

# Vector Semantics

Jelke Bloem & Giovanni Colavizza

Text Mining  
Amsterdam University College

February 28, 2023

# Announcements

- **Individual assignment 1 deadline: 03/03, 23:59**
- Reading assignment 2 (Word2Vec) next week: Monday 06/03, 23:59

# Overview

- 1 Vector semantics
- 2 Matrix representations
- 3 Calculating vectors
- 4 Similarity measures
- 5 Evaluation

## **Vector semantics**

# The quest for meaning

- The realm of *lexical semantics*.
- A **lemma** can be associated with multiple **word senses** (e.g., “mouse (N)”).
- The many facets of ‘meaning’:
  - ▶ *Propositional synonymy*: two words (senses) are synonyms if the truth conditions of a sentence do not change when we swap them.
  - ▶ *Word similarity*: some features are shared, but no synonyms. E.g., ‘cat’ and ‘dog’.
  - ▶ *Word relatedness*: no features are shared, but there is a relationship. E.g., ‘water’ and ‘bottle’.
  - ▶ Semantic frames or *topics*: topical structure in documents. E.g., ‘sport’ and ‘politics’.
  - ▶ *Connotations*, e.g., sentiment (positive, negative), tone (formal or not).

# Denotational vs distributional approaches

- Denotational approach: define (dictionary) meaning then apply definition. Meaning as dictionary index.
- Distributional approach: look at data to come up with meaning.  
**Distributional hypothesis:** “the amount of meaning difference between two words corresponds roughly to the amount of difference in their environments” (contexts of appearance). Harris, 1954.

# Vector Semantics

“The meaning of a word is its use in the language.” Wittgenstein, 1953.

- ① *??? is best when cooked just right.*
- ② *I prefer my ??? with abundant tomato sauce.*
- ③ *I eat ??? for lunch.*
- ④ *Would you like tomato sauce on your pizza?*
- ⑤ *We went to a pizza place for lunch.*
- ⑥ *Pizza should not be too cooked: it burns!*

Can you guess ???

# Vector Semantics

“The meaning of a word is its use in the language.” Wittgenstein, 1953.

- ① *Pasta is best when cooked just right.*
- ② *Pizza should not be too cooked: it burns!*
- ③ Vector semantics combines two intuitions:
  - ▶ **Distributional approach**: define a word by the contexts it occurs into.
  - ▶ **Vectorize it**: use vectors to represent word meaning, as a point in space.
- ④ **Feature engineering** for NLP: word vectors are increasingly used as features for other tasks.
- ⑤ (Word) vectors are usually referred as (word) **embeddings** in modern neural network literature.



