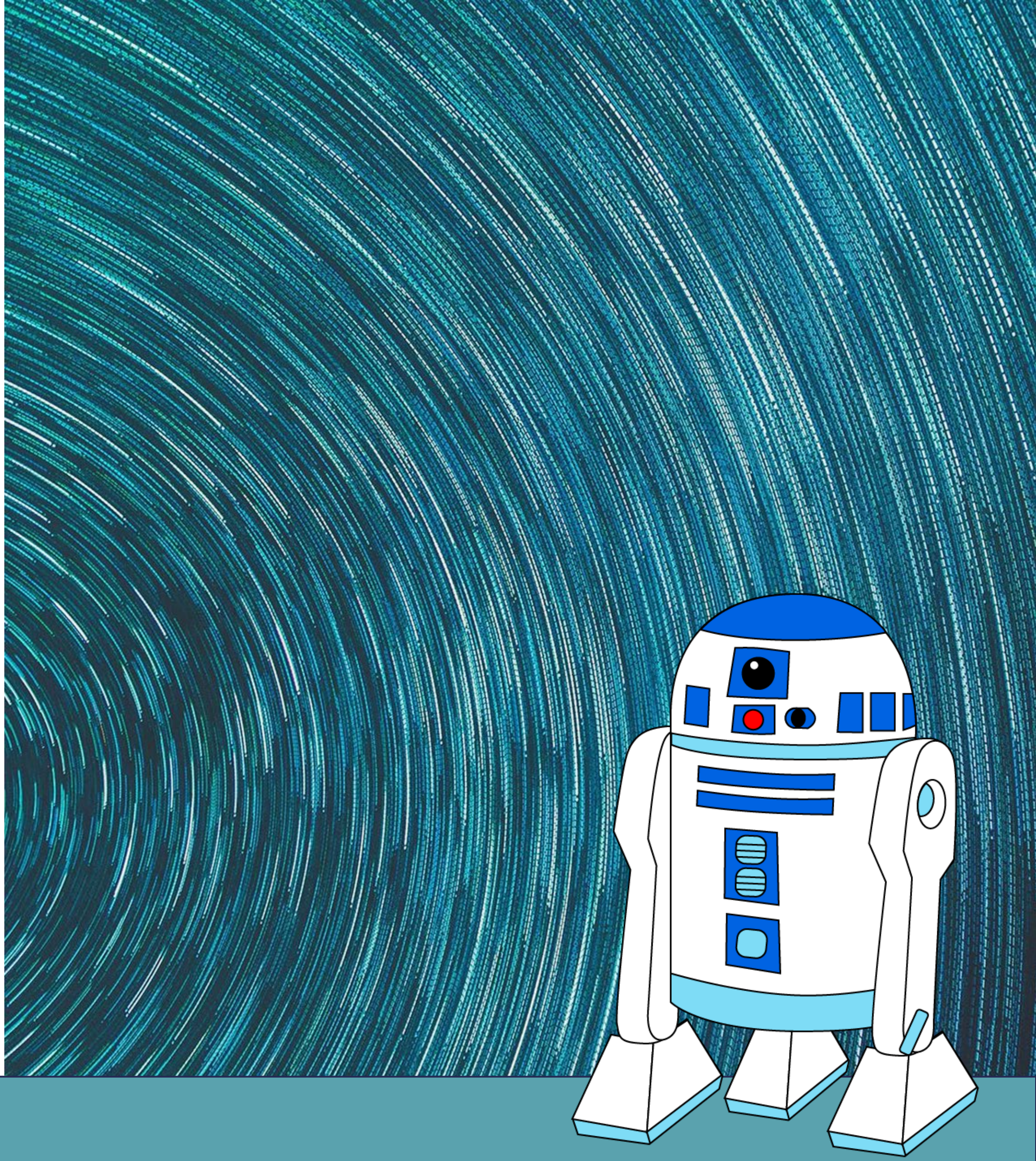
CIS 421/521:
ARTIFICIAL INTELLIGENCE

NLP: Vector Semantics

Jurafsky and Martin Chapter 6



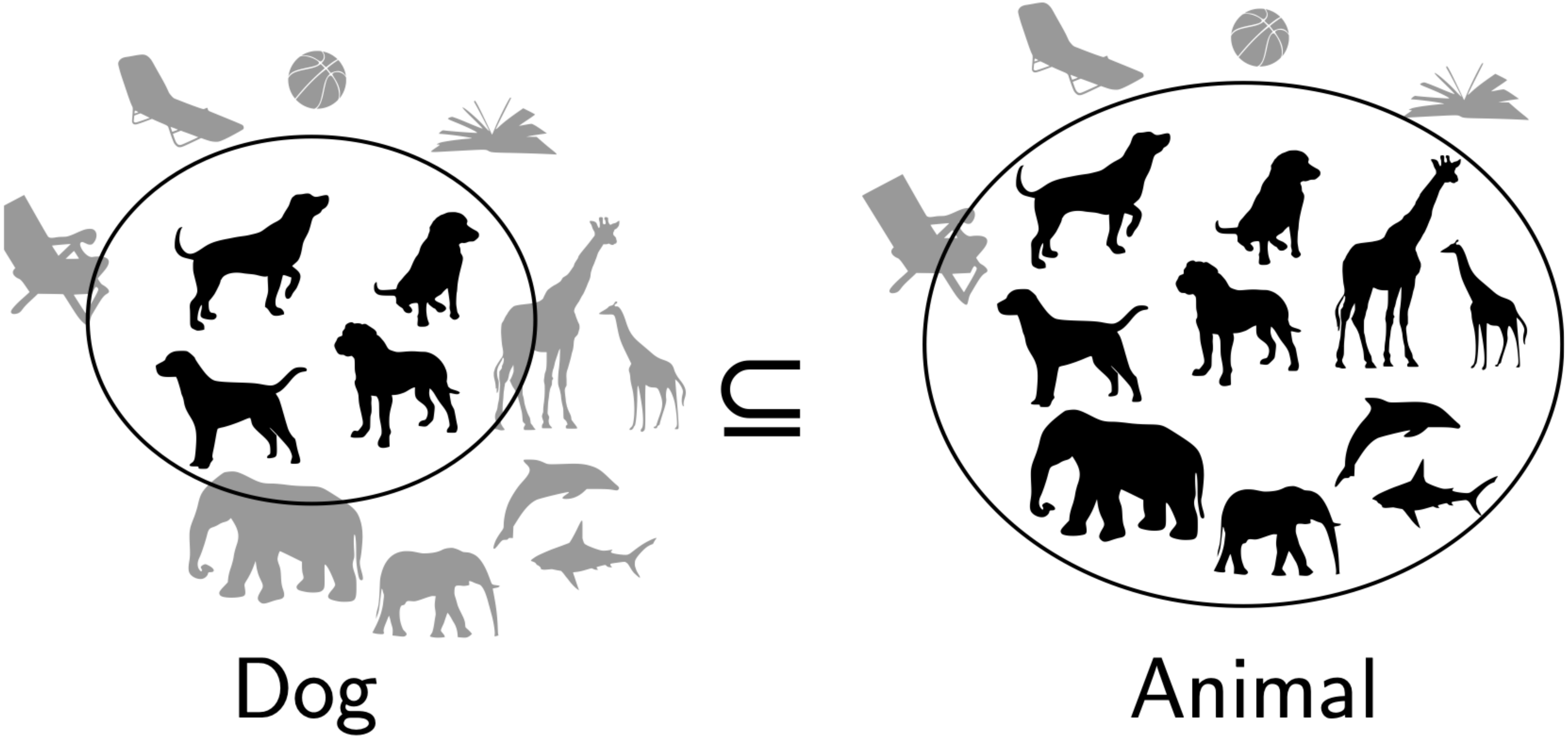
Penn Engineering
UNIVERSITY of PENNSYLVANIA

Word Meaning

- How should we **represent** the **meaning** of a word?
- In N-gram LMs we represented words as a string of letters or as an index in a vocabulary list.
- Ideally, we want a meaning representation to encode:

1. **Synonyms** – words that have similar meanings
2. **Antonyms** – words that have opposite meanings
3. **Connotations** – words that are positive or negative
4. **Semantic Roles** – *buy, sell*, and *pay* are different parts of the same underlying *purchasing* event
5. Support for **entailment**

Entailment in formal semantics



Entailment in formal semantics

All animals have an ulnar artery

$\Rightarrow$

All dogs have an ulnar artery

+ Mathematically well-understood
+ Powerful machinery for handling logical operations
- Knowledge must come from somewhere else

Noun

- S: (n) **dog**, domestic dog, Canis familiaris (a member of the genus Canis (probably descended from the common wolf) that has been domesticated by man since prehistoric times; occurs in many breeds) *"the dog barked all night"*
- S: (n) frump, **dog** (a dull unattractive unpleasant girl or woman) *"she got a reputation as a frump"; "she's a real dog"*
- S: (n) **dog** (informal term for a man) *"you lucky dog"*
- S: (n) cad, bounder, blackguard, **dog**, hound, heel (someone who is morally reprehensible) *"you dirty dog"*
- S: (n) frank, frankfurter, hotdog, hot dog, **dog**, wiener, wienerwurst, weenie (a smooth-textured sausage of minced beef or pork usually smoked; often served on a bread roll)
- S: (n) pawl, detent, click, **dog** (a hinged catch that fits into a notch of a ratchet to move a wheel forward or prevent it from moving backward)
- S: (n) andiron, firedog, **dog**, dog-iron (metal supports for logs in a fireplace) *"the andirons were too hot to touch"*

