KI-Campus
Die Lernplattform
für Künstliche Intelligenz

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Word Embedding

Salar Mohtaj | DFKI

Word embedding

- What is word embedding?
- One-hot word representation
- Distributional word vectors
  - Frequency based
  - Prediction based
- Word embedding evaluation
- Word embedding in Python

Word embedding

- What is word embedding?
- One-hot word representation
- Distributional word vectors
  - Frequency based
  - Prediction based
- Word embedding evaluation
- Word embedding in Python

What is word embedding?

- ***Word vectors*** are simply vectors of numbers that represent the ***meaning*** of a word
- Vector models are also called **embeddings** (i.e., word embedding)
- The objective is to represent words in vectors in a way that those with similar meaning have similar representation

What is word embedding?

Text vectors

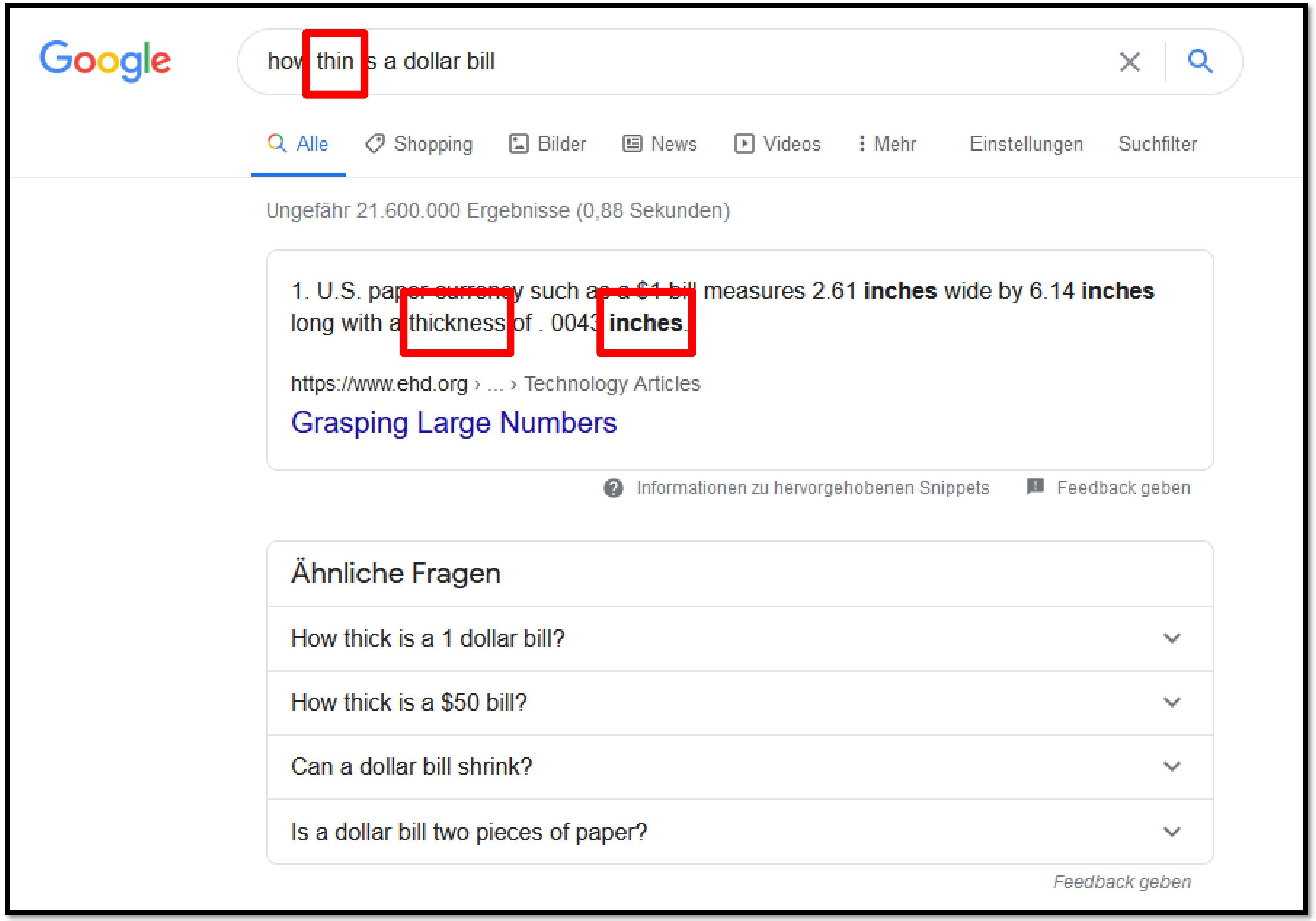
1	-1	0	1	0	0	…	-1

What is word embedding?

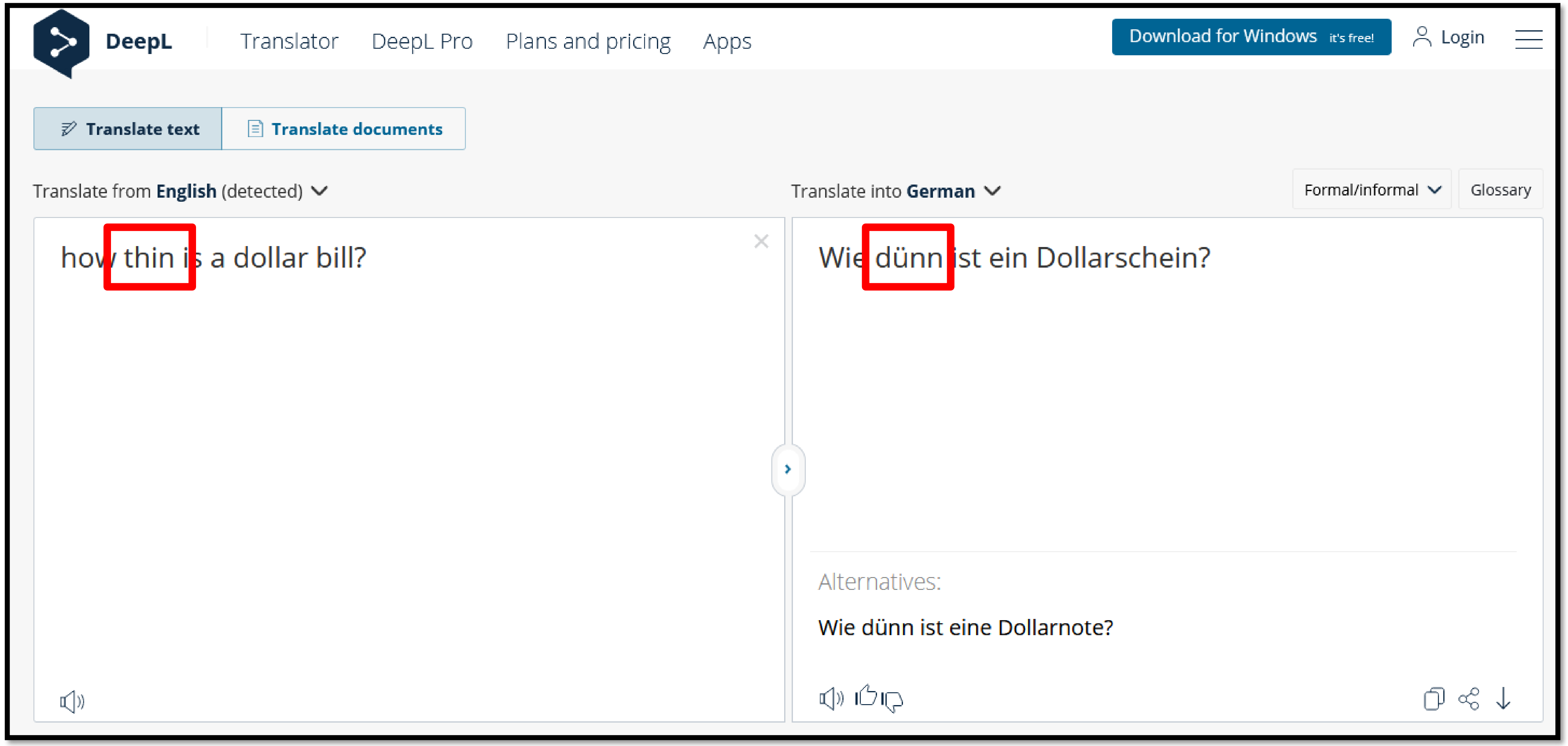
Word vectors

women	1	0	0	…	-1
history	-1	1	0	…	1
program	1	1	1	…	0
…	0	0	0	…	0

Word similarity, why does it matter?



Word similarity, why does it matter?



Word embedding

- What is word embedding?
- One-hot word representation
- Distributional word vectors
  - Frequency based
  - Prediction based
- Word embedding evaluation
- Word embedding in Python

One-hot word representation

- In **one-hot** representation each word is represented with a large vector of size |V| (v is vocabulary's size for the given corpus)
- There is just one element of ***1*** for each word in the corpus

v = [book, machine, artificial, NLP, code]

machine	**0**	**1**	**0**	**0**	**0**
artificial	**0**	**0**	**1**	**0**	**0**
code	**0**	**0**	**0**	**0**	**1**

One-hot word representation

- Pros
  - Simple and easy to understand
- Cons
  - The resulting vectors are long (|V|) and sparse
  - We represent each word as a completely independent entity
  - The vector representation is in binary form, therefore no frequency information is taken into account
  - This word representation does not give us directly any notion of similarity

One-hot word representation

$$\cos(\theta) = \frac{A.B}{\|A\|\|B\|} = \frac{\sum_1^n A_i B_i}{\sqrt{\sum_{1=1}^n A_i^2}\sqrt{\sum_{1=1}^n B_i^2}}$$

v = [book, machine, artificial, NLP, code]

machine	**0**	**1**	**0**	**0**	**0**
artificial	**0**	**0**	**1**	**0**	**0**

Word embedding

- What is word embedding?
- One-hot word representation
- Distributional word vectors
  - Frequency based
  - Prediction based
- Word embedding evaluation
- Word embedding in Python

Distributional word vectors

- It aims to quantify and categorize semantic ***similarities*** between words based on their ***distributional properties*** in large data
- Two words are similar if they have similar ***word contexts***
  - Football and basketball have similar context words (run, ball, referee, …)
- Humans also can guess the meaning of an unknown word from context words

Memes generally replicate through exposure to humans,
who have evolved as efficient copiers of information and behavior.

Distributional word vectors

- Frequency based
  - Document-term matrix
  - Term-term matrix
  - Pointwise mutual information (PMI)
- Prediction based
  - Word2Vec

Document-term matrix

- **Similar words** tend to occur together in the **same documents**
- It describes the frequency of terms that occur in a collection of documents
- In a **document-term** matrix, rows correspond to documents in the collection and columns correspond to terms

Document-term matrix

- D_1 = "Text is a complex human language representation."
- D_2 = "Natural human language is complex and also is diverse."
- D_3 = "Natural human body clock is complex."
- D_4 = "Text representation differs from human to human."