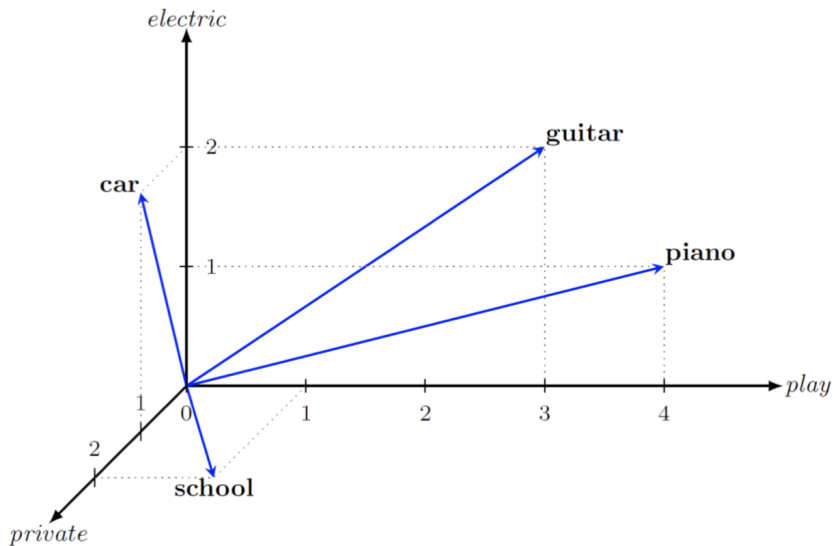
# Example

...ound and sonic power of a [new electric guitar played through] a guitar amp has play...  
...[Some electric guitar models feature] piezoelectric pickups...  
...[Playing guitar with a] pick produces a bright sound ...  
...ings, he is known for [playing fretless guitar in his] performances...  
...the neck of [a classical guitar is too] wide and the normal position ...  
...t in the centre of Bristol [playing the piano , I was] punched in the head while, a...  
...r in Houston, Texanstagram [playing the piano in his] flooded home after Hurrican H...  
... some supplies, he stopped to [play the piano that was] sitting in knee-high water ...  
...te and one black, who [played classical pianos together]...  
...The [first electric from the] late 1920s used metal strin...  
...technologies, for example [the electric and the] integration of mobile commun...  
...study had each driver of [each electric drive unimpeded], perform a task whil...  
...Honda to commence testing of [their new and the] American was no doubt more t...  
...many design considerations for [the new car were "safety] innovations, performanc...  
...would be possible if almost [all private cars requiring drivers], which are not in ...  
... who donate to groups [providing private school scholarships have] written pieces att...  
... that students participating [in private school choice programs] graduate high school...  
...s in the establishment of this [new high school , named the] Gavirate Business School...  
...Anna heads into her [final high school year before] university wanting somet...  
... but he can prevent them from [playing at school]

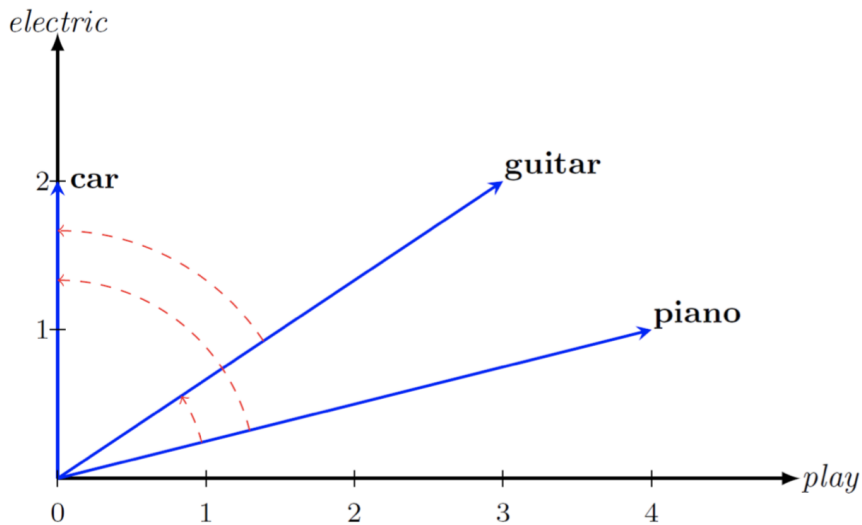
# Example

	play	electric	classical	private	high	...	the	new
guitar	3	2	1	0	0	...	0	1
piano	4	1	1	0	0	...	4	0
car	0	2	0	1	0	...	4	2
school	1	0	0	2	2	...	1	1

## Example



## Example



## **Matrix representations**

# Word-Document matrix

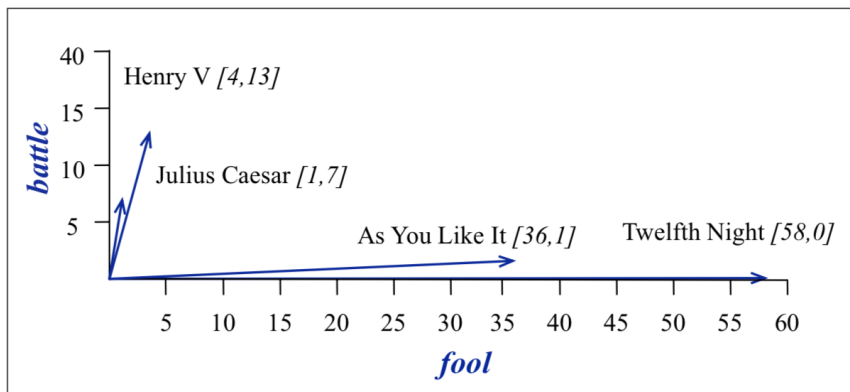
- We have a set of documents  $D$  and a vocabulary  $V$ .  $X$  is a  $|V| \times |D|$  matrix with word occurrences in documents.

	As You Like It	Twelfth Night	Julius Caesar	Henry V
<b>battle</b>	1	0	7	13
<b>good</b>	114	80	62	89
<b>fool</b>	36	58	1	4
<b>wit</b>	20	15	2	3

**Figure 6.2** The term-document matrix for four words in four Shakespeare plays. Each cell contains the number of times the (row) word occurs in the (column) document.