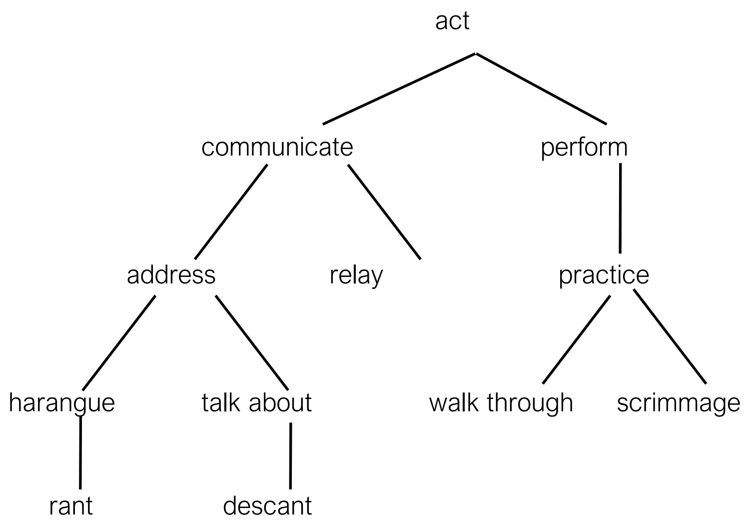
Verb

Noun

- S: (n) **dog**, domestic dog, Canis familiaris (a member of the genus Canis (probably descended from the common wolf) that has been domesticated by man since prehistoric times; occurs in many breeds) *"the dog barked all night"*
  - *direct hyponym* / *full hyponym*
  - *part meronym*
  - *member holonym*
  - ***direct hypernym*** / *inherited hypernym* / *sister term*
    - S: (n) canine, canid (any of various fissiped mammals with nonretractile claws and typically long muzzles)
    - S: (n) domestic animal, domesticated animal (any of various animals that have been tamed and made fit for a human environment)
- S: (n) frump, **dog** (a dull unattractive unpleasant girl or woman) *"she got a reputation as a frump"; "she's a real dog"*
- S: (n) **dog** (informal term for a man) *"you lucky dog"*
- S: (n) cad, bounder, blackguard, **dog**, hound, heel (someone who is morally reprehensible) *"you dirty dog"*
- S: (n) frank, frankfurter, hotdog, hot dog, **dog**, wiener, wienerwurst, weenie

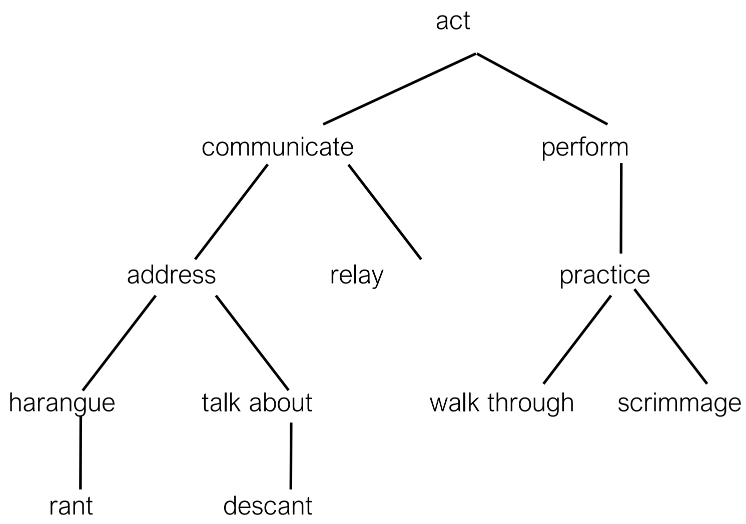
- S: (n) canine, canid (any of various fissiped mammals with nonretractile claws and typically long muzzles)
    - S: (n) carnivore (a terrestrial or aquatic flesh-eating mammal) *"terrestrial carnivores have four or five clawed digits on each limb"*
        - S: (n) placental, placental mammal, eutherian, eutherian mammal (mammals having a placenta; all mammals except monotremes and marsupials)
            - S: (n) mammal, mammalian (any warm-blooded vertebrate having the skin more or less covered with hair; young are born alive except for the small subclass of monotremes and nourished with milk)
                - S: (n) vertebrate, craniate (animals having a bony or cartilaginous skeleton with a segmented spinal column and a large brain enclosed in a skull or cranium)
                    - S: (n) chordate (any animal of the phylum Chordata having a notochord or spinal column)
                        - S: (n) animal, animate being, beast, brute, creature, fauna (a living

