

# BC COMS 2710: Computational Text Analysis

BARNARD COLLEGE OF COLUMBIA UNIVERSITY

## Lecture 20 Word Representation

# Announcements – Assignments



- Readings 05:
  - No more this semester – congrats!
- HW04
  - Due Thursday
- Course Evaluations
  - Due 06/14



# Word Representations



# Document-Term Matrix

DMT:

- Rows represent a document
- Columns represent a word
- Values represent some feature of word  $w_i$  in document  $d_j$

	$w_1$	$w_2$	$w_3$	$w_4$	...	...	...	...	$w_v$
$d_1$									
$d_1$									
...									
$d_n$									10

# Document-Term Matrix



We represent each word in our vocabulary as ...  
an index in our matrix

	$w_1$	$w_2$	$w_3$	$w_4$	...	...	...	...	$w_v$
$d_1$									
$d_2$									
...									
$d_n$									

# One Hot Vector



- Unique vector for each word
- $n-1$  elements in vector are 0
- One element in vector is 1

# One hot vector example



*a pioneer of computer science for work combining statistics and linguistics, and an advocate for women in the field*





# One hot vector example

*a pioneer of computer science for work combining statistics and linguistics, and an advocate for women in the field*

<b>a</b>	?	...	?	...	?	...	?
<b>pioneer</b>	?	...	?	...	?	...	?
<b>science</b>	?	...	?	...	?	...	?
...	?	...	?	...	?	...	?
<b>advocate</b>	?	...	?	...	?	...	?



# One hot vector example

*a pioneer of computer science for work combining statistics and linguistics, and an advocate for women in the field*

<b>a</b>	1	...	0	...	0	...	0
<b>pioneer</b>	?	...	?	...	?	...	?
<b>science</b>	?	...	?	...	?	...	?
...	?	...	?	...	?	...	?
<b>advocate</b>	?	...	?	...	?	...	?



# One hot vector example

*a pioneer of computer science for work combining statistics and linguistics, and an advocate for women in the field*

<b>a</b>	1	...	0	...	0	...	0
<b>pioneer</b>	0	...	1	...	0	...	0
<b>science</b>	?	...	?	...	?	...	?
...	?	...	?	...	?	...	?
<b>advocate</b>	?	...	?	...	?	...	?



# One hot vector example

*a pioneer of computer science for work combining statistics and linguistics, and an advocate for women in the field*

<b>a</b>	1	...	0	...	0	...	0
<b>pioneer</b>	0	...	1	...	0	...	0
<b>science</b>	0	...	0	...	1	...	0
<b>...</b>	?	...	?	...	?	...	?
<b>advocate</b>	?	...	?	...	?	...	?



# One hot vector example

*a pioneer of computer science for work combining statistics and linguistics, and an advocate for women in the field*

<b>a</b>	1	...	0	...	0	...	0
<b>pioneer</b>	0	...	1	...	0	...	0
<b>science</b>	0	...	0	...	1	...	0
<b>...</b>	0	...	0	...	0	...	1
<b>advocate</b>	0	...	0	...	0	...	1