*Credit: J&M, ch. 6.*

# Word-Document matrix



**Figure 6.4** A spatial visualization of the document vectors for the four Shakespeare play documents, showing just two of the dimensions, corresponding to the words *battle* and *fool*. The comedies have high values for the *fool* dimension and low values for the *battle* dimension.

Credit: J&M, ch. 6.

# Word-Context matrix

- We have a set of words  $V$  and a set of contexts they occur into  $C$ , taken from our corpus of documents.  $X$  in this case is a  $|V| \times |C|$  matrix with word occurrences in contexts.
- The most intuitive context are co-occurrences with other words in  $V$ , within a certain **window**. In this case,  $X$  would be a  $|V| \times |V|$  matrix.

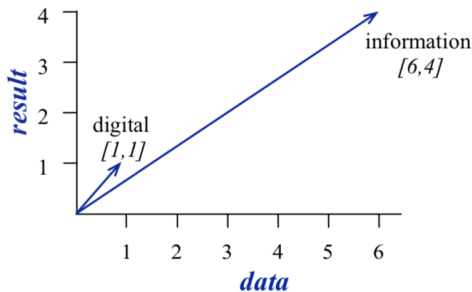
	aardvark	...	computer	data	pinch	result	sugar	...
apricot	0	...	0	0	1	0	1	
pineapple	0	...	0	0	1	0	1	
digital	0	...	2	1	0	1	0	
information	0	...	1	6	0	4	0	

**Figure 6.5** Co-occurrence vectors for four words, computed from the Brown corpus, showing only six of the dimensions (hand-picked for pedagogical purposes). The vector for the word *digital* is outlined in red. Note that a real vector would have vastly more dimensions and thus be much sparser.

Credit: J&M, ch. 6.



# Word-Context matrix



**Figure 6.6** A spatial visualization of word vectors for *digital* and *information*, showing just two of the dimensions, corresponding to the words *data* and *result*.

Credit: J&M, ch. 6.

# Types of co-occurrences

- **Surface co-occurrence:**

- ▶ a contextual word co-occurs with the target word as many times as the former appears in a collocational span (window) surrounding the latter.
- ▶ The span may be symmetric ( $[-5w, +5w]$ ) or asymmetric ( $[-5w, 0]$ ).

- **Textual co-occurrence:**

- ▶ words co-occur if they appear in the same text segment (e.g., a sentence, a paragraph, a web page ...).
- ▶ It usually does not matter how many times each word occur in each document.

- **Syntactic co-occurrence:**

- ▶ count word co-occurrences in a specific syntactic relation (e.g., verb-object, adjective-noun ...).

## Contingency tables

- Tabular representation of the observed frequencies between the variable whose values are reported in rows and the variable whose values are reported in columns.
- Intermediate step for some calculations we will see.
- If  $u$  is our target word and  $v$  is a contextual word:
  - ▶  $O_{11}$  – observed frequency of  $u$  and  $v$  (i.e.,  $f(u, v)$ ).
  - ▶  $R_1, R_2, C_1, C_2$  – marginal frequencies.
  - ▶  $R_1$  – absolute frequency of  $u$  (i.e.  $f(u)$ ).
  - ▶  $C_1$  – absolute frequency of  $v$  (i.e.  $f(v)$ ).
  - ▶  $N = O_{11} + O_{12} + O_{21} + O_{22}$  – sample size.

	$V = v$	$V \neq v$	
$U = u$	$O_{11}$	$O_{12}$	$= R_1$
$U \neq u$	$O_{21}$	$O_{22}$	$= R_2$
	$= C_1$	$= C_2$	$= N$

# Contingency tables – Surface co-occurrences

- $w1$  = “hat”;  $w2$  = “roll”.
- Collocational span:  $\pm 4$  words (i.e.  $[-4w, +4w]$ ).
- Spans cannot cross sentence boundaries.

A vast deal of coolness and a peculiar degree of judgement, are [requisite in catching a hat]. A man must not be precipitate, or he runs over it ; he must not rush into the opposite extreme, or he loses it altogether. There was a fine gentle [wind, and Mr. Pickwick's hat rolled sportively before it] . The wind puffed, and Mr. [Pickwick puffed, and the hat rolled over and over] as merrily as a lively porpoise in a strong tide ; and on it might have rolled, far beyond Mr. Pickwick's reach, had not its course been providentially stopped, just as that gentleman was on the point of resigning it to its fate.

