How to turn words into vectors?

Sandeep Soni

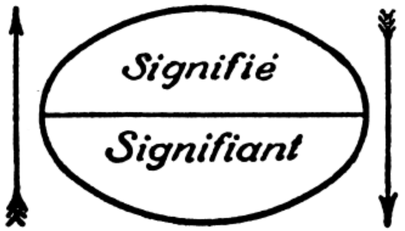
09/05/2024

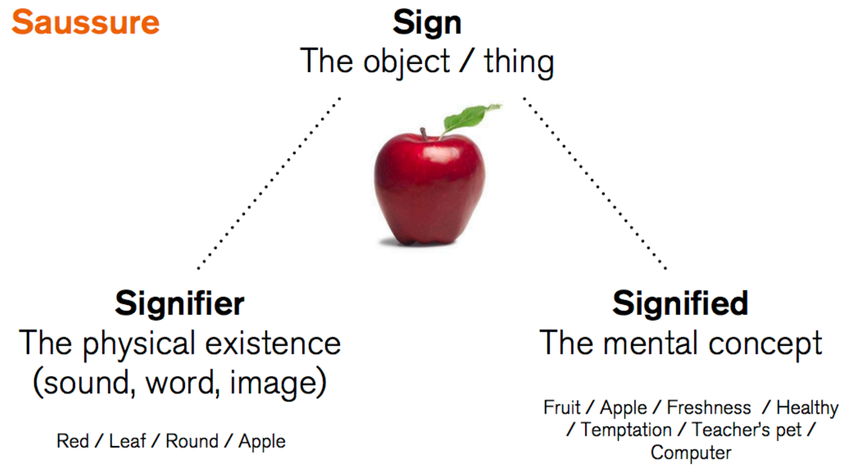
Story so far

- Words/Types/Tokens/Stems/Lemmas
- Tokenization
- Frequency based statistics to compare corpora/groups
  - χ^2 test
  - Pointwise mutual information (PMI)

Meaning

- Saussure proposed two characteristics of signs
  - Signified or the “plane of content”
  - Signifier or the “plane of expression”
- Words are a surface representation of concepts





What is the meaning of the word tezgüino?

A bottle of	_____	is on the table
Everybody	likes	_____
Don’t have	_____	before you drive
We make	_____	out of corn

Example taken from Eisenstein 2018, which was in turn taken from Lin, 1998

Is tezgüino similar to loud?

A bottle of	_____	is on the table
Everybody	likes	_____
Don’t have	_____	before you drive
We make	_____	out of corn

Example taken from Eisenstein, 2018, which was in turn taken by Lin, 1998

Is tezgüino similar to oil?

A bottle of	_____	is on the table
Everybody	likes	_____
Don’t have	_____	before you drive
We make	_____	out of corn

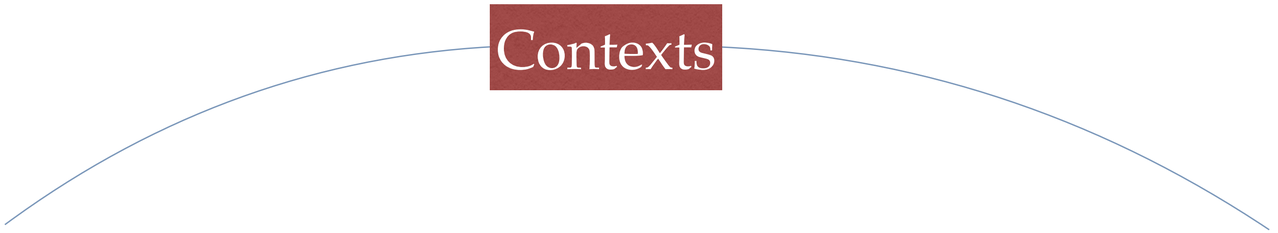
Example taken from Eisenstein, 2018, which was in turn taken by Lin, 1998

Is tezgüino similar to tortilla?

A bottle of	_____	is on the table
Everybody	likes	_____
Don’t have	_____	before you drive
We make	_____	out of corn

Example taken from Eisenstein, 2018, which was in turn taken by Lin, 1998

Is tezgüino similar to beer?



A bottle of	_____	is on the table
Everybody	likes	_____
Don’t have	_____	before you drive
We make	_____	out of corn

Example taken from Eisenstein, 2018, which was in turn taken by Lin, 1998

Question for the day

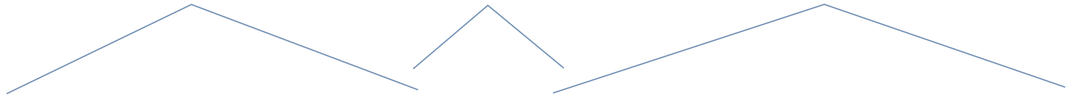
"How do we represent word meaning?"

Semantics

- Three perspectives in semantics
  - Relational
  - Compositional
  - Distributional

Semantics

Relational



Compositional

compose —> composition —> compositional

Distributional

The paint is still wet
Paint that wall red
The old paint is coming off

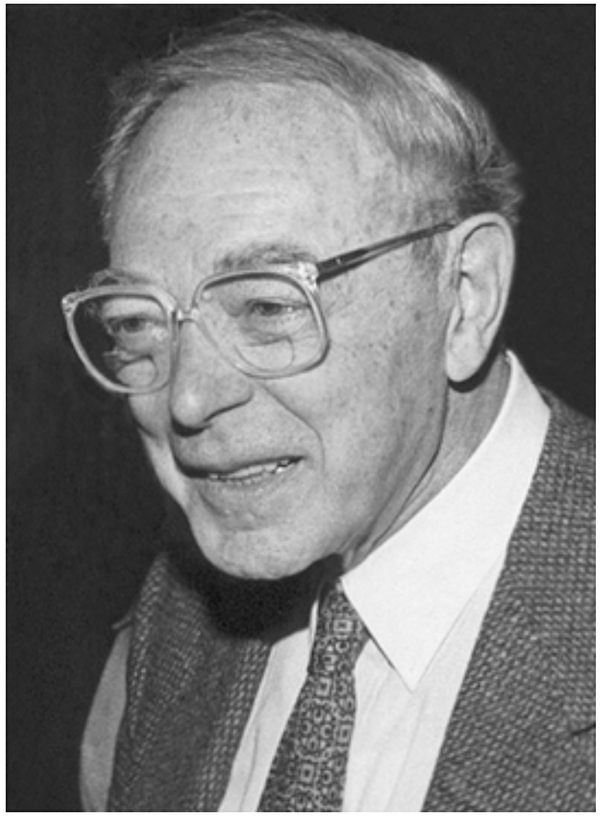
“a word is characterized by the company it keeps”

–J.R.Firth

Distributional semantics

- Word meanings can be learned from the the contexts in which they appear
- A foundational block in modern NLP

J.R.Firth



Zellig Harris

Distributed representations

- Distributed representations are vector representations that summarize the distribution of the contexts in which a word appears.
- Words that appear in similar contexts should have similar representations (distributional hypothesis)
- Key questions:
  - What type of contexts?
  - What type of vectors?
  - How to systematically do this?

