Recurrent Neural Networks III

Dr. Parlett-Pelleriti

Text Processing

Standardization

The cat in the hat sat at the table and ate a bat.

→

the cat in the hat sat at the table and ate a bat

Tokens

The cat in the hat sat at the table and ate a bat.

the cat in the hat sat at the table and ate a bat

['the', 'cat', 'in', 'the', 'hat', 'sat', 'at', 'the', 'table', 'and', 'ate', 'a', 'bat']

Tokens

the	1
cat	0
in	0
hat	0
sat	0
at	0
table	0
and	0
ate	0
a	0
bat	0

the cat in the hat sat at the table and ate a bat

['the', 'cat', 'in', 'the', 'hat', 'sat', 'at', 'the', 'table', 'and', 'ate', 'a', 'bat']

Tokens

the	0
cat	0
in	0
hat	0
sat	1
at	0
table	0
and	0
ate	0
a	0
bat	0

the cat in the hat sat at the table and ate a bat

['the', 'cat', 'in', 'the', 'hat', 'sat', 'at', 'the', 'table', 'and', 'ate', 'a', 'bat']

Tokens

the	0
cat	0
in	0
hat	0
sat	0
at	0
table	0
and	0
ate	0
a	0
bat	1

the cat in the hat sat at the table and ate a bat

['the', 'cat', 'in', 'the', 'hat', 'sat', 'at', 'the', 'table', 'and', 'ate', 'a', 'bat']

Out of Vocab Token

the cat in the hat sat at the table and ate a **platypus**

Out of Vocab Token

the cat in the hat sat at the
table and ate a **platypus**

Tokens

the	0
cat	0
in	0
hat	0
sat	0
at	0
table	0
and	0
ate	0
a	0
bat	0

the cat in the hat sat at the table and ate a **platypus**

['the', 'cat', 'in', 'the', 'hat', 'sat', 'at', 'the', 'table', 'and', 'ate', 'a', 'bat']

Tokens

the	0
cat	0
in	0
hat	0
sat	0
at	0
table	0
and	0
ate	0
a	0
bat	0
[UNK]	1

the cat in the hat sat at the table and ate a **platypus**

['the', 'cat', 'in', 'the', 'hat', 'sat', 'at', 'the', 'table', 'and', 'ate', 'a', 'bat']

Tokens

The cat in the hat sat at the table and ate a bat.

the cat in the hat sat at the table and ate a bat

['the', 'cat', 'in', 'the', 'hat', 'sat', 'at', 'the', 'table', 'and', 'ate', 'a', 'bat']

```
[1,0,0,0,0,0,0,0,0,0,0]
[0,1,0,0,0,0,0,0,0,0,0]
[0,0,1,0,0,0,0,0,0,0,0]
…
[0,0,0,0,0,0,0,0,0,0,1]
```

Miscellaneous Text Processing

Stems

I was sitting on the bench and I thought that it was a nice place to sit and think about life.

I was **[sit]** on the bench and I **[think]** that it was a nice place to **[sit]** and **[think]** about life.

Bag-Of-Words

The cat in the hat sat at the table and ate a bat

Set vs. Sequence

The cat in the hat sat at the table and ate a bat

Sequence

N-Grams

I was sitting on the bench and I thought that it was a nice place to sit and think about life.

N-Grams

I $\boxed{\text{was sitting}}$ on the bench and I thought that it was a nice place to sit and think about life.

N-Grams

I was sitting on the bench and I thought that it was a nice place to sit and think about life.

N-Grams

I was sitting on the bench and I thought that it was a nice place to sit and think about life.

N-Grams

I was sitting on the bench and I thought that it was a nice place to sit and think about life.

N-Grams

I was sitting on the bench and I thought that it was a nice place to sit and think about life.

N-Grams

I was sitting on the bench and I thought that it was a nice place to sit and think about life.

N-Grams

I was sitting on the bench and I thought that it was a nice place to sit and think about life.

N-Grams

I was sitting on the bench and I thought that it was a nice place to sit and think about life.

```
['i', 'i was', 'was', 'was sitting', 'sitting', 'sitting on', 'on', 'on the',
'the', 'the bench', 'bench', 'bench and', 'and', 'and i', 'i', 'i thought',
'thought', 'thought that', 'that', 'that it', 'it', 'it was', 'was', 'was a',
'a', 'a nice', 'nice', 'nice place', 'place', 'place to', 'to', 'to sit',
'sit', 'sit and', 'and', 'and think', 'think', 'think about', 'about', 'about
life']
```

TF-IDF

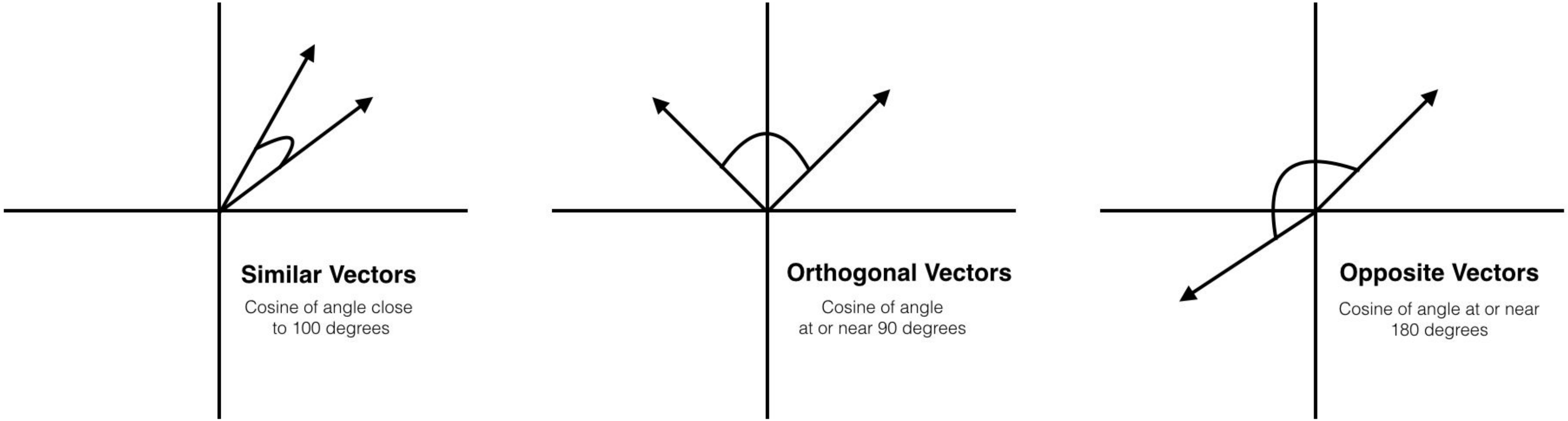
term **f**requency, **i**nverse **d**ocument **f**requency