# Contingency tables – Surface co-occurrences

- $w1 = \text{"hat"}; w2 = \text{"roll"}.$
- Collocational span:  $\pm 4$  words (i.e.  $[-4w, +4w]$ ).
- Spans cannot cross sentence boundaries.

*observed frequencies*

	<i>roll</i>	$\neg$ <i>roll</i>	
<i>hat</i>	2	18	20
$\neg$ <i>hat</i>	1	90	91
	3	108	111

**NOTE:**  $N$  equals the number of tokens in the corpus

# Contingency tables – Textual co-occurrences

- $w1 = \text{"hat"}; w2 = \text{"over"}$ .
- Unit is sentence (multiple occurrence is the same unit are ignored).

A vast deal of coolness and a peculiar degree of judgement, are requisite in catching a <b>hat</b>	hat	---
A man must not be precipitate, or he runs <b>over</b> it	---	over
he must not rush into the opposite extreme, or he loses it altogether	---	---
There was a fine gentle wind, and Mr. Pickwick's <b>hat</b> rolled sportively before it	hat	---
The wind puffed, and Mr. Pickwick puffed, and the <b>hat</b> rolled <b>over</b> and <b>over</b> as merrily as a lively porpoise in a strong tide	hat	over

## Contingency tables – Textual co-occurrences

- $w1 = \text{"hat"}; w2 = \text{"over"}$ .
- Unit is sentence (multiple occurrence of the same unit are ignored).

*observed frequencies*

	<i>over</i>	$\neg$ <i>over</i>	
<i>hat</i>	1	2	3
$\neg$ <i>hat</i>	1	1	2
	2	3	5

**NOTE:**  $N$  equals the number of text units

## (Discussion) questions: Phil. of science and linguistics

- Chomsky believes that only the deep why question matter in science, and only the insight these questions provide measure scientific success. However, since with every why question answered we only end up with a more fundamental why questions, do we ever experience true insight? Or are we just gaining more accurately describing models of reality without ever describing the fundamentals of reality itself?
- Does the success of a certain scientific approach immediately mean that it is the one we should take? (in reference to Norvig's essay – just because statistical/probabilistic approaches have shown success, does that mean they are more fruitful than mathematical ones?)
- I definitely learn something about learning or speaking a language by looking at statistical models, would that mean language acquisition can be not innate as well?
- How can the extent of contextualization for a single word be determined? In other words, how does one establish a criterion or goal for language models to stop contextualizing a word, and how is the adaptability and retention of learned vocabulary addressed? (based on the line: "It is difficult to obtain widespread agreement on how many senses should be allocated to different words" from one of the readings)



## (Discussion) questions: Models

- Norvig and Smith suggest that it is possible to combine statistical, contextual and discrete models into a sort of “hybrid” form, how exactly is that implemented? What aspects of language processing would be covered by which model?
- What is the future of language modelling in matter of non-verbal communication?
- Does Norvig agree or disagrees with Chomsky's view that a probabilistic model cannot be made to generate language?
- why is algorithmic modeling culture subscribed by only 2% of statisticians, since it can give a more detailed approximation of real language use? (algorithmic vs data statistical modeling)
- Norvig paper is over 12 years old, how far has this field advanced today?

## (Discussion) questions: Vectors

- So I am going with the assumption that the point of having vector based word storage is to have essentially a  $n$ th dimensional map of related words near each other that can then be located by the program and referenced. Are there normally a significant number of different planes, and what determines what a plane should be?
- I didn't understand the concept of \*dimensionality reduction\* in the section 'Words as Distributional Vectors: Context as Meaning'; could you go over this in class? -> ...
- why are computers not able to handle full vectors efficiently?

## Calculating vectors

# Families of vectors

- **Sparse vectors:** many zero values and high-dimensional spaces. E.g., weighted co-occurrence matrices (this class).
- **Dense vectors:** no zero values and smaller-dimensional spaces.
  - ▶ Dimensionality reduction (Latent Semantic Analysis or truncated Singular Value Decomposition, Principal Component Analysis, Non-negative Matrix Factorization, and many more): mostly we skip, there is a little in the next lab.
  - ▶ Neural-networks (Skip-gram, CBOW, GloVe): next class.

## (Better) quantifying association

**Raw co-occurrence frequency** is often not the optimal measure of association between a word and a context:

- we need a way to estimate to what extent a context word is particularly informative about a target word;
- frequencies are very skewed.
- A couple of solutions: **tf-idf** and **PPMI**.