Corpus

Imagine that you have a bunch of reviews

	Reviews
1	The temperature was cold and the food was cold … horrible service.
2	Wine was great. Beer was OK. Go here if you are into wine Wine WINE
3	I like the restaurant because their service is good … service is key!
4	I mostly order a single wine or a wine and cold beer
5	I ordered a beer. The service was pathetic. The beer was not even cold.

In general, such a collection is called a text corpus

Vocabulary

	Reviews
1	The temperature was cold and the food was cold … horrible service.
2	Wine was great. Beer was OK. Go here if you are into wine Wine WINE
3	I like the restaurant because their service is good … service is key!
4	I mostly order a single wine or a wine and cold beer
5	I ordered a beer. The service was pathetic. The beer was not even cold.

You can lowercase all reviews and then tokenize to construct a vocabulary

Vocabulary

	Reviews
1	The temperature was cold and the food was cold … horrible service.
2	Wine was great. Beer was OK. Go here if you are into wine Wine WINE
3	I like the restaurant because their service is good … service is key!
4	I mostly order a single wine or a wine and cold beer
5	I ordered a beer. The service was pathetic. The beer was not even cold.

the
temperature
was
cold
and
food
horrible
service
wine
great
beer
ok
…
not
even

You can lowercase all reviews and then tokenize to construct a vocabulary

Vocabulary

	Reviews
1	The temperature was cold and the food was cold … horrible service.
2	Wine was great. Beer was OK. Go here if you are into wine Wine WINE
3	I like the restaurant because their service is good … service is key!
4	I mostly order a single wine or a wine and cold beer
5	I ordered a beer. The service was pathetic. The beer was not even cold.

the	5
temperature	1
was	6
cold	4
and	1
food	1
horrible	1
service	4
wine	6
great	1
beer	4
ok	1
…	…
not	1
even	1

You can lowercase all reviews and then tokenize to construct a vocabulary

Term-document matrix

Vocabulary

Reviews

	Reviews
1	The temperature was cold and the food was cold … horrible service.
2	Wine was great. Beer was OK. Go here if you are into wine Wine WINE
3	I like the restaurant because their service is good … service is key!
4	I mostly order a single wine or a wine and cold beer
5	I ordered a beer. The service was pathetic. The beer was not even cold.

Term-document matrix

Vocabulary

Reviews

	Reviews
1	The temperature was cold and the food was cold … horrible service.
2	Wine was great. Beer was OK. Go here if you are into wine Wine WINE
3	I like the restaurant because their service is good … service is key!
4	I mostly order a single wine or a wine and cold beer
5	I ordered a beer. The service was pathetic. The beer was not even cold.

Term-document matrix

	Reviews
1	The temperature was cold and the food was cold … horrible service.
2	Wine was great. Beer was OK. Go here if you are into wine Wine WINE
3	I like the restaurant because their service is good … service is key!
4	I mostly order a single wine or a wine and cold beer
5	I ordered a beer. The service was pathetic. The beer was not even cold.

Vocabulary	Reviews
the	5
temperature	1
was	6
cold	4
and	1
food	1
horrible	1
service	4
wine	6
great	1
beer	4
ok	1
…	…
not	1
even	1

Term-document matrix

	Reviews
1	The temperature was cold and the food was cold … horrible service.
2	Wine was great. Beer was OK. Go here if you are into wine Wine WINE
3	I like the restaurant because their service is good … service is key!
4	I mostly order a single wine or a wine and cold beer
5	I ordered a beer. The service was pathetic. The beer was not even cold.

Vocabulary

the	5
temperature	1
was	6
cold	4
and	1
food	1
horrible	1
service	4
wine	6
great	1
beer	4
ok	1
…	…
not	1
even	1

Reviews

Term-document matrix

	Reviews
1	The temperature was cold and the food was cold … horrible service.
2	Wine was great. Beer was OK. Go here if you are into wine Wine WINE
3	I like the restaurant because their service is good … service is key!
4	I mostly order a single wine or a wine and cold beer
5	I ordered a beer. The service was pathetic. The beer was not even cold.

Vocabulary

the	5
temperature	1
was	6
cold	4
and	1
food	1
horrible	1
service	4
wine	6
great	1
beer	4
ok	1
…	…
not	1
even	1

Reviews

Terms	*R1*	*R2*	*R3*	*R4*	*R5*
the	2		1		2
temperature	1				
was	2	2			2
cold	1			2	1
and	1				
food	1				
horrible	1				
service	1		2		1
wine		4		2	
great		1			
beer		1		1	2
ok		1			
…	…	…	…	…	…
not					1
even					1

Term-document matrix

- Every term in the vocabulary is mapped to a row in this matrix
- Each row is a vector encoding a distribution of counts over reviews/ documents
- Context = entire document

Term-document matrix

Terms	*R1*	*R2*	*R3*	*R4*	*R5*
the	2		1		2
temperature	1				
was	2	2			2
cold	1			2	1
and	1				
food	1				
horrible	1				
service	1		2		1
wine		4		2	
great		1			
beer		1		1	2
ok		1			
…	…	…	…	…	…
not					1
even					1

- Every term in the vocabulary is mapped to a row in this matrix
- Each row is a vector encoding a distribution of counts over reviews/documents
- Context = entire document

Term-document matrix

- Blank entries in this matrix correspond to zero, so vectors are sparse!
- The size of each vector is equal to the number of documents

Term-document matrix

Terms	*R1*	*R2*	*R3*	*R4*	*R5*
the	2		1		2
temperature	1				
was	2	2			2
cold	1			2	1
and	1				
food	1				
horrible	1				
service	1		2		1
wine		4		2	
great		1			
beer		1		1	2
ok		1			
…	…	…	…	…	…
not					1
even					1

- Blank entries in this matrix correspond to zero, so vectors are sparse!
- The size of each vector is equal to the number of documents

Corpus

	Reviews
1	The temperature was cold and the food was cold … horrible service.
2	Wine was great. Beer was OK. Go here if you are into wine Wine WINE
3	I like the restaurant because their service is good … service is key!
4	I mostly order a single wine or a wine and cold beer
5	I ordered a beer. The service was pathetic. The beer was not even cold

Let's focus on these words for now. Is wine more similar to beer or cold?

Vector similarity

- The similarity between two vectors can be calculated by taking the dot product between them.
- If vectors are normalized, then dot products can be interpreted as the cosine of the angle between them.

$$a \cdot b \qquad \text{dot}(a, b) \qquad \sum_i a_i b_i$$

$$< a, b > \qquad a^T b$$

$$b \cdot a \qquad < b, a >$$

$$b^T a \qquad \|a\| \|b\| \cos\theta$$

These are all the same!

