Basic Processing

COMS4054A

COMS7066A

Natural Language Processing

What is a “word”?

What is a “word”?

“A matter of debate” ~ Yoav Goldberg

“Sequence of letters separated by whitespace or punctuation”

AN ATOMIC UNIT OF MEANING

Tokenization

Tokenization is the task of **segmenting** running text into **words**

In English it's Simple: Process of breaking up sentence into words/tokens splitting on white space and punctuation

The children are selling the clothes to each other.

[“The”, “children”, “are”, “selling”, “the”, “clothes”, “to”, “each”, “other”, “.”]

Tokenization

For most **African languages** tokenization is **NON-trivial** because many African languages are **AGGLUTINATIVE**

children

clothes

Abashana bayazithengiselana izimpahla

tense

buy

Aba-shana ba-ya-zi-theng-is-el-an-a izimpahla

subject marker

reciprocal

[“Aba”,”-shana”, “ba”, “-ya”, “-zi”,”-theng”,”-is”, “-el”, “-an”, “-a”, “izimpahla”]

Lemmatization & Stemming

“The goal of both stemming and lemmatization is to:
reduce inflectional forms and sometimes derivationally related forms of a word to a common base form”

democracy, *democratic*, and *democratization*

car, cars, car's, cars' => car

Stemming

Stemming = “crude heuristic process that chops off the ends of words in the hope of achieving this goal correctly most of the time, and often includes the removal of derivational affixes.”

Sample text: Such an analysis can reveal features that are not easily visible from the variations in the individual genes and can lead to a picture of expression that is more biologically transparent and accessible to interpretation

Lovins stemmer: such an analys can reve featur that ar not eas vis from th vari in th individu gen and can lead to a pictur of expres that is mor biolog transpar and acces to interpres

Porter stemmer: such an analysi can reveal featur that ar not easili visibl from the variat in the individu gene and can lead to a pictur of express that is more biolog transpar and access to interpret

Paice stemmer: such an analys can rev feat that are not easy vis from the vary in the individ gen and can lead to a pict of express that is mor biolog transp and access to interpret

https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html

Terminology

- Morphemes: Smallest meaningful unit of grammar. Can be “bound” or “unbound”.
- Morphology: the study of how morphemes are placed together.
- Syntax: the rules of how to combine words to form sentences
- Grammar: syntax and morphology – sentence and word rules.

Terminology

- Bound Morpheme: Morpheme which does not occur naturally on its own in language.
- Unbound Morpheme: Morpheme which can be found on its own in language.
- These are “morpheme types”. Others include “root”, “stem” and “affix”.

Terminology

- Affix: A bound morpheme that is combined with a root or stem

- Root: the portion of the word with all affixes removed – it carries the principle portion of meaning in a word (morphologically simple).

- Stem: Root + derivational affixes added (without inflectional affixes).

Terminology

- Derivation: Formation of new word or inflectable stem from another word or stem. Obtained by adding an affix.
- Derivation usually changes the word class (eg: noun, verb).
- Example: kind (root/word) + ness (bound affix) =
  kindness (stem/word)

Terminology

- Inflection: Formation of a new word or inflectable stem by adding an affix.
- Usually does not change the word class. Changes meaning in a predictable way and is invoked by grammar (obligatory).
- Example: come (root/word) + s (affix) = comes (word) (note how it is still a verb).

Terminology

Comparison:
Kinds of Affixes

Here is a table showing some kinds of affixes with examples:

Affix	Relationship to root or stem	Example
prefix	Occurs in the front of a root or stem	*un*happy
suffix	Occurs at the end of a root or stem	happi*ness*
infix	Occurs inside of a root or stem	b*um*ili 'buy' (Tagalog, Philippines)
circumfix	Occurs in two parts on both outer edges of a root or stem	*ka*baddang*an* 'help' (Tuwali Ifugao, Philippines)
simulfix	Replaces one or more phonemes in the root or stem	m*a*n + plural > m*e*n
suprafix	Superimposed on one or more syllables in the root or stem as a suprasegmental	stress in the words 'produce, *n.* and pro'duce, *v.*