# Term frequency - Inverse document frequency

- Tf-idf is the standard weighting scheme for term-document matrices.
- It is likely the most used weighting scheme in Information Retrieval.
- **Term frequency**  $tf(t, d)$  = the number of times term  $t$  occurs in document  $d$ . Many variants exist (e.g., using a log transform). It accounts for how frequent  $t$  is within the document collection.
- **Inverse document frequency**  $idf(t) = \log\left(\frac{|D|}{|D_t|}\right)$ ; where  $D$  is the collection of documents and  $D_t$  is the subset where term  $t$  appears once or more. It accounts for how 'discriminative'  $t$  is with respect to the document collection.

$$tfidf(t, d) = tf(t, d) \times idf(t)$$

## Term frequency - Inverse document frequency

	As You Like It	Twelfth Night	Julius Caesar	Henry V
<b>battle</b>	1	0	7	13
<b>good</b>	114	80	62	89
<b>fool</b>	36	58	1	4
<b>wit</b>	20	15	2	3

**Figure 6.2** The term-document matrix for four words in four Shakespeare plays. Each cell contains the number of times the (row) word occurs in the (column) document.

Word	df	idf
Romeo	1	1.57
salad	2	1.27
Falstaff	4	0.967
forest	12	0.489
battle	21	0.074
fool	36	0.012
good	37	0
sweet	37	0

## Term frequency - Inverse document frequency

	As You Like It	Twelfth Night	Julius Caesar	Henry V
<b>battle</b>	1	0	7	13
<b>good</b>	114	80	62	89
<b>fool</b>	36	58	1	4
<b>wit</b>	20	15	2	3

**Figure 6.2** The term-document matrix for four words in four Shakespeare plays. Each cell contains the number of times the (row) word occurs in the (column) document.

	As You Like It	Twelfth Night	Julius Caesar	Henry V
<b>battle</b>	0.074	0	0.22	0.28
<b>good</b>	0	0	0	0
<b>fool</b>	0.019	0.021	0.0036	0.0083
<b>wit</b>	0.049	0.044	0.018	0.022

**Figure 6.8** A tf-idf weighted term-document matrix for four words in four Shakespeare plays, using the counts in Fig. 6.2. Note that the idf weighting has eliminated the importance of the ubiquitous word *good* and vastly reduced the impact of the almost-ubiquitous word *fool*.



## (Positive) Pointwise Mutual Information

Intuition: in order to discriminate between informative and uninformative word-context associations, let us **take the expected frequency into account as a baseline**.

- The **expected frequency** of a (word, context) pair is a measure of how often a word would occur in a context if the two linguistic entities were statistically independent (i.e., if they were occurring by chance).
- The expected frequency can be estimated from the marginals in the contingency table:

$$E_{11} = \frac{f(u)f(v)}{N}$$

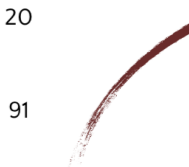
	$V = v$	$V \neq v$
$U = u$	$E_{11} = \frac{R_1 C_1}{N}$	$E_{12} = \frac{R_1 C_2}{N}$
$U \neq u$	$E_{21} = \frac{R_2 C_1}{N}$	$E_{22} = \frac{R_2 C_2}{N}$

# (Positive) Pointwise Mutual Information

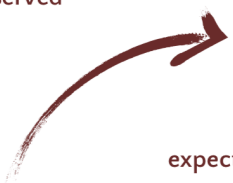
	<i>roll</i>	$\neg roll$
<i>hat</i>	2	18
$\neg hat$	1	90
	3	108

111

observed



# (Positive) Pointwise Mutual Information

	<i>roll</i>	$\neg$ <i>roll</i>			
<i>hat</i>	2	18	20		
$\neg$ <i>hat</i>	1	90	91		
	3	108	111	expected	

	<i>roll</i>	$\neg$ <i>roll</i>	
<i>hat</i>	0.54	19.46	
$\neg$ <i>hat</i>	2.46	88.54	

## (Positive) Pointwise Mutual Information

- Mutual Information provides a measure of independence of two random variables  $X$  and  $Y$ :

$$MI(X, Y) = \sum_{x \in X} \sum_{y \in Y} P(x, y) \log \frac{P(x, y)}{P(x)P(y)}$$