$$tf = \frac{\text{frequency of word}}{\text{\# of words in document}}$$

$$idf = \frac{\text{\# of Documents}}{\text{\# of Documents that contain word}}$$

$$tfidf = tf * idf$$

Cosine Similarity of Word Counts



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

Image from: https://deepai.org/machine-learning-glossary-and-terms/cosine-similarity

Word Embeddings

Embeddings

$$\begin{bmatrix} 0.1 \\ 0.2 \\ 0.4 \\ 0.2 \\ 0.6 \end{bmatrix}$$

gorgeous

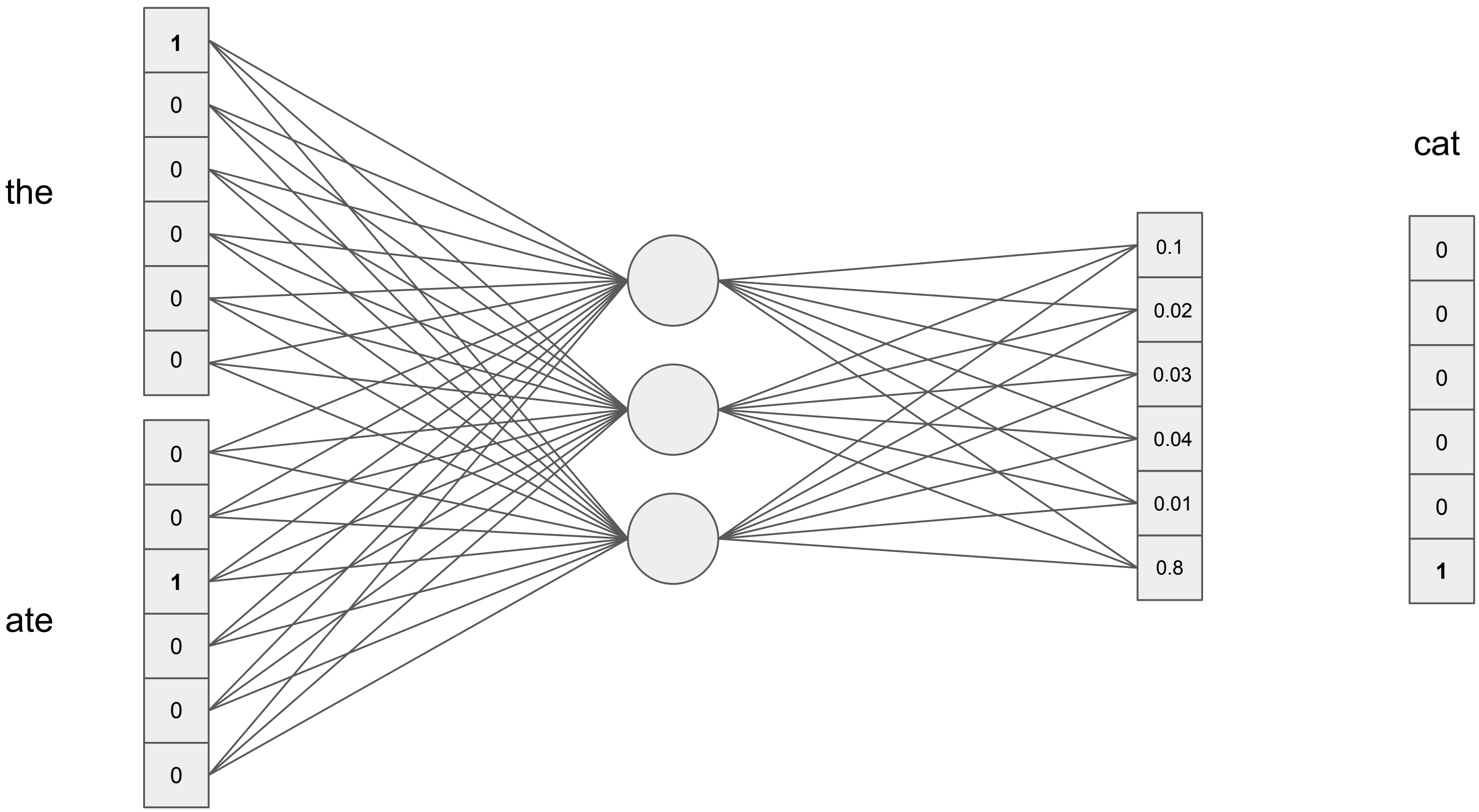
$$\begin{bmatrix} 0.2 \\ 0.4 \\ 0.5 \\ 0.9 \\ 0.3 \end{bmatrix}$$