Lexical Semantics

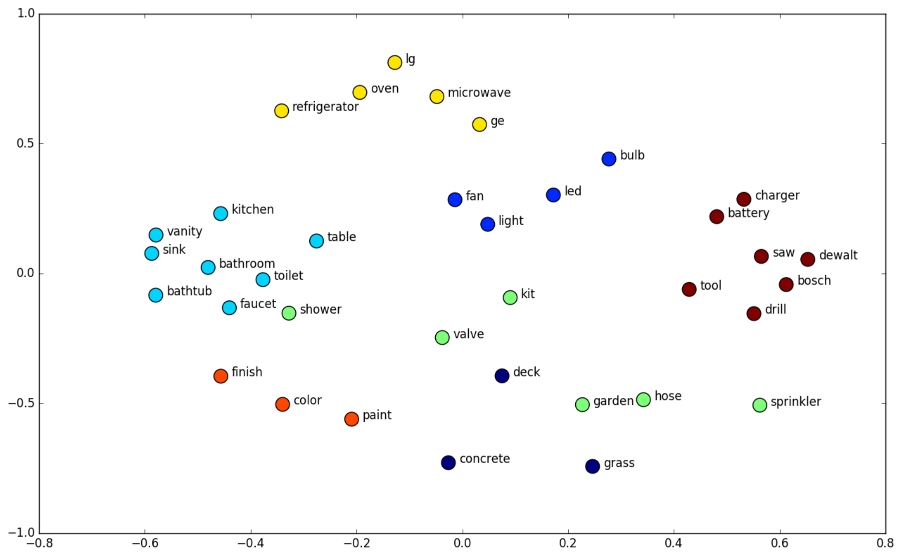


WordNet

Lexical Semantics

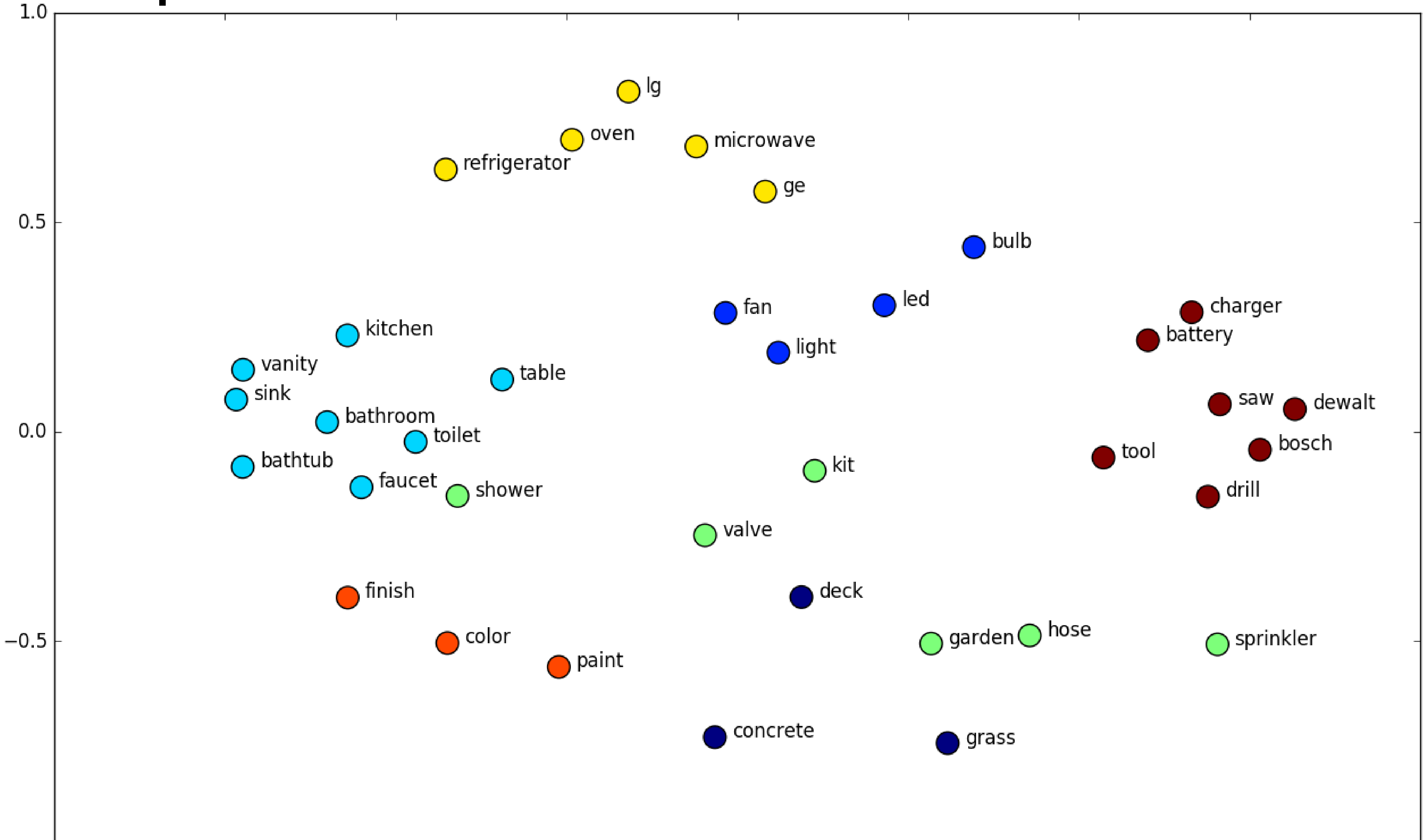


WordNet



Vector Space Models

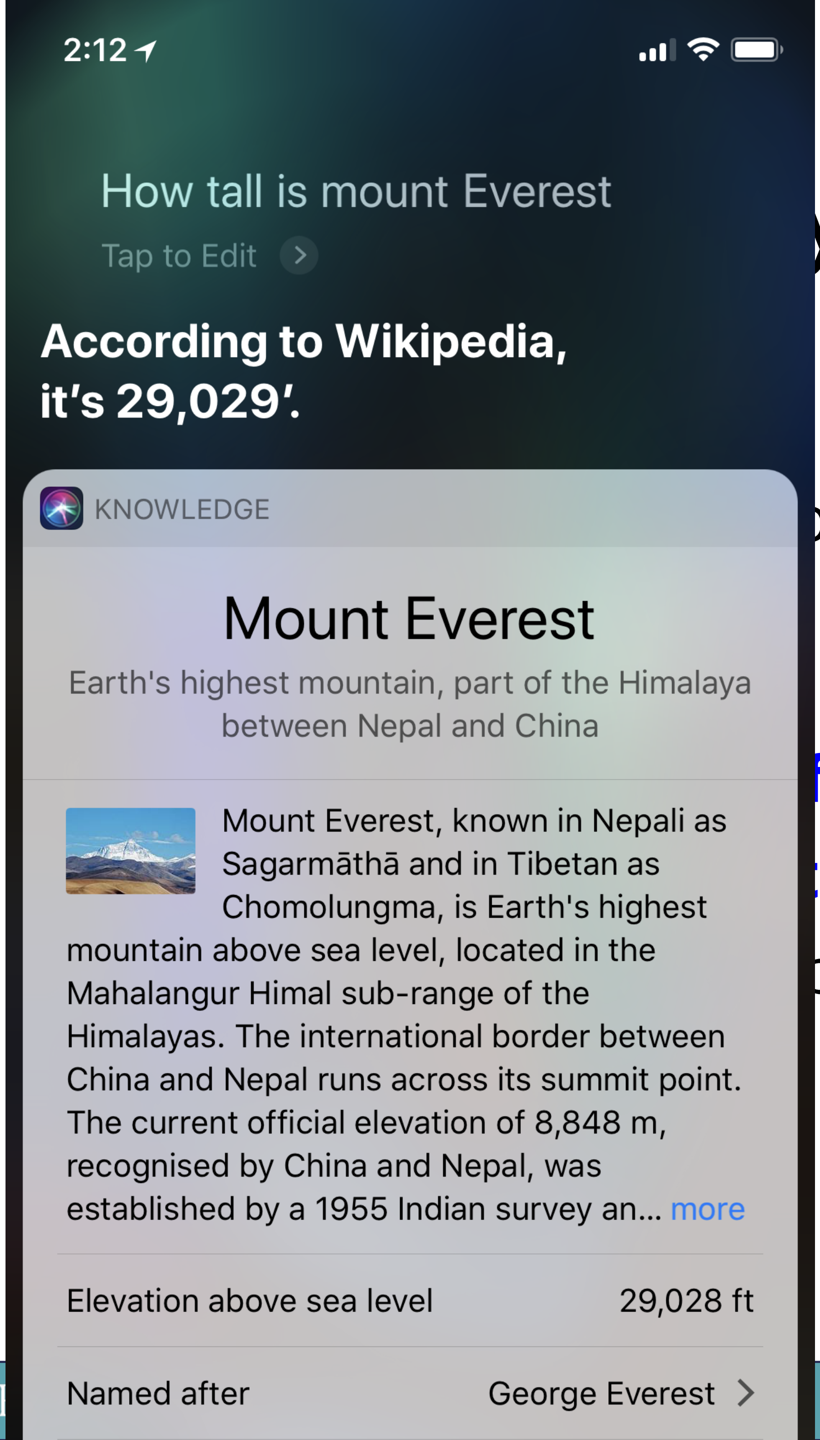
Vector Space Models

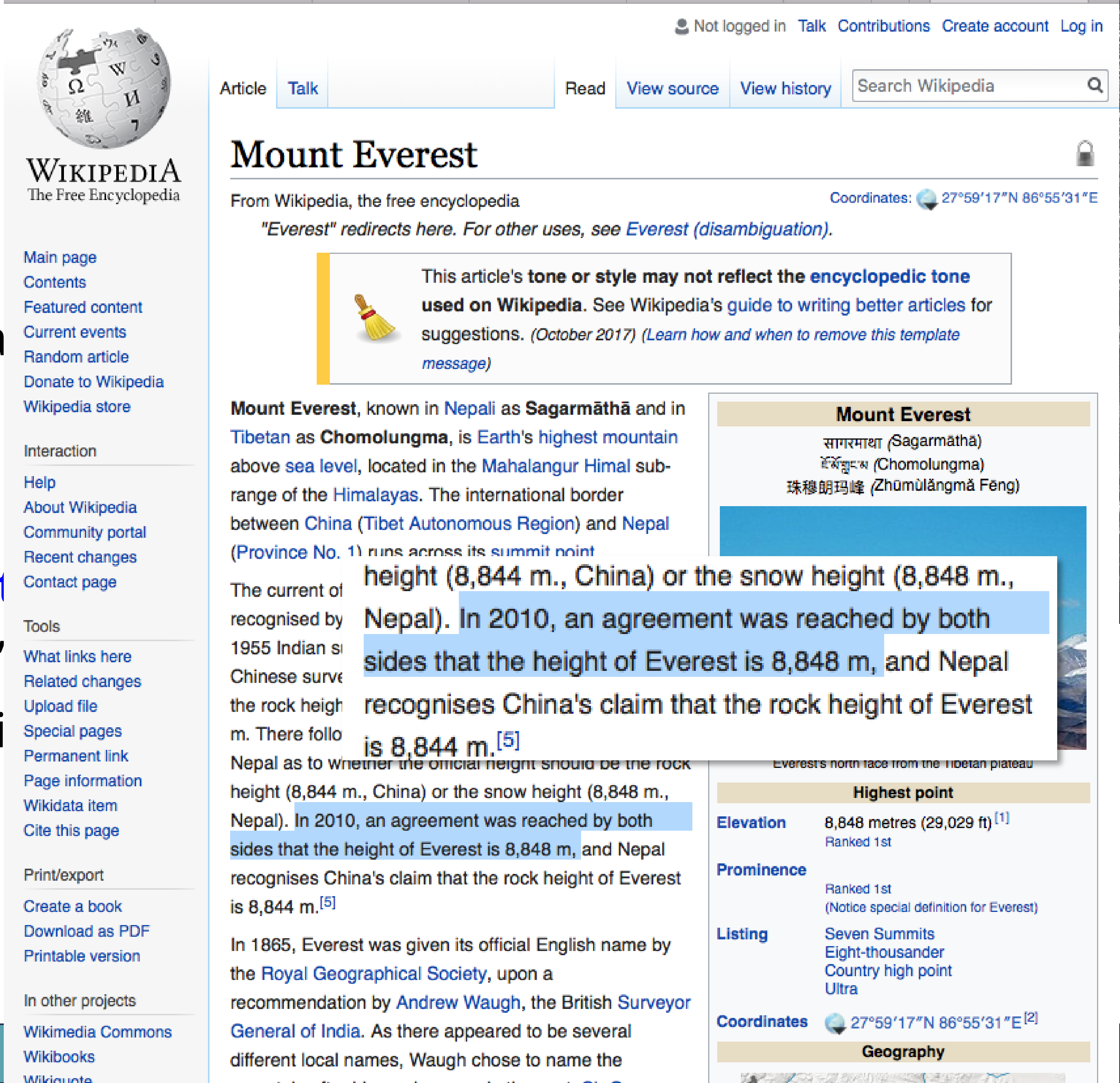


Word similarity

- Most words don’t have many **synonyms**, but they do have a lot of **similar** words. *Cat* is not a synonym of *dog*, but *cats* and *dogs* are certainly similar words.