- https://glossary.sil.org/term/affix-linguistics

Lemmatization

Lemmatization usually refers to doing things properly with the use of a vocabulary and **morphological analysis** of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma”

runs, running, ran => run

https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html

Corpus

corpus (plural corpora), **a computer-readable collection of text or speech.**

Helvetica

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

Vector Representations

COMS4054A
COMS7066A
Natural Language Processing

Word Vectors

Word Embeddings

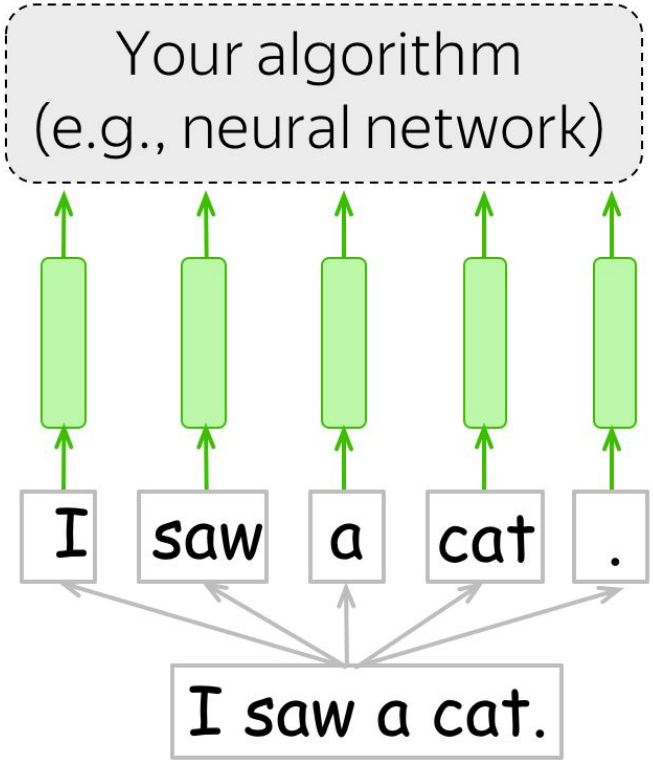
Semantic Embeddings

Semantic Representations

Thought Exercise

How would you think
about representing this?

“Snowboarding in Lesotho is magical”