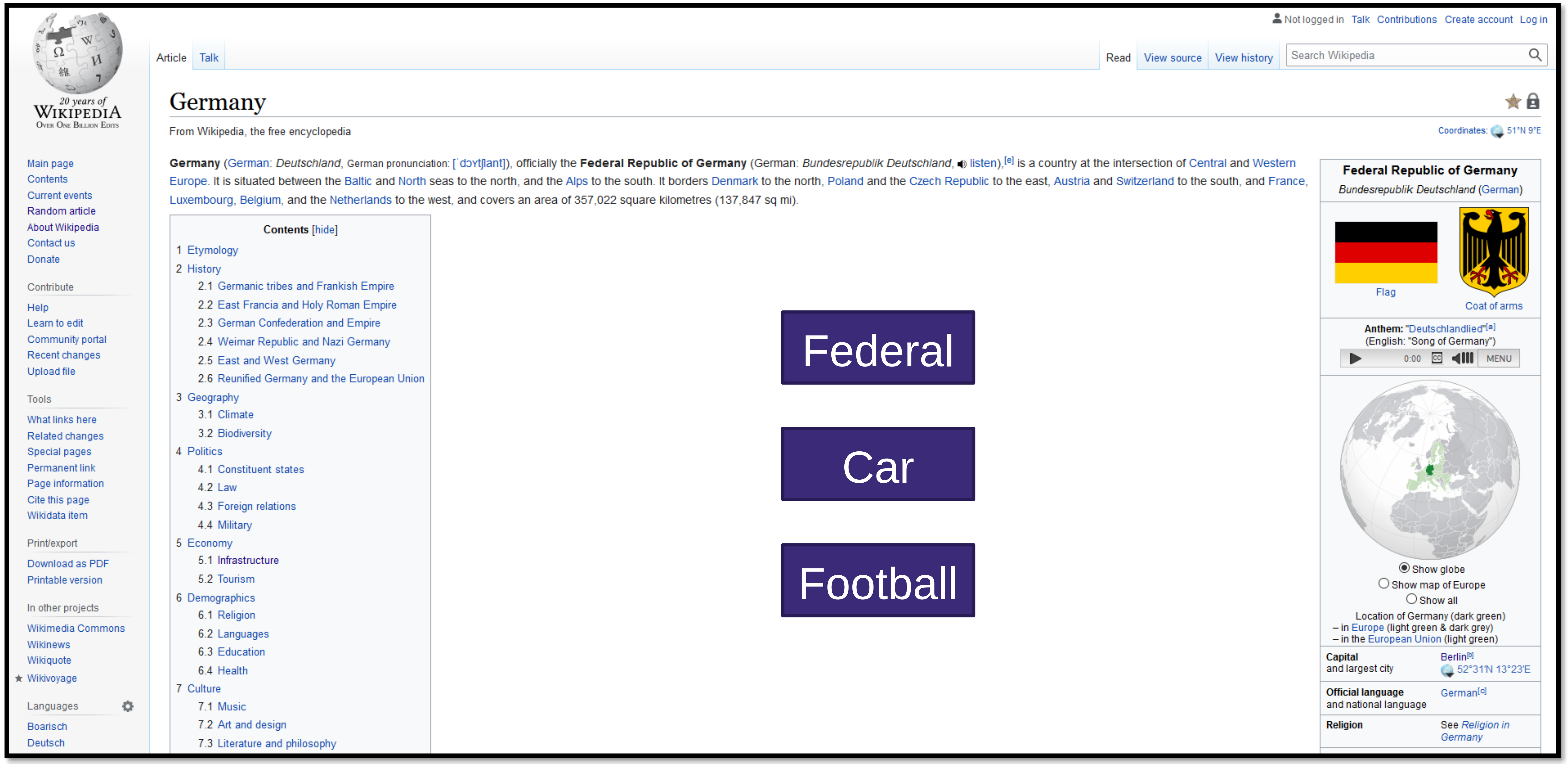
clock = [0,0,1,0]
human = [1,1,1,2]

	clock	is	human	language	natural	diverse	text	differs	represen tation	complex	body
D_1	0	1	1	1	0	0	1	0	1	1	0
D_2	0	2	1	1	1	1	0	0	0	1	0
D_3	1	1	1	0	1	0	0	0	0	1	1
D_4	0	0	2	0	0	0	1	1	1	0	0

Document-term matrix

- Pros
  - Simple
  - Fast to implement
- Cons
  - The resulting vectors are long (|D|) and sparse
  - It capture relatedness than similarity
  - It’s not a good idea in very long documents

Document-Term matrix



Distributional word vectors

- Frequency based
  - Document-Term matrix
  - Term-term matrix
  - Pointwise Mutual Information (PMI)
- Prediction based
  - Word2Vec

Term-term matrix

- Term-document does not work well, especially in the case of long documents
- Instead of entire documents, use smaller contexts
  - Paragraph
  - Window of surrounding words (e.g., ±3 words)
- Context words refers to surrounding words (i.e., Term-context matrix)
- The vector length is |V|

Term-term matrix

- D_1 = “Text is a **complex** human **language** representation.”
- D_2 = “Natural human **language** is **complex** and also is diverse.”

	text	is	a	complex	human	language	representation	natural	and	also	diverse
text											
is											
a											
complex		1	1		1	1					
human											
language				1	1		1				
representation											
natural											
and											
also											
diverse											

Context

±2

Term-term matrix

- D_1 = "Text is a **complex** human **language** representation."
- D_2 = "Natural human **language** is **complex** and also is diverse."

Context

	text	is	a	complex	human	language	represen tation	natural	and	also	diverse
text											
is											
a											
complex		2	1		1	2			1	1	
human											
language		1		2	2		1	1			
representation											
natural											
and											
also											
diverse											

±2

Term-term matrix

- D_1 = "Text is a **complex** human **language** representation."
- D_2 = "Natural human **language** is **complex** and also is diverse."

	text	is	a	complex	human	language	representation	natural	and	also	diverse
text	0	1	1	0	0	0	0	0	0	0	0
is	1	0	1	2	1	1	0	0	1	0	0
a	1	1	0	1	1	0	0	0	0	0	0
complex	0	2	1	0	1	2	0	0	1	1	0
human	0	1	1	1	0	2	1	1	0	0	0
language	0	1	0	2	2	0	1	1	0	0	0
representation	0	0	0	0	1	1	0	0	0	0	0
natural	0	0	0	0	1	1	0	0	0	0	0
and	0	2	0	1	0	0	0	0	0	1	0
also	0	1	0	1	0	0	0	0	1	0	1
diverse	0	1	0	0	0	0	0	0	0	1	0

Term-term matrix

- How to set the window size? (e.g., ±n)
  - n = 1, 2, 3, …

Natural human language is **complex** and also is diverse.

Natural human language is **complex** and also is diverse.

Natural human language is **complex** and also is diverse.

Natural human language is **complex** and also is diverse.

Shorter windows
More syntactic representation
±1-3 → Very syntactically

Longer windows
More semantic representation
> ±4 → More semantically (i.e., meaning)

First/second order co-occurrence

- Syntagmatic association (first order co-occurrence)
  - Words that are typically nearby each other
- Paradigmatic association (second order co-occurrence)
  - Words that have similar neighbors

Why is the water in the glass?
Drinking a glass of milk is part of maintaining a healthy diet

First/second order co-occurrence

- D_1 = "Text is a **complex** human **language** representation."
- D_2 = "Natural human **language** is **complex** and also is diverse."

	text	is	a	complex	human	language	represen tation	natural	and	also	diverse
text	0	1	1	0	0	0	0	0	0	0	0
is	1	0	1	2	1	1	0	0	1	0	0
a	1	1	0	1	1	0	0	0	0	0	0
complex	0	2	1	0	1	2	0	0	1	1	0
human	0	1	1	1	0	2	1	1	0	0	0
language	0	1	0	2	2	0	1	1	0	0	0
representation	0	0	0	0	1	1	0	0	0	0	0
natural	0	0	0	0	1	1	0	0	0	0	0
and	0	2	0	1	0	0	0	0	0	1	0
also	0	1	0	1	0	0	0	0	1	0	1
diverse	0	1	0	0	0	0	0	0	0	1	0

First/second order co-occurrence

- D_1 = “Text is a **complex** human **language** representation.”
- D_2 = “Natural human **language** is **complex** and also is diverse.”

	text	is	a	complex	human	language	representation	natural	and	also	diverse
text	0	1	1	0	0	0	0	0	0	0	0
is	1	0	1	2	1	1	0	0	1	0	0
a	1	1	0	1	1	0	0	0	0	0	0
complex	0	2	1	0	1	2	0	0	1	1	0
human	0	1	1	1	0	2	1	1	0	0	0
language	0	1	0	2	2	0	1	1	0	0	0
representation	0	0	0	0	1	1	0	0	0	0	0
natural	0	0									0
and	0	2									0
also	0	1									1
diverse	0	1	0	0	0	0	0	0	0	1	0

Syntagmatic Association (First order co-occurrence)

- Word that are typically nearby each other

First/second order co-occurrence

- D_1 = “Text is a **complex** human **language** representation.”
- D_2 = “Natural human **language** is **complex** and also is diverse.”

	text	is	a	complex	human	language	representation	natural	and	also	diverse
text	0	1	1	0	0	0	0	0	0	0	0
is	1	0	1	2	1	1	0	0	1	0	0
a	1	1	0	1	1	0	0	0	0	0	0
complex	0	2	1	0	1	2	0	0	1	1	0
human	0	1	1	1	0	2	1	1	0	0	0
language	0	1	0	2	2	0	1	1	0	0	0
representation	0	0	0	0	1	1	0	0	0	0	0
natural	0	0									
and	0	2									
also	0	1									
diverse	0	1	0	0	0	0	0	0	0	1	0

Paradigmatic Association (Second order co-occurrence)

- Word that have similar neighbors

Term-term matrix

- Pros
  - Simple to understand
  - Better capture word meaning than the term-document matrix
- Cons
  - The resulting vectors are long (|V|) and sparse
  - Some common words (e.g., “is”) relate some unrelated words to each other

Distributional word vectors

- Frequency based
  - Document-Term matrix
  - Term-Term matrix
  - Pointwise mutual information (PMI)
- Prediction based
  - Word2Vec

Pointwise mutual information (PMI)

- Problem with raw counts (e.g., term-term matrix)
  - Some words (like “is”) are very frequent, but maybe not the most **discriminative**
- We try to measure whether a context word is **informative**

$$PMI(W_1, W_2) = \log_2 \frac{P(W_1, W_2)}{P(W_1)P(W_2)}$$

- Do words W_1 and W_2 co-occur more than if they were independent?

PMI

$$PMI(W_1, W_2) = \log_2 \frac{P(W_1, W_2)}{P(W_1)P(W_2)}$$

Two events W_1, W_2 are independent if their joint probability is equal to the product of their individual probabilities

$$P(W_1, W_2) = P(W_1)P(W_2)$$

$$\frac{P(W_1, W_2)}{P(W_1)P(W_2)} = 1$$

$$\log_2 1 = 0$$

PMI

- D_1 = "Text is a complex human language representation."
- D_2 = "Natural human language is complex and also is diverse."

	text	is	a	complex	human	language	represen tation	natural	and	also	diverse
text	0	1	1	0	0	0	0	0	0	0	0
is	1	0	1	2	1	1	0	0	1	0	0
a	1	1	0	1	1	0	0	0	0	0	0
complex	0	2	1	0	1	2	0	0	1	1	0
human	0	1	1	1	0	2	1	1	0	0	0
language	0	1	0	2	2	0	1	1	0	0	0
representation	0	0	0	0	1	1	0	0	0	0	0
natural	0	0	0	0	1	1	0	0	0	0	0
and	0	2	0	1	0	0	0	0	0	1	0
also	0	1	0	1	0	0	0	0	1	0	1
diverse	0	1	0	0	0	0	0	0	0	1	0

PMI

- D_1 = “Text is a complex human language representation.”
- D_2 = “Natural human language is complex and also is diverse.”

	text	is	a	complex	human	language	representation	natural	and	also	diverse
text	0	1	1	0	0	0	0	0	0	0	0
is	1	0	1	2	1	1	0	0	1	0	0
a	1	1	0	1	1	0	0	0	0	0	0
complex	0	2	1	0	1	2	0	0	1	1	0
human	0	1	1	1	0	2	1	1	0	0	0
language	0	1	0	2	2	0	1	1	0	0	0
representation	0	0	0	0	1	1	0	0	0	0	0
natural	0	0	0	0	1	1	0	0	0	0	0
and	0	2	0	1	0	0	0	0	0	1	0
also	0	1	0	1	0	0	0	0	1	0	1
diverse	0	1	0	0	0	0	0	0	0	1	0

PMI

$$PMI(W_1, W_2) = \log_2 \frac{P(W_1, W_2)}{P(W_1)P(W_2)}$$

- $P(W_1, W_2) = \frac{\#\ of\ times\ W_1\ occurs\ in\ context\ of\ W_2}{\#\ of\ times\ all\ words\ occur\ in\ context\ of all\ the\ other\ words}$
- $P(W_1) = \frac{\#\ of\ times\ W_1\ occurs\ in\ context\ of\ all\ context\ words}{\#\ of\ times\ all\ words\ occur\ in\ context\ of all\ the\ other\ words}$
- $P(W_2) = \frac{\#\ of\ times\ that\ all\ the\ words\ occurs\ in\ context\ of\ W_2}{\#\ of\ times\ all\ words\ occur\ in\ context\ of all\ the\ other\ words}$