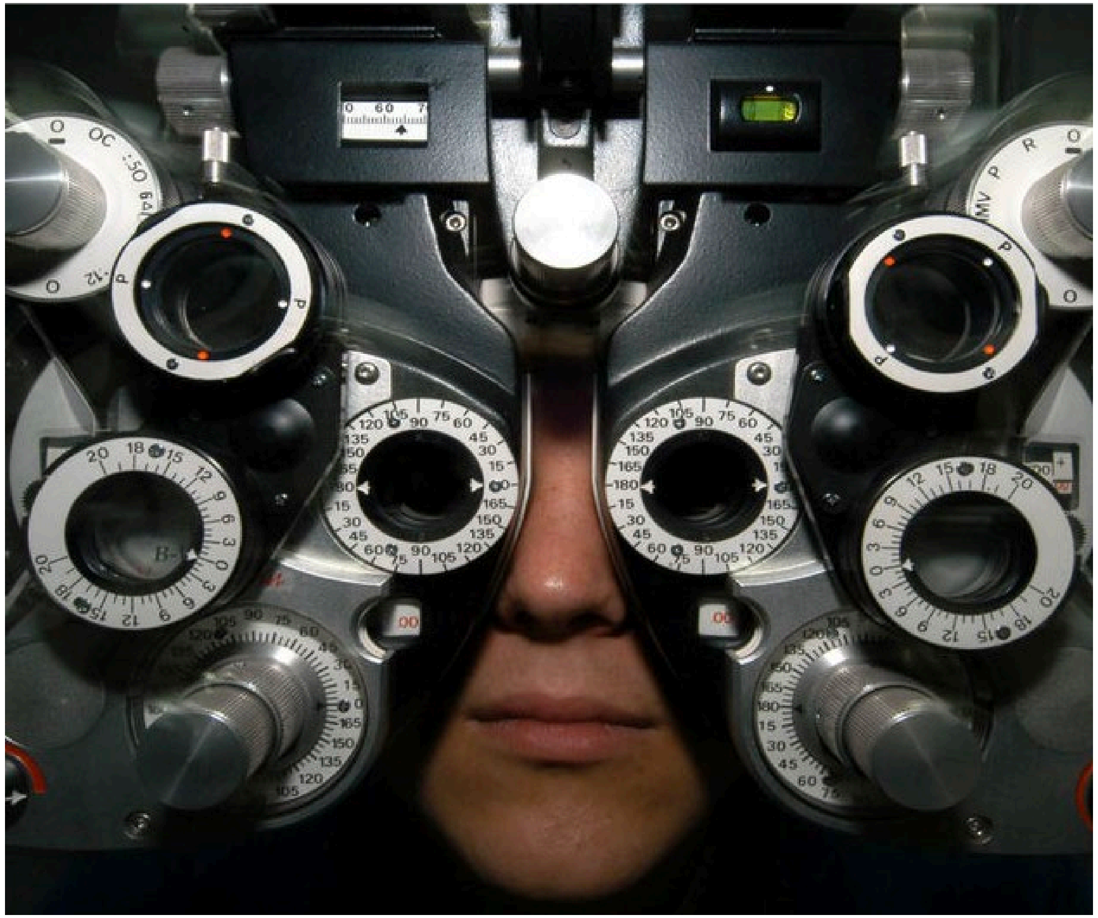
  - “**fast**” is similar to “**rapid**”
  - “**tall**” is similar to “**height**”
  - Useful for applications like question answering


2:12
How tall is mount Everest
Tap to Edit
According to Wikipedia, it's 29,029′.
KNOWLEDGE
Mount Everest
Earth's highest mountain, part of the Himalaya between Nepal and China
Mount Everest, known in Nepali as Sagarmāthā and in Tibetan as Chomolungma, is Earth's highest mountain above sea level, located in the Mahalangur Himal sub-range of the Himalayas. The international border between China and Nepal runs across its summit point. The current official elevation of 8,848 m, recognised by China and Nepal, was established by a 1955 Indian survey an... more
Elevation above sea level 29,028 ft
Named after George Everest


Not logged in Talk Contributions Create account Log in
Wikipedia
The Free Encyclopedia
Article Talk Read View source View history Search Wikipedia
Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store
Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page
Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item
Cite this page
Print/export
Create a book
Download as PDF
Printable version
In other projects
Wikimedia Commons
Wikibooks
Mount Everest
From Wikipedia, the free encyclopedia
Coordinates: 27°59′17″N 86°55′31″E
"Everest" redirects here. For other uses, see Everest (disambiguation).
This article's tone or style may not reflect the encyclopedic tone used on Wikipedia. See Wikipedia's guide to writing better articles for suggestions. (October 2017) (Learn how and when to remove this template message)
Mount Everest, known in Nepali as Sagarmāthā and in Tibetan as Chomolungma, is Earth's highest mountain above sea level, located in the Mahalangur Himal sub-range of the Himalayas. The international border between China (Tibet Autonomous Region) and Nepal (Province No. 1) runs across its summit point.
height (8,844 m., China) or the snow height (8,848 m., Nepal). In 2010, an agreement was reached by both sides that the height of Everest is 8,848 m, and Nepal recognises China's claim that the rock height of Everest is 8,844 m.[5]
Nepal as to whether the official height should be the rock height (8,844 m., China) or the snow height (8,848 m., Nepal). In 2010, an agreement was reached by both sides that the height of Everest is 8,848 m, and Nepal recognises China's claim that the rock height of Everest is 8,844 m.[5]
In 1865, Everest was given its official English name by the Royal Geographical Society, upon a recommendation by Andrew Waugh, the British Surveyor General of India. As there appeared to be several different local names, Waugh chose to name the
Mount Everest
सगरमाथा (Sagarmāthā)
(Chomolungma)
珠穆朗玛峰 (Zhūmùlǎngmǎ Fēng)
Everest's north face from the Tibetan plateau
Highest point
Elevation 8,848 metres (29,029 ft)[1]
Ranked 1st
Prominence
Ranked 1st
(Notice special definition for Everest)
Listing Seven Summits
Eight-thousander
Country high point
Ultra
Coordinates 27°59′17″N 86°55′31″E[2]
Geography

Distributional Hypothesis

- If we consider *optometrist* and *eye-doctor* we find that, as our corpus of utterances grows, these two occur in almost the same environments. In contrast, there are many sentence environments in which *optometrist* occurs but *lawyer* does not...
- It is a question of the relative frequency of such environments, and of what we will obtain if we ask an informant to substitute any word he wishes for optometrist (not asking what words have the same meaning).
- These and similar tests all measure the probability of particular environments occurring with particular elements... If A and B have almost identical environments we say that they are synonyms.
- –Zellig Harris (1954)



Intuition of distributional word similarity

- Nida (1975) example:

  A bottle of **tesgüino** is on the table
  Everybody likes **tesgüino**
  Tesgüino makes you drunk
  We make **tesgüino** out of corn.
- From context words humans can guess **tesgüino** means *an alcoholic beverage like beer*
- Intuition for algorithm:
  Two words are similar if they have similar word contexts.

History of Vector Space Models

- Vector Space Models were initially developed in the SMART information retrieval system (Salton, 1971)
- Each document in a collection is represented as point in a space (a vector in a vector space)
- A user's query is a pseudo-document and is represented as a point in the same space as the documents
- Perform IR by retrieving documents whose vectors are close together in this space to the query vector

Term-Document Matrix

	D1	D2	D3	D4	D5
abandon					
abdicate					
abhor					
academic					
…					
zygodactyl					
zymurgy					

Term-Document Matrix

	D1	D2	D3	D4	D5
abandon					
abdicate					
abhor					
academic					
…					
zygodactyl					
zymurgy					

Each column vector
represents a Document

Term-Document Matrix

	D1	D2	D3	D4	D5
abandon					
abdicate					
abhor					
academic					
…					
zygodactyl					
zymurgy					

Each row vector represents a Term

Penn Engineering

Term-Document Matrix

	D1	D2	D3	D4	D5
abandon					
abdicate					
abhor					
academic					
…					
zygodactyl					
zymurgy					

The value in a cell is
based on how often that term
occurred in that document

Term-Document Matrix

	D1	D2	D3	D4	D5
abandon					
abdicate					
abhor					
academic					
…					
zygodactyl					
zymurgy					

The length of the
document vectors
is the size of the vocabulary

Term-Document Matrix

	D1	D2	D3	D4	D5
abandon					
abdicate					
abhor					
academic					
…					
zygodactyl					
zymurgy					

Document vectors
can be sparse
(most values are 0)

Term-Document Matrix

	D1	D2	D3	D4	D5
abandon					
abdicate					
abhor					
academic					
…					
zygodactyl					
zymurgy					

We can measure
how similar two
documents are
by comparing their
column vectors

What can document similarity let you do?

W ection

MAINFRAMES

Mainframes are primarily referred to large computers with rapid, advanced processing capabilities that can execute and perform tasks equivalent to many Personal Computers (PCs) machines networked together. It is characterized with high quantity Random Access Memory (RAM), very large secondary storage devices, and high-speed processors to cater for the needs of the computers under its service.