# One hot vector example

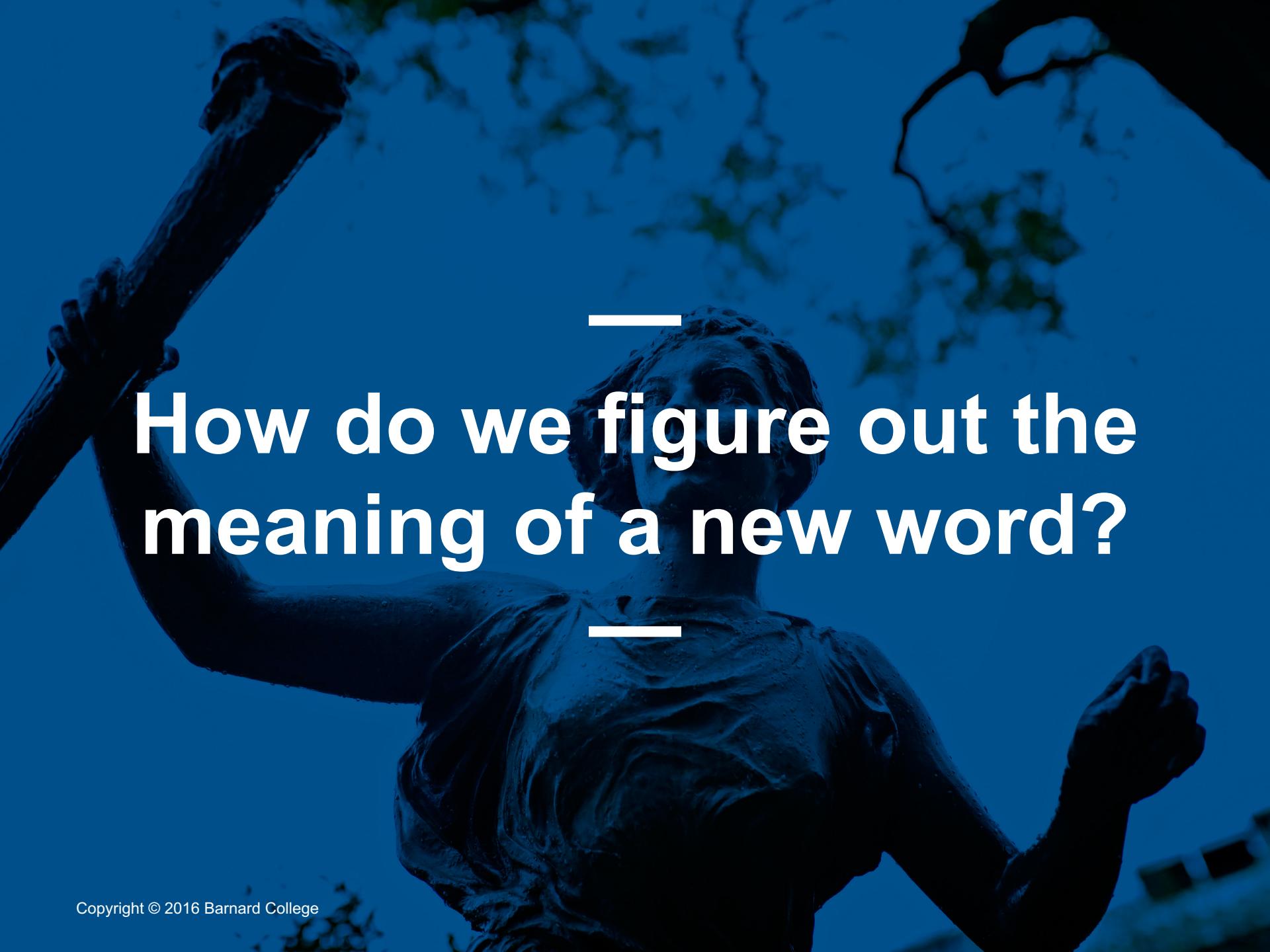
*a pioneer of computer science for work combining statistics and linguistics, and an advocate for women in the field*

	a	...	pioneer	...	science	...	advocate
a	1	...	0	...	0	...	0
pioneer	0	...	1	...	0	...	0
science	0	...	0	...	1	...	0
...	0	...	0	...	0	...	1
advocate	0	...	0	...	0	...	1



# Issues with one-hot vector

- Sparse
  - Lots of 0's
- Very big
  - As big as vocabulary
- Doesn't capture any meaning of the word
  - DTM actually captures some aspects of the documents' meaning
  - We'd like the same for our word representations



—

# How do we figure out the meaning of a new word?

—

# Meaning from Context: Tezguino



A bottle of *tezgüino* is on the table.  
Everyone likes *tezgüino*.  
*Tezgüino* makes you drunk.  
We make *tezgüino* out of corn.

Lin, ACL 1998; Nida, 1975 p.167

# Meaning from Context: Tezguino



A bottle of *tezgüino* is on the table.  
Everyone likes *tezgüino*.  
*Tezgüino* makes you drunk.  
We make *tezgüino* out of corn.



