Word Embeddings

Sandeep Soni

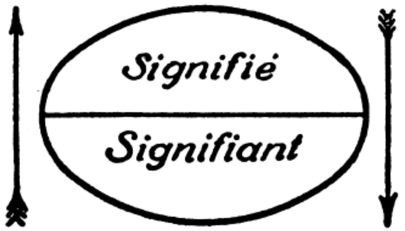
01/30/2024

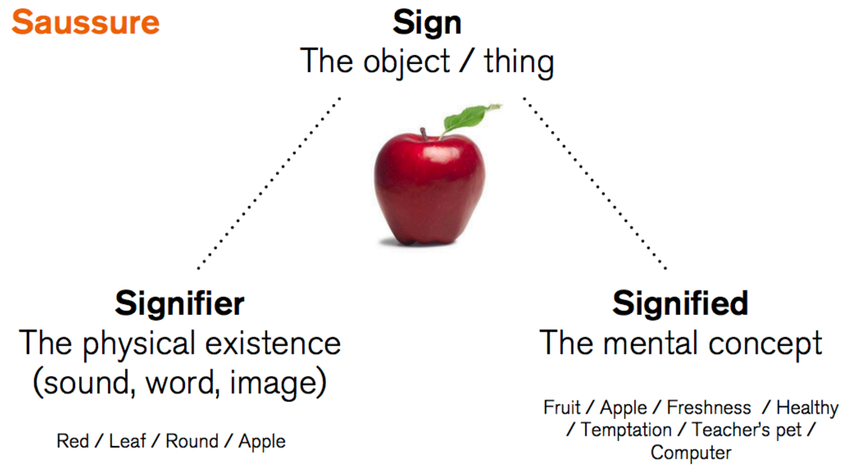
Story so far

- 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

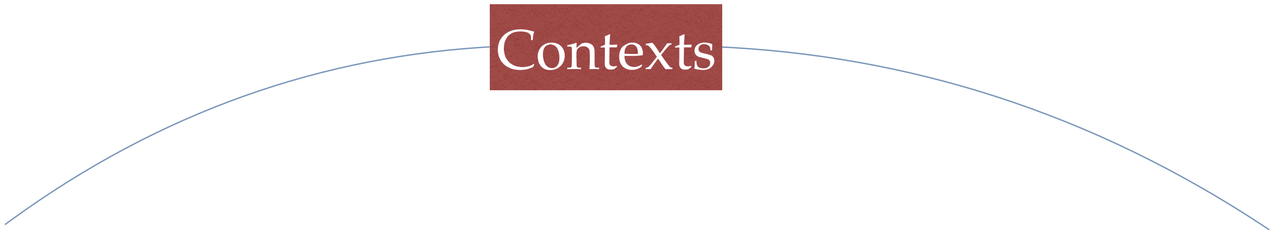
Example taken from Eisenstein, 2018, which was in turn taken by Lin, 1998

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

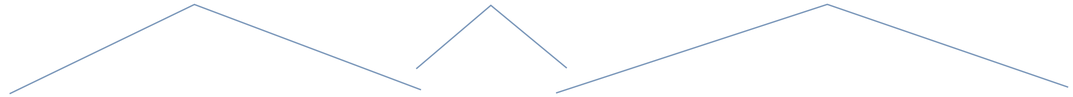
"How do we represent word meaning?"

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

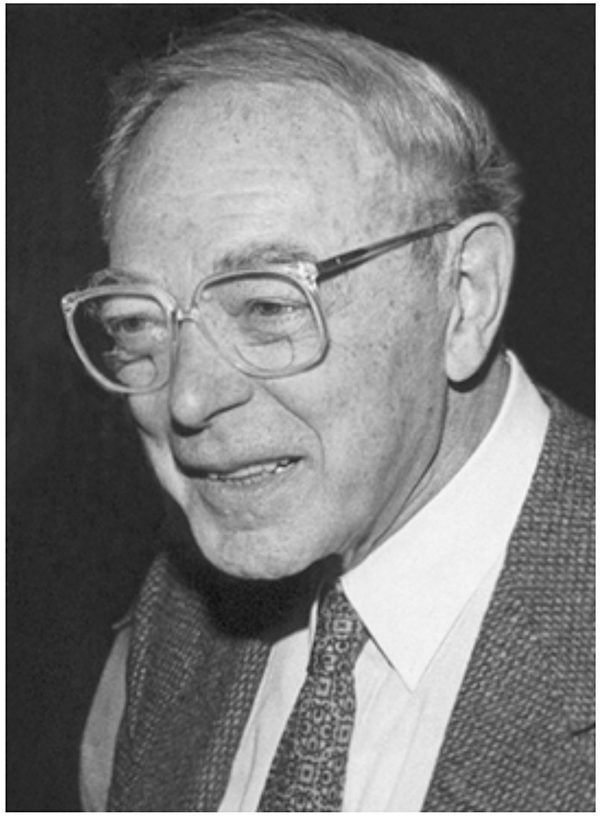
“a word is characterized by the company it keeps”

