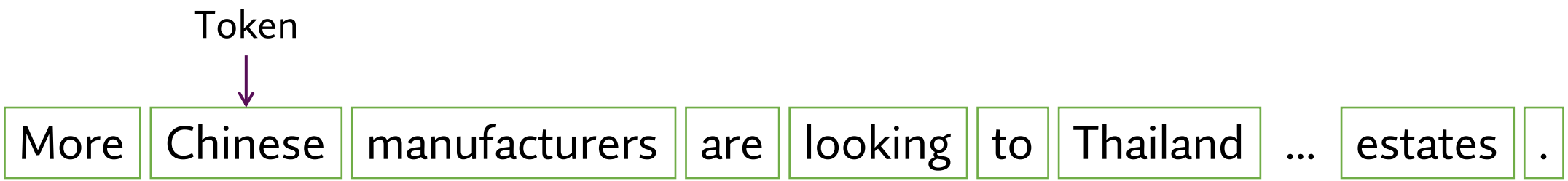
Feature Engineering

Papangkorn Inkeaw, Ph.D.

Feature Extraction

Chapter 5 (Part II) - Feature Extraction for Text Data

Text – Basic Knowledge



Text is a human-readable sequence of word(s).

Text – Basic Knowledge

n-gram A contiguous sequence of *n* tokens from a given sample of text.

More | Chinese | manufacturers | are | looking | to | Thailand | ... | estates | .

Unigram

More | Chinese | manufacturers | are | looking | to | Thailand | ...

Bigram

More Chinese | Chinese manufacturers | manufacturers are

are looking | looking to | to Thailand | ...

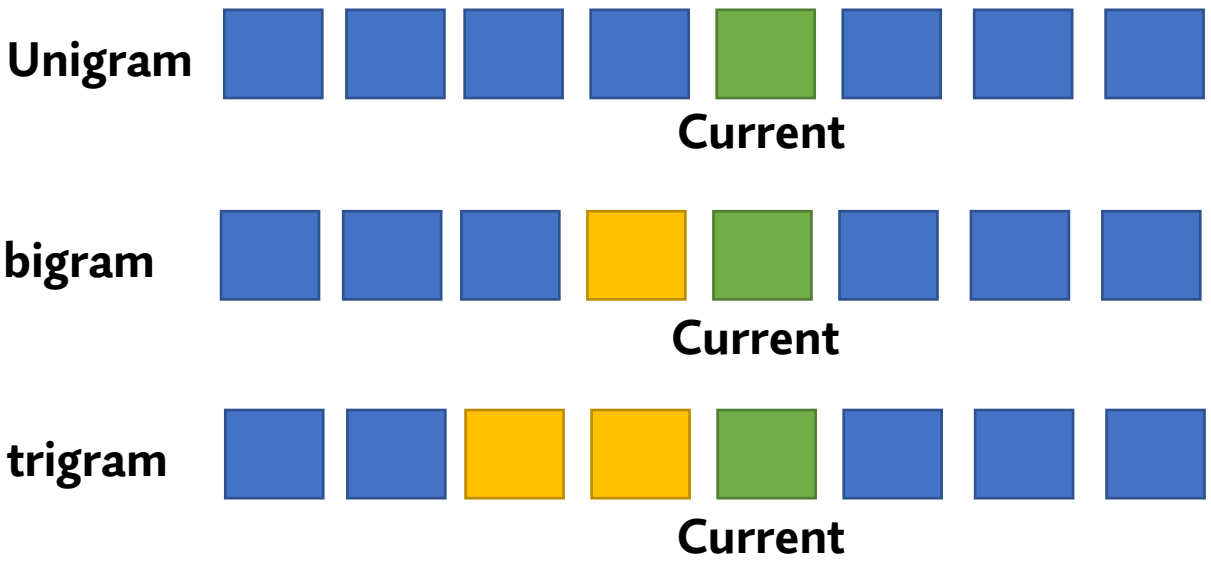
Trigram

More Chinese manufacturers | Chinese manufacturers are

manufacturers are looking | are looking to | looking to Thailand | ...

Text – Basic Knowledge

n-gram model predicts the occurrence of a word based on the occurrence of its N – 1 previous words.



A word is independent from its previous words.

The occurrence of a word depends on the occurrence of it's a previous words.