python

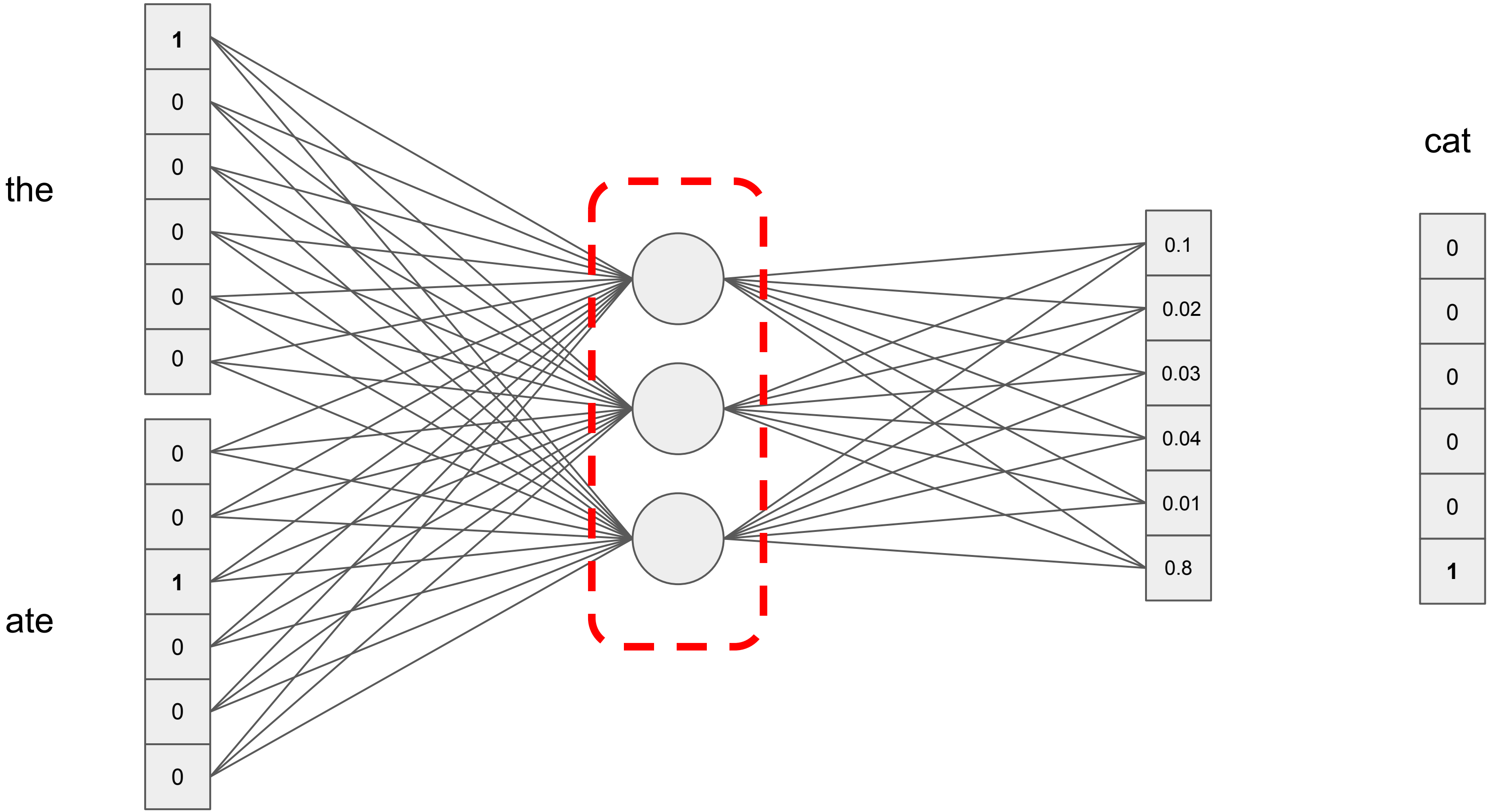
Word2Vec NN

- **Problem**: Fill in the Blank
- Use context to learn words' meanings
- **Side Effect**: Word Embeddings!