Vector similarity

Reviews

	R1	*R2*	*R3*	*R4*	*R5*
…	…	…	…	…	…
wine		4		2	
cold	2			1	1
beer		2		1	2
…	…	…	…	…	…

```
dot(wine, cold)

= 0*2 + 4*0 + 0*0 + 2*1 + 0*1

= 2
```

Vector similarity

Reviews

	R1	*R2*	*R3*	*R4*	*R5*
…	…	…	…	…	…
wine		4		2	
cold	2			1	1
beer		2		1	2
…	…	…	…	…	…

```
dot(wine, beer)

= 0*0 + 4*2 + 0*0 + 2*1 + 0*2

= 10
```

Cooccurence matrix

- Instead of taking the entire document as context, we can define a context window of some size
- We can construct a matrix with number of rows and columns equal to vocabulary size
- Each cell's value is the number of times a word pair corresponding to the row and column appear together.

Dataset

The old wine tastes good

The bottled beer is stale

The red wine is stale

Store the beer for long

Assume context window size of 2

All words that co-appear with wine are:

{the, old, tastes, good, red, is, stale}

Dataset

The old wine tastes good

The bottled beer is stale

The red wine is stale

Store the beer for long

Assume context window size of 2

All words that co-appear with wine are:

{the, bottled, is, stale, store, for, long}

Coccurrence matrix

Contexts

	the	*bottled*	*tastes*	*good*	…	*stale*	*long*
…							
wine	2		1	1		1	
beer	2	1				1	
…							

- Every term in the vocabulary is mapped to a row in this matrix
- Each row is a vector encoding a distribution of counts over its cooccurring words
- Context = window around the term

Vectors

Contexts (c)

Terms (t)

	the	*bottled*	*tastes*	*good*	…	*stale*	*long*
…							
wine	2		1	1		1	
beer	2	1				1	
…							

- The vectors are still sparse!
- The matrix grows quadratically with the size of the vocabulary

Weighting

- Many words cooccur with functional words such as “the”, “a”, etc
- So can you weigh the dimensions based on their informativeness?

TFIDF

- Idea: Upweight the cooccurrence counts for terms that cooccur frequently and exclusively.
- $tfidf(t, c) = tf(t, c) \times \frac{N}{d_t}$
- N is the total number of contexts, $tf(t, c)$ is the number of times t co-appears with c, d_t is the number of times t appears in any context

PMI

- Idea: Upweight the concurrence counts if the the pair is more likely to cooccur than chance
- $PMI(t,c) = \log_2 \frac{P(t,c)}{P(t) \cdot P(c)}$
- P(t,c) is the probability of cooccurrence; P(t) and P(c) is probability of independent occurrence
- $PPMI(t,c) = max(0, \log_2 \frac{P(t,c)}{P(t) \cdot P(c)})$

Dense representations

- Ideally, we want to learn characteristic dimensions of a word, instead of distributions over all words/contexts
- We learn a low-dimensional but compact/dense vectors by formulating the coccurrence events as prediction tasks

Word	Ball game	Non-square arena	Olympic Game	Indoor
Soccer	1	0	1	0
Javelin Throw	0	1	1	0
Squash	1	0	0	1
Chess	0	1	0	1

So far!

- Vector representations are sparse and long
- They’re based on counting

Matrix factorization

- You can factorize the coccurrence matrix
- The vectors in the factors are low-dimensional and dense

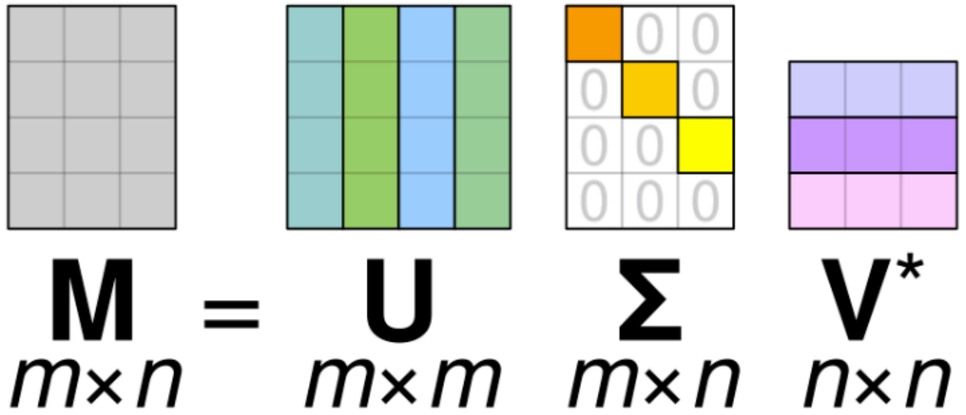


Image taken from Wikipedia

Word2vec

- Skipgram (Mikolov et. al. 2013) predicts a context word given a target word
- Think of this as a huge multi-class classification task (classes are words) for every word

Cooccurrence Matrix

x	y
…	…
wine	*cold*
wine	*sweet*
wine	*spirit*
wine	*drink*
…	…

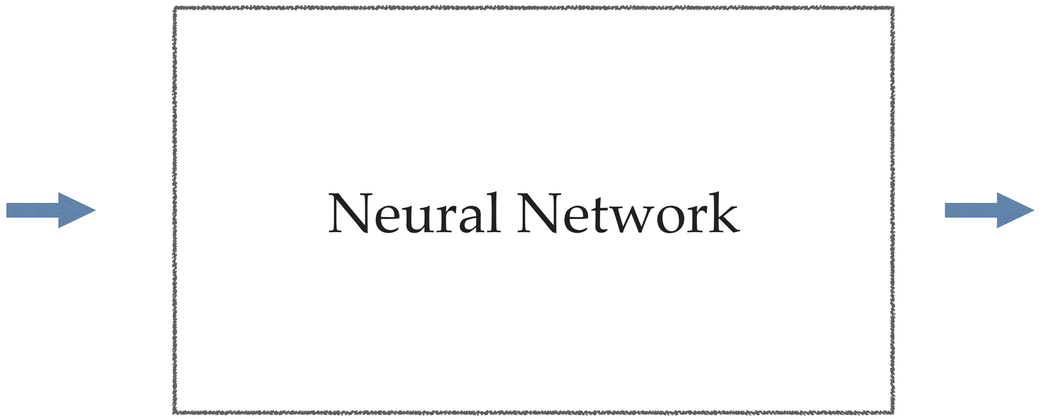
Word2vec

- We can estimate the empirical probability as $P(y \mid x) \propto \exp(\mathbf{y} \cdot \mathbf{x})$
- $\mathbf{y}$ is a vector representation of a word when it appears in the context; $\mathbf{x}$ is a vector representation of a word when it appears in the input