PMI

	text	is	a	complex	human	language	representation	natural	and	also	diverse
text	0	1	1	0	0	0	0	0	0	0	0
is	1	0	1	2	1	1	0	0	1	0	0
a	1	1	0	1	1	0	0	0	0	0	0
complex	0	2	1	0	1	2	0	0	1	1	0
human	0	1	1	1	0	2	1	1	0	0	0
language	0	1	0	2	2	0	1	1	0	0	0
representation	0	0	0	0	1	1	0	0	0	0	0
natural	0	0	0	0	1	1	0	0	0	0	0
and	0	2	0	1	0	0	0	0	0	1	0
also	0	1	0	1	0	0	0	0	1	0	1
diverse	0	1	0	0	0	0	0	0	0	1	0

PMI

$$\boldsymbol{PMI(human, is)}$$

$$p(human, is) = {}^{1}/_{49} \;\mid\; p = (human) = {}^{7}/_{49} \mid p(is) = {}^{10}/_{49}$$

	text	$\boxed{\text{is}}$	a	complex	human	language	represen tation	natural	and	also	diverse
text	0	1	1	0	0	0	0	0	0	0	0
is	1	0	1	2	1	1	0	0	1	0	0
a	1	1	0	1	1	0	0	0	0	0	0
complex	0	2	1	0	1	2	0	0	1	1	0
$\boxed{\text{human}}$	0	$\boxed{1}$	1	1	0	2	1	1	0	0	0
language	0	1	0	2	2	0	1	1	0	0	0
representation	0	0	0	0	1	1	0	0	0	0	0
natural	0	0	0	0	1	1	0	0	0	0	0
and	0	2	0	1	0	0	0	0	0	1	0
also	0	1	0	1	0	0	0	0	1	0	1
diverse	0	1	0	0	0	0	0	0	0	1	0

PMI

$$\boldsymbol{PMI(human, is)}$$

$$p(human, is) = {}^{1}/_{49} \ \mid \ p = (human) = {}^{7}/_{49} \mid p(is) = {}^{10}/_{49}$$

	text	is	a	complex	human	language	representation	natural	and	also	diverse
text	0	1	1	0	0	0	0	0	0	0	0
is	1	0	1	2	1	1	0	0	1	0	0
a	1	1	0	1	1	0	0	0	0	0	0
complex	0	2	1	0	1	2	0	0	1	1	0
human	0	1	1	1	0	2	1	1	0	0	0
language	0	1	0	2	2	0	1	1	0	0	0
representation	0	0	0	0	1	1	0	0	0	0	0
natural	0	0	0	0	1	1	0	0	0	0	0
and	0	2	0	1	0	0	0	0	0	1	0
also	0	1	0	1	0	0	0	0	1	0	1
diverse	0	1	0	0	0	0	0	0	0	1	0

PMI

$$\boldsymbol{PMI(human, is)}$$

$$p(human, is) = {}^{1}/_{49} \mid p = (human) = {}^{7}/_{49} \mid p(is) = {}^{10}/_{49}$$

	text	is	a	complex	human	language	represen tation	natural	and	also	diverse
text	0	1	1	0	0	0	0	0	0	0	0
is	1	0	1	2	1	1	0	0	1	0	0
a	1	1	0	1	1	0	0	0	0	0	0
complex	0	2	1	0	1	2	0	0	1	1	0
human	0	1	1	1	0	2	1	1	0	0	0
language	0	1	0	2	2	0	1	1	0	0	0
representation	0	0	0	0	1	1	0	0	0	0	0
natural	0	0	0	0	1	1	0	0	0	0	0
and	0	2	0	1	0	0	0	0	0	1	0
also	0	1	0	1	0	0	0	0	1	0	1
diverse	0	1	0	0	0	0	0	0	0	1	0

PMI

$$\boldsymbol{PMI(human, is)}$$

$$p(human, is) = {}^{1}/_{49} \;\mid\; p = (human) = {}^{7}/_{49} \mid p(is) = {}^{10}/_{49}$$

$$PMI(human, is) = \log_2 \frac{{}^{1}/_{49}}{{}^{7}/_{49} * {}^{10}/_{49}} = -0.51$$

	text	is	a	complex	human	language	represen tation	natural	and	also	diverse
text	0	1	1	0	0	0	0	0	0	0	0
is	1	0	1	2	1	1	0	0	1	0	0
a	1	1	0	1	1	0	0	0	0	0	0
complex	0	2	1	0	1	2	0	0	1	1	0
human	0	1	1	1	0	2	1	1	0	0	0
language	0	1	0	2	2	0	1	1	0	0	0
representation	0	0	0	0	1	1	0	0	0	0	0
natural	0	0	0	0	1	1	0	0	0	0	0
and	0	2	0	1	0	0	0	0	0	1	0
also	0	1	0	1	0	0	0	0	1	0	1
diverse	0	1	0	0	0	0	0	0	0	1	0

PMI

$$\boldsymbol{PMI(human, natural)}$$

$$p(human, natural) = {}^{1}/_{49} \ \mid p = (human) = {}^{7}/_{49} \ \mid p(natural) = {}^{2}/_{49}$$

	text	is	a	complex	human	language	representation	natural	and	also	diverse
text	0	1	1	0	0	0	0	0	0	0	0
is	1	0	1	2	1	1	0	0	1	0	0
a	1	1	0	1	1	0	0	0	0	0	0
complex	0	2	1	0	1	2	0	0	1	1	0
human	0	1	1	1	0	2	1	1	0	0	0
language	0	1	0	2	2	0	1	1	0	0	0
representation	0	0	0	0	1	1	0	0	0	0	0
natural	0	0	0	0	1	1	0	0	0	0	0
and	0	2	0	1	0	0	0	0	0	1	0
also	0	1	0	1	0	0	0	0	1	0	1
diverse	0	1	0	0	0	0	0	0	0	1	0

PMI

$$\boldsymbol{PMI(human, natural)}$$

$$p(human, natural) = {}^{1}\!/_{49} \;\mid\; p = (human) = {}^{7}\!/_{49} \;\mid\; p(natural) = {}^{2}\!/_{49}$$

	text	is	a	complex	human	language	representation	natural	and	also	diverse
text	0	1	1	0	0	0	0	0	0	0	0
is	1	0	1	2	1	1	0	0	1	0	0
a	1	1	0	1	1	0	0	0	0	0	0
complex	0	2	1	0	1	2	0	0	1	1	0
human	0	1	1	1	0	2	1	1	0	0	0
language	0	1	0	2	2	0	1	1	0	0	0
representation	0	0	0	0	1	1	0	0	0	0	0
natural	0	0	0	0	1	1	0	0	0	0	0
and	0	2	0	1	0	0	0	0	0	1	0
also	0	1	0	1	0	0	0	0	1	0	1
diverse	0	1	0	0	0	0	0	0	0	1	0

PMI

$$\boldsymbol{PMI(human, natural)}$$

$$p(human, natural) = {}^{1}/_{49} \mid p = (human) = {}^{7}/_{49} \mid p(natural) = {}^{2}/_{49}$$

	text	is	a	complex	human	language	representation	natural	and	also	diverse
text	0	1	1	0	0	0	0	0	0	0	0
is	1	0	1	2	1	1	0	0	1	0	0
a	1	1	0	1	1	0	0	0	0	0	0
complex	0	2	1	0	1	2	0	0	1	1	0
human	0	1	1	1	0	2	1	1	0	0	0
language	0	1	0	2	2	0	1	1	0	0	0
representation	0	0	0	0	1	1	0	0	0	0	0
natural	0	0	0	0	1	1	0	0	0	0	0
and	0	2	0	1	0	0	0	0	0	1	0
also	0	1	0	1	0	0	0	0	1	0	1
diverse	0	1	0	0	0	0	0	0	0	1	0

PMI

$$\boldsymbol{PMI(human, natural)}$$

$$p(human, natural) = {}^{1}/_{49} \ \mid p = (human) = {}^{7}/_{49} \ \mid p(natural) = {}^{2}/_{49}$$

$$PMI(human, natural) = \log_2 \frac{1/_{49}}{7/_{49} * 2/_{49}} = 1.8$$

	text	is	a	complex	human	language	representation	natural	and	also	diverse
text	0	1	1	0	0	0	0	0	0	0	0
is	1	0	1	2	1	1	0	0	1	0	0
a	1	1	0	1	1	0	0	0	0	0	0
complex	0	2	1	0	1	2	0	0	1	1	0
human	0	1	1	1	0	2	1	1	0	0	0
language	0	1	0	2	2	0	1	1	0	0	0
representation	0	0	0	0	1	1	0	0	0	0	0
natural	0	0	0	0	1	1	0	0	0	0	0
and	0	2	0	1	0	0	0	0	0	1	0
also	0	1	0	1	0	0	0	0	1	0	1
diverse	0	1	0	0	0	0	0	0	0	1	0

PMI

	text	is	a	complex	human	language	represen tation	natural	and	also	diverse
text	0	1	1	0	0	0	0	0	0	0	0
is	1	0	[illegible]	[illegible]	1	1	0	0	1	0	0
a	1	1	[illegible]	[illegible]	1	0	0	0	0	0	0
complex	0	2	[illegible]	[illegible]	1	2	0	0	1	1	0
human	0	**1**	1	1	0	2	1	**1**	0	0	0
language	0	1	0	2	2	[illegible]	[illegible]	1	0	0	0
representation	0	0	0	0	1	[illegible]	[illegible]	[illegible]	0	0	0
natural	0	0	0	0	1	[illegible]	[illegible]	[illegible]	0	0	0
and	0	2	0	1	0	[illegible]	[illegible]	0	0	1	0
also	0	1	0	1	0	0	0	0	1	0	1
diverse	0	1	0	0	0	0	0	0	0	1	0

$$PMI = -0.51$$

$$PMI = +1.8$$

Positive pointwise mutual information (PPMI)

$$PPMI(W_1, W_2) = max\left(\log_2 \frac{P(W_1, W_2)}{P(W_1)P(W_1)}, 0\right)$$

- The values should be counted on a huge corpus to be sure if two terms are really unrelated
- It’s also difficult to interpret if larger negative value means more un-relatedness

PMI

- PMI is biased toward infrequent events
  - Very rare words have very high PMI values
- Possible solution
  - Use add-one smoothing (Laplace smoothing)

Use add-one smoothing

	text	is	a	complex	human	language	represen tation	natural	and	also	diverse
text	0	1	1	0	0	0	0	0	0	0	0
is	1	0	1	2	1	1	0	0	1	0	0
a	1	1	0	1	1	0	0	0	0	0	0
complex	0	2	1	0	1	2	0	0	1	1	0
human	0	1	1	1	0	2	1	1	0	0	0
language	0	1	0	2	2	0	1	1	0	0	0
representation	0	0	0	0	1	1	0	0	0	0	0
natural	0	0	0	0	1	1	0	0	0	0	0
and	0	2	0	1	0	0	0	0	0	1	0
also	0	1	0	1	0	0	0	0	1	0	1
diverse	0	1	0	0	0	0	0	0	0	1	0

Use add-one smoothing

	text	is	a	complex	human	language	represen tation	natural	and	also	diverse
text	2	3	3	2	2	2	2	2	2	2	2
is	3	2	3	4	3	3	2	2	3	2	2
a	3	3	2	3	3	2	2	2	2	2	2
complex	2	4	3	2	3	4	2	2	3	3	2
human	2	3	3	3	2	4	3	3	2	2	2
language	2	3	2	4	4	2	3	3	2	2	2
representation	2	2	2	2	3	3	2	2	2	2	2
natural	2	2	2	2	3	3	2	2	2	2	2
and	2	4	2	3	2	2	2	2	2	3	2
also	2	3	2	3	2	2	2	2	3	2	3
diverse	2	3	2	2	2	2	2	2	2	3	2