Lin, ACL 1998; Nida, 1975 p.167

# Distributional Hypothesis



*words with similar contexts  
share similar meanings*

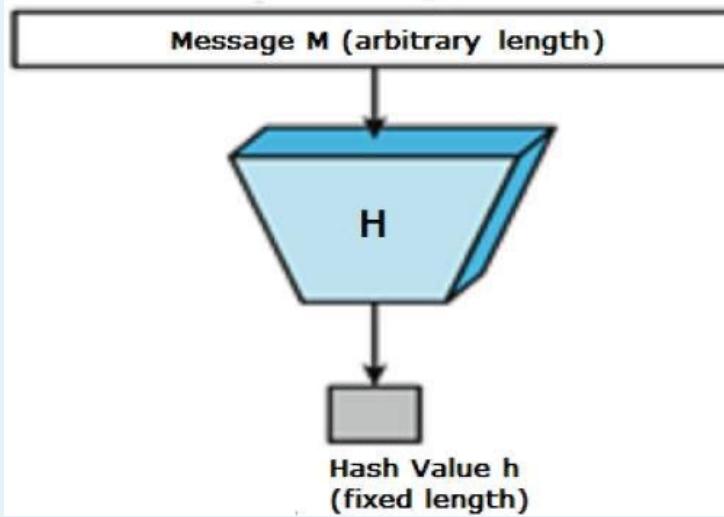
(Harris, 1954)

*you shall know a word by  
the company it keeps*

(Firth 1957)



# Meaning from Context: *Hash*



# Meaning from Context: *Hash*



*about to get my hands on some top shelf **hash** but I have no idea what the **hash** price is in my area. There is no one that sells **hash** in my area actually.*



# Co-occurrence matrix



*about to get my hands on some top shelf hash but  
I have no idea what the hash price is in my area.  
There is no one that sells hash in my area actually.*

	on	hands	hash	price	actually	area	my
on							
hands							
hash							
price							
actually							
area							
my							

v x v matrix

# Co-occurrence matrix



*about to get my hands on some top shelf hash but  
I have no idea what the hash price is in my area.  
There is no one that sells hash in my area actually.*

	on	hands	hash	price	actually	area	my
on							
hands							
hash							
price							
actually							
area							
my						????	

Window  
size of 2

# Co-occurrence matrix



*about to get my hands on some top shelf hash but  
I have no idea what the hash price is in my area.  
There is no one that sells hash in my area actually.*

	on	hands	hash	price	actually	area	my
on							
hands							
hash							
price							
actually							
area							
my						????	

Window size of 2

# Co-occurrence matrix



*about to get my hands on some top shelf hash but  
I have no idea what the hash price is in my area.  
There is no one that sells hash in my area actually.*

	on	hands	hash	price	actually	area	my
on							
hands							
hash							
price							
actually							
area							
my						????	

Window size of 2

# Co-occurrence matrix



*about to get my hands on some top shelf hash but  
I have no idea what the hash price is in my area.  
There is no one that sells hash in my area actually.*

	on	hands	hash	price	actually	area	my
on							
hands							
hash							
price							
actually							
area							
my							2

Window size of 2

# Co-occurrence matrix



*about to get my hands on some top shelf hash but  
I have no idea what the hash price is in my area.  
There is no one that sells hash in my area actually.*

	on	hands	hash	price	actually	area	my
on							
hands							
hash							
price							
actually							
area							2
my						2	

Window size of 2

# Co-occurrence matrix



*about to get my hands on some top shelf hash but  
I have no idea what the hash price is in my area.  
There is no one that sells hash in my area actually.*

	on	hands	hash	price	actually	area	my
on							
hands							
hash							
price			????				
actually							
area							2
my						2	

Window size of 2

# Co-occurrence matrix



*about to get my hands on some top shelf hash but  
I have no idea what the hash price is in my area.  
There is no one that sells hash in my area actually.*

	on	hands	hash	price	actually	area	my
on							
hands							
hash							
price			2				
actually							
area							2
my						2	

Window size of 2



# Co-occurrence matrix

*about to get my hands on some top shelf hash but  
I have no idea what the hash price is in my area.  
There is no one that sells hash in my area actually.*

	on	hands	hash	price	actually	area	my
on	0	1	0	0	0	0	0
hands	1	0	0	0	0	0	1
hash	0	0	0	1	0	0	1
price	0	0	1	0	0	0	0
actually	0	0	0	0	0	1	0
area	0	0	0	0	1	0	2
my	0	1	1	0	1	2	0