–J.R.Firth

Distributional semantics

- Distributional semantics allows us to learn the meanings from the contexts in which the words appear
- The ability to learn meanings in this way has been instrumental in recent progress

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

	Reviews
1	The temperature was cold and the food was cold
2	Wine was great. Beer was OK. Go here if you are into wine Wine WINE
3	I like the electrician because his 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

	R1	R2	R3	R4	R5
wine		4		2	
cold	2			1	1
beer		2		1	2
service			2		1

Every term is a vector encoding the distribution over documents

Context=entire document

Vectors

Reviews

	R1	R2	R3	R4	R5
wine		4		2	
cold	2			1	1
beer		2		1	2
service			2		1

The vectors are sparse

Their size is the number of documents

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
service			2		1

Cosine similarity between the vectors for wine and cold is less than the cosine similarity between the vectors for wine and beer

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 rows as terms, columns as the contexts, and values as the number of times the words appear in contexts

Dataset

The old wine tastes good

The bottled beer is stale

The red wine is stale

Store the beer for long

Words in the context window of size 2 for wine:

[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

Words in the context window of size 2 for beer:

[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 is a vector encoding the distribution over other terms

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	

Vectors are sparse

Size=Terms in vocabulary

Weighting

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

TFIDF

- The contexts for a term are weighted higher if they are frequent and specific
- $tfidf(t,c) = tf(t,c) \times \frac{N}{d_t}$
- 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

- The contexts for a term are weighted higher if they cooccur more than they would if they had no relation
- $PMI(t,c) = \log_2 \frac{P(t,c)}{P(t) \cdot P(c)}$
- 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