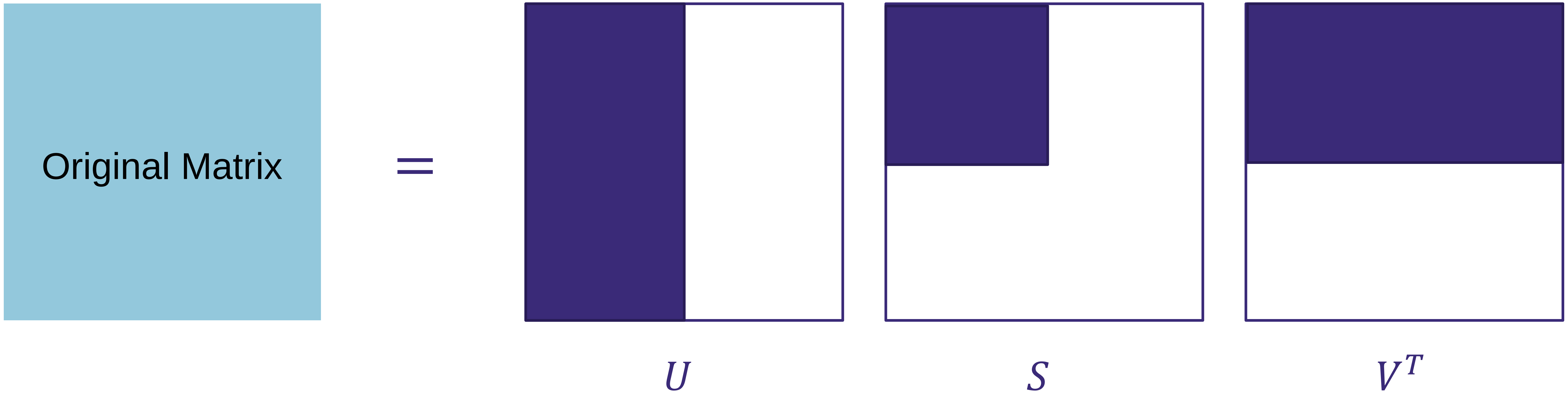
+2

PMI

- Pros
  - Better capture word meaning then the term-term matrix
  - Penalize scores by the common words
- Cons
  - The resulting vectors are long (|V|) and sparse

PMI

- How to resolve the sparsity issue in PMI
- Matrix factorization
  - Singular value decomposition (SVD)



Word embedding

- What is word embedding?
- One-hot word representation
- Distributional word vectors
  - Frequency based
  - Prediction based
- Word embedding evaluation
- Word embedding in Python

Distributional word vectors

- Frequency based
  - Document-Term matrix
  - Term-Term matrix
  - Pointwise Mutual Information (PMI)
- Prediction based
  - Word2Vec

From sparse to dense vectors

- Frequency based embedding
  - Long (~10,000 to 50,000)
  - Sparse (most elements are 0)
- Prediction based embedding (word embedding)
  - Short (~100 to 1,000)
  - Dense (most element are non-zero)

Why dense vectors

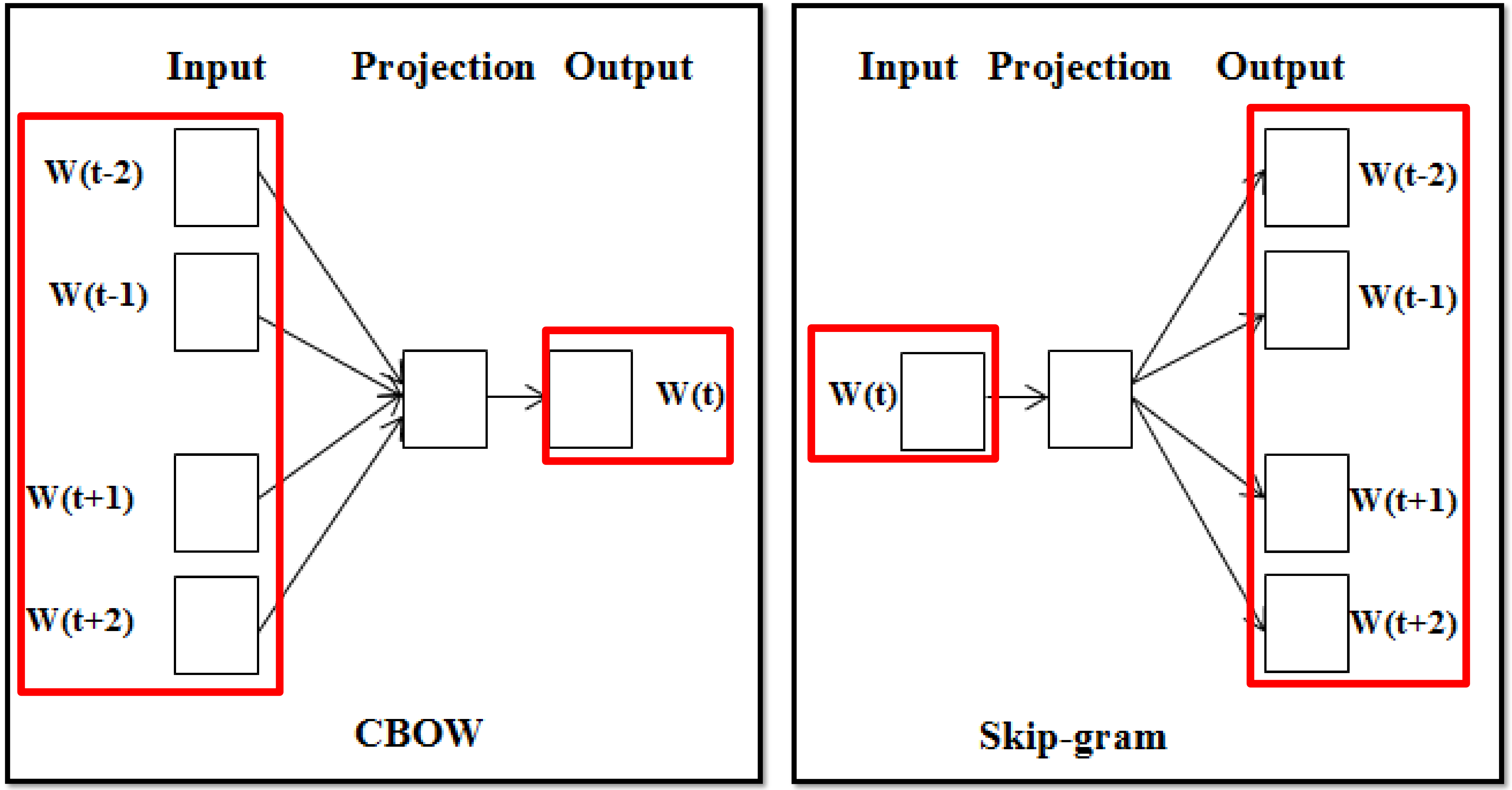
- They usually better capture meaning (e.g., work better in finding synonyms)
- Leads to less weights to trains in machine learning models

Word2Vec

- The ***word2vec*** model uses a ***neural network*** architecture (**two-layer** neural net) to learn word associations from a ***large corpus of text***
- ***Word2vec*** was created and published in ***2013*** by a team of researchers led by ***Tomas Mikolov*** at ***Google*** over two papers
- While word2vec ***is not a deep neural network***, it turns text into a numerical form that deep neural networks can understand
- Two word2Vec models:
  - continuous bag-of-words (CBOW)
  - skip-gram

Word2Vec

- Given context words
- Predict the probability of a target word



- Given a target word
- Predict the probability of context words

Word2Vec

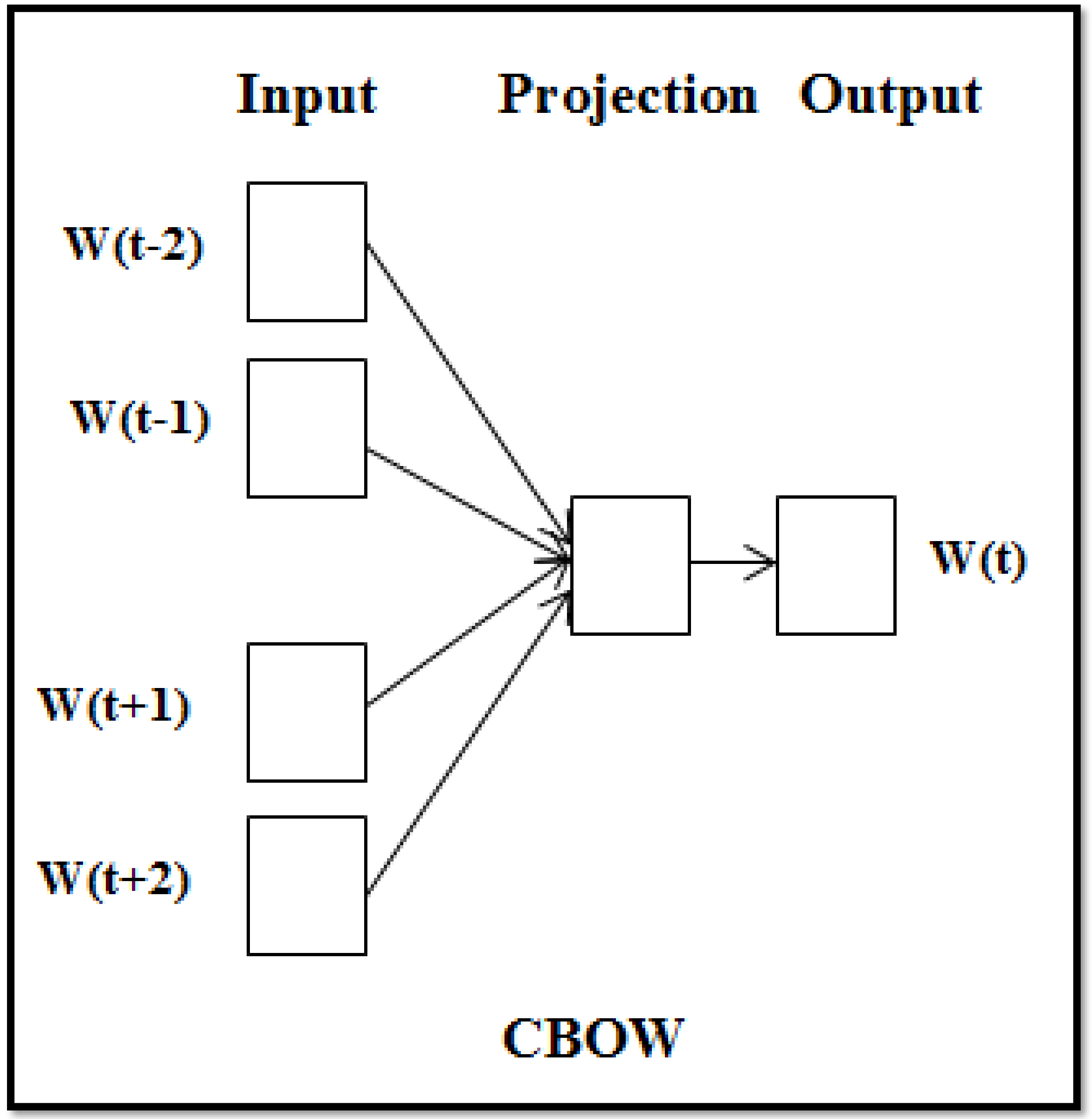
- We won’t be interested in the ***inputs*** and ***outputs*** of this network
- Rather the goal is actually just to learn the weights of the hidden layer that are actually the ***word vectors*** that we’re trying to learn