Window  
size of 2



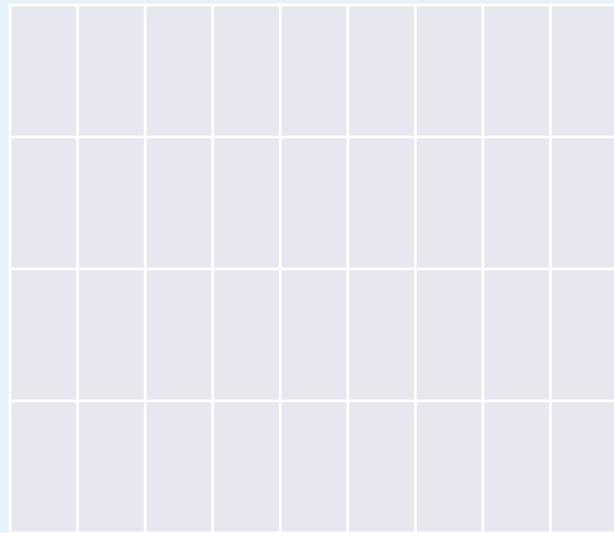
# Issues with co-occurrence matrix

- Large dimensions
- Still sparse
  - Not as much as one-hot but still sparse
- Is meaning captured?
- Solution:
  - Dimensionality Reduction to the rescue

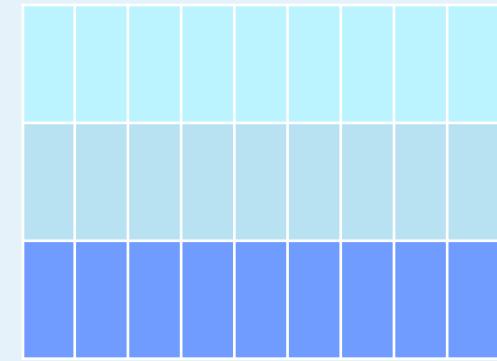


# Singular Value Decomposition

## Document Term Matrix



$s_1$	0	0
0	$s_2$	0
0	0	$s_2$



$$M = U \quad S \quad V$$

$n \times v$        $n \times k$        $k \times k$        $k \times v$



# Singular Value Decomposition

## Co-occurrence matrix

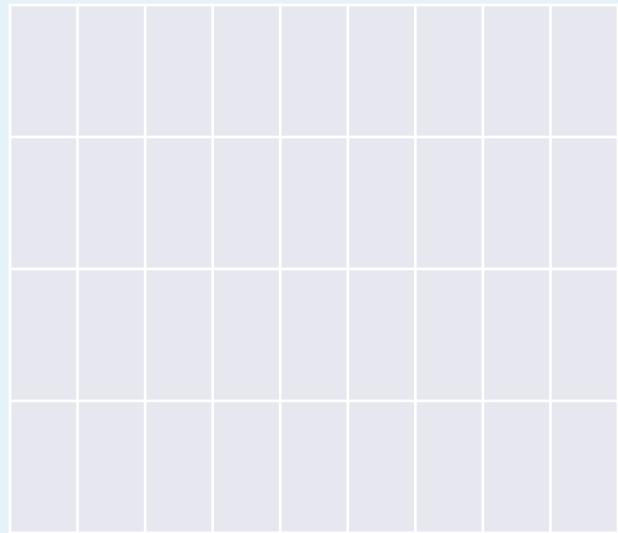

M

V X V



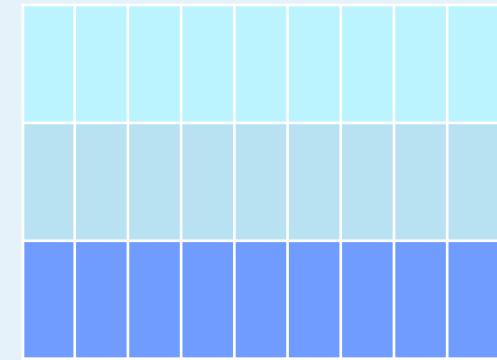
# Singular Value Decomposition

## Co-occurrence matrix



$$=: \boxed{U} \quad v \times k$$

$s_1$	0	0
0	$s_2$	0
0	0	$s_2$



$$\begin{matrix} M \\ v \times v \end{matrix} = \begin{matrix} U \\ v \times k \end{matrix} \begin{matrix} S \\ k \times k \end{matrix} \begin{matrix} V \\ k \times v \end{matrix}$$