x	y
…	…
wine	*cold*
wine	*sweet*
wine	*spirit*
wine	*drink*
…	…

Word2vec

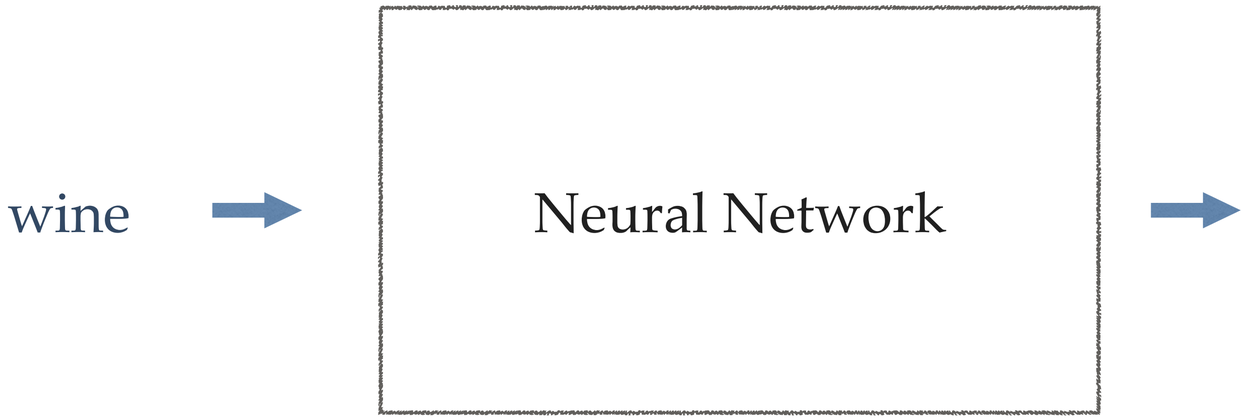
x	y
…	…
wine	*cold*
wine	*sweet*
wine	*spirit*
wine	*drink*
…	…



Take every entry on the left and guess the corresponding entry on the right. Update the network if guess is wrong.

Word2vec

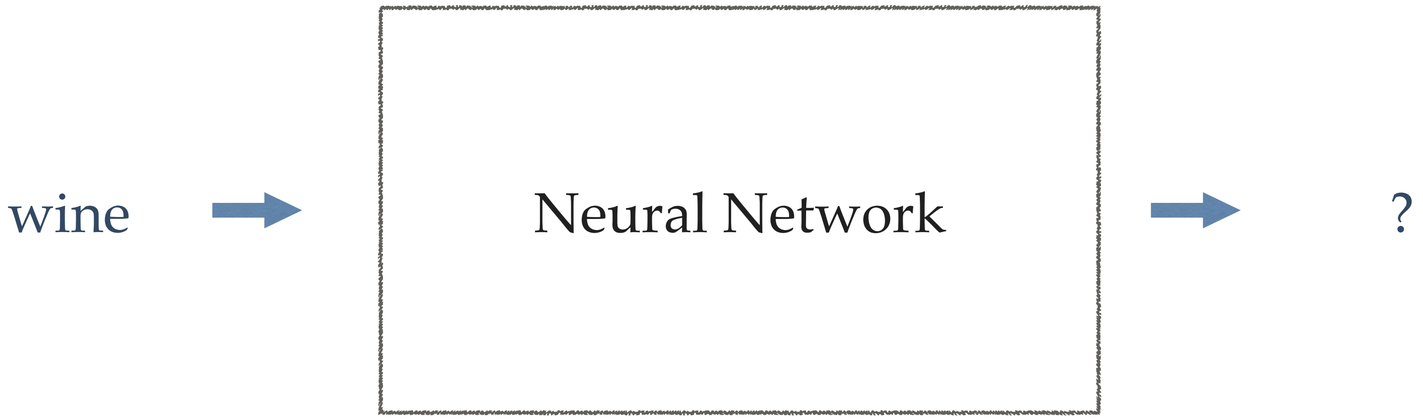
x	y
…	…
wine	*cold*
wine	*sweet*
wine	*spirit*
wine	*drink*
…	…



Take every entry on the left and guess the corresponding entry on the right. Update the network if guess is wrong.

Word2vec

x	y
…	…
wine	*cold*
wine	*sweet*
wine	*spirit*
wine	*drink*
…	…



Take every entry on the left and guess the corresponding entry on the right. Update the network if guess is wrong.