CBOW

Natural human **language** is complex and also is diverse

- Window size: ±2 (hyperparameter)
- Vocabulary size: 8
- Vector size: 5 (hyperparameter)

natural
human
language
is
complex
and
also
diverse



CBOW

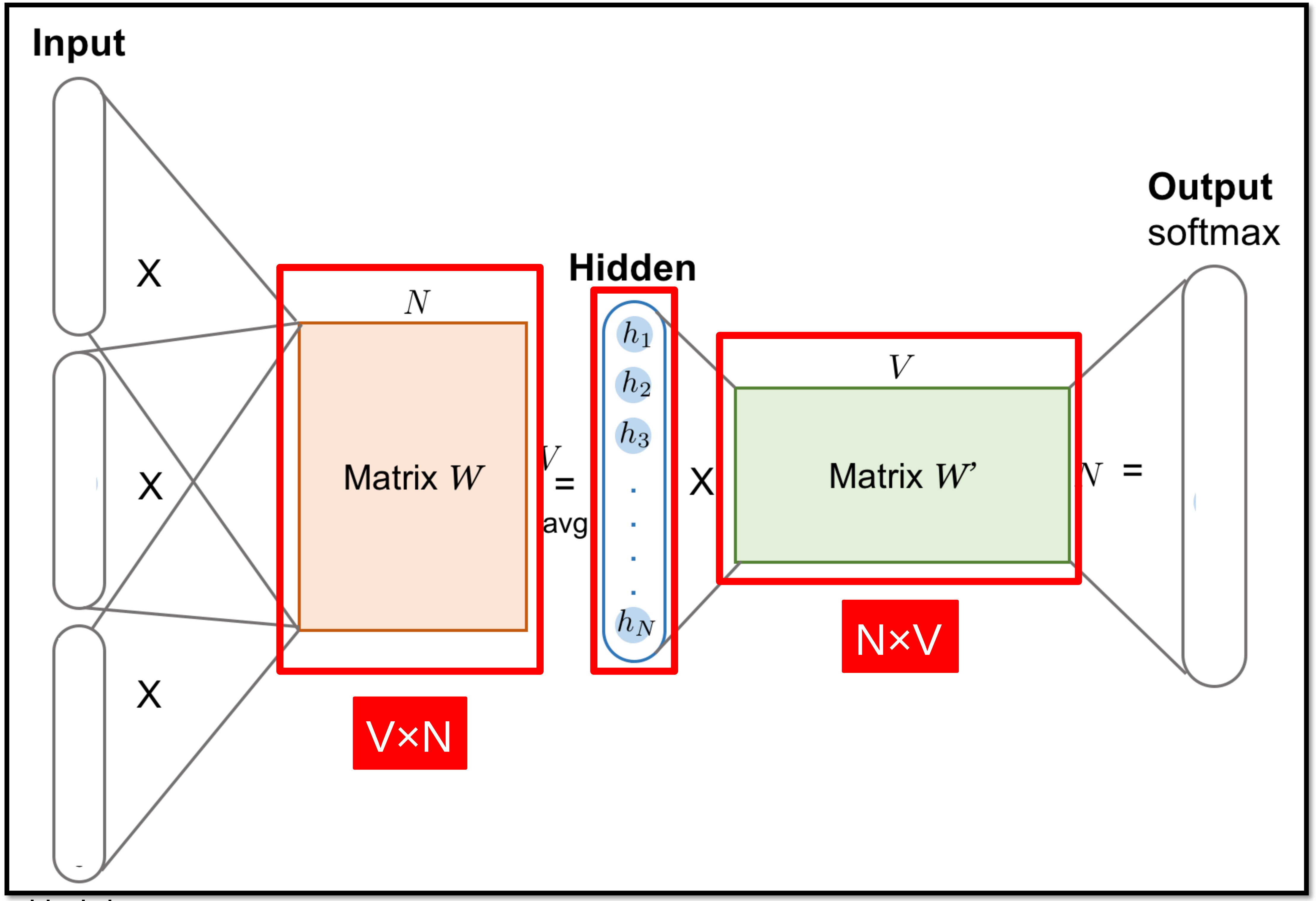


Image from www.lilianweng.github.io

CBOW

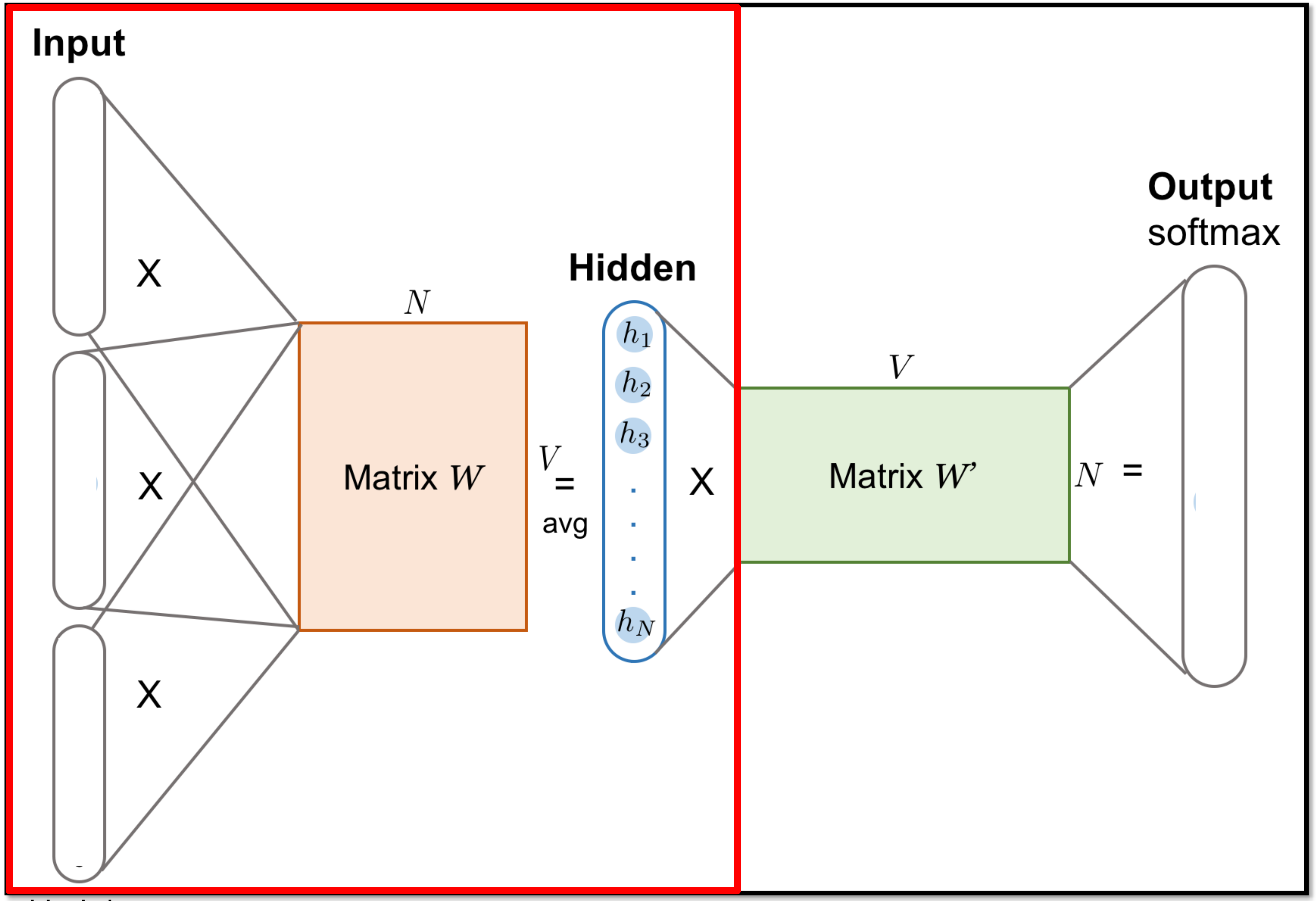


Image from www.lilianweng.github.io

CBOW

Window size: ±2 (hyperparameter)
Vocabulary size: 8
Vector size: 5 (hyperparameter)

Natural human **language** is complex and also is diverse

natural	1	0	0	0	0	0	0	0
human	0	1	0	0	0	0	0	0
is	0	0	0	1	0	0	0	0
complex	0	0	0	0	1	0	0	0