The **cat** ate

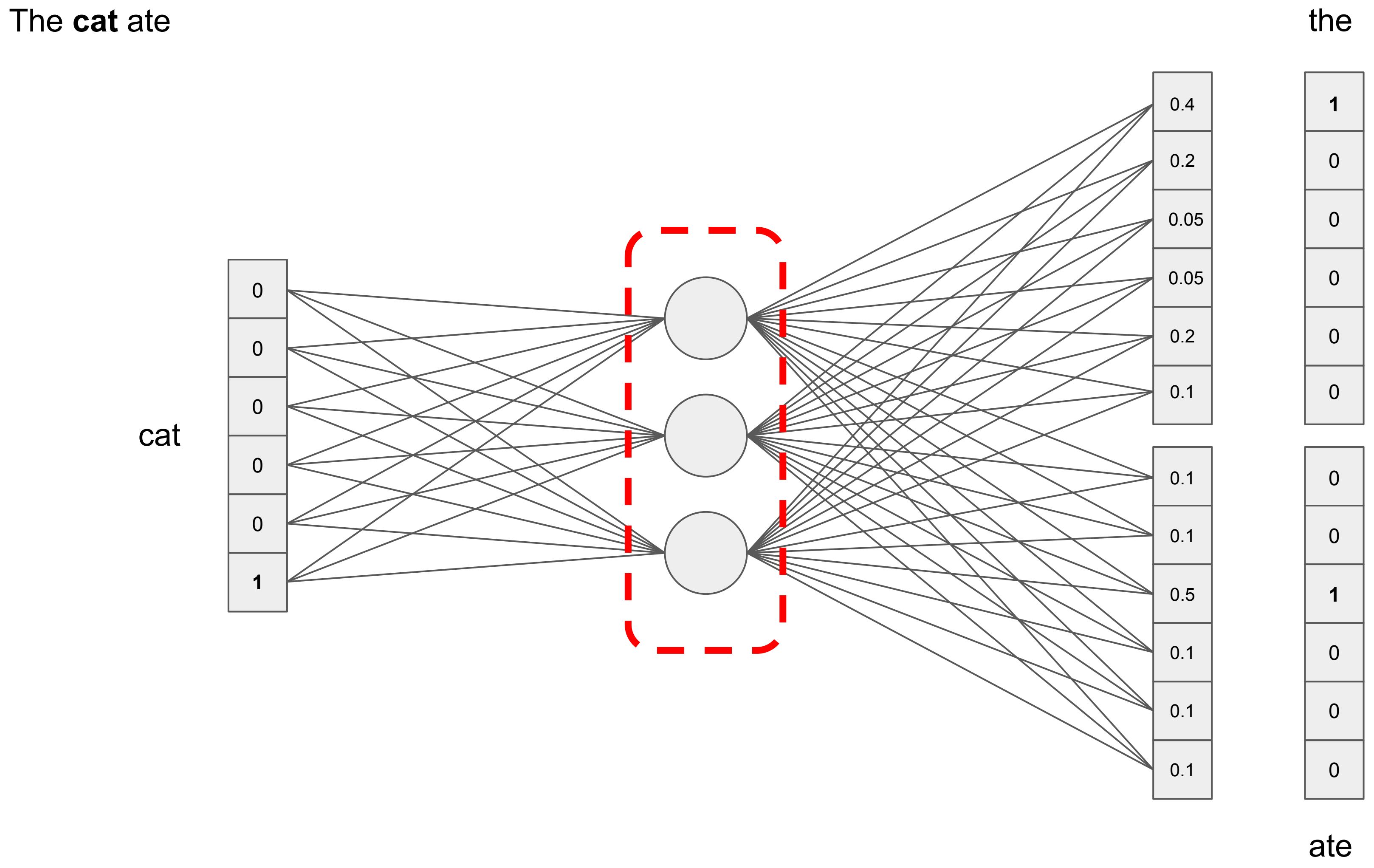


The **cat** ate



CBOW vs. Skip Gram

- **Continuous Bag of Words**: Predict target word from context words
- **Skip Gram**: Predict context words from target word



The cat ate
the
cat
0
0
0
0
0
1
0.4
0.2
0.05
0.05
0.2
0.1
0.1
0.1
0.5
0.1
0.1
0.1
1
0
0
0
0
0
0
0
1
0
0
0
ate

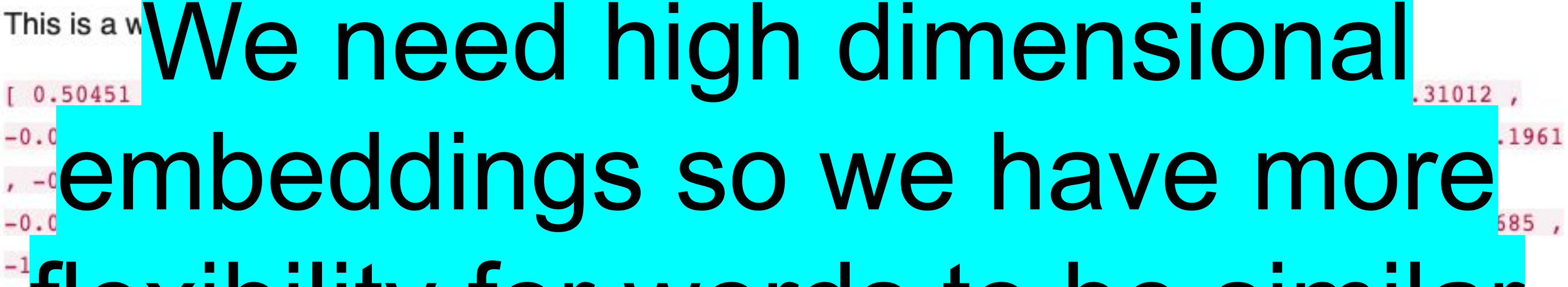
Word Embedding Example

This is a word embedding for the word "king" (GloVe vector trained on Wikipedia):

```
[ 0.50451 , 0.68607 , -0.59517 , -0.022801, 0.60046 , -0.13498 , -0.08813 , 0.47377 , -0.61798 , -0.31012 ,
-0.076666, 1.493 , -0.034189, -0.98173 , 0.68229 , 0.81722 , -0.51874 , -0.31503 , -0.55809 , 0.66421 , 0.1961
, -0.13495 , -0.11476 , -0.30344 , 0.41177 , -2.223 , -1.0756 , -1.0783 , -0.34354 , 0.33505 , 1.9927 ,
-0.04234 , -0.64319 , 0.71125 , 0.49159 , 0.16754 , 0.34344 , -0.25663 , -0.8523 , 0.1661 , 0.40102 , 1.1685 ,
-1.0137 , -0.21585 , -0.15155 , 0.78321 , -0.91241 , -1.6106 , -0.64426 , -0.51042 ]
```