x	y
…	…
wine	*cold*
wine	*sweet*
wine	*spirit*
wine	*drink*
…	…

Vocab

x	y
…	…
wine	*cold*
wine	*sweet*
wine	*spirit*
wine	*drink*
…	…

service
wine
cold
great

Vocab

wine

Vocab
service
wine
cold
great

wine

x
0
1
0
0

x	y
…	…
wine	*cold*
wine	*sweet*
wine	*spirit*
wine	*drink*
…	…

service
wine
cold
great

Vocab

wine

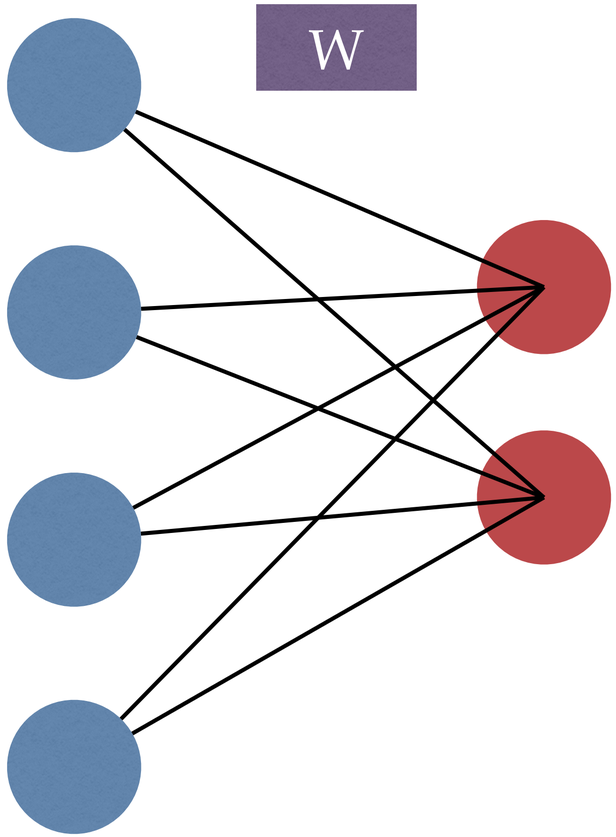
x

0

1

0

0

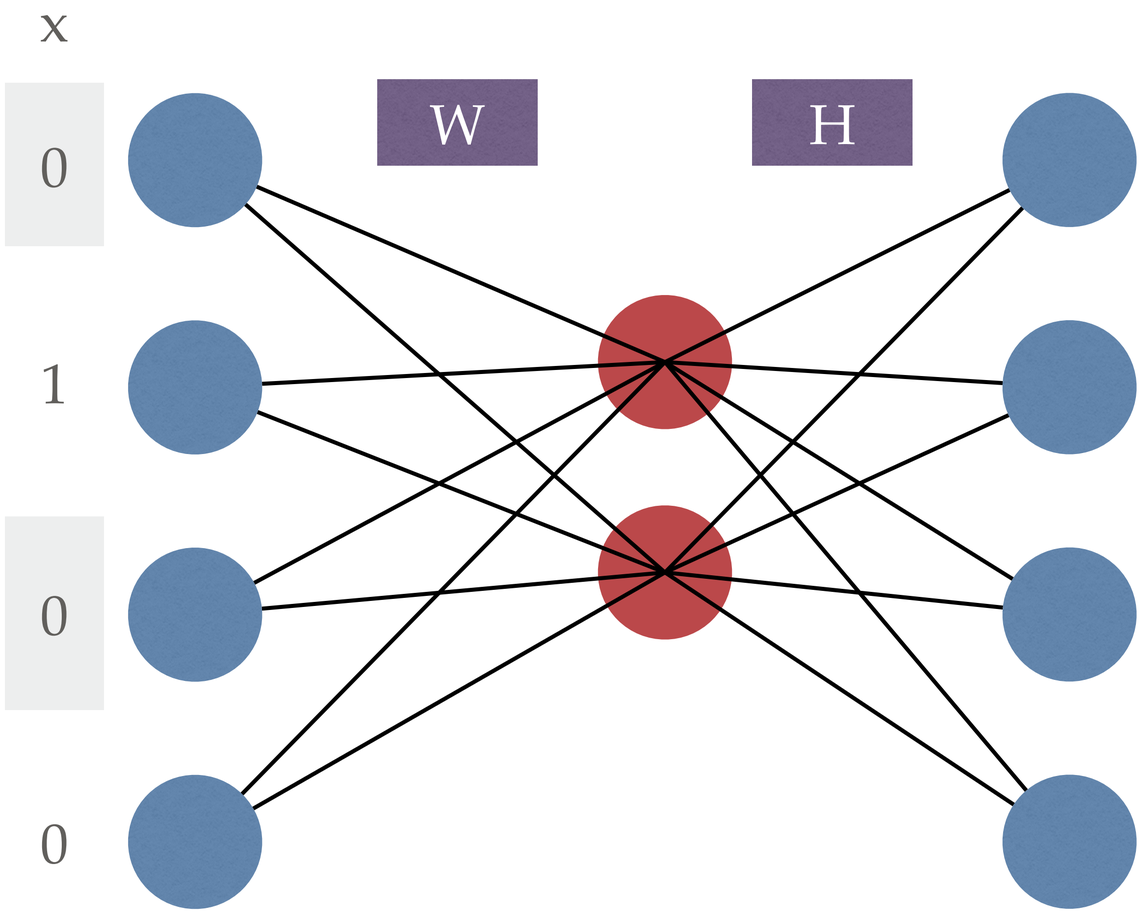


x	y
…	…
wine	*cold*
wine	*sweet*
wine	*spirit*
wine	*drink*
…	…

service
wine
cold
great

Vocab

wine

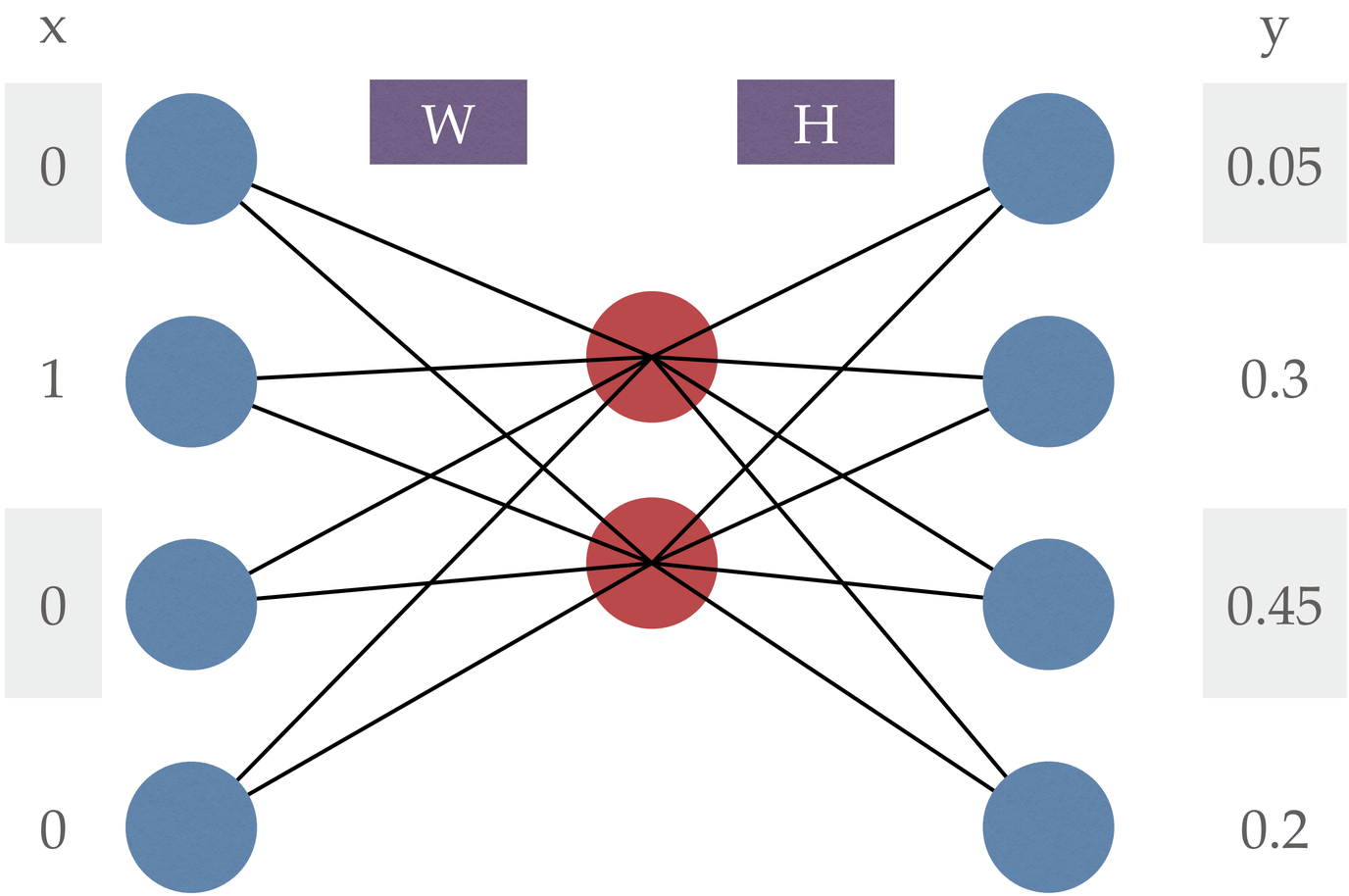