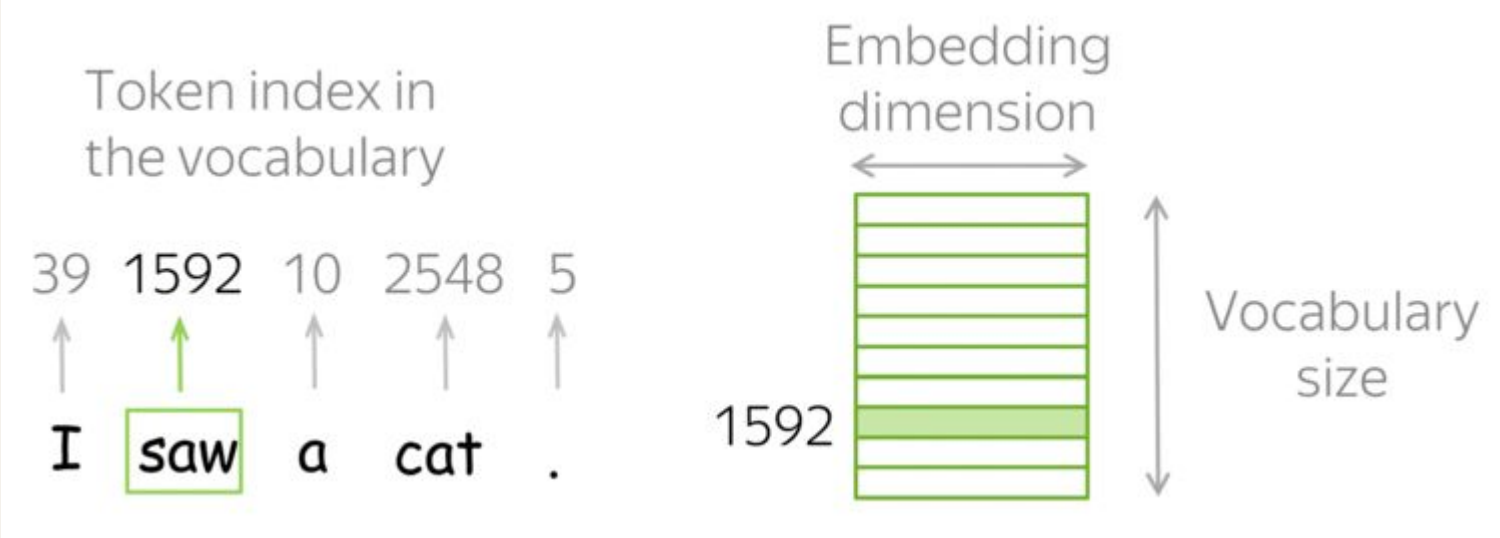


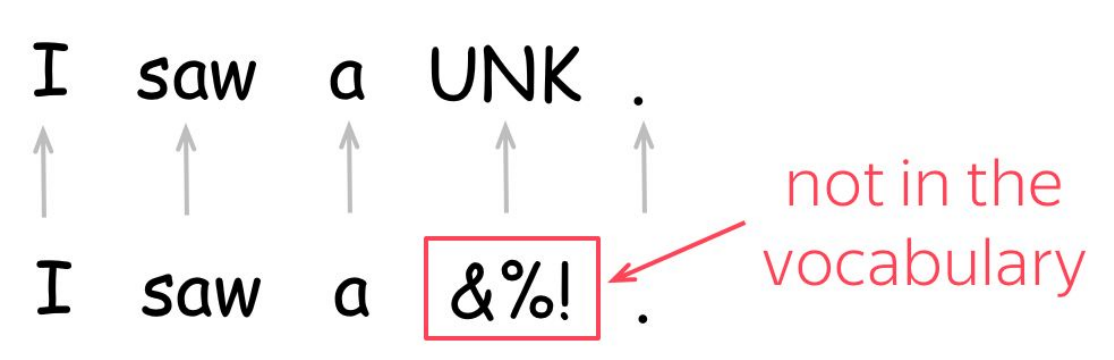
Any algorithm for solving a task

Word representation - vector
(input for your model/algorithm)

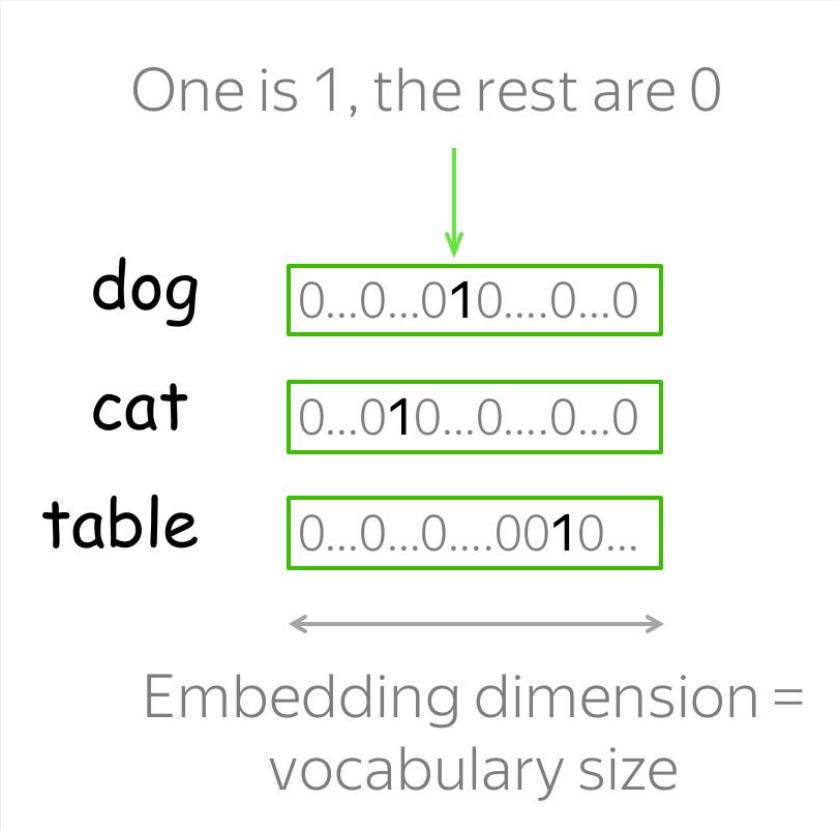
Sequence of tokens

Text (your input)