Image from: https://jalammar.github.io/illustrated-word2vec/

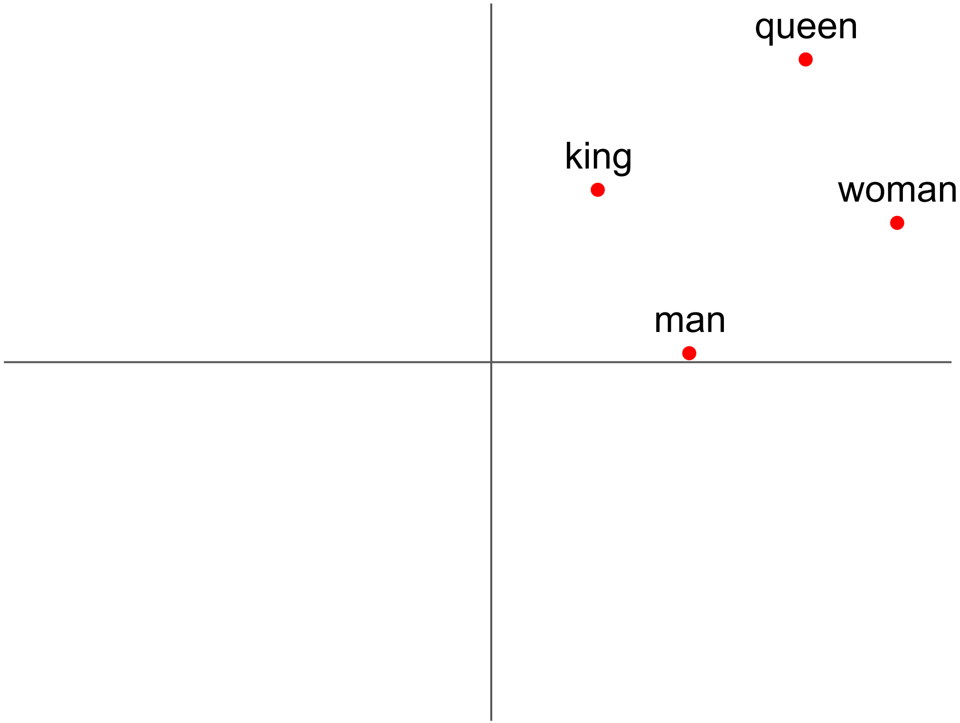
Word Embedding Example



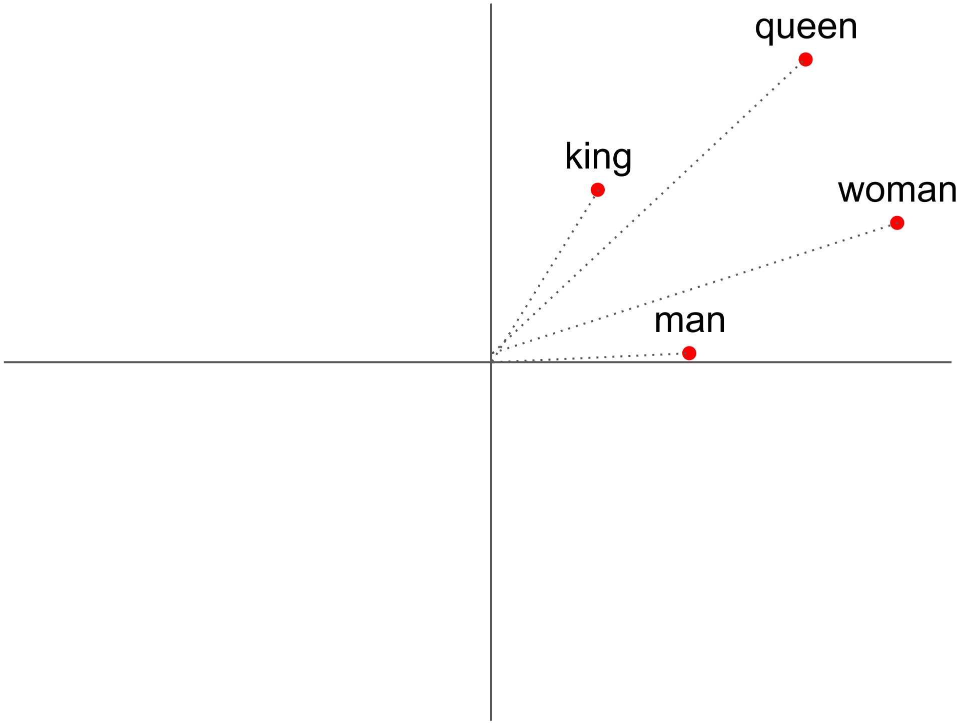
We need high dimensional embeddings so we have more flexibility for words to be similar in different dimensions

