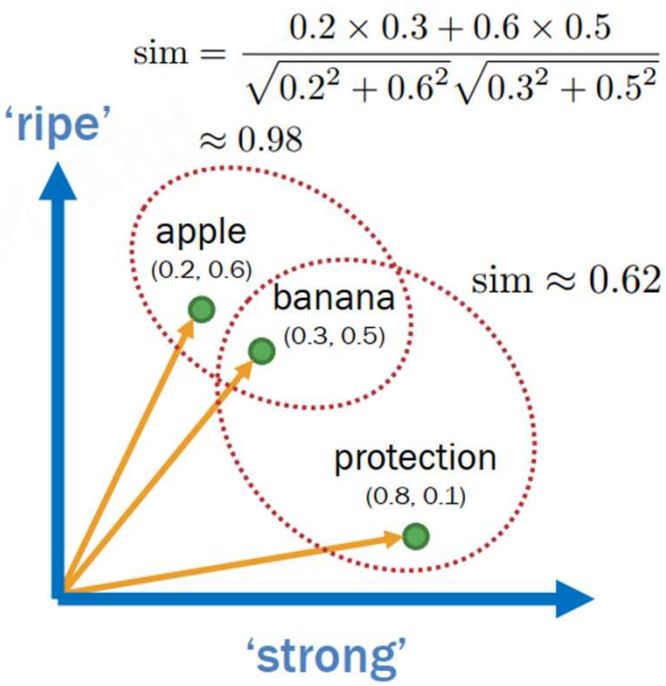
1. find the vector representation of each word
2. add all the vectors together

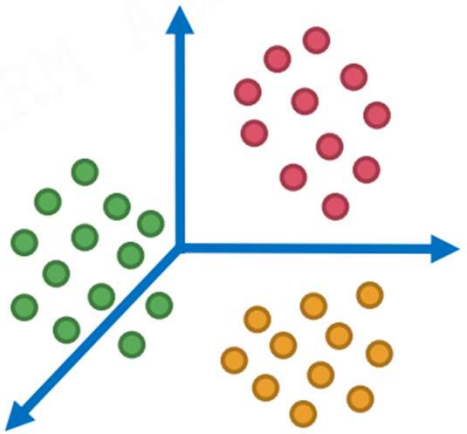
- E.g. I like fried chicken
  - I -> v1
  - like -> v2
  - fried -> v3
  - chicken -> v4
  - I like fried chicken -> v1 + v2 + v3 + v4

Example: simple chatbot

- prepare a (long) list of query-reply examples. E.g.
  - hello : hello
  - how are you today : I'm fine
  - I want more information about product X : Here is the information you requested <link>
- Sentence embed all the queries
- When get a new query:
  - sentence embed the new query
  - find the most similar query using cosine similarity
  - return the reply of that matched query
  - return "I don't understand" if similarity is below a certain threshold

Representing Words





Vector Space Model

Example of query-reply

B	C
Message	**Answer**
วันนี้ปวดขา	ลองนวดขาเบา ๆ ดูค่ะ
น่าเบื่อจังเลย	มีอะไรระบายกับไออุ่นได้นะคะ
ช่วงนี้เจ็บคออยู่บ่อยๆ	พยายามลดการดื่มน้ำเย็นนะคะ// เจ็บคอแบบนี้ควรดื่มน้ำอุ่นๆค่ะ
ปวดหัวจัง	พักสักครู่ ถ้าไม่หาย //ลองปรึกษาหมอไหมคะ
วันนี้รู้สึกมึนๆหัว	พักสักครู่ ถ้าไม่หาย ลองปรึกษาหมอไหมคะ
วันนี้เหมือนฝนจะตก	ออกไปข้างนอกอย่าลืมพกร่มด้วยนะคะ
ร้อนมากๆเลยวันนี้	ใส่เสื้อผ้าที่ปลอดโปร่ง สบายๆ จิบน้ำเปล่าบ่อยๆนะคะ
เป็นตะคิวทำไงดี	ลองเหยียดเข่าและกระดกปลายเท้าขึ้นดูค่ะ
เมื่อไหร่โควิดจะหมด	ไออุ่นก็ยังให้คำตอบไม่ได้ แต่ก็ยังคงต้องดูแลตัวเองให้ปลอดภัยอยู่เสมอนะคะ
ดูแลตัวเองยังไงในช่วงโควิด	พยายามล้างมือบ่อยๆ สวมหน้ากากอนามัย //และเว้นระยะห่างนะคะ
ไม่อยากไปหาหมอเลย	ให้ไออุ่นไปเป็นเพื่อนไหมคะ
พรุ่งนี้ต้องไปหาหมออีกแล้ว	พาไออุ่นไปด้วยก็ได้นะคะ
ง่วงนอน	ง่วงก็นอนพักนะคะ //ไออุ่นจะอยู่เป็นเพื่อนค่ะ
ง่วง	ง่วงก็นอนพักนะคะ// ไออุ่นจะอยู่เป็นเพื่อนค่ะ
รู้สึกเหมือนจะไม่สบาย	พักผ่อนมากๆ //ถ้าไม่ดีขึ้นไปหาหมอนะคะ
เป็นไข้	ลองวัดไข้ดีไหมคะ
ช่วงนี้หงุดหงิดบ่อยจัง	หงุดหงิดเรื่องอะไร //ลองระบายให้ไออุ่นฟังได้นะคะ

Assignment

- word2vec can also be used for document searching
  - embed each document as vector
  - embed query
  - find the most similar document

- get 10 news headlines from any news website(s)
- input query
- find the most similar headline
- print the headline text