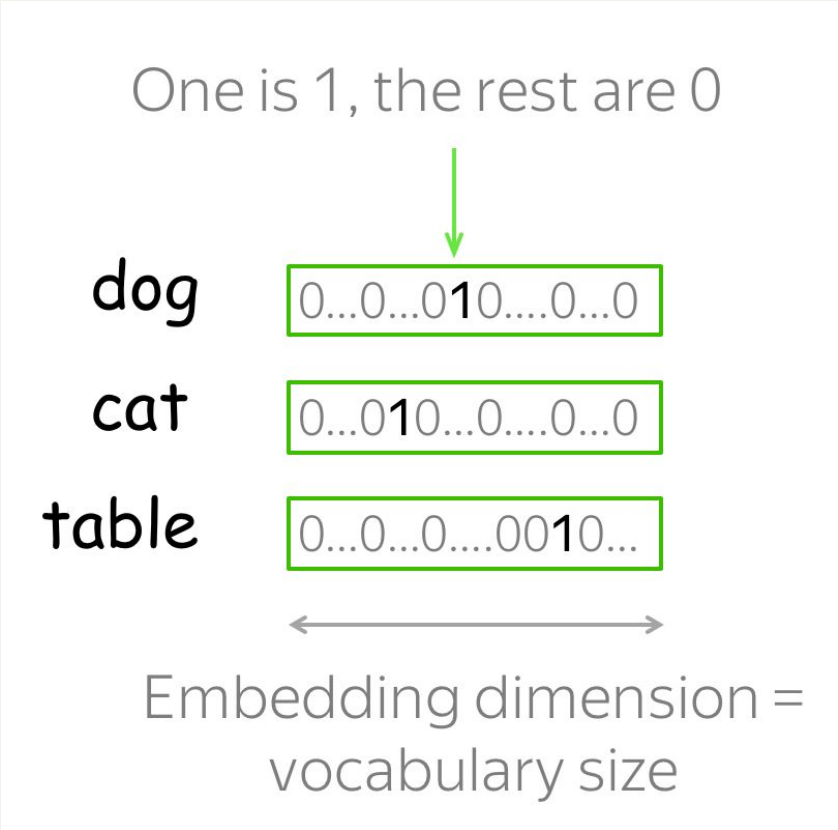

Token index in
the vocabulary
39 1592 10 2548 5
I saw a cat .
Embedding
dimension
1592
Vocabulary
size


I saw a UNK .
I saw a &%! .
not in the
vocabulary

Represent as Discrete Symbols: One-hot Vectors



Represent as Discrete Symbols: One-hot Vectors



We can say that <u>one-hot vectors do not capture meaning.</u>

What is MEANING?!

Let’s think about how WE determine meaning

How do WE know which words have similar meaning?

Do you know what the word “Tej” means?

Let’s look how this word is used in different contexts

A bottle of **Tej** is on the table

Everyone likes **Tej**

Tej makes you drunk

We make **Tej** out of honey

Can you understand what Tej means?

Let’s look how this word is used in different contexts

Tej is an alcoholic beverage made from honey!

A bottle of **Tej** is on the table

Everyone likes **Tej**

Tej makes you drunk

We make **Tej** out of honey

With context, you can understand the MEANING!

How did your brain do this?

(1) A bottle of ___ is on the table

(2) Everyone likes ___

(3) ___ makes you drunk

(4) We make __ out of honey

What other words fit into these contexts?

	(1)	(2)	(3)	(4)
Tej	1	1	1	1
loud	0	0	0	0
lip balm	0	0	0	1
wine	1	1	1	0
pancakes	0	1	0	1

<- contexts

<- rows show contextual properties. 1 if a word can appear in the context; 0 if not

How did your brain do this?