1×8

×

4	5	2	3	2
3	5	3	4	0
1	3	4	3	1
3	5	1	3	4
5	5	5	2	2
3	2	1	2	4
5	3	2	1	4
1	4	1	5	4

8×5

=

4	5	2	3	2
3	5	3	4	0
3	5	1	3	4
5	5	5	2	2

Average

3	5	4	3	2

CBOW

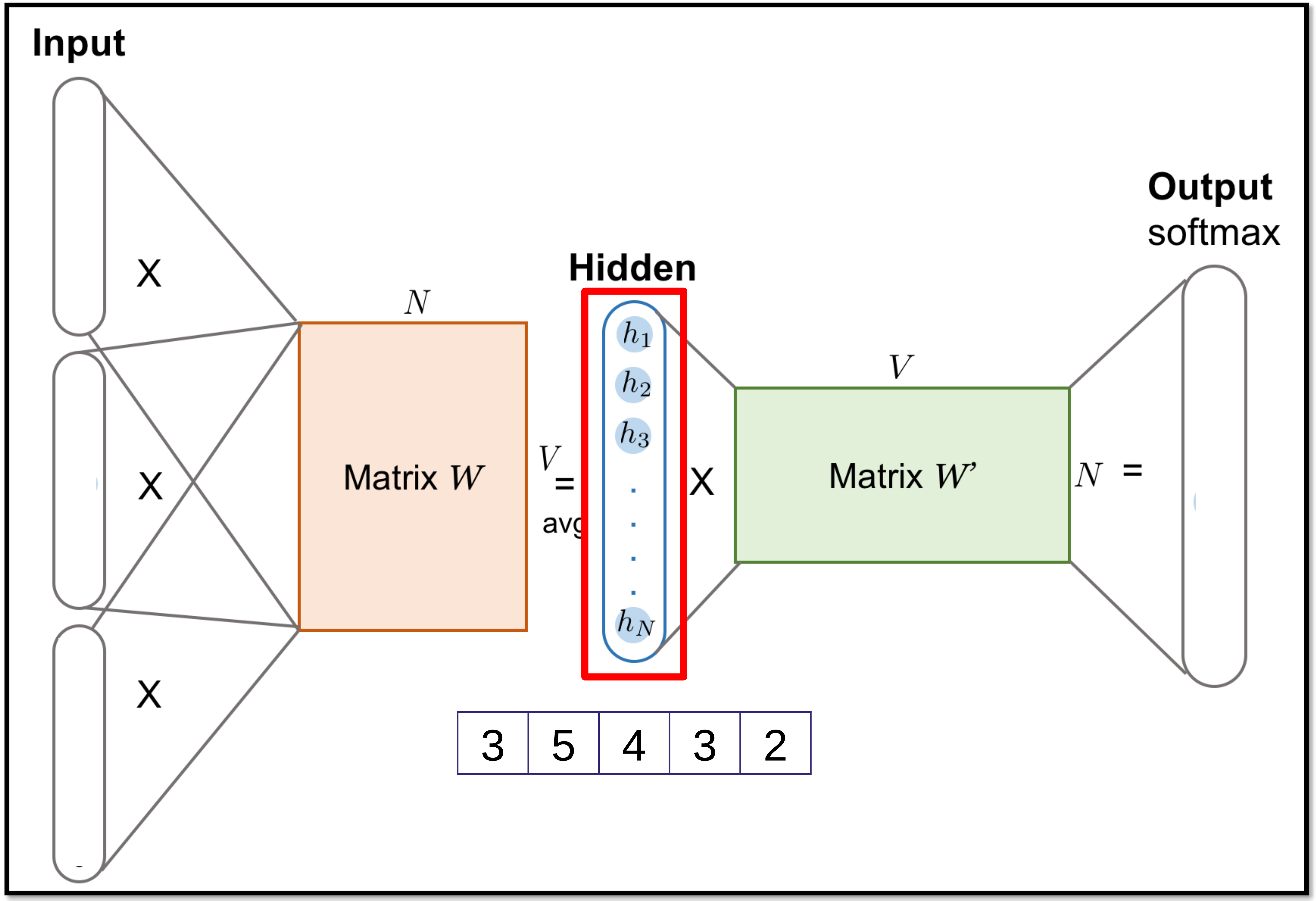


Image from www.lilianweng.github.io

CBOW

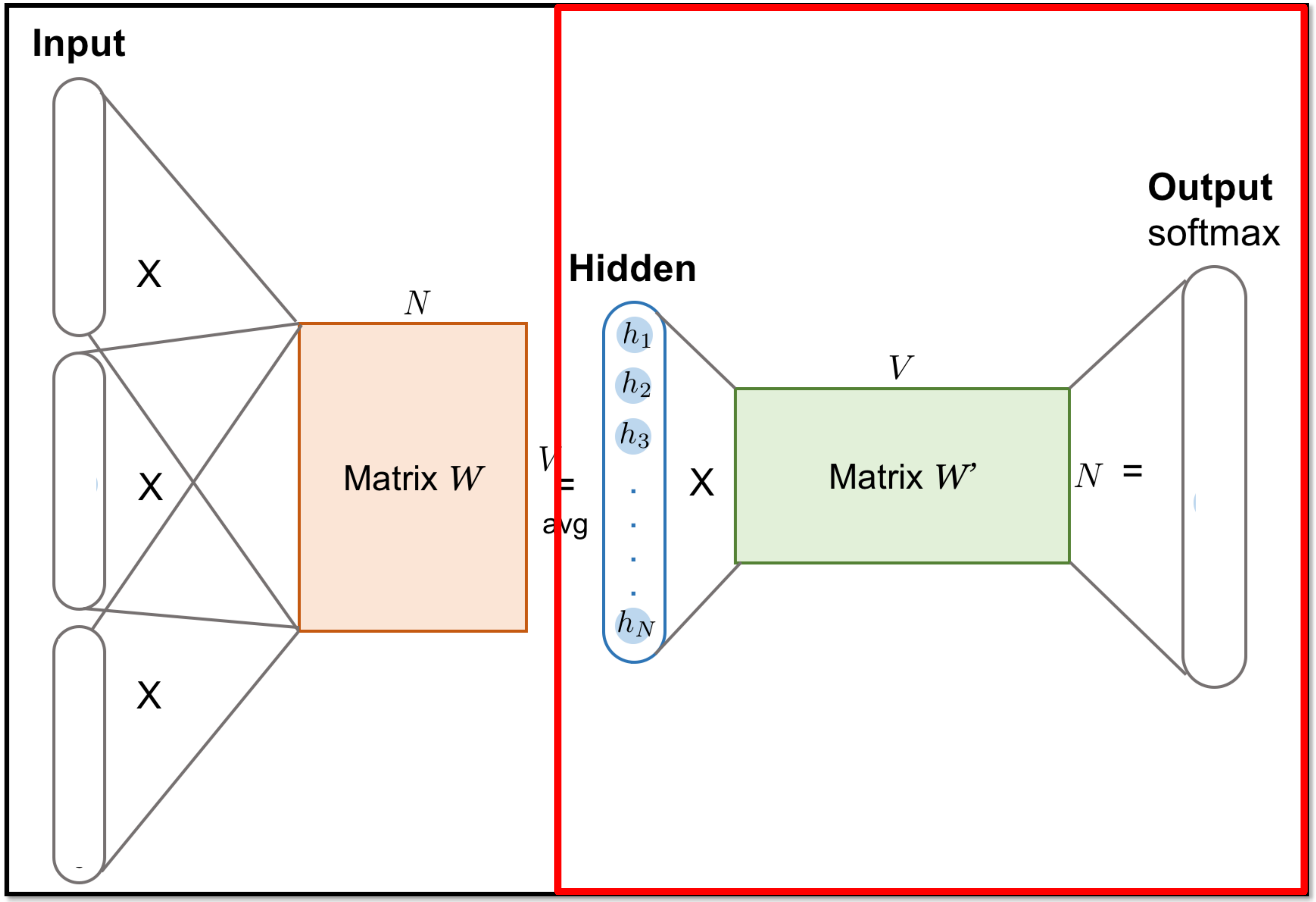


Image from www.lilianweng.github.io

CBOW

Window size: ±2 (hyperparameter)
Vocabulary size: 8
Vector size: 5 (hyperparameter)

$$\begin{bmatrix} 3 & 5 & 4 & 3 & 2 \end{bmatrix} \times \underbrace{\begin{bmatrix} 3 & 2 & 3 & 1 & 5 & 2 & 3 & 0 \\ 3 & 5 & 1 & 0 & 3 & 5 & 5 & 4 \\ 2 & 0 & 0 & 1 & 5 & 2 & 2 & 2 \\ 2 & 4 & 3 & 1 & 0 & 5 & 3 & 4 \\ 3 & 3 & 5 & 2 & 5 & 3 & 5 & 5 \end{bmatrix}}_{5\times 8} = \underbrace{\begin{bmatrix} 44 & 49 & 33 & 14 & 60 & 60 & 61 & 50 \end{bmatrix}}_{1\times 8}$$

Output layer

CBOW

Natural human **language** is complex and also is diverse

language

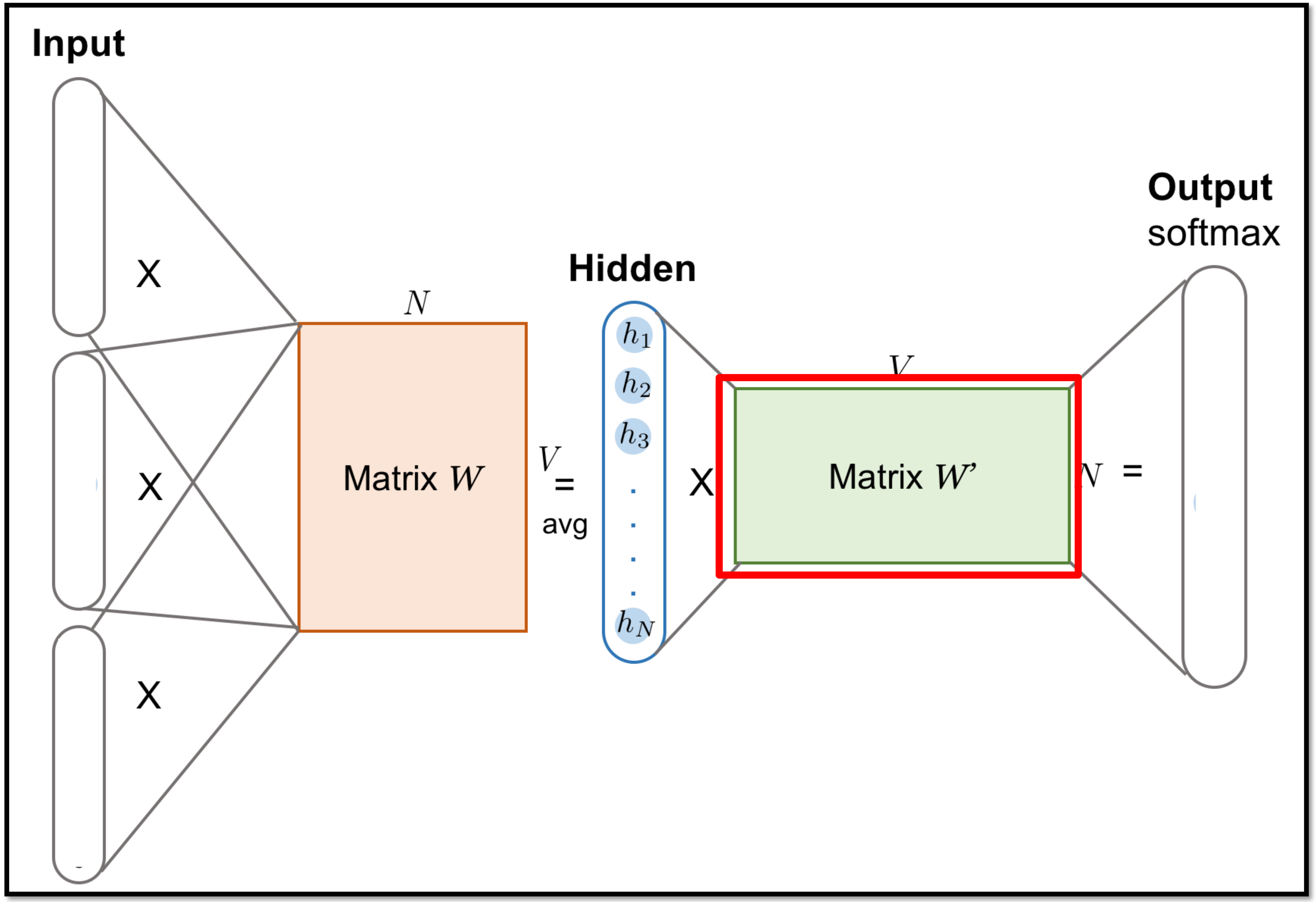
0	0	1	0	0	0	0	0

44	49	33	14	60	60	61	50
natural	human	language	is	complex	and	also	diverse

softmax

0.00	0.00	0.00	0.00	0.21	0.21	0.58	0.00
natural	human	language	is	complex	and	also	diverse

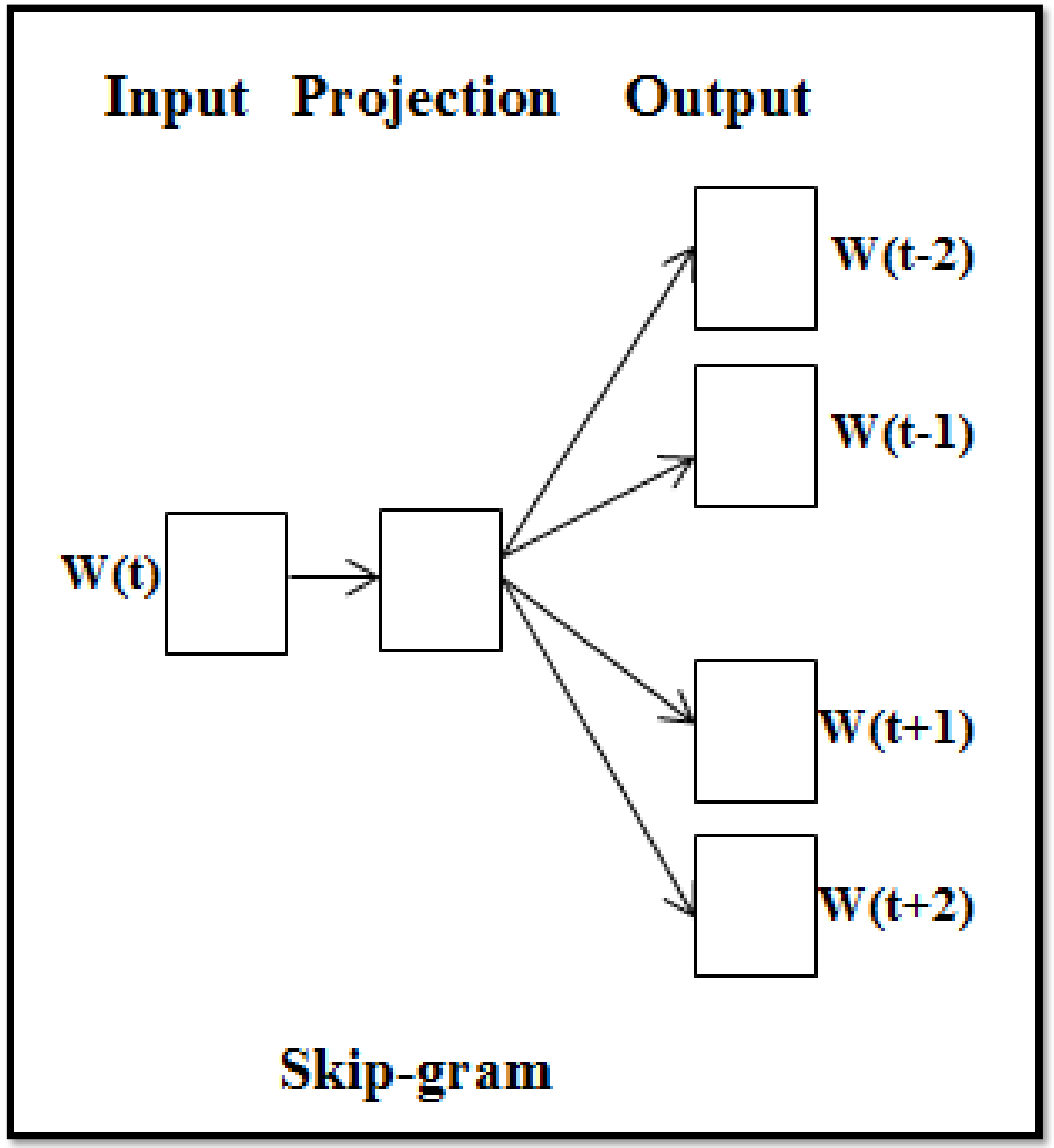
CBOW



3	2	3	1	5	2	3	0
3	5	1	0	3	5	5	4
2	0	0	1	5	2	2	2
2	4	3	1	0	5	3	4
3	3	5	2	5	3	5	5

Skip-gram

- The calculations up to hidden layer activations are the same as CBOW
- The difference will be in the target variable
  - Considering a context window of 2 words in each side, there will be ***4*** one hot encoded target variables and ***4*** corresponding outputs
  - So we calculate ***4*** errors and the error vectors obtained are added element-wise to obtain a final error vector which is propagated back to update the weights