Consisting of advanced components, mainframes have the capability of running multiple large applications required by many and most enterprises and organizations. This is one of its advantages. Mainframes are also suitable to cater for those applications (programs) or files that are of very high demand by its users (clients). Examples of such organizations and enterprises using mainframes are online shopping websites such as Ebay, Amazon, and computing-giant

MAINFRAMES

Mainframes usually are referred those computers with fast, advanced processing capabilities that could perform by itself tasks that may require a lot of Personal Computers (PC) Machines. Usually mainframes would have lots of RAMs, very large secondary storage devices, and very fast processors to cater for the needs of those computers under its service.

Due to the advanced components mainframes have, these computers have the capability of running multiple large applications required by most enterprises, which is one of its advantage. Mainframes are also suitable to cater for those applications or files that are of very large demand by its users (clients). Examples of these include the large online shopping websites -i.e. : Ebay, Amazon, Microsoft, etc.

Term-Document Matrix

	D1	D2	D3	D4	D5
abandon					
abdicate					
abhor					
academic					
...					
zygodactyl					
zymurgy					

What does comparing two row vectors do?

Vector comparisons

	doc_X	doc_Y
A	2	4
B	10	15
C	14	10

Vector comparisons

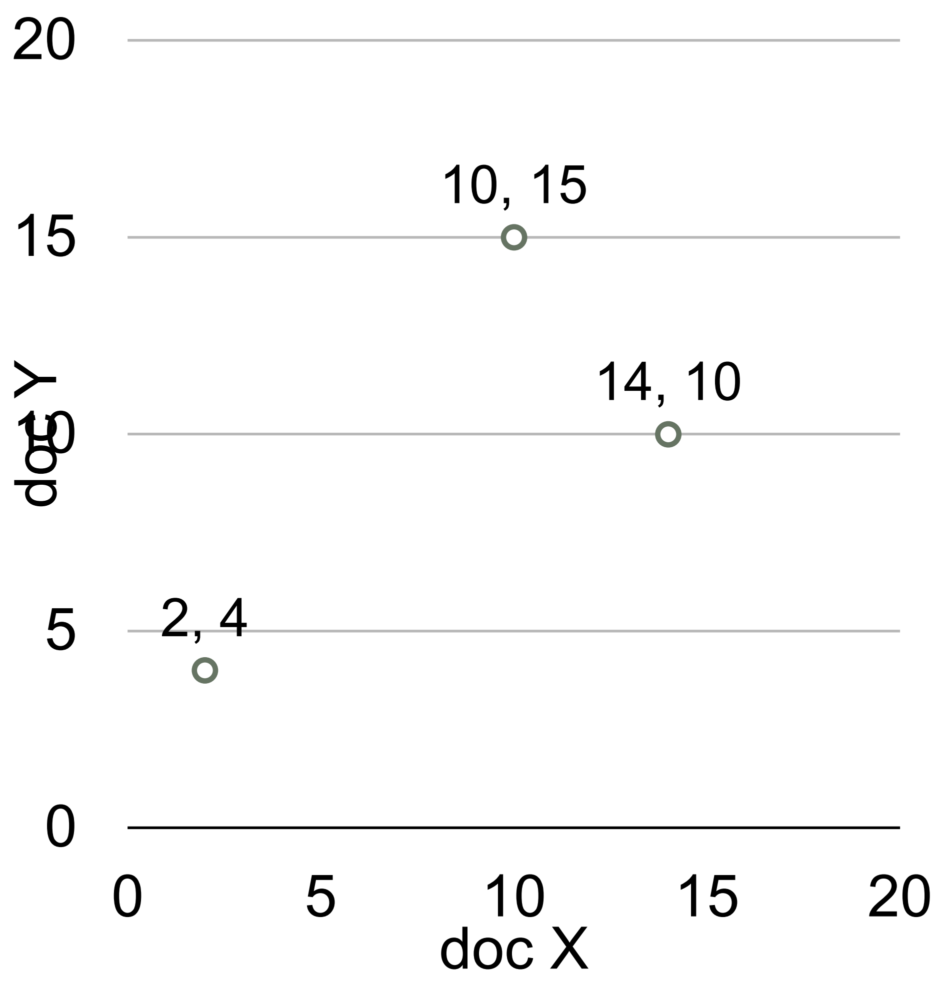
	doc_X	doc_Y
A	2	4
B	10	15
C	14	10

doc_Y is a positive movie review
doc_X is a less positive movie review

A = "superb" positive / low frequency
B = "good" positive / high frequency
C = "disappointing" negative / high frequency