(1) A bottle of ___ is on the table

(2) Everyone likes ___

(3) ___ makes you drunk

(4) We make __ out of honey

rows are similar → **meanings** of the words are similar

Is this true?

	(1)	(2)	(3)	(4)
Tej	1	1	1	1
loud	0	0	0	0
lip balm	0	0	0	1
wine	1	1	1	0
pancakes	0	1	0	1

How did your brain do this?

(1) A bottle of ___ is on the table

(2) Everyone likes ___

(3) ___ makes you drunk

(4) We make __ out of honey

	(1)	(2)	(3)	(4)
Tej	1	1	1	1
loud	0	0	0	0
lip balm	0	0	0	1
wine	1	1	1	0
pancakes	0	1	0	1

rows are similar → **meanings** of the words are similar

THIS is the DISTRIBUTIONAL HYPOTHESIS

The Distributional Hypothesis

Words which frequently appear in **similar contexts** have **similar meaning**.

According to the hypothesis **"to capture meaning"** and **"to capture contexts"** are inherently the same.

Main idea: We need to put information about word contexts into word representation.

COUNT-BASED METHODS

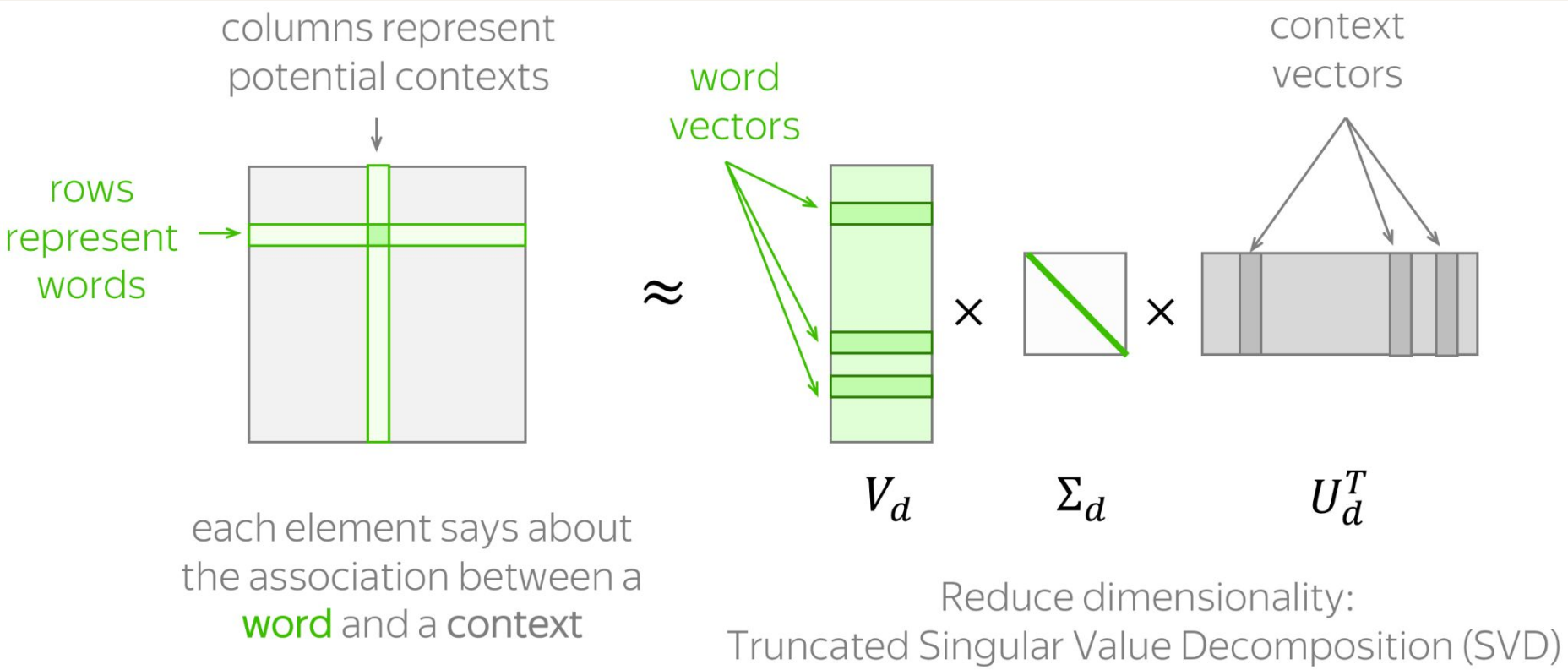
Main idea: We need to put information about word contexts into word representation.