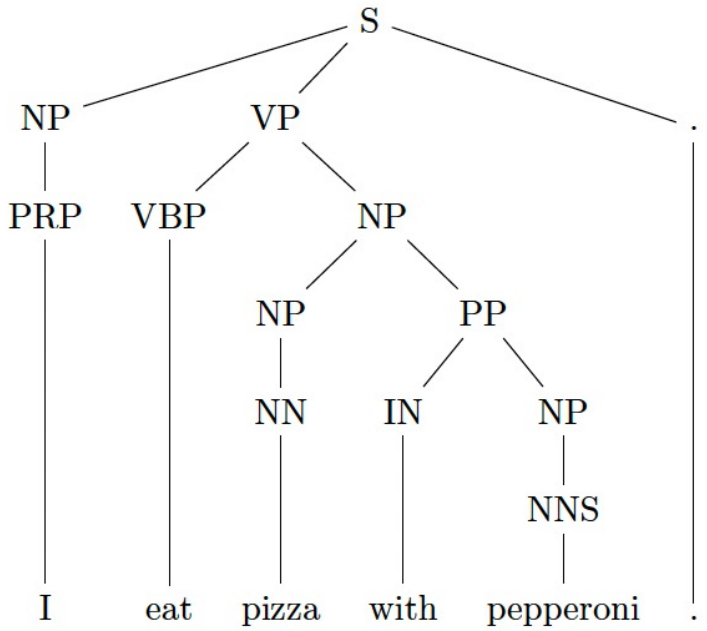
The occurrence of a word depends on the occurrence of it's two previous words.

Text – Basic Knowledge

Part-of-speech (POS) The grammatical function of each word in a sentence

We can use NLP tool to to categorizing words in a text in correspondence with a particular part of speech, depending on the definition of the word and its context.

I	eat	pizza	with	pepperoni	.
PRP	VBP	NNS	IN	NNS	.



From Text to Feature Vector

General Feature Extraction Process

- POS tagging
- Stemming
- Lemmatization
- Name-entity recognition

- Remove stop words
- Remove irrelevant words
- Create dictionary

One-hot Encoding

- A representation of categorical variables as binary vectors.
- Each word is represented as a binary vector that is:
  - All zero values
  - Except the index of the word, which is marked with a 1.

All possible words

Word	.	It	a	cat	is
It =	0.	1.	0.	0.	0.
is =	0.	0.	0.	0.	1.
a =	0.	0.	1.	0.	0.
cat =	0.	0.	0.	1.	0.
. =	1.	0.	0.	0.	0.

Bag of Words

- What about full texts instead of single words?
- The vector representation of a text is simply the vector sum of all the words it contains:

All possible words

. It a cat is

Word

It = [0., 1., 0., 0., 0.],
is = [0., 0., 0., 0., 1.],
a = [0., 0., 1., 0., 0.],
cat = [0., 0., 0., 1., 0.],
. = [1., 0., 0., 0., 0.]

[1., 1., 1., 1., 1.] ← Vector sum of all words represents the text "It is a cat."

Bag of Words

- In practice, it's much more convenient to use a dictionary instead of an actual vector.
- This is known as a **bag-of-words**, and word order is discarded.

Dictionary

Word	cat	dog	bird	panda
They =	[0.,	0.,	0.,	0.],
are =	[0.,	0.,	0.,	0.],
cat =	[1.,	0.,	0.,	0.],
and =	[0.,	0.,	0.,	0.],
dog =	[0.,	1.,	0.,	0.]
They are cat and dog =	[1.,	1.,	0.,	0.]

Bag of words represents the text "They are cat and dog"

TF-IDF

Students present this topic in the classroom.

Word Embeddings

“ Words with similar meanings should occur in similar contexts. ”

- The distributional hypothesis in linguistics

- From a word we can get some idea about the context where it might appear.