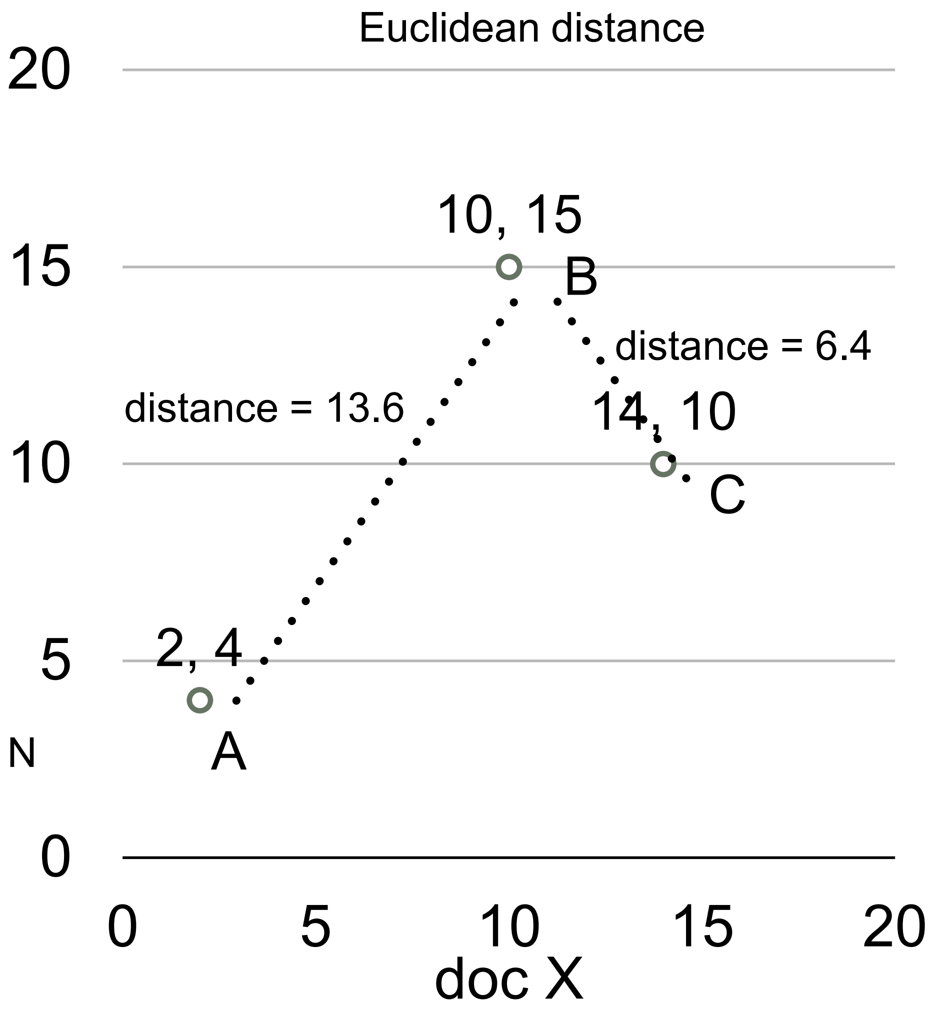
Vector comparisons

	doc_X	doc_Y
A	2	4
B	10	15
C	14	10



Vector comparisons

	docX	docY
A	2	4
B	10	15
C	14	10



Euclidean distance : vectors u, v of dimension N

$$\sqrt{\sum_{i=1}^{N} |u_i - v_i|^2}$$

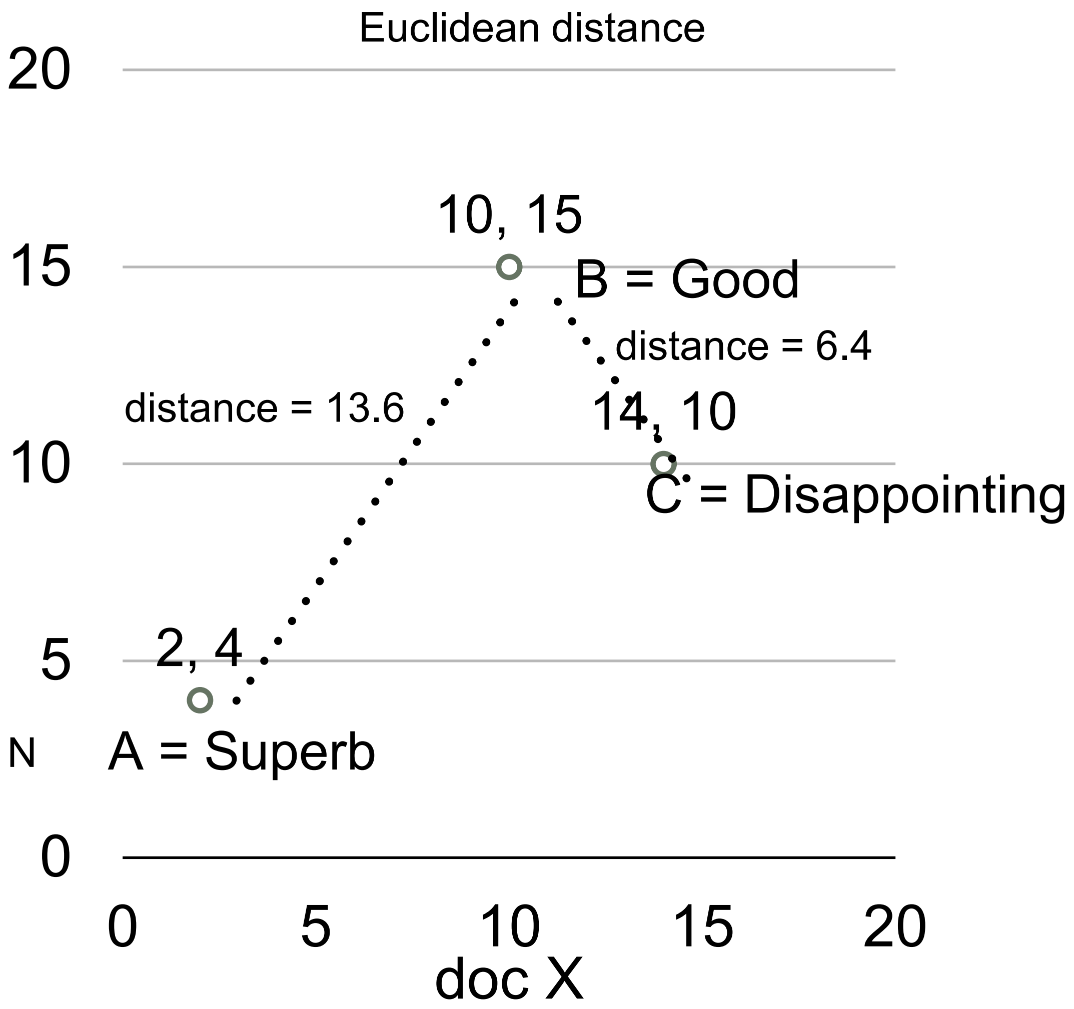
Vector comparisons

Oh no! Good is closer to Disappointing than to Superb.