How: Put this information **manually** based on global corpus statistics.

COUNT-BASED METHODS

(1) construct a word-context matrix

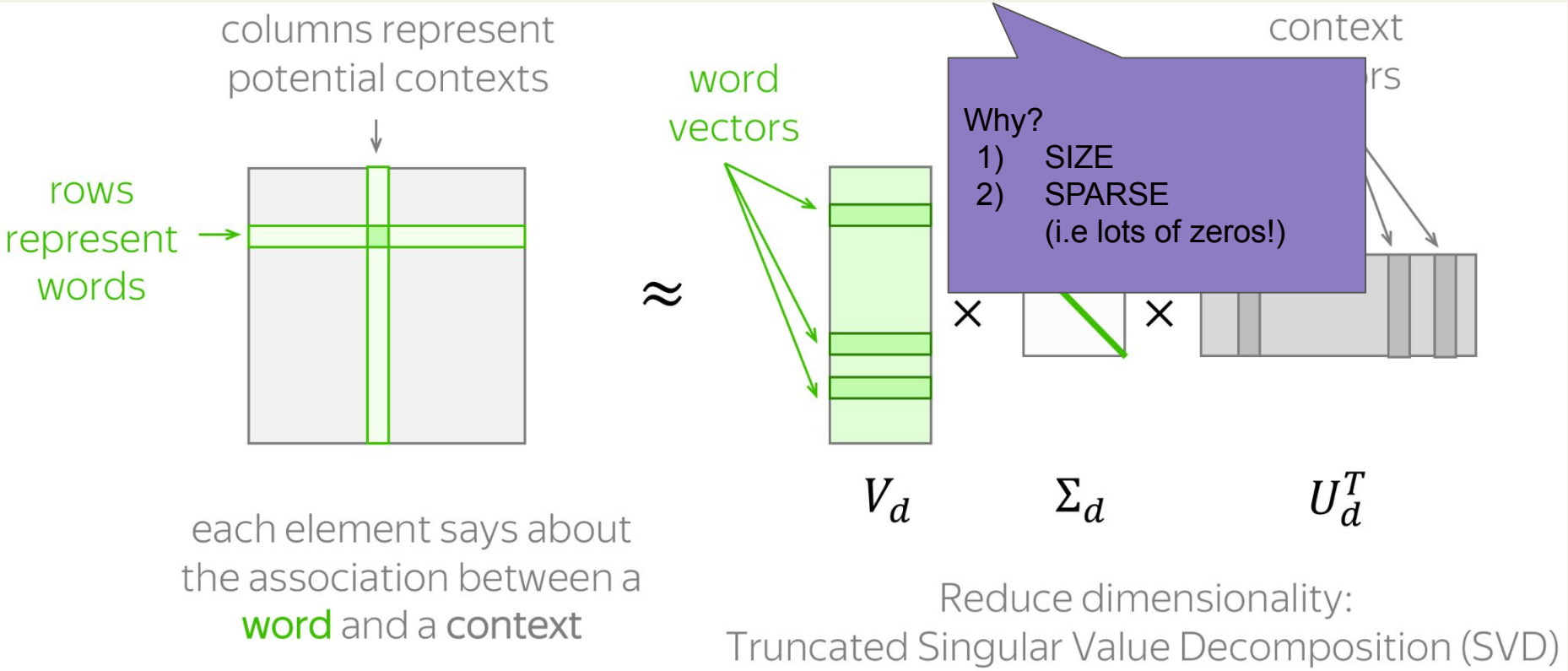
(2) reduce its dimensionality.