# Word Embeddings



# Initialize random vectors



## Embedding



This is a look-up table where each row indicates the list of numbers for a word

# Update word embeddings by reading a corpus



## Embedding





# Example

↑  
0  
↓

Posted by u/SaltyPositive 1 year ago

## Ziip Disposable Device

Where are all the ziip device posts at?! I recently bought the ziip refilled disposable device and I'm so so unsure on what to make of it, because there is NO hit, but the cloud is dense upon exhaling, but I don't feel a rush and I'm not sure how hard you have to pull(????) it really doesn't feel like I'm pulling at anything at all. I'm posting here because I bought this pod for 7 cad as a substitute for the Juul ones but don't know if I just got a faulty device? Any other similar experiences?

2 Comments Share Save Hide Report

50% Upvoted



↑ Posted by u/SaltyPositive 1 year ago 

0  
↓

## Ziip Disposable Device

Where are all the ziip device posts at?! I recently bought the [ ? ] refilled disposable device and I'm so so unsure on what to make of it, because there is NO hit, but the cloud is dense upon exhaling, but I don't feel a rush and I'm not sure how hard you have to pull(????) it really doesn't feel like I'm pulling at anything at all. I'm posting here because I bought this pod for 7 cad as a substitute for the Juul ones but don't know if I just got a faulty device? Any other similar experiences?