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

x	y
wine	cold
wine	sweet
wine	spirit
wine	drink

Word2vec

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

x	y
wine	cold
wine	sweet
wine	spirit
wine	drink

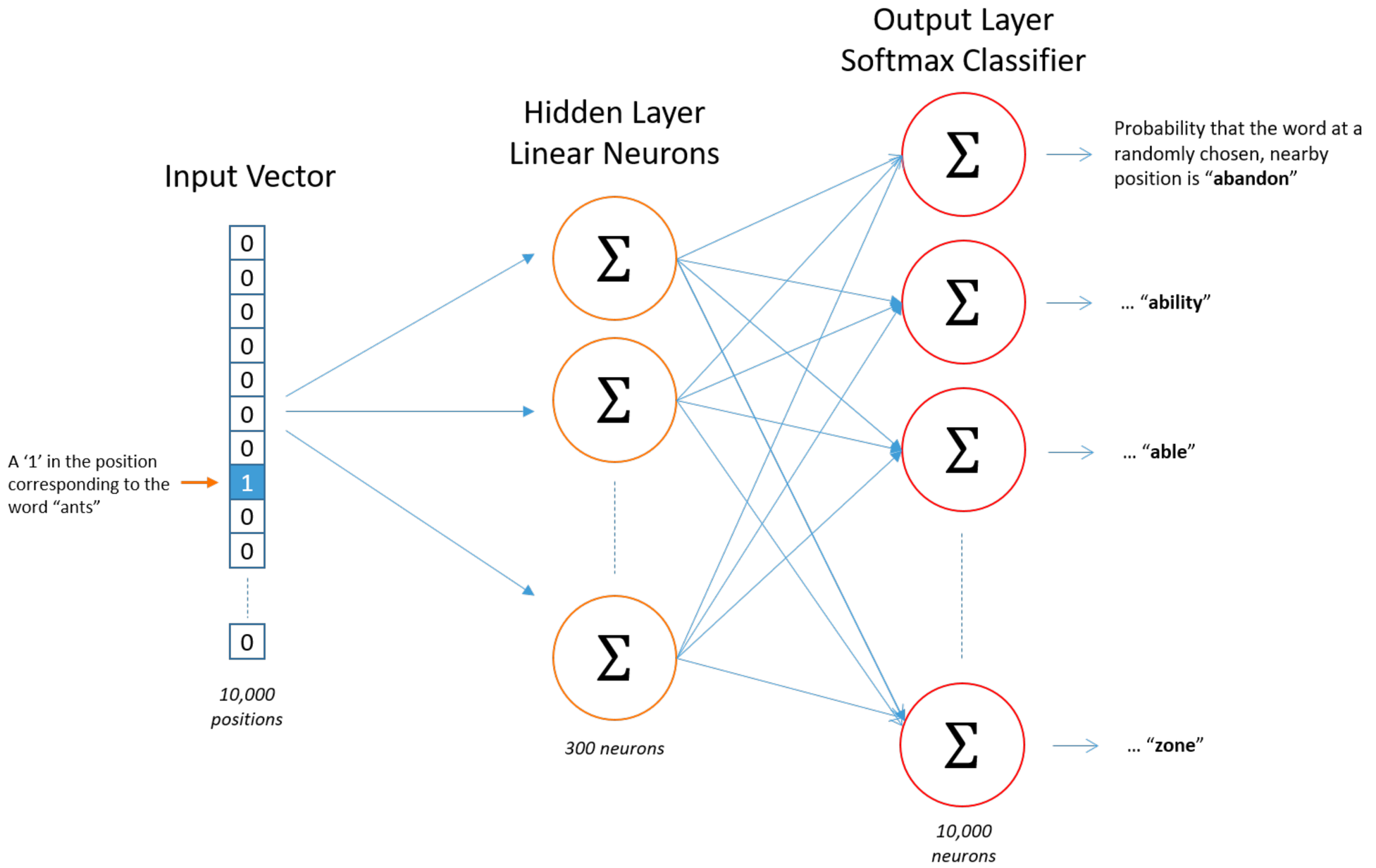
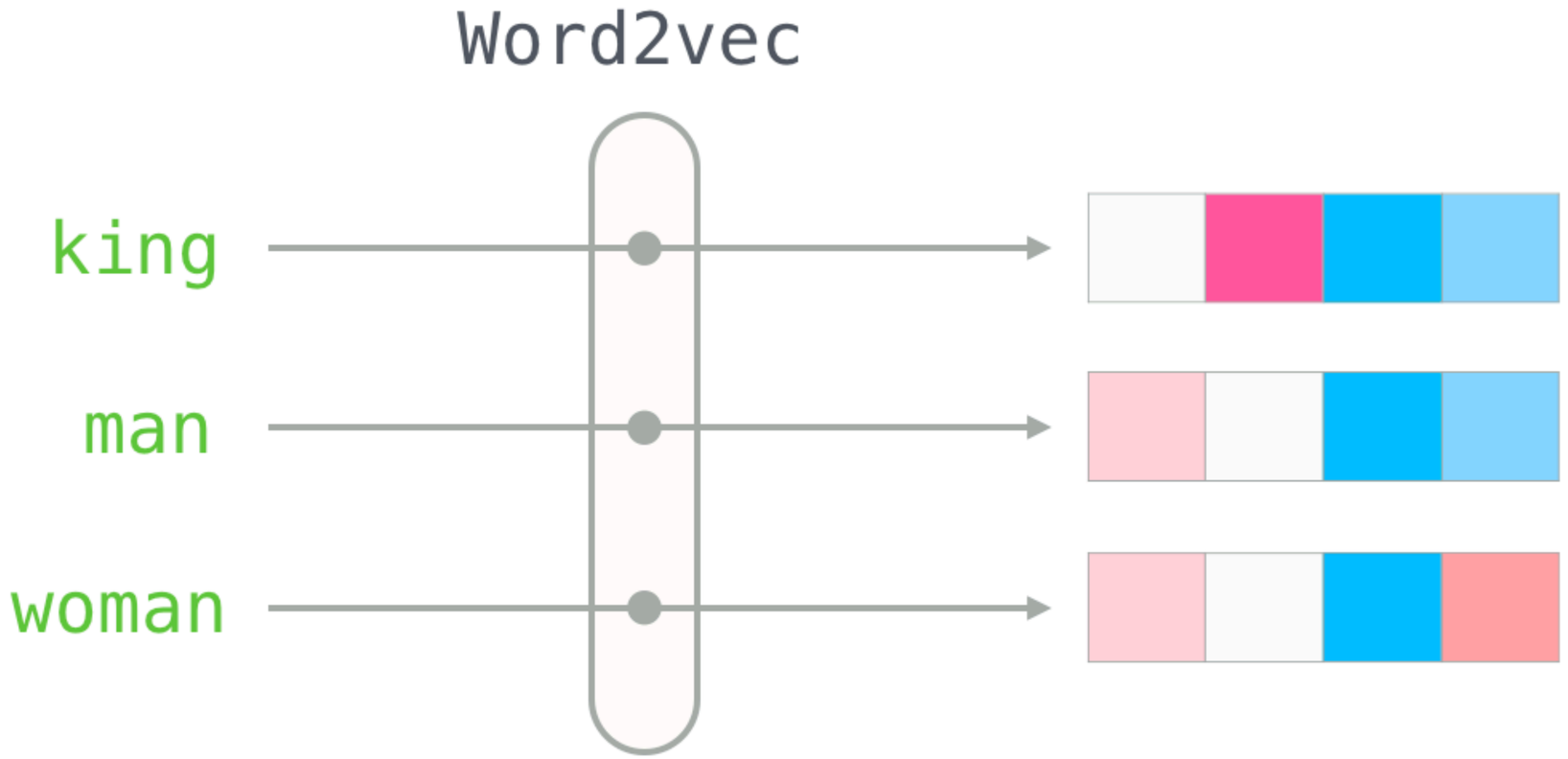
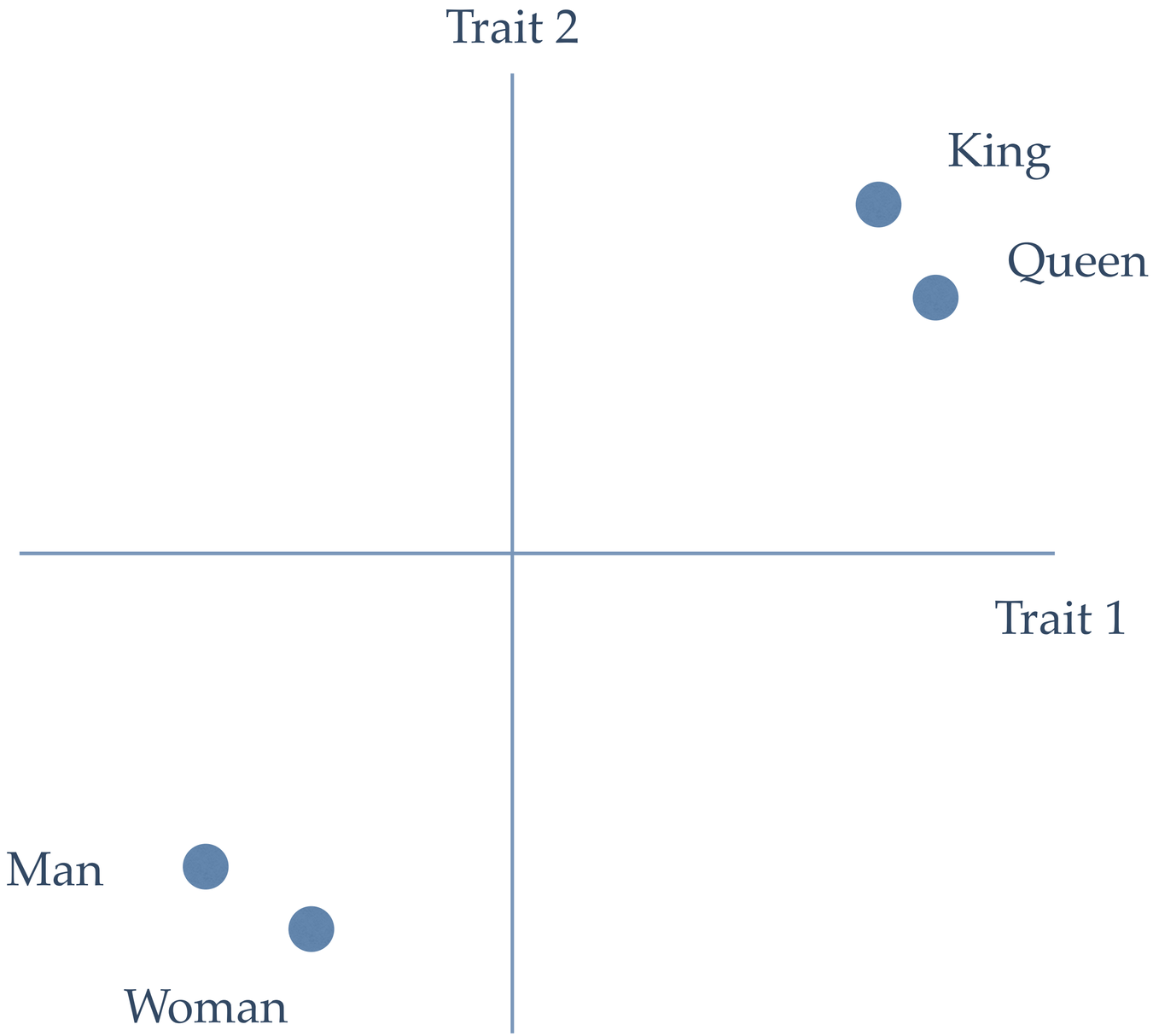


Image courtesy of Chris McCormick



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