COUNT-BASED METHODS

(1) construct a word-context matrix

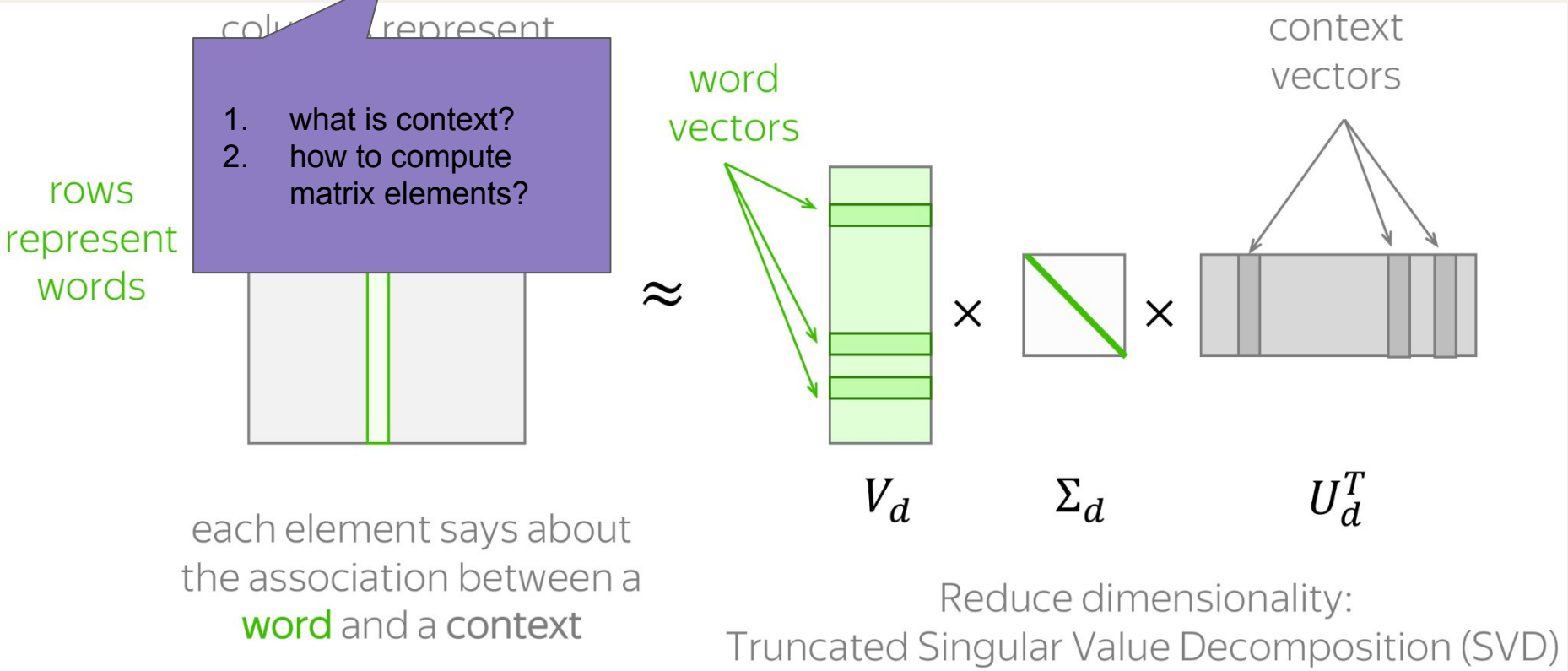
(2) reduce its dimensionality.