- Pointwise Mutual Information is the part related to two outcomes  $x$  and  $y$ :

$$PMI(x, y) = \log \frac{P(x, y)}{P(x)P(y)}$$

- Us, we are interested in a word-context pair,  $w$  and  $c$ :

$$PMI(w, c) = \log \frac{P(w, c)}{P(w)P(c)}$$

# Positive Pointwise Mutual Information

- We are not interested in joint events more unlikely than independent ones, thus we usually just consider the positive values of **PPMI**:

$$PPMI(w, c) = \max\left(0, \log \frac{P(w, c)}{P(w)P(c)}\right)$$

- Many variants to account for minor issues:
  - ▶ **Positive Local Mutual Information**: deals with the tendency of PPMI to emphasize rare events over frequent ones:

$$PLMI(w, c) = \max\left(0, f(w, c) \times PMI(w, c)\right)$$

	p(w,context)					p(w)
	computer	data	pinch	result	sugar	p(w)
apricot	0	0	0.05	0	0.05	0.11
pineapple	0	0	0.05	0	0.05	0.11
digital	0.11	0.05	0	0.05	0	0.21
information	0.05	.32	0	0.21	0	0.58
p(context)	0.16	0.37	0.11	0.26	0.11	

**Figure 6.9** Replacing the counts in Fig. 6.5 with joint probabilities, showing the marginals around the outside.

	computer	data	pinch	result	sugar
apricot	0	0	2.25	0	2.25
pineapple	0	0	2.25	0	2.25
digital	1.66	0	0	0	0
information	0	0.57	0	0.47	0

**Figure 6.10** The PPMI matrix showing the association between words and context words, computed from the counts in Fig. 6.5 again showing five dimensions. Note that the 0 ppmi values are ones that had a negative pmi; for example  $\text{pmi}(\text{information}, \text{computer}) = \log_2(.05/ (.16 * .58)) = -0.618$ , meaning that *information* and *computer* co-occur in this mini-corpus slightly less often than we would expect by chance, and with ppmi we replace negative values by zero. Many of the zero ppmi values had a pmi of  $-\infty$ , like  $\text{pmi}(\text{apricot}, \text{computer}) = \log_2(0/ (.16 * 0.11)) = \log_2(0) = -\infty$ .

## Example: Collostructions in linguistics

an accident	waiting to happen	- Idioms by The Free Dictio
accident	waiting to happen	Definition from Wiktionary,
, News 'This is a Ferguson	waiting to happen'	' Activists speak out again:
ce > Idioms > A > Accident	waiting to happen	Idiom: Accident waiting to
Trouble	Waiting to Happen	From Wikipedia, the free enc
26 April 2015. 'Nightmare	Waiting to Happen'	Quake Experts Gathered in
rt Start reading Accidents	Waiting to Happen	on the free Kindle Reading
Earthquake Was 'Nightmare	Waiting to Happen"	Slate Sign In Sign Up Slat
ions Disasters 'Nightmare	waiting to happen'	: Experts gathered in Nepal
an</span></div> A tragedy	waiting to happen	By Colin Stark Updated 1759
cal quake was a "nightmare	waiting to happen"	Breaking News US Secretary
Nepal quake was 'nightmare	waiting to happen'	Major earthquakes hit Nepa
umph of Life Contact Tarot	Waiting to Happen -	A Tarot Deck by Andrew Mc
berg Solution or Accident	Waiting to Happen:	The Public and Nuclear Pow
/ Angela Weight and Sanity	Waiting to Happen	don't live here anymore. Cor
of Things Is a Revolution	Waiting to Happen	The challenge of the IoT is
earthquake was a 'nightmare	waiting to happen'	says lead scientist Expert
Angela Weight Sanity	Waiting to Happen	Skip to content Home Book(s
cos Games Music En Deadman	Waiting To Happen	by nProcess Random Animatio
an accident	waiting to happen	- definition in the British
Bhuj-temblorA catastrophe	waiting to happen?	The author has posted comm
grade Winter Meeting trade	waiting to happen	between M's, Rockies JP Mor
: Author Says Disaster Was	Waiting to Happen	Don't Miss Out - Follow us
earthquake was a disaster	waiting to happen	Sitting on one of the most
gallery 10 Sports Injuries	Waiting to Happen	of Advertisement Skip this
en Noun (plural disasters	waiting to happen)	Something potentially very
st Classic FAILs accident	waiting to happen	Share on Facebook- Featur

## What is generally waiting to happen?

## Example: Collostructions in linguistics

	<i>accident</i>	$\neg$ <i>accident</i>	Row totals
[N waiting to happen]	14	21	35
$\neg$ [N waiting to happen]	8,606	10,197,659	10,206,265
Column totals	8,620	10,197,680	10,206,300