Image from: https://jalammar.github.io/illustrated-word2vec/

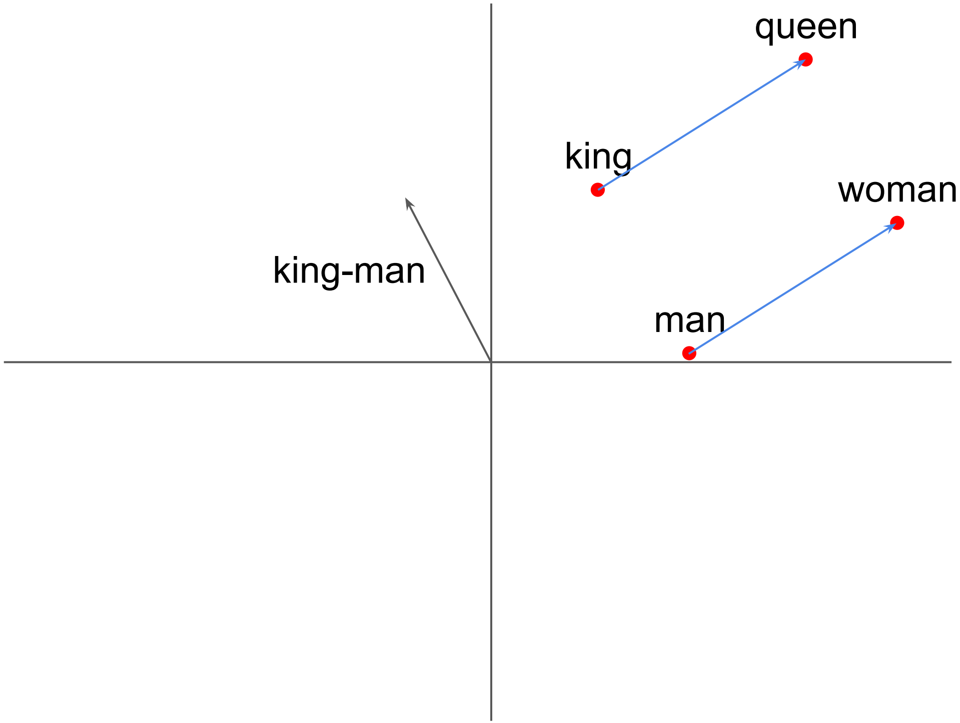
Vector Algebra



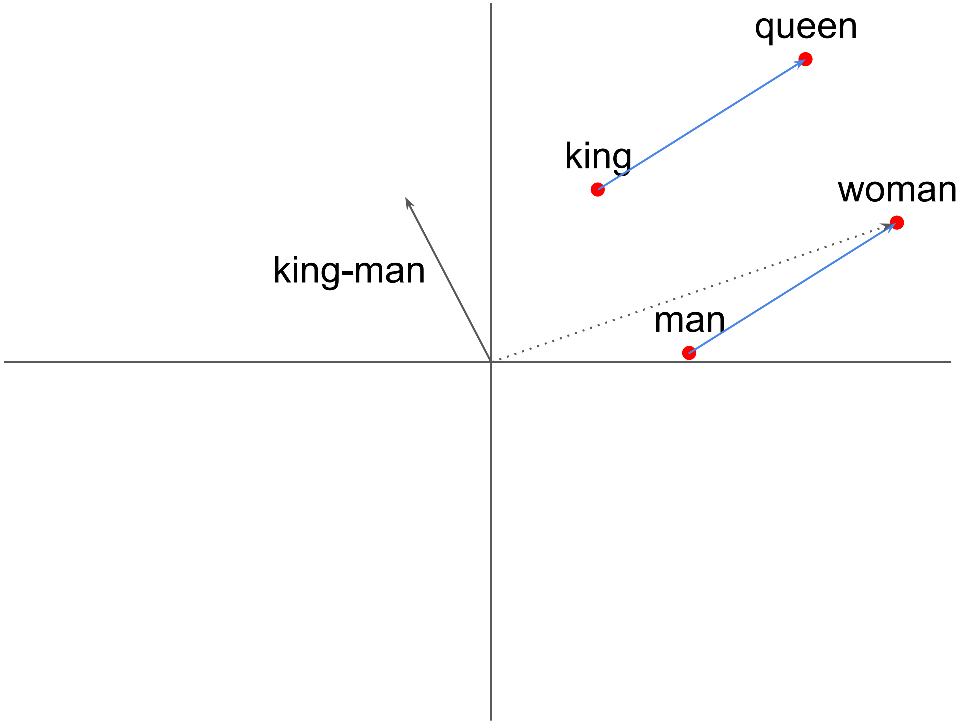
Vector Algebra



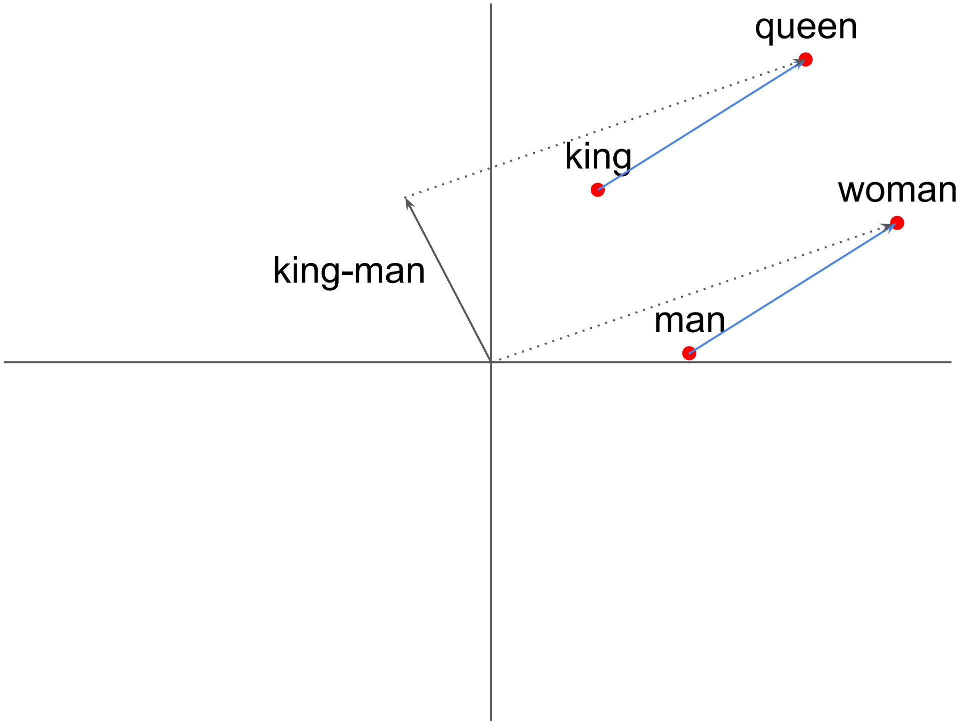
Vector Algebra