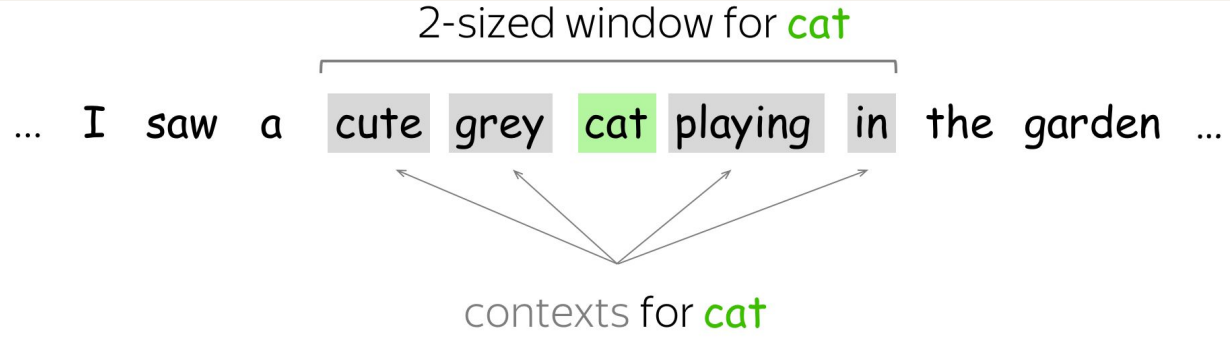
COUNT-BASED METHODS

(1) construct a word-context matrix

(2) reduce its dimensionality.



SIMPLE: CO-OCCURRENCE COUNTS



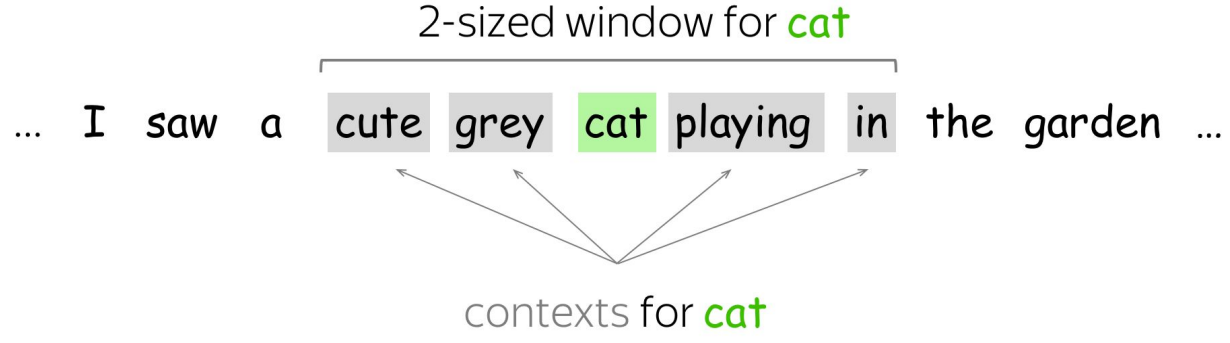
Context:

- surrounding words in a L-sized window

Matrix element:

- N(w, c) – number of times word w appears in context c

SIMPLE: CO-OCCURRENCE COUNTS



Context:

- surrounding words in a L-sized window

Matrix element:

- N(w, c) – number of times word w appears in context c

A Toy Example

Vocab: a, cat, cute, clever, grey, loving, I, saw

window size of 2

Corpus: I saw a cute grey cat. a loving, cute cat

	a	cat	cute	clever	grey	loving	I	saw
a								
cat								
cute								
clever								
grey								
loving								
I								
saw								

Vocab: a, cat, cute, clever, grey, loving, I, saw

window size of 2

Corpus: I saw a cute grey cat. a loving, cute cat

	a	cat	cute	clever	grey	loving	I	saw
a			1		1		1	1
cat								
cute								
clever								
grey								
loving								
I								
saw								

Vocab: a, cat, cute, clever, grey, loving, I, saw

window size of 2

Corpus: I saw a cute grey cat. a loving, cute cat

	a	cat	cute	clever	grey	loving	I	saw
a			1		1		1	1
cat								
cute	1	1			1			1
clever								
grey	2	1	1					
loving								
I								
saw								

Vocab: a, cat, cute, clever, grey, loving, I, saw

window size of 2

Corpus: I saw a cute grey cat. a loving, cute cat

	a	cat	cute	clever	grey	loving	I	saw
a		1	2		2	1	1	1
cat	1		1		1	1		
cute	2	2			1	1		1
clever								
grey	2	1	1					
loving	1	2	1					
I	1							1
saw	1		1				1	