	docX	docY
A	2	4
B	10	15
C	14	10

Euclidean distance : vectors u, v of dimension N

$$\sqrt{\sum_{i=1}^{N} |u_i - v_i|^2}$$



Penn Engineering

Vector L2 (length) Normalization

	doc_X	doc_Y	$\|\|u\|\|$
A	2	4	4.47
B	10	15	18.02
C	14	10	17.20

$$||u|| = \sqrt{\Sigma_{i=1}^{n} u_i^2}$$

Vector L2 (length) Normalization

	doc_X	doc_Y
A	2/4.47	4/4.47
B	10/18.02	15/18.02
C	14/17.2	10/17.2

$\|\|u\|\|$
4.47
18.02
17.20

$$||u|| = \sqrt{\Sigma_{i=1}^{n} u_i^2}$$

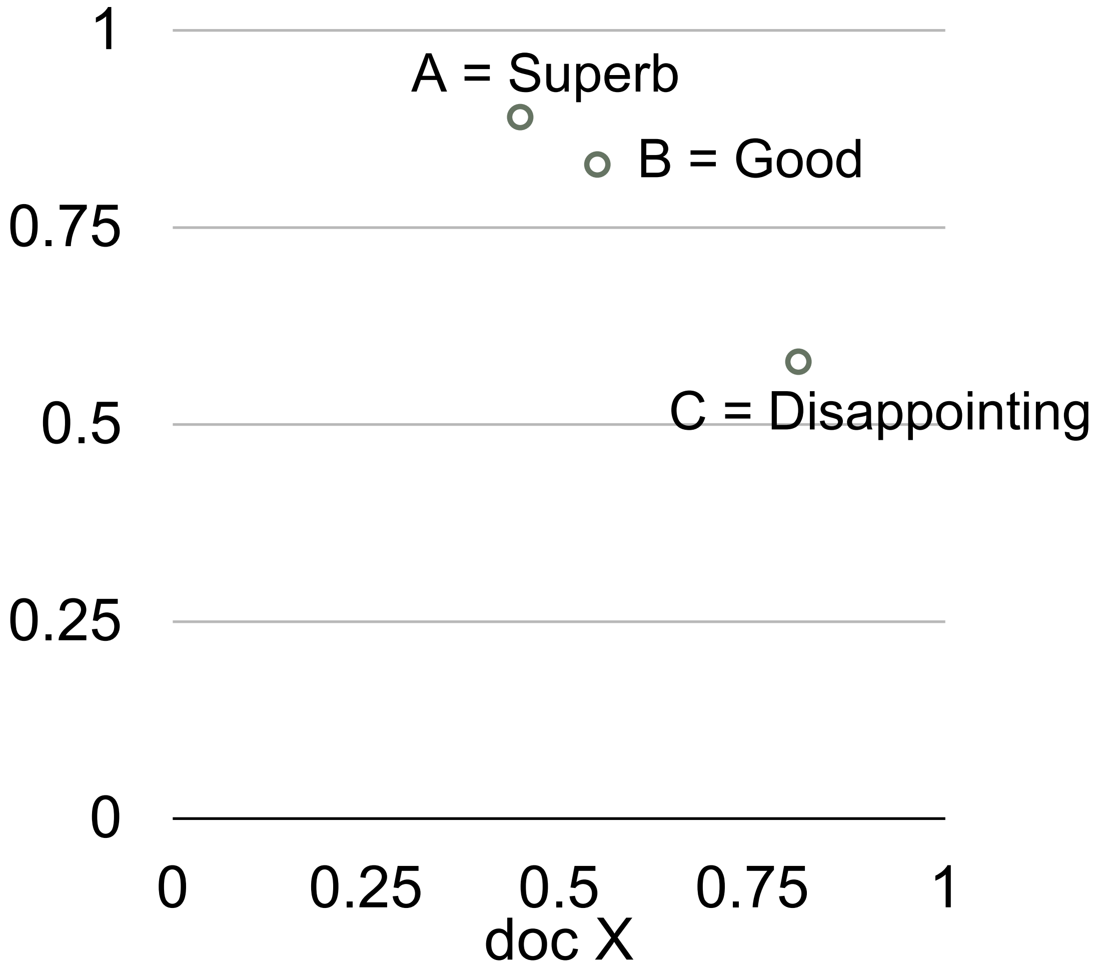
Divide each vector by its L2 length

Penn Engineering

Vector L2 (length) Normalization

	doc_X	doc_Y
$\dot{A}$	0.45	0.89
$\dot{B}$	0.55	0.83
$\dot{C}$	0.81	0.58

Now Good is closer to Superb than to Disappointing

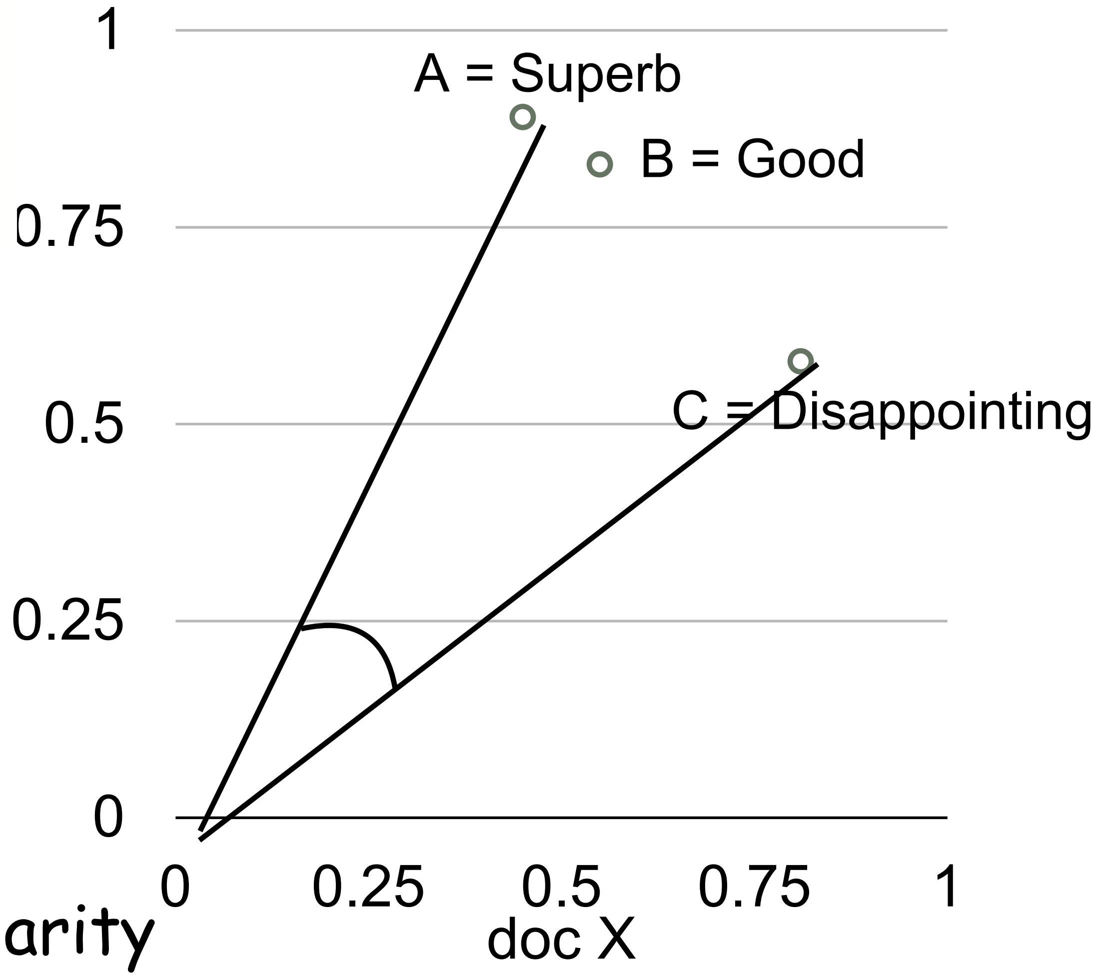


Cosine Distance

$$1 - \frac{\sum_{i=1}^{n} u_i \times v_i}{\sqrt{\sum_{i=1}^{n} u_i^2} \times \sqrt{\sum_{i=1}^{n} v_i^2}}$$

Cosine does the L2 normalization too

Cosine angle between vectors tells us their similarity



Term-Term Matrix

	abandon	abdicate	abhor	…	zymurgy
abandon					
abdicate					
abhor					
academic					
…					
zygodactyl					
zymurgy					

Term-Term Matrix

AKA
Term-Context
Matrix

	abandon	abdicate	abhor	…	zymurgy
abandon					
abdicate					
abhor					
academic					
…					
zygodactyl					
zymurgy					

Length of the vector is now |V|
instead of number of documents

Term-Term Matrix

AKA
Term-Context
Matrix

	abandon	abdicate	abhor	…	zymurgy
abandon					
abdicate	$\boxed{}$				
abhor					
academic					
…					
zygodactyl					
zymurgy					

The value in a cell indicates how often abandon appears in a context window surrounding abdicate

Context windows

w-2, w-1 **target_word** w+1 w+2

The government most not **abdicate** responsibility to non-elected
it has led men to **abdicate** their family responsibilities
other demands, but declining to **abdicate** his responsibility
leaders **abdicate** their role and present people with no plans

	his	leaders	not	responsibility	to
abdicate	1	1	1	2	3

Context windows

Occur in a window of +/- 2 words, in the same sentence, in the same document

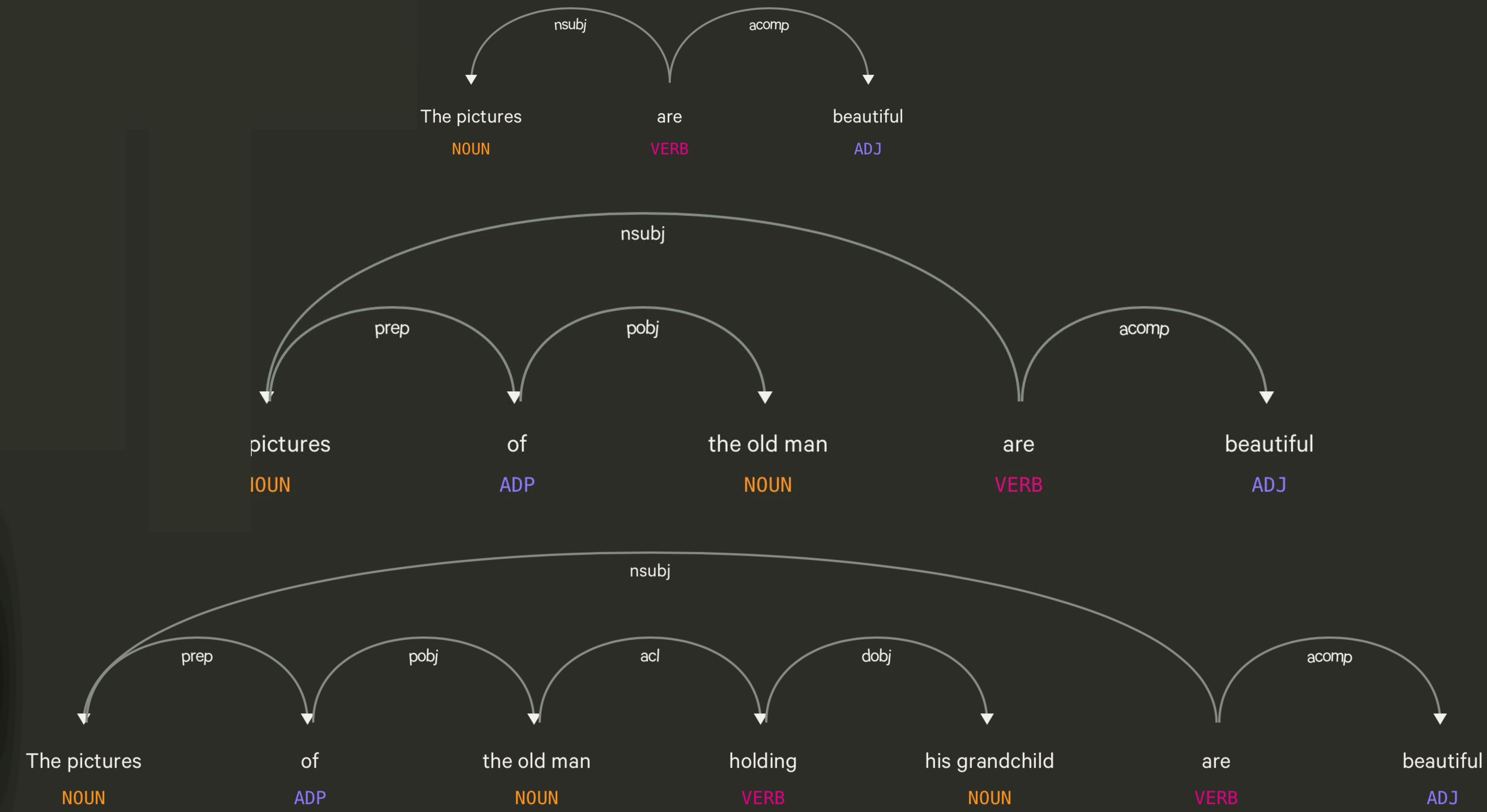
Instead of window of words use more complex contexts: dependency patters. Subj-of-verb, adj-mod, obj-of-verb

Languages have long distance dependencies

*The **pictures are** beautiful.*

*The **pictures** of the old man **are** beautiful.*

*The **pictures** of the old man holding his grandchild **are** beautiful.*


nsubj
acomp
The pictures
NOUN
are
VERB
beautiful
ADJ
nsubj
prep
pobj
acomp
pictures
IOUN
of
ADP
the old man
NOUN
are
VERB
beautiful
ADJ
nsubj
prep
pobj
acl
dobj
acomp
The pictures
NOUN
of
ADP
the old man
NOUN
holding
VERB
his grandchild
NOUN
are
VERB
beautiful
ADJ

Using syntax to define a word's context

- Zellig Harris (1968) "The meaning of entities, and the meaning of grammatical relations among them, is related to the restriction of combinations of these entities relative to other entities"

- Duty and Responsibility have similar syntactic distributions

Modified by adjectives	additional, administrative, assumed, collective, congressional, constitutional …
Object of verbs	assert, assign, assume, attend to, avoid, become, breach..

Alternates to counts

- Raw word frequency is not a great measure of association between words. It’s very skewed “the” and “of” are very frequent, but maybe not the most discriminative
- We’d rather have a measure that asks whether a context word is particularly informative about the target word.
- Instead of raw counts, it’s common to transform vectors using TF-IDF or PPMI

TF-IDF

- *Term frequency * inverse document frequency*

How often a word occurred in a document

1 over the number of documents that it occurred in

Sparse v. Dense Vectors

- Co-occurrence matrix (weighted by TF-IDF or mutual information)
  - **Long** (length |V| = 50,000+)
  - **Sparse** (most elements are zeros)
- Alternative: learn vectors that are
  - **Short** (length 200-1000)
  - **Dense** (most elements are non-zero)

How do we get dense vectors?

- One recipe: train a classifier!

- 1. Treat the target word and a neighboring context word as positive examples.
- 2. Randomly sample other words in the lexicon to get negative samples.
- 3. Use logistic regression to train a classifier to distinguish those two cases.
- 4. Use the weights as the embeddings.

Word2Vec

Mikolov et al. 2013

- Learn embeddings as part of the process of word prediction.
- Train a classifier to predict neighboring words
- Inspired by neural net language models.
- In so doing, learn dense embeddings for the words in the training corpus.

- Advantages:
- Fast, easy to train (much faster than SVD)
- Available online in the word2vec package Including sets of pretrained embeddings!