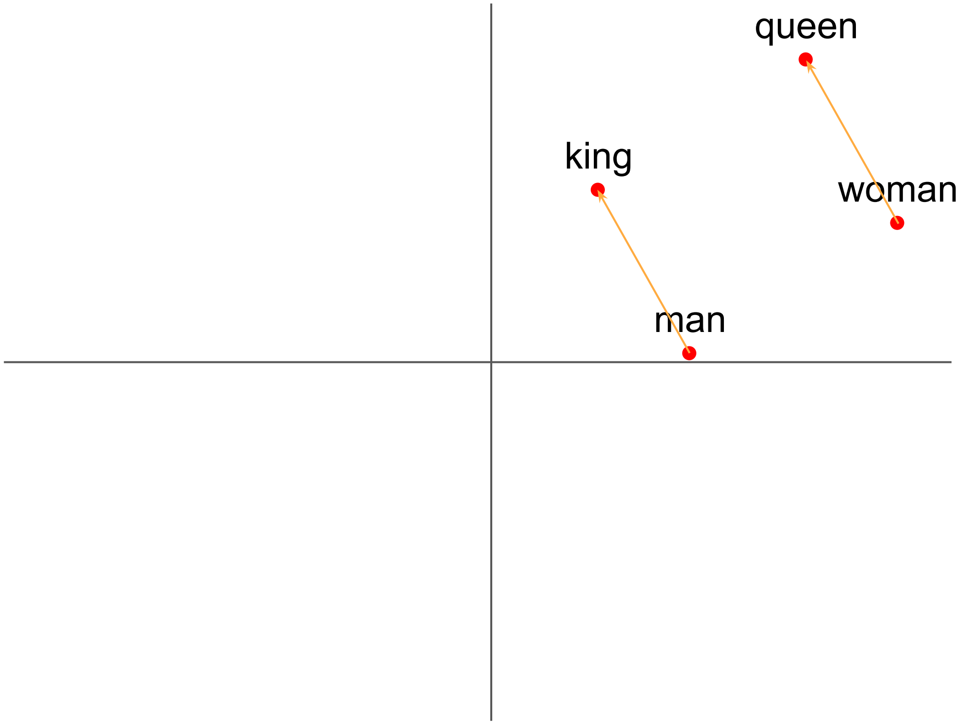
Vector Algebra



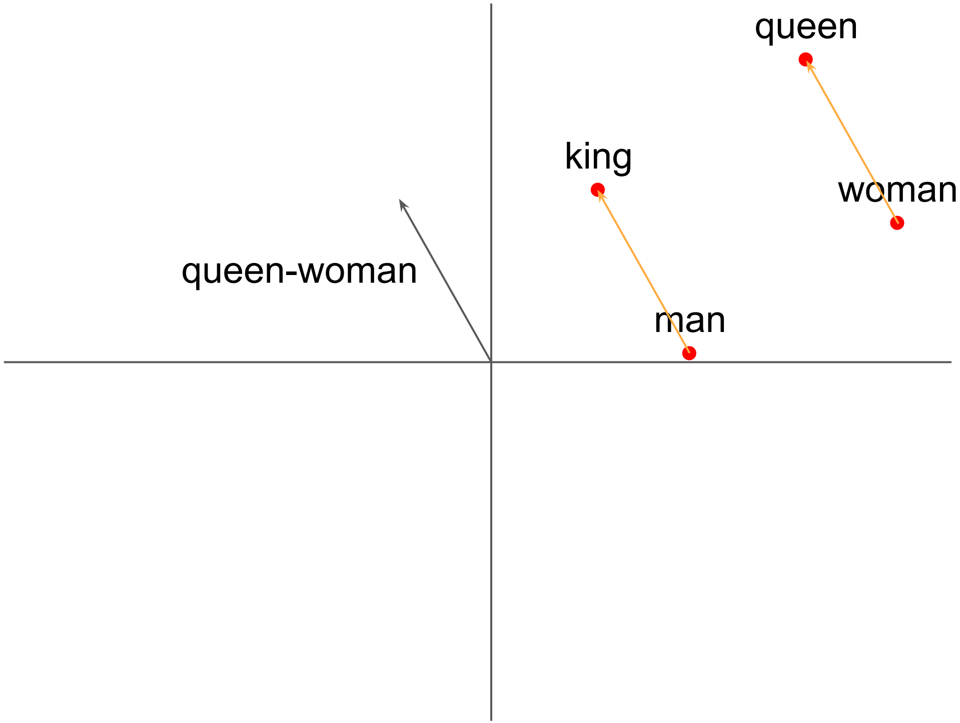
Vector Algebra



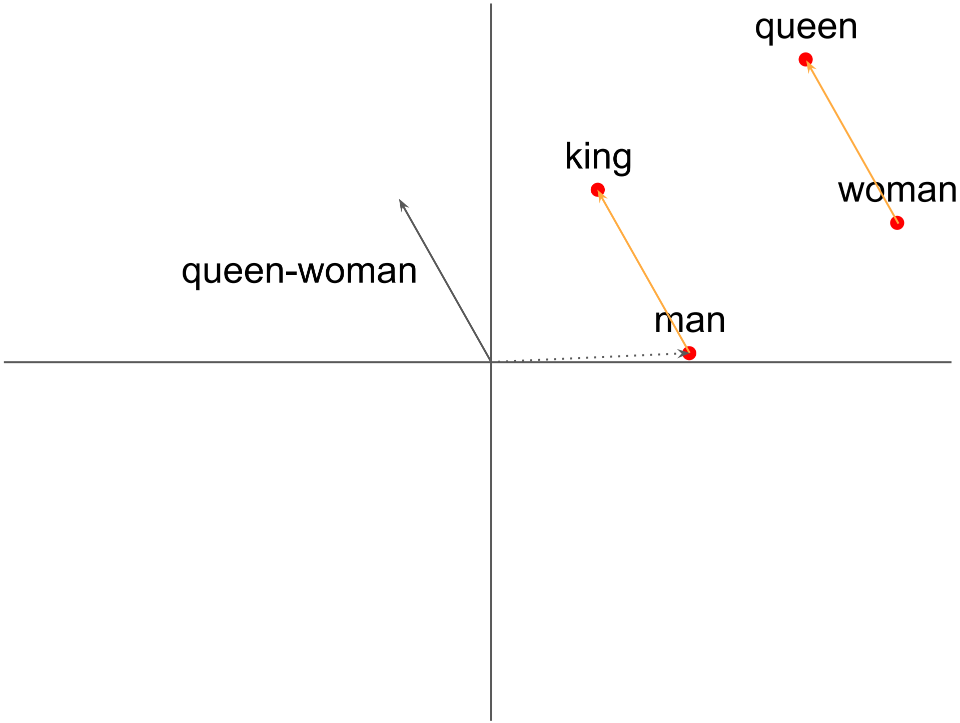
Vector Algebra



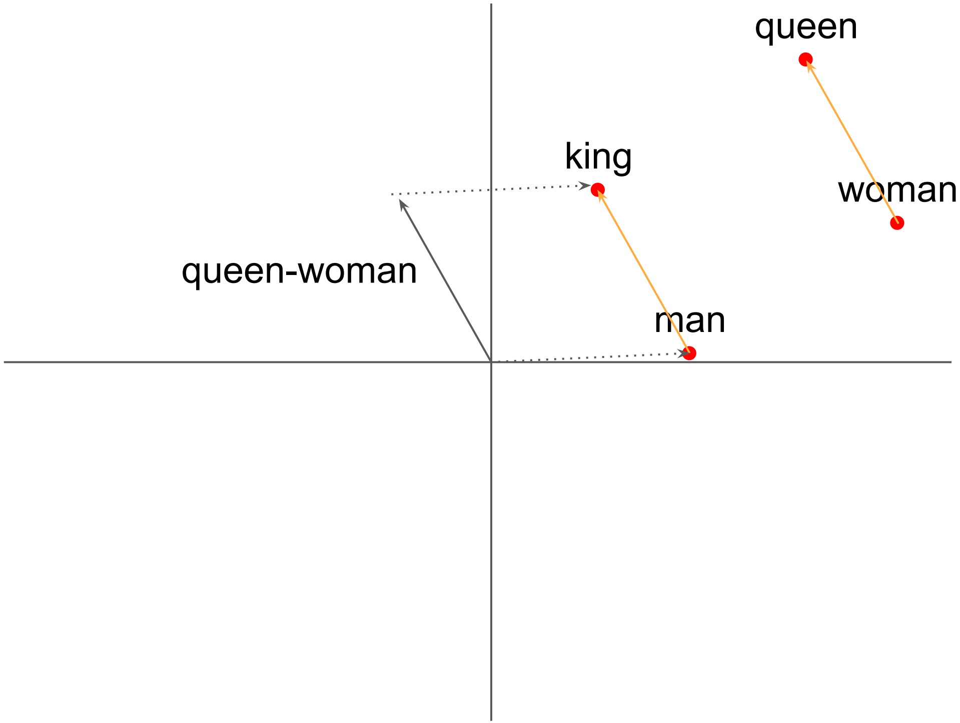
Vector Algebra