__ bird __ __

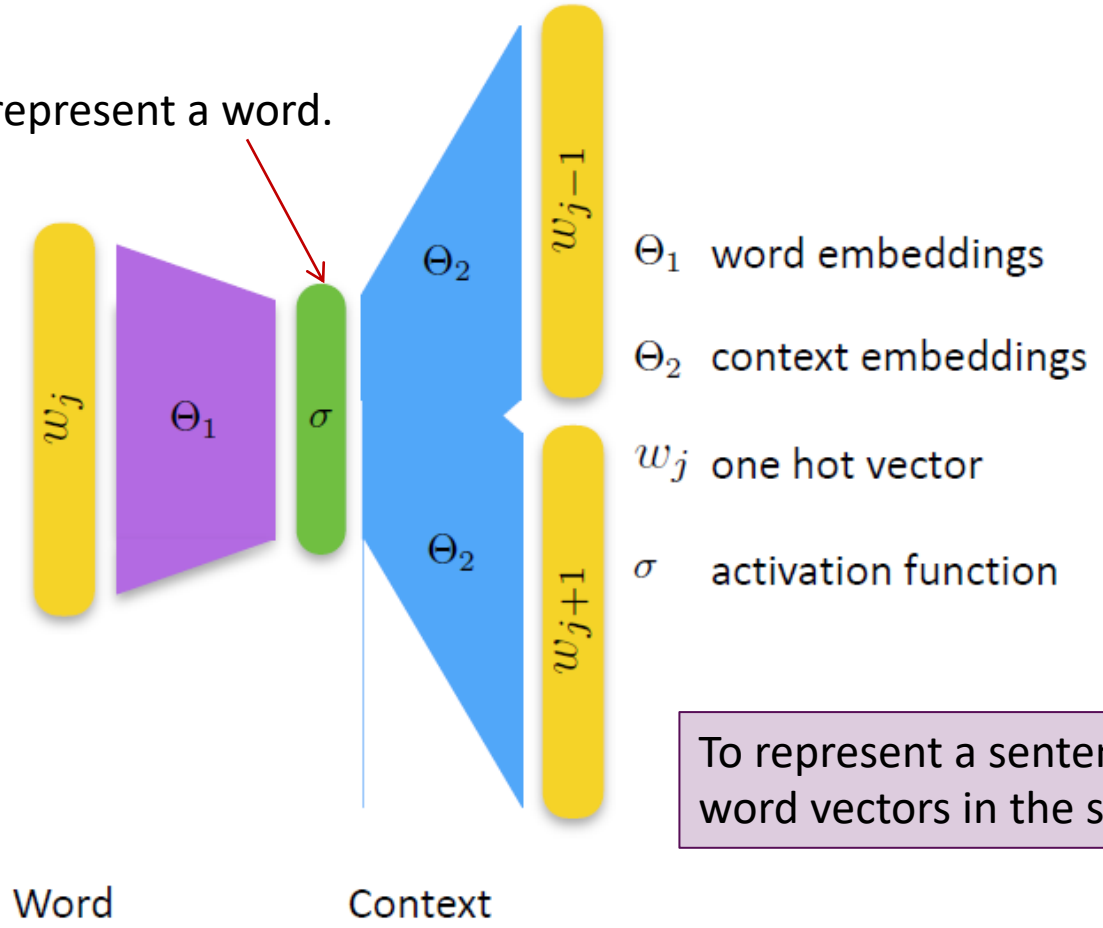
- From the context, we have some idea about possible words.

The red __ is nice

Word Embeddings

Skipgram

Use the output of this layer to represent a word.



Θ_1 word embeddings

Θ_2 context embeddings

w_j one hot vector

σ activation function

To represent a sentence, we can average all word vectors in the sentence.

Source: https://data4sci.com/nlp

Word Embeddings

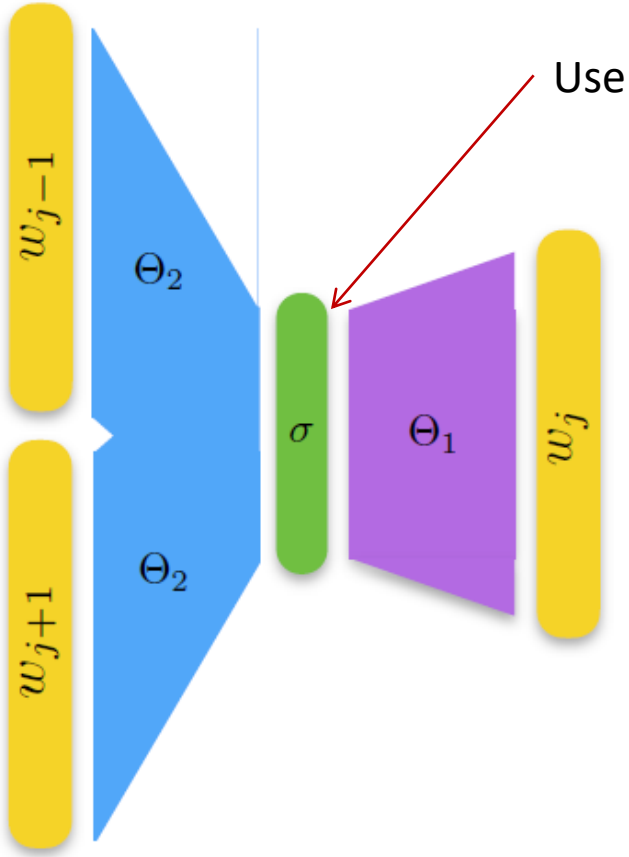
Continuous Bag of Words

Θ_1 word embeddings

Θ_2 context embeddings

w_j one hot vector

σ activation function



Use the output of this layer to represent a word.

To represent a sentence, we can average all word vectors in the sentence.

Source: https://data4sci.com/nlp

References & Study Resources

- Guozhu Dong and Huan Liu. (2020). *Feature Engineering for Machine Learning and Data Analytics*. CRC Press.
- https://data4sci.com/nlp