Problems with CBOW/Skip-gram

1. For each training sample, **only the weights corresponding to the target word might get a significant update**.
   - The weight corresponding to non-target words would receive a marginal or no change at all
2. For every training sample, **the calculation of the final probabilities using the softmax is quite an expensive operation**

- Possible solutions
  - Negative sampling
  - Sub sampling

Problems with CBOW/Skip-gram

- Negative Sampling
  - Instead of trying to predict the probability of being a nearby word for all the words in the vocabulary, we try to predict the probability that our training sample words are neighbors or not
  - Referring to our previous example of ***(human, language)***, we don't try to predict the probability for human to be a nearby word, we try to predict whether ***(human, language)*** are nearby words or not
  - Modifying the problem from a ***multi-class classification*** with ***N*** classes into ***N binary classification*** problem

Problems with CBOW/Skip-gram

- Sub Sampling
  - The distribution of words in a corpus is not uniform. Some words occur more frequently than the other

- Analyzing the occurrence of words with "**the**" doesn't tell us much about the meaning of words. "**the**" appears in the context of pretty much every word.
- We will have many more samples of ("**the**", …) than we need to learn a good vector for "**the**".

  - In sub-sampling, we limit the number of samples for a word by capping their frequency of occurrence. For frequently occurring words, we remove a few of their instances both as a neighboring word and as the input word

CBOW vs. Skip-gram

Skip-gram

- Works well with a small training data
- Represents well for rare words or phrases

CBOW

- Several times faster
- Better accuracy for the frequent words

Word Embedding

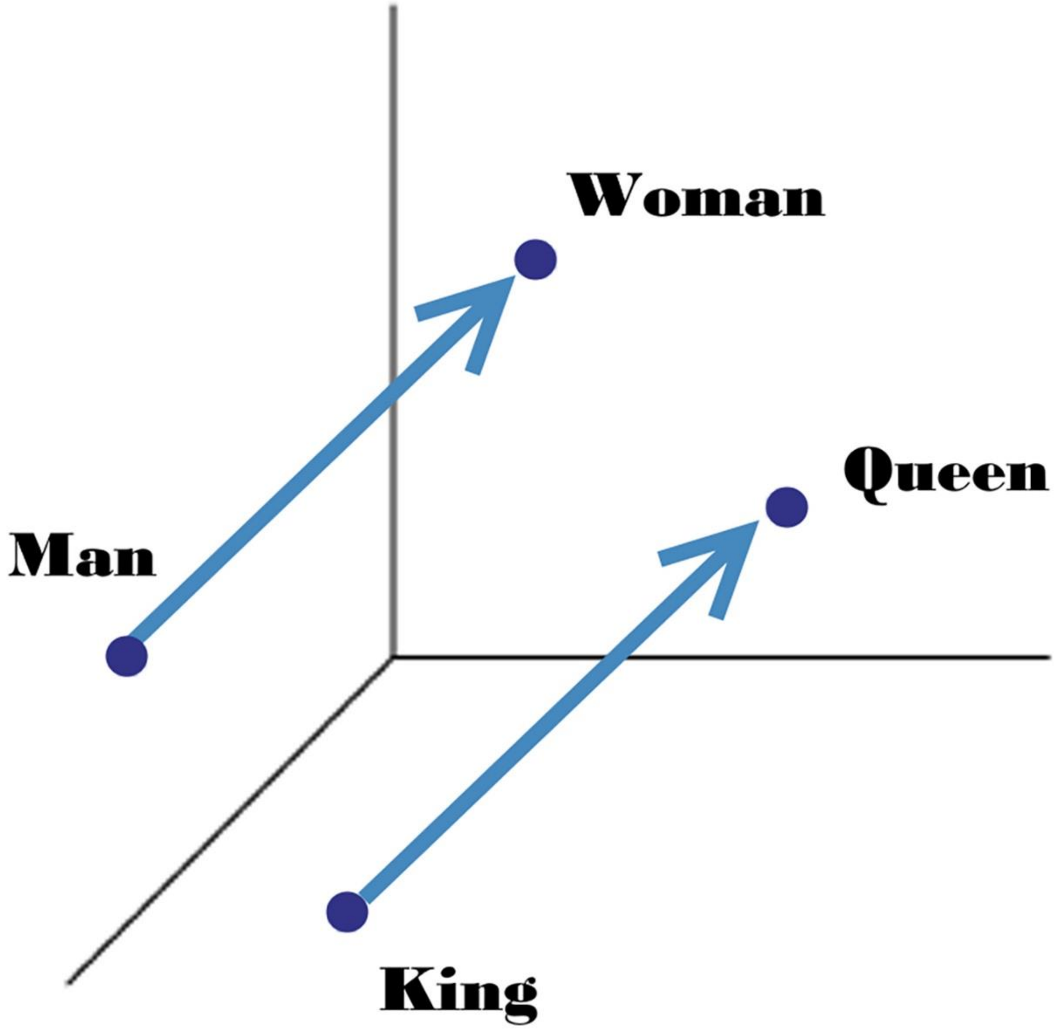
- What is word embedding?
- One-hot word representation
- Distributional word vectors
  - Frequency based
  - Prediction based
- Word embedding evaluation
- Word embedding in Python

Word embedding arithmetic properties

- A surprising property of word vectors is that word analogies can often be solved with vector arithmetic

King — Man + Woman = Queen

Rome - Italy = Berlin - Germany



Word embedding evaluation

Explicit methods

- Human evaluation
- Semantic analogies

Implicit methods

- Downstream task

Word embedding evaluation

- Explicit methods
  - Human evaluation
  - Semantic analogies

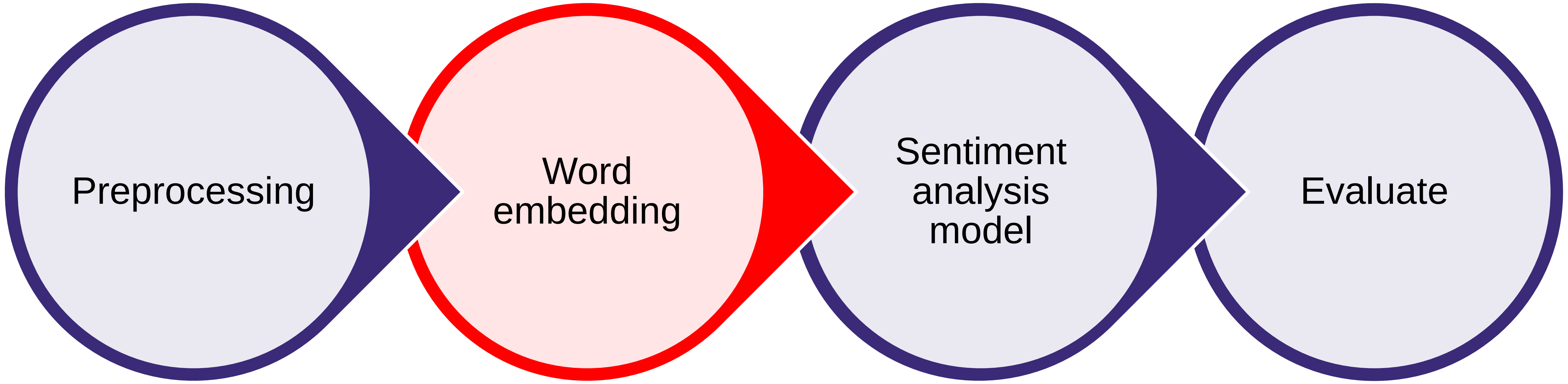
(Man , King) = (Woman , ?)

(Germany , Berlin) = (France , ?)

doctor	nurse	**7.00**
professor	doctor	**6.62**
stock	jaguar	**0.92**
stock	market	**8.08**
company	stock	**7.08**

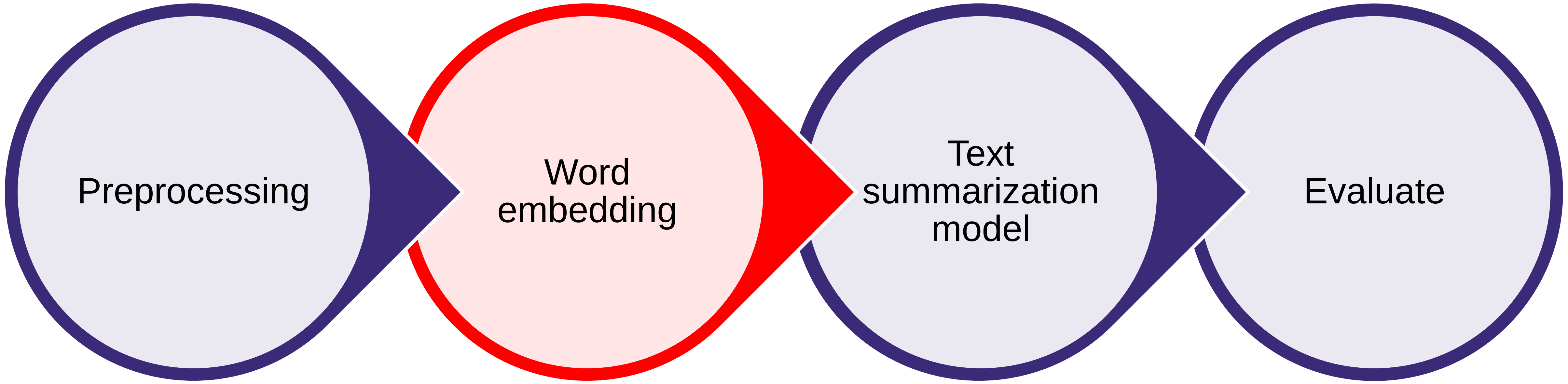
Word embedding evaluation

- Implicit methods
  - Measure performance in a downstream task



Word embedding evaluation

- Implicit methods
  - Measure performance in a downstream task



Word Embedding

- What is word embedding?
- One-hot word representation
- Distributional word vectors
  - Frequency based
  - Prediction based
- Word embedding evaluation
- Word embedding in Python