x	y
…	…
wine	*cold*
wine	*sweet*
wine	*spirit*
wine	*drink*
…	…

service
wine
cold
great

Vocab

wine

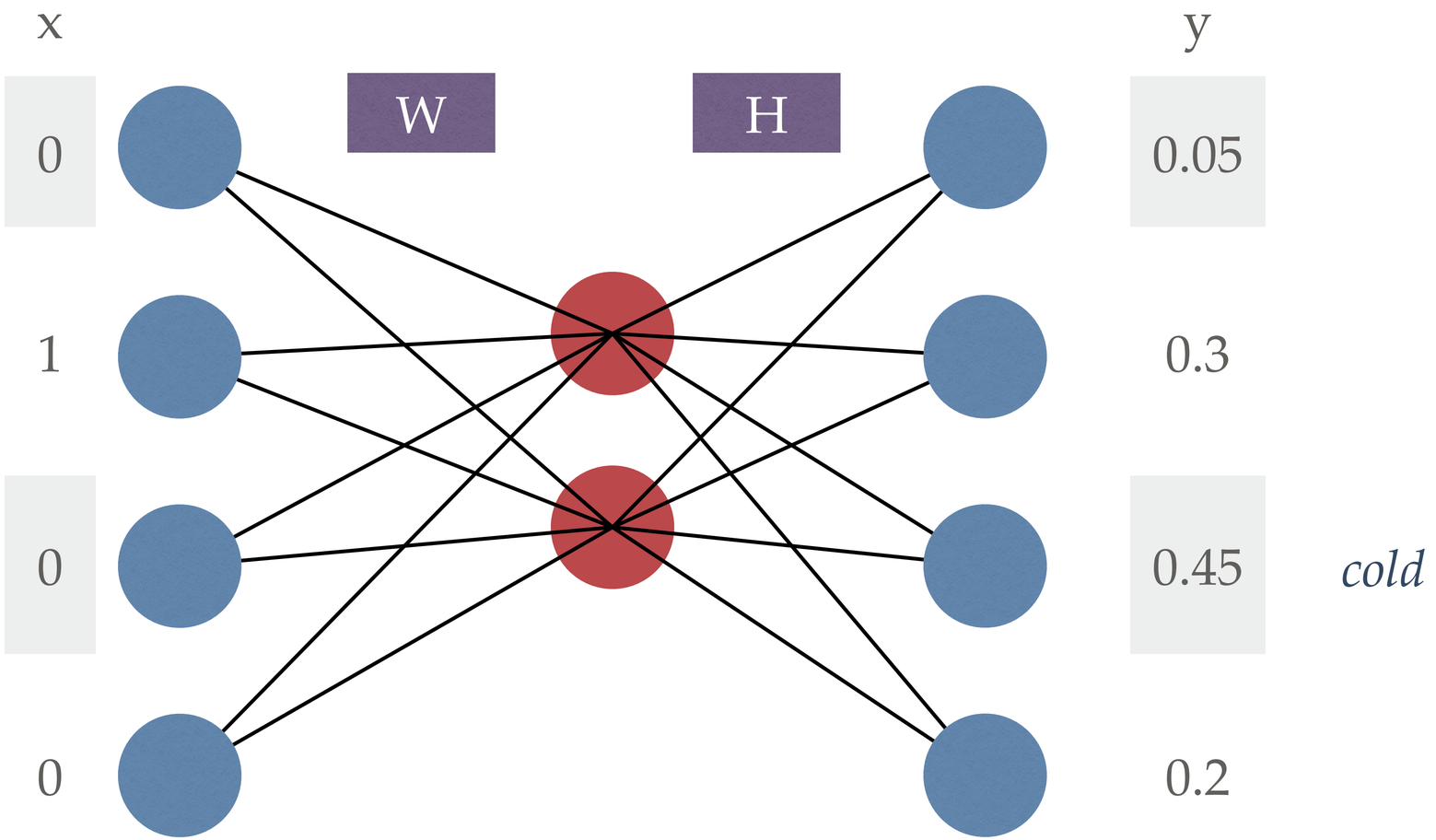


x	y
…	…
wine	*cold*
wine	*sweet*
wine	*spirit*
wine	*drink*
…	…

service
wine
cold
great

Vocab

wine



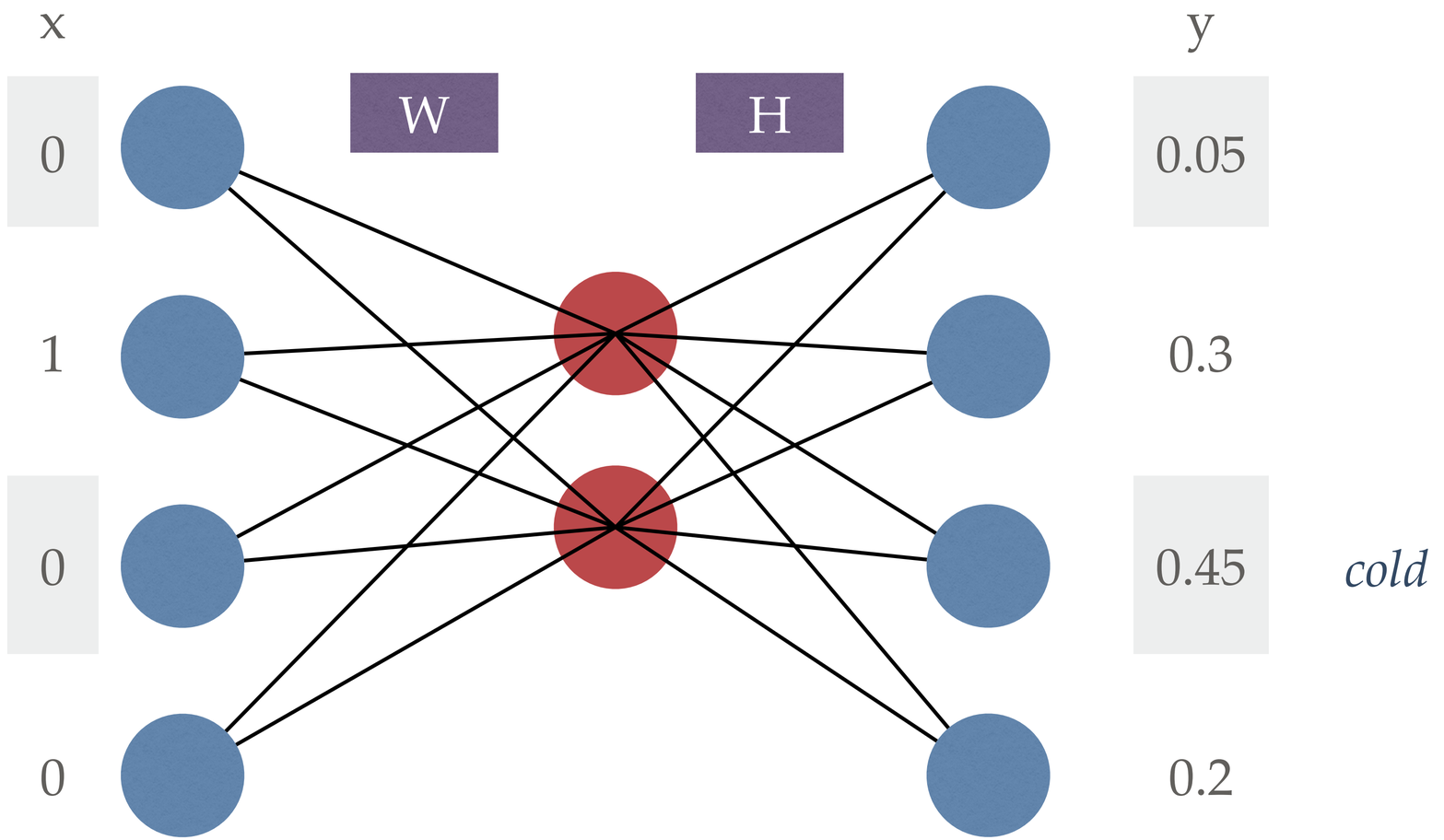
cold

x	y
…	…
wine	*cold*
wine	*sweet*
wine	*spirit*
wine	*drink*
…	…

service
wine
cold
great

Vocab

wine



cold

x	y
…	…
wine	*cold*
wine	*sweet*
wine	*spirit*
wine	*drink*
…	…

W

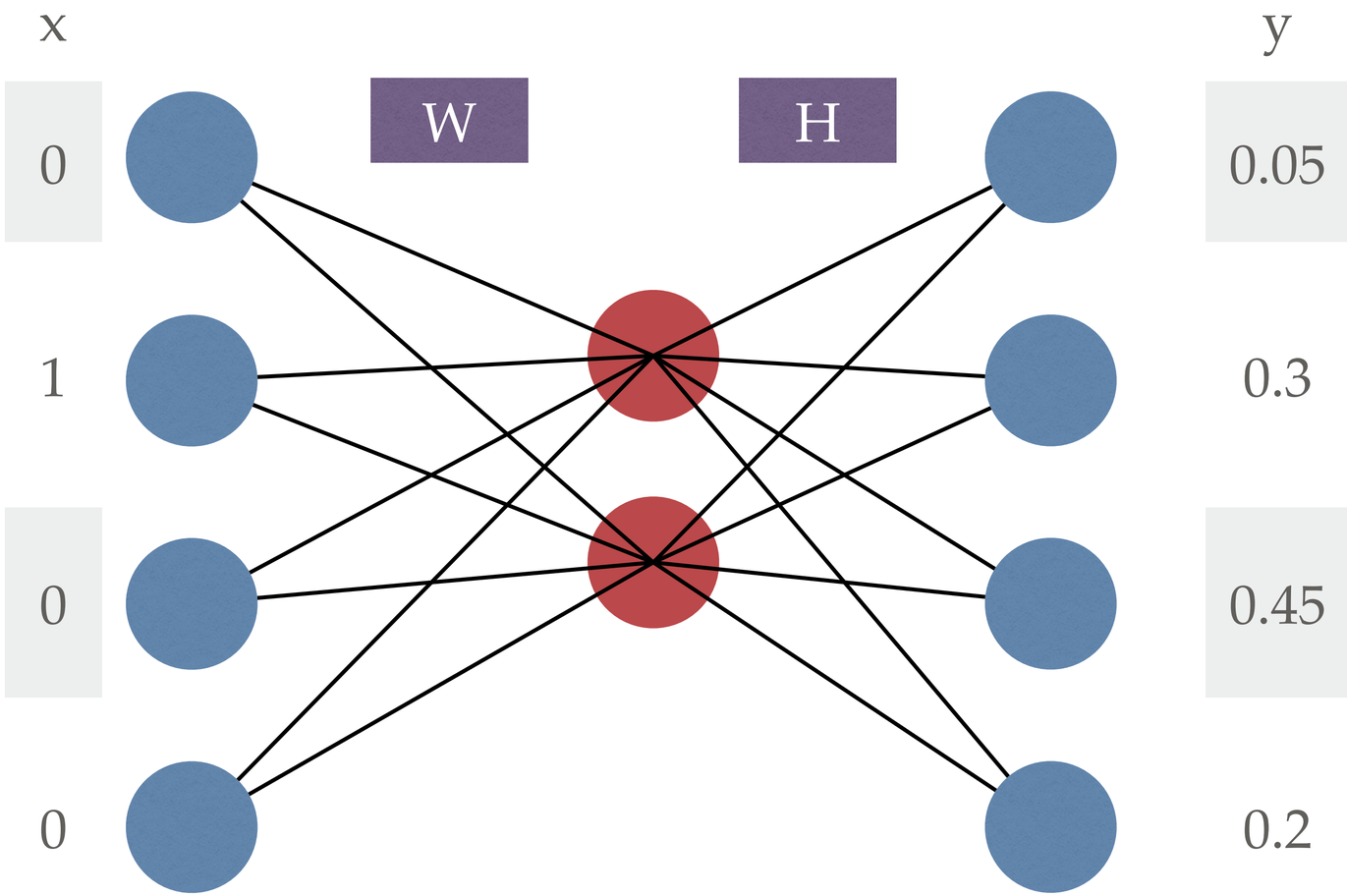
Word embeddings as columns