 2 Comments

 Share

 Save

 Hide

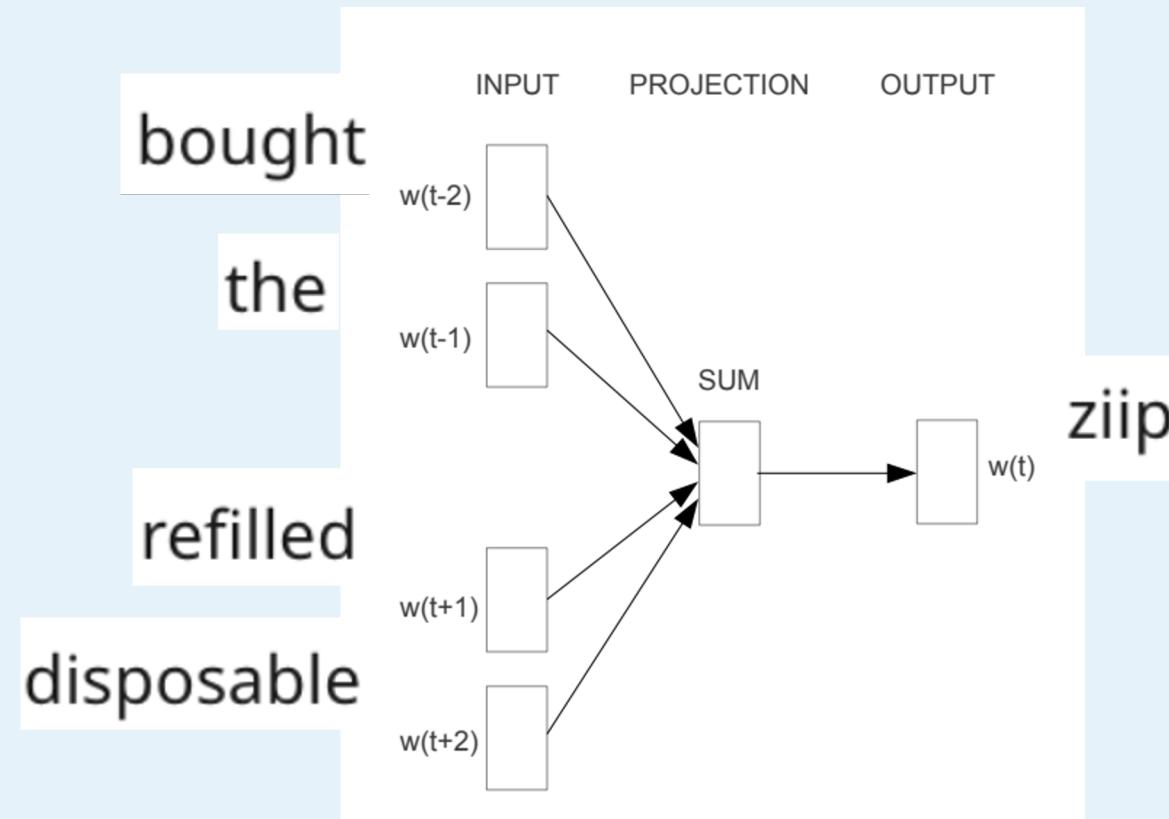
 Report

50% Upvoted

# Continuous Bag of Words (CBOW) (Mikolov et al. 2013)



- Predict a word given its context





↑ Posted by u/SaltyPositive 1 year ago 

0  
↓

## Ziip Disposable Device

Where are all the ziip device posts at?! I recently bought the ziip refilled disposable device and I'm so so unsure on what to make of it, because there is NO hit, but the cloud is dense upon exhaling, but I don't feel a rush and I'm not sure how hard you have to pull(????) it really doesn't feel like I'm pulling at anything at all. I'm posting here because I bought this pod for 7 cad as a substitute for the Juul ones but don't know if I just got a faulty device? Any other similar experiences?

 2 Comments

 Share

 Save

 Hide

 Report

50% Upvoted



Posted by u/SaltyPositive 1 year ago 

0  
↑  
↓

## Ziip Disposable Device

Where are all the ziip device posts at?! I recently    device and I'm so so unsure on what to make of it, because there is NO hit, but the cloud is dense upon exhaling, but I don't feel a rush and I'm not sure how hard you have to pull(????) it really doesn't feel like I'm pulling at anything at all. I'm posting here because I bought this pod for 7 cad as a substitute for the Juul ones but don't know if I just got a faulty device? Any other similar experiences?