**Table:** Contingency table from Stefanowitsch & Gries (2003) for the noun *accident* in the [N waiting to happen] construction

$$P(\text{accident}, \text{waitingto happen}) = \frac{14}{10,206,300}$$

$$P(\text{accident}) = \frac{8,606}{10,206,300}$$

$$P(\text{waitingto happen}) = \frac{35}{10,206,300}$$

$$PMI(\text{accident}, \text{waitingto happen}) = \log \frac{0.00000137}{0.000843 * 0.00000342} = 474.38$$

Independence when PMI is 0, association when  $PMI > 0$



## Example: Collostructions in linguistics

Prediction	Prob.	Prob. Top-K
<b>event</b>	0.097763	0.2703
<b>disaster</b>	0.064560	0.1611
<b>accident</b>	0.059664	0.1394
explosion	0.049361	0.1390
<b>invasion</b>	0.016694	0.0486
<b>earthquake</b>	0.016525	0.0478
action	0.016206	0.0422
emergency	0.014662	0.0417
attack	0.013799	0.0403
miracle	0.013404	0.0371
adventure	0.011491	0.0324

The top eleven words from the average distribution of BERT's predictions over a series of X waiting to happen construction sentences from the British National Corpus

Viewing word associations in Dutch Google N-grams corpus:  
[https://urd2.let.rug.nl/~gosse/bin/Web1T5\\_assoc.perl](https://urd2.let.rug.nl/~gosse/bin/Web1T5_assoc.perl)

## Similarity measures

## Comparing vectors: the dot product

- Now we know how to calculate the first-order associations between words and how to use this information to create a *distributional representation* of each word.
- **Similarity measures** can be used to quantify the distance between two vectors in a space, and this can be used to estimate how similar the represented words are.
- Most vector similarity measure are based on the **dot (inner) product**:

$$\vec{u} \cdot \vec{v} = \sum_{i=1}^d u_i v_i = u_1 v_1 + u_2 v_2 + \cdots + u_d v_d$$

## Comparing vectors: Euclidean distance

- When used as a similarity metric, the dot product has a problem: it favors vectors with higher values (e.g., frequencies).
- It is the same issue you have with the Euclidean norm:

$$\|\vec{v}\| = \sqrt{\sum_{i=1}^d v_i^2}$$

- from which the Euclidean distance stems:

$$d(\vec{u}, \vec{v}) = \|\vec{u} - \vec{v}\| = \sqrt{\sum_{i=1}^d (u_i - v_i)^2}$$

- A similarity measure that is sensitive to frequency can sometimes work, yet other times it is the direction of vectors which is more important.

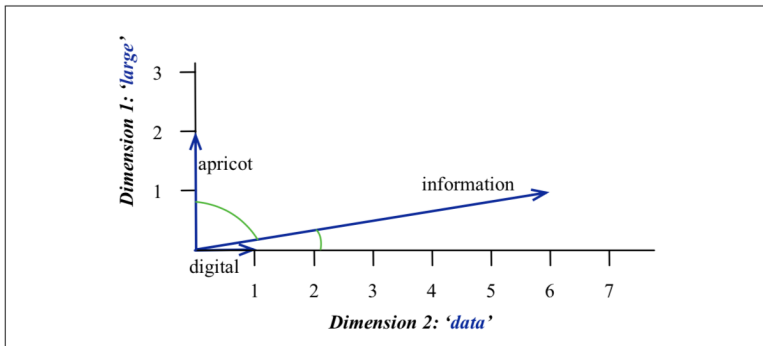
## Comparing vectors: Cosine distance

- A solution is to use the **cosine of the angle between the two vectors**:

$$\text{cosine}(\vec{u}, \vec{v}) = \frac{\vec{u} \cdot \vec{v}}{\|\vec{u}\| \|\vec{v}\|} = \frac{\sum_{i=1}^d u_i v_i}{\sqrt{\sum_{i=1}^d u_i^2} \sqrt{\sum_{i=1}^d v_i^2}}$$

- The cosine ranges between 1 and -1, taking value 0 for orthogonal vectors. Due to the fact that PPMIs and other frequencies are always non-negative, cosine ranges from 0 to 1 (identically directed vectors).