-0.3	1.2	0.5	-0.6
0.2	0.9	0.1	-0.4

service
wine
cold
great

Vocab

wine



cold

x	y
…	…
wine	*cold*
wine	*sweet*
wine	*spirit*
wine	*drink*
…	…

W

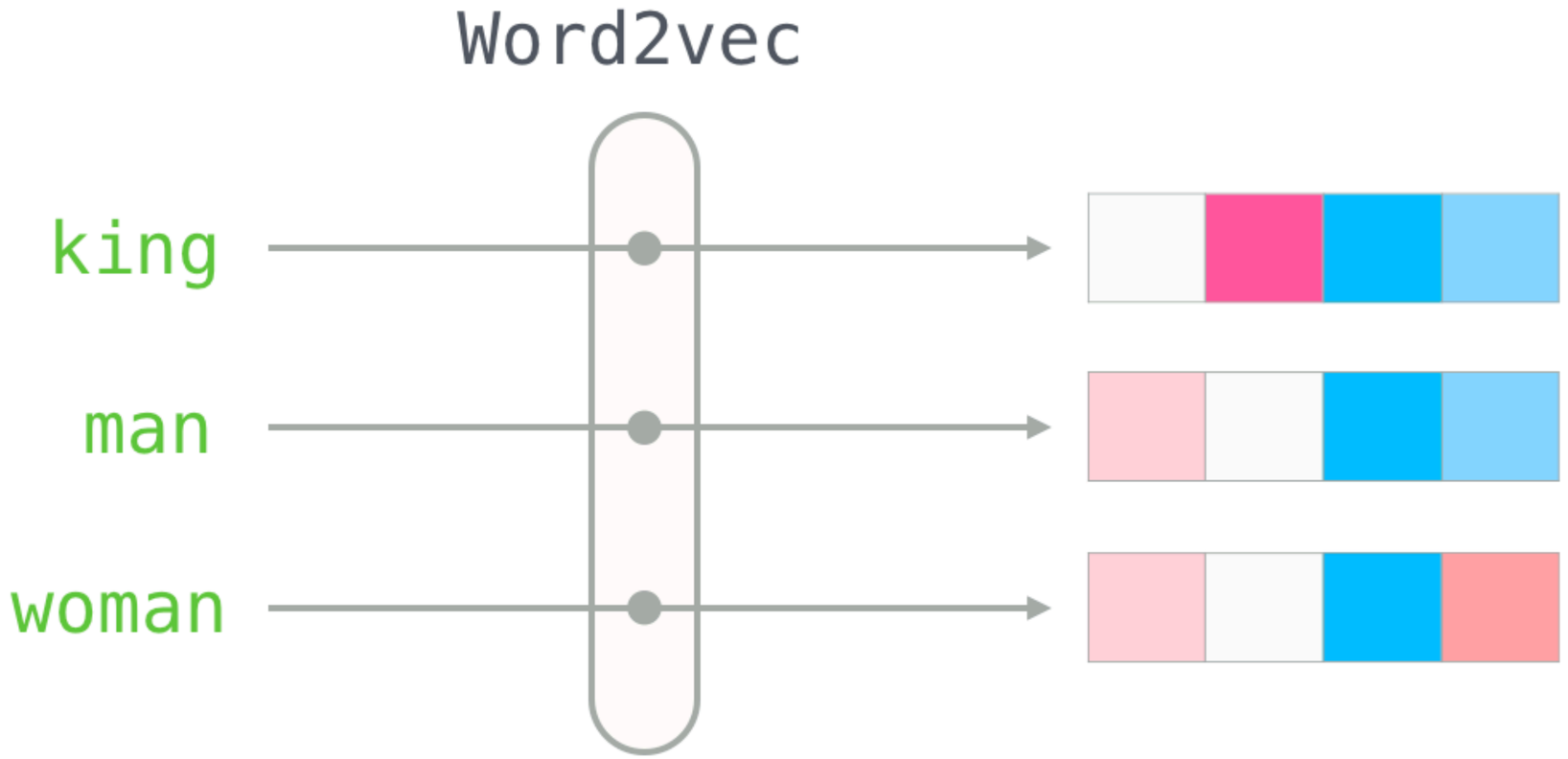
Word embeddings as columns

-0.3	1.2	0.5	-0.6
0.2	0.9	0.1	-0.4

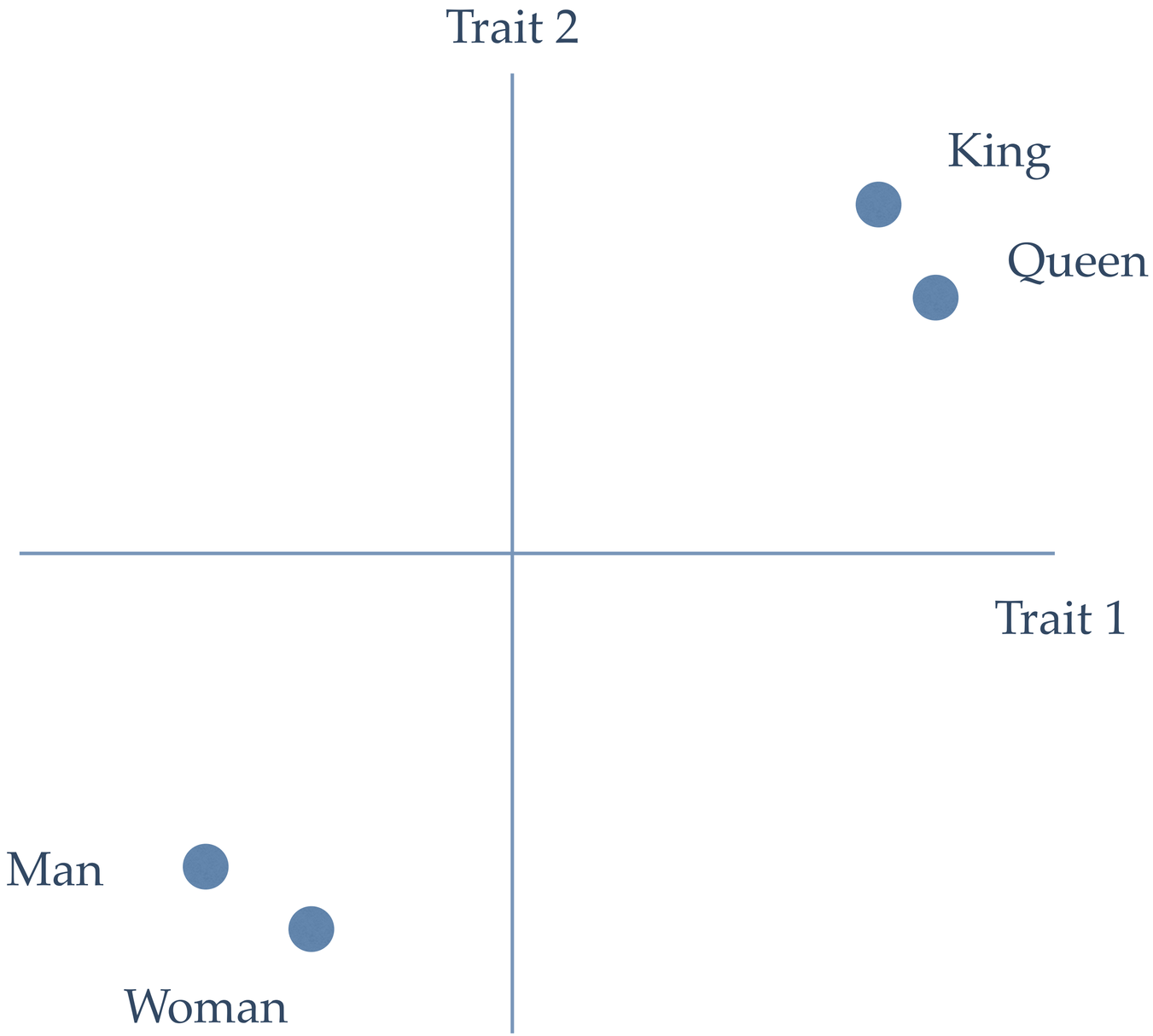
H

Context embeddings as rows

0.1	-0.4
0.4	-0.5
0.3	-0.1
0.2	0.1



http://jalammar.github.io/illustrated-word2vec/


Trait 2
King
Queen
Trait 1
Man
Woman

Glove

- Pennington et. al. 2014 used a similar idea but directly model some statistic of the cooccurrence
- Think of this as a regression task

x	y	statistic
wine	cold	3
wine	sweet	5
wine	spirit	1
wine	drink	2

Evaluation

Evaluation

- How do we know if the embeddings we learned are any good?

Evaluation

Evaluation

- Intrinsic evaluation
  - Word relatedness: the similarity between vector representations should correlate with human judgments of relatedness of pairs of words
  - Analogical reasoning: King:Queen::Man:?
- Extrinsic evaluation
  - Plug in the embeddings as features in some downstream task

In class

- Word2Vec demo