 2 Comments

 Share

 Save

 Hide

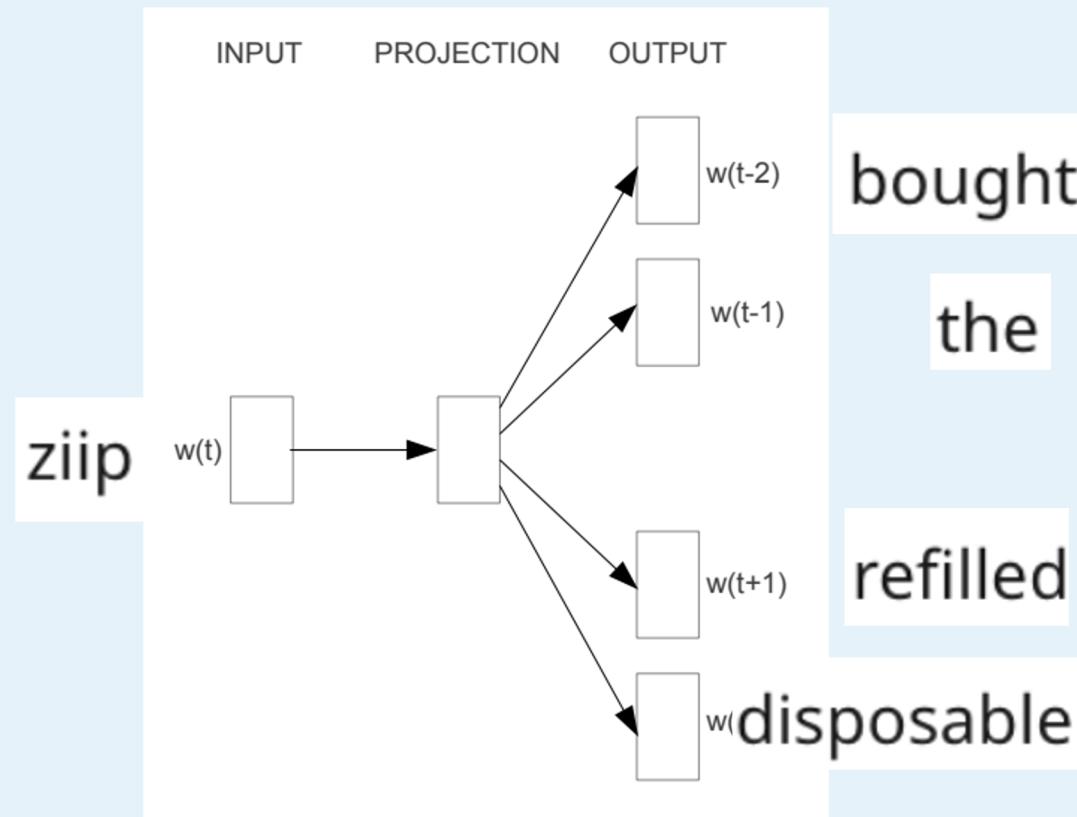
 Report

50% Upvoted

# Skip-Gram



- Predict the context around a word



# Updated Word Embeddings as byproduct of training



## Embedding



After training the neural network, we have updated values in our look-up table

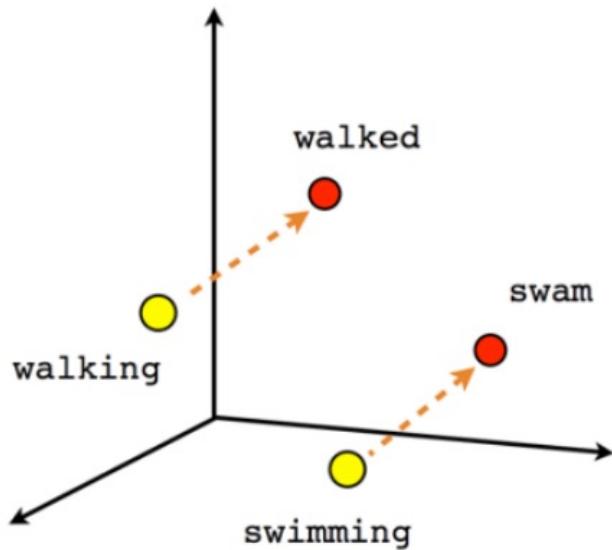


# Word Embeddings

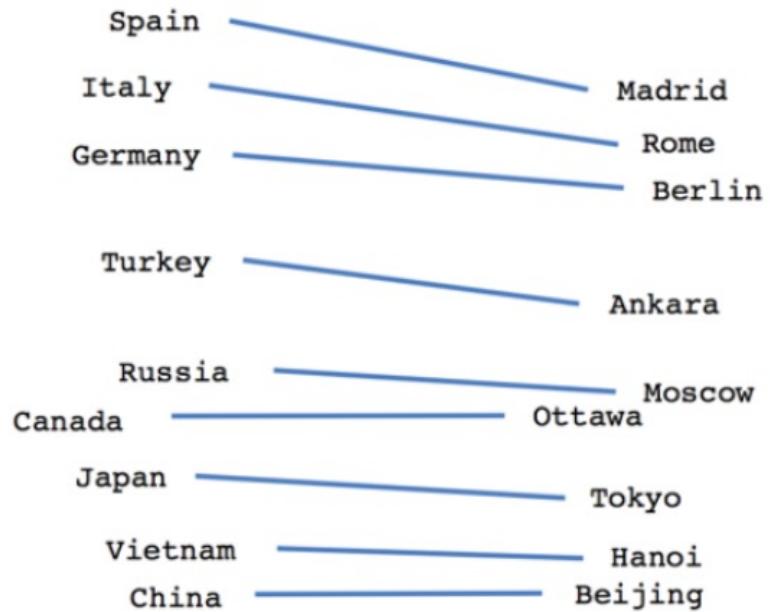
<b>a</b>	0.4420	...	0.167	...	0.4838	...	0.2314
<b>pioneer</b>	0.2401	...	0.3732	...	0.9653	...	0.6366
<b>science</b>	0.7532	...	0.3245	...	0.5893	...	0.7772
...	0.2032	...	0.5792	...	0.9302	...	0.4924
<b>advocate</b>	0.3424	...	0.2944	...	0.3923	...	0.3492



# Word Embeddings Preserve Meaning



Verb tense



Country-Capital



- Wednesday 06/09 – Guest Lecture
  - Attendance required
- Thursday 06/10 – ngrams & phrases
- Monday 06/14 – Project Presentations