Word embedding in Python

- Gensim

GENSIM
topic modelling for humans

Gensim

- Train a model

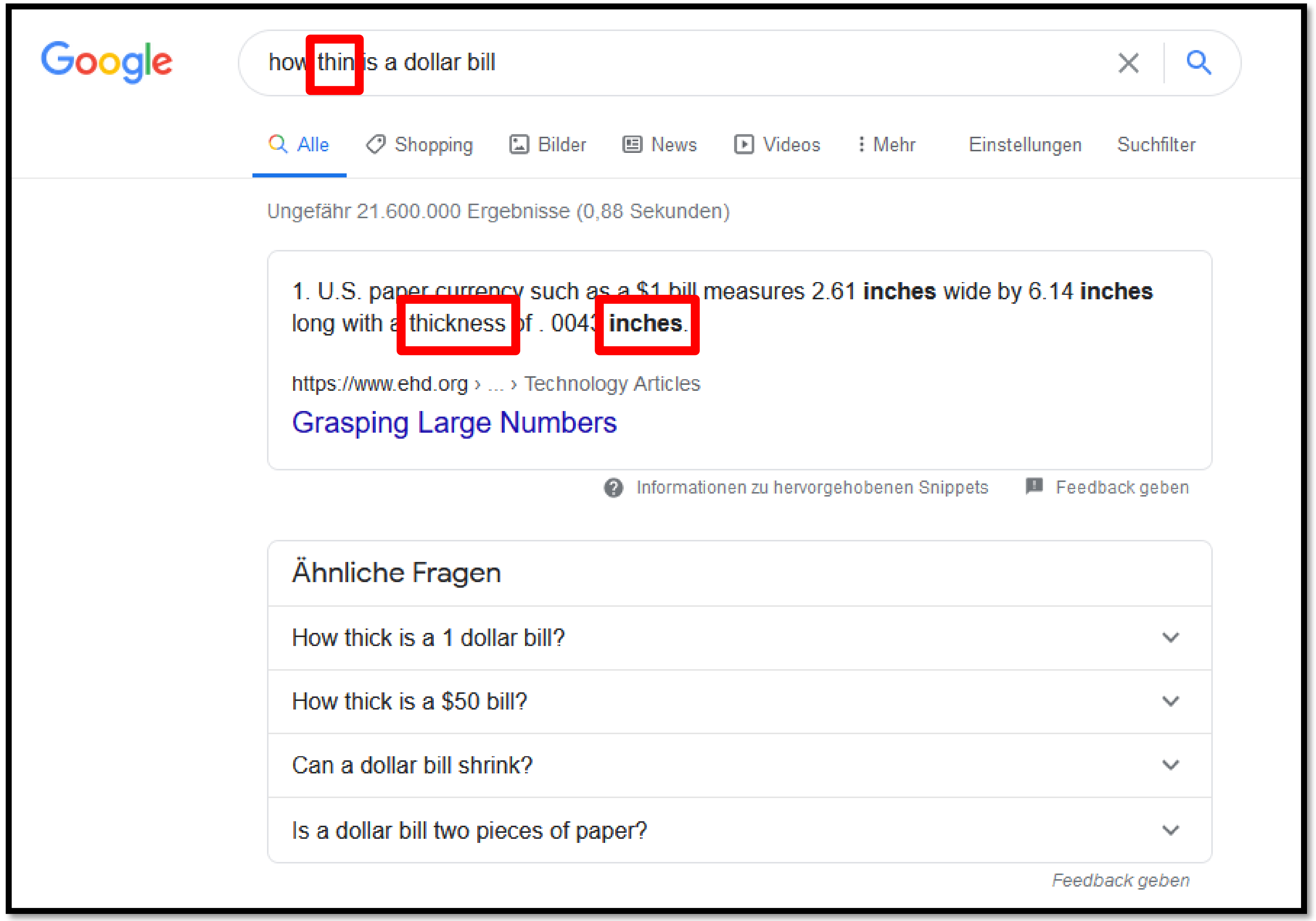
```
>>> from gensim.models import Word2Vec
>>> model = Word2Vec(sentences=sample_texts, vector_size=100,
window=5)
>>> vector = model.wv['computer’]

>>> vector.most_similar('computer')
[(‘laptop', 0.948005199432373),
 (‘mouse', 0.9403423070907593)]
```

Gensim

- Load a model

```
>>> import gensim.downloader
>>> print(list(gensim.downloader.info()['models'].keys()))
['word2vec-ruscorpora-300',
 'word2vec-google-news-300',
 'glove-wiki-gigaword-50',
 'glove-wiki-gigaword-100’,
…
 'glove-twitter-100',
 'glove-twitter-200']
>>> word2vec_vectors = gensim.downloader.load('word2vec-google-news-300')
```

Summary



v = [book, machine, artificial, NLP, code]

machine	0	1	0	0	0
artificial	0	0	1	0	0
code	0	0	0	0	1

Summary

- Distributional word vectors
  - Frequency based
  - Prediction based

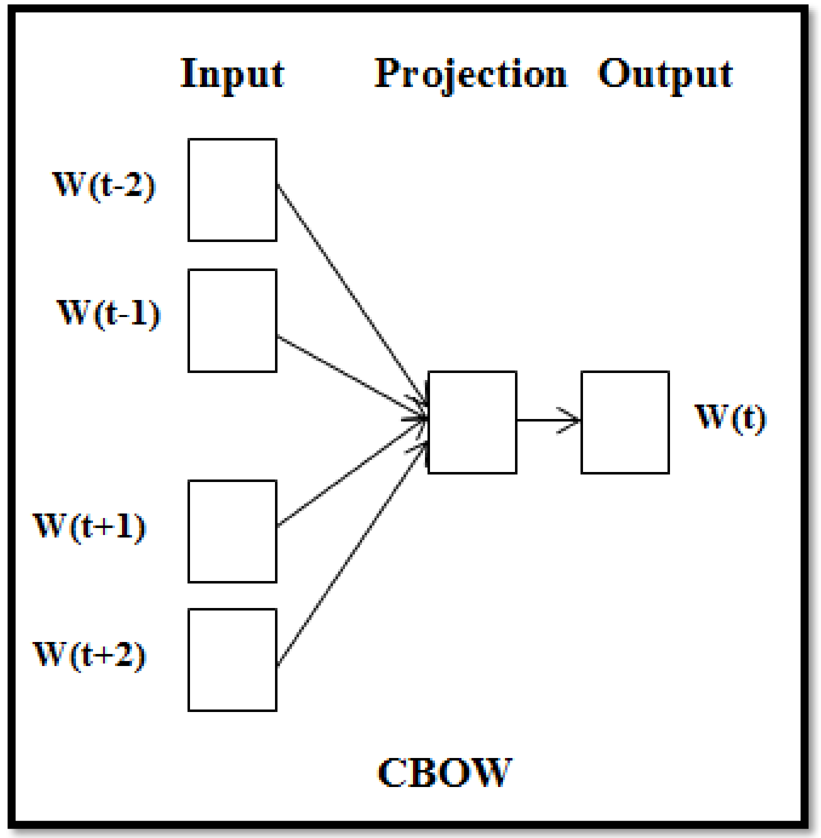
> **Memes** generally replicate through exposure to humans, who have evolved as efficient copiers of information and behavior.

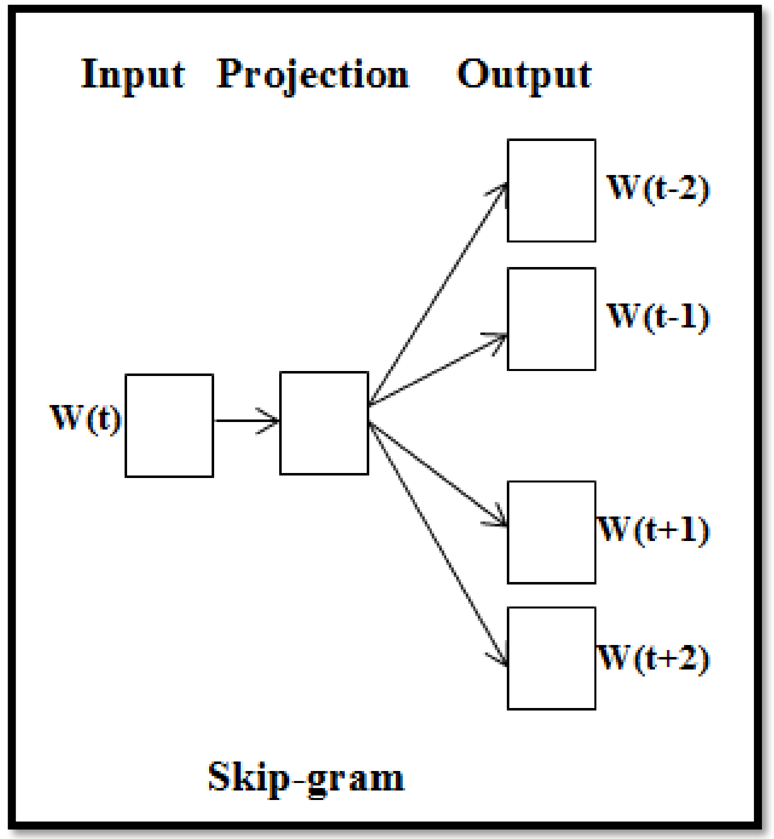
- Frequency based
  - Document-Term matrix
  - Term-Term matrix
  - PMI

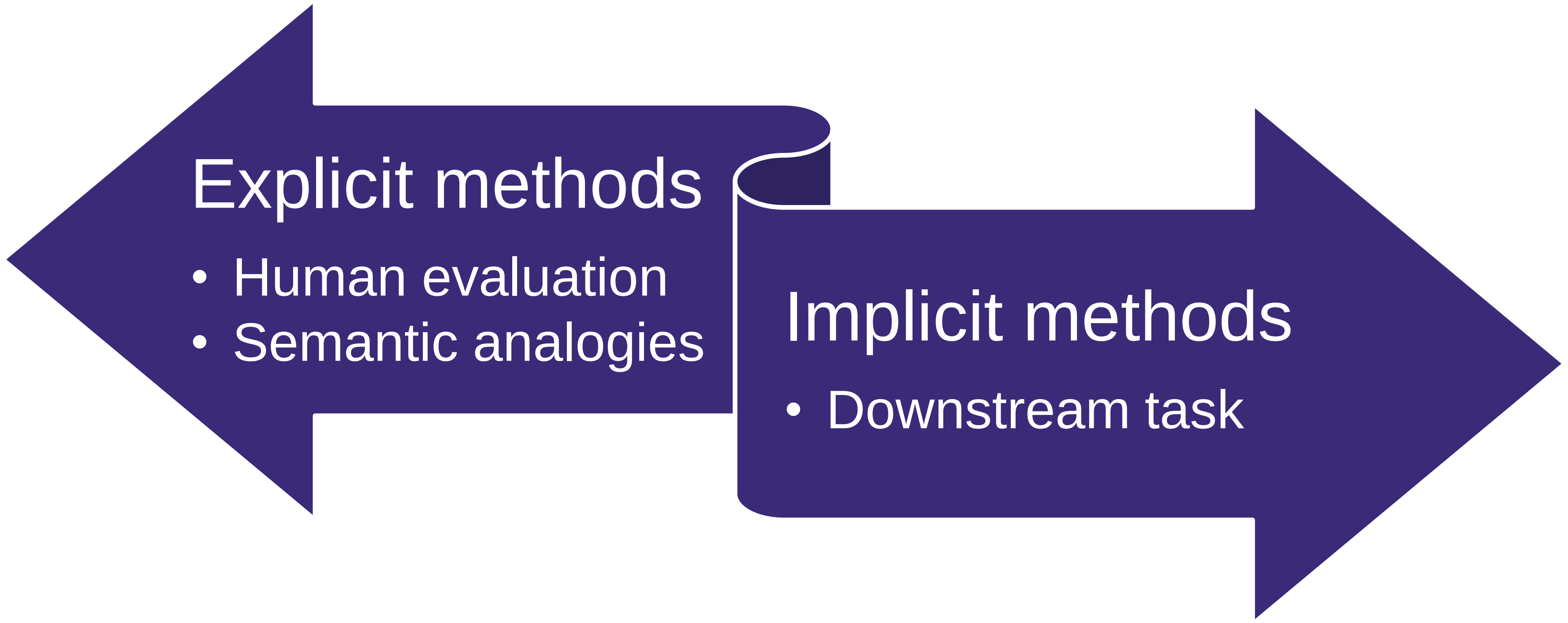
> Why is the water in the glass?
> Drinking a glass of milk is part of maintaining a healthy diet

Summary

- Prediction based (dense word embedding)
  - ***Word2vec***







GENSIM
topic modelling for humans

„KI-Campus – Die Lernplattform für Künstliche Intelligenz“ ist ein Projekt von

STIFTERVERBAND

HPI Hasso Plattner Institut
Digital Engineering • Universität Potsdam

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

NEOCOSMO

mmb Institut

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

www.ki-campus.org