# Cosine



**Figure 6.7** A graphical demonstration of cosine similarity, showing vectors for three words (*apricot*, *digital*, and *information*) in the two dimensional space defined by counts of the words *data* and *large* in the neighborhood. Note that the angle between *digital* and *information* is smaller than the angle between *apricot* and *information*. When two vectors are more similar, the cosine is larger but the angle is smaller; the cosine has its maximum (1) when the angle between two vectors is smallest ( $0^\circ$ ); the cosine of all other angles is less than 1.

Credit: J&M, ch. 6.

# Cosine similarity

Similarity in Quine philosophy corpus

- ('epistemology', 1.0)
- ('naturalized', 0.6133719682693481)
- ('normative', 0.4007303714752197)
- ('doctrinal', 0.39382827281951904)
- ('epistemological', 0.37758421897888184)
- ('evolutionary', 0.34609323740005493)
- ('dominates', 0.3314409852027893)
- ('traces', 0.3263438642024994)
- ('irreducibility', 0.3175702691078186)
- ('psychological', 0.3145177364349365)



## Evaluation

# Association/Similarity

Two words or a word and a context, may have two kinds of associations:

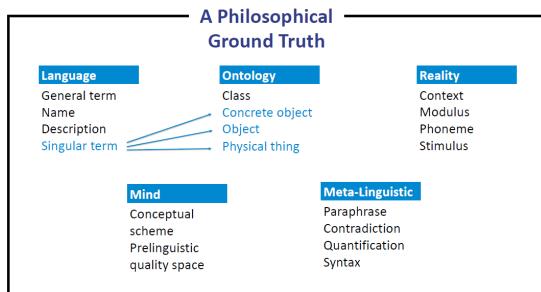
- **Syntagmatic associations** (first-order co-occurrence): how much two words appear one next to the other.
  - ① E.g., the association between a verb and its typical complements;
  - ② what is represented in a co-occurrence matrix.
- **Paradigmatic associations** (second-order co-occurrence) is similarity of context: how similar are the neighbors of the two words.
  - ① E.g., the association between two synonyms, or “wrote”, “said”, “remarked”.
  - ② what is estimated by calculating (first-order) vectors similarity.

# Evaluation of vectors

- The most common evaluation is to test their performance on similarity tasks.
- Correlation between algorithm and human word similarity ratings:
  - ① *WordSim-353*: 353 noun pairs rated on a 0-10 scale.  $\text{sim}(\text{"plane"}, \text{"car"}) = 5.77$ .
  - ② *SimLex-999*: similarity of noun, adjectives and adjectives pairs. *Assignment 3*.
- Taking *TOEFL* multiple-choice vocabulary tests
  - ① Levied is closest in meaning to: imposed, believed, requested, correlated.
- Judgments in context
  - ① *Stanford Contextual Word Similarity (SCWS)* dataset: human judgments on 2,003 pairs of nouns, verbs, and adjectives in their sentence context.

# Intrinsic evaluation of vectors

- Comparison to human word similarity ratings
- Comparison to gold standard



- Post-hoc evaluation

# Intrinsic evaluation of vectors

- Comparison to human word similarity ratings
- Comparison to gold standard
- Post-hoc evaluation
  - ▶ Synonym detection task

---

**What word is most related to 'Information' ?**

---

- |                 |   |
|-----------------|---|
| a) Learning     | b) Reductions                                   |
| c) Collateral   | d) Application                                  |
| e) Ordered Pair | f) None of these words is even remotely related |
- 

- ▶ Coherence task

---

**What word does not belong to the group?**

---

- |   |             |
|---|-------------|
| a) Numbers  | b) Pronouns |
| c) Subtraction                                      | d) Actually |
| e) No coherent group can be formed from these words |             |
-

# Extrinsic evaluation of vectors

- Use vectors to perform some other task
- For example: Information retrieval

4. (0.8002) 9783772804557\_sect62\_sent1247 context query

[einfach 0.33][einfachen 0.16][zusammengesetzt 0.17]

Alle Begriffe sind entweder einfach, oder aus mehreren einfachen zusammengesetzt.

5. (0.7994) Wissenschaftslehre\_sect370\_para2287\_sent14029 context query

[zusammengesetzt 0.17][einfach 0.33]

Ein solcher Beweis mag zusammengesetzt, ein anderer einfach heißen.

- Use cosine similarity to rank results and evaluate those results

# Lab

- Extracting Co-occurrences
- Sparse Vectors
- Calculating Cosine Similarities
- Visualizing Similarities
- Evaluating your Model
- Working with Dense Vectors