Word2Vec

- Predict each neighboring word in a context window of 2C of surrounding words
- So for C=2, we are given a word w_t and we try to predict its 4 surrounding words
  - $[w_{t-2}, w_{t-1}, w_{t+1}, w_{t+2}]$
- Uses "negative sampling" for training

Negative sampling

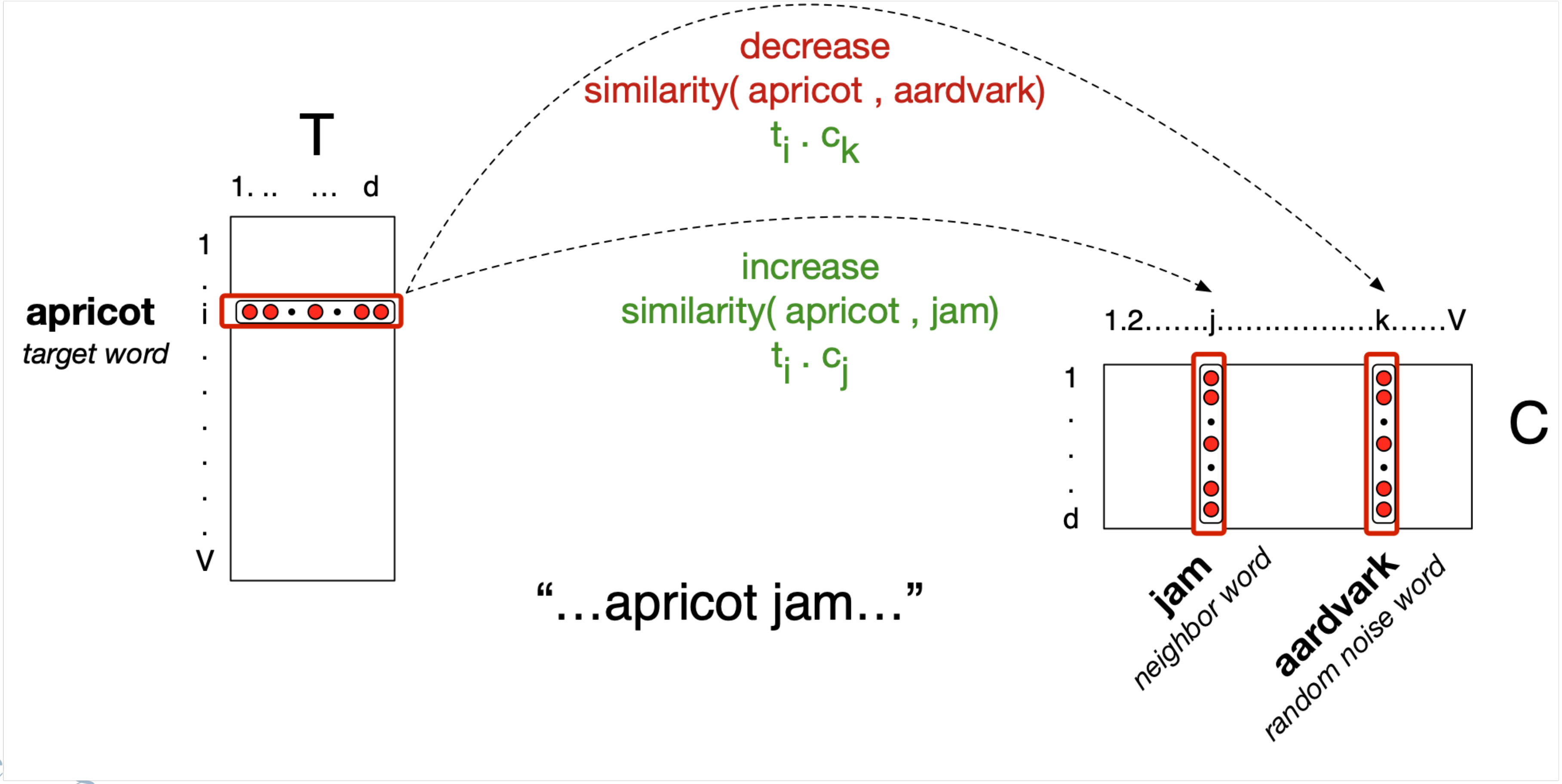
```
lemon,  a [tablespoon of apricot preserves or] jam
          c1          c2   w     c3         c4
```

We want predictions
of these words to be high

And these words to be low

```
[cement metaphysical dear coaxial    apricot attendant whence forever puddle]
 n1     n2           n3   n4                 n5        n6     n7      n8
```

Neural Network



Penn Engineering

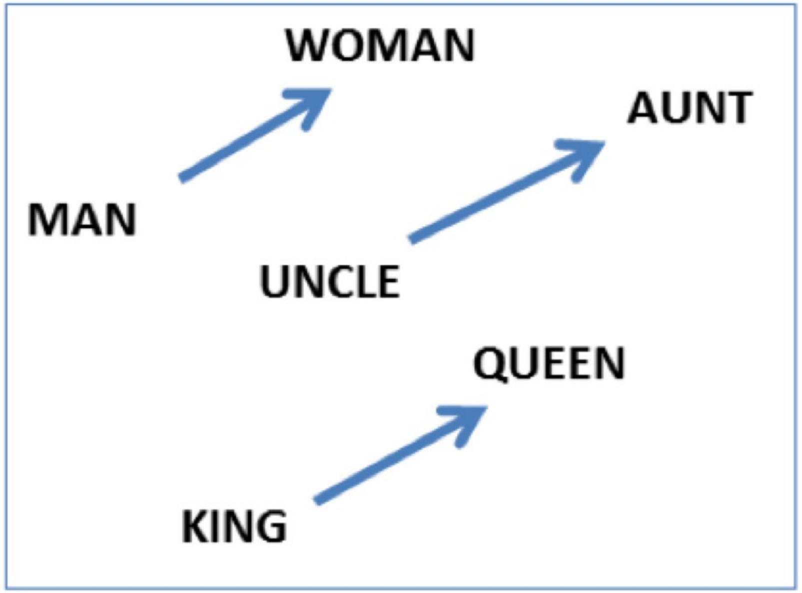
Properties of Embeddings

- Nearest Neighbors are surprisingly good

target:	Redmond	Havel	ninjutsu	graffiti	capitulate
	Redmond Wash.	Vaclav Havel	ninja	spray paint	capitulation
	Redmond Washington	president Vaclav Havel	martial arts	grafitti	capitulated
	Microsoft	Velvet Revolution	swordsmanship	taggers	capitulating

Embeddings capture relational meanings

- .vector('king') - vector('man') + vector('queen') ≅ vector('woman')



Magnitude: A Fast, Efficient Universal Vector Embedding Utility Package

Ajay Patel
Plasticity Inc.
San Francisco, CA
ajay@plasticity.ai

Alexander Sands
Plasticity Inc.
San Francisco, CA
alex@plasticity.ai

Chris Callison-Burch
Computer and Information
Science Department
University of Pennsylvania
ccb@upenn.edu

Marianna Apidianaki
LIMSI, CNRS
Université Paris-Saclay
91403 Orsay, France
marapi@seas.upenn.edu

Abstract

Vector space embedding models like word2vec, GloVe, and fastText are extremely popular representations in natural language processing (NLP) applications. We present Magnitude, a fast, lightweight tool for utilizing and processing embeddings. Magnitude is an open source Python package with a compact vector storage file format that allows for efficient manipulation of huge numbers of embeddings. Magnitude performs common operations up to 60 to 6,000 times faster than Gensim. Magnitude introduces several

Metric	Cold	Warm
Initial load time	97x	–
Single key query	1x	110x
Multiple key query (n=25)	68x	3x
k-NN search query (k=10)	1x	5,935x

Table 1: Speed comparison of Magnitude versus Gensim for common operations. The 'cold' column represents the first time the operation is called. The 'warm' column indicates a subsequent call with the same keys.

file, a 97x speed-up. Gensim uses 5GB of RAM versus 18KB for Magnitude.

Penn Engineering

Demo of word vectors

```
# Install Magnitude
pip3 install pymagnitude

# Download Google's word2vec vectors
wget http://magnitude.plasticity.ai/word2vec+approx/GoogleNews-vectors-negative300.magnitude
# Warning it's 11GB large

# Start Python, and try the commands
# on the next slide
python3
```

Demo of word vectors

```
from pymagnitude import *
vectors = Magnitude("GoogleNews-vectors-negative300.magnitude")

queen = vectors.query('queen')
king = vectors.query("king")
vectors.similarity(king, queen)
# 0.6510958

vectors.most_similar_approx(king, topn=5)
#[('king', 1.0), ('kings', 0.72), ('prince', 0.62), ('sultan', 0.59), ('ruler', 0.58)]
```

Many possible models

Matrix type
Term-document
Term-context
Pattern-pair

Dim. Reduction
word2vec
GloVe
PCA
LDA
LSA

Reweighting
length norm.
TF-IDF
PPMI
probabilities

How many dimensions?

Comparisons
cosine
Manhattan
Jaccard
KL divergence
JS distance
DICE

What modifications should we make to the input?

Penn Engineering