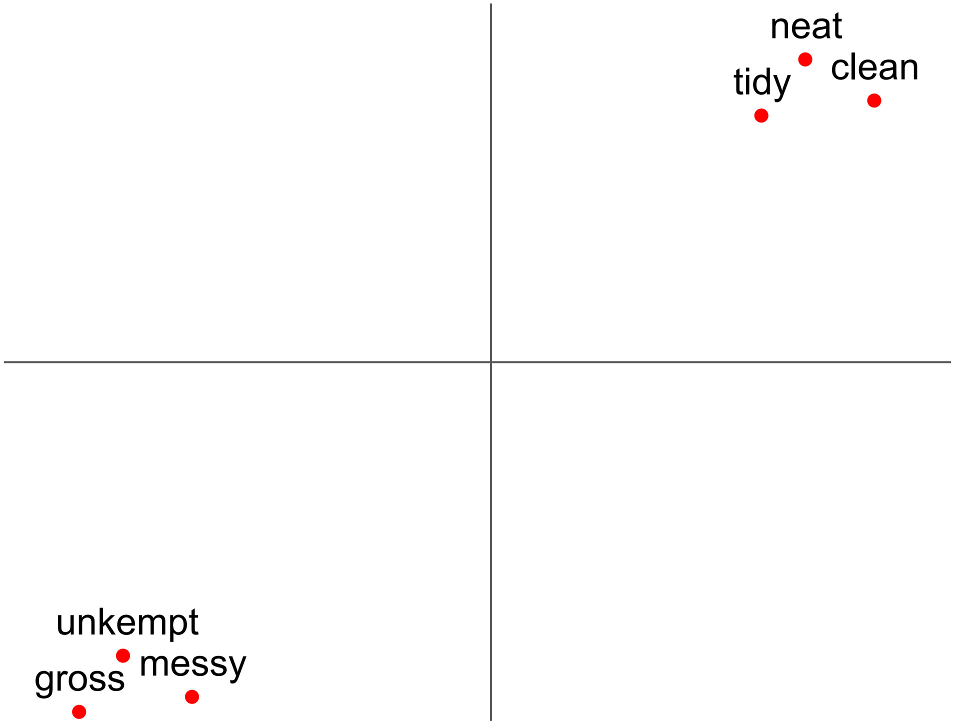
Vector Algebra



Vector Algebra



Vector Algebra



Other Word Embeddings

- Word2vec
- GloVe
- Train your own during GD

Transformer Models You Might Know

- BERT
- GPT
- Language Translations