Text similarity

311.

Introduction to Text Similarity

Text Similarity

313

Spelling Similarity:

Edit Distance

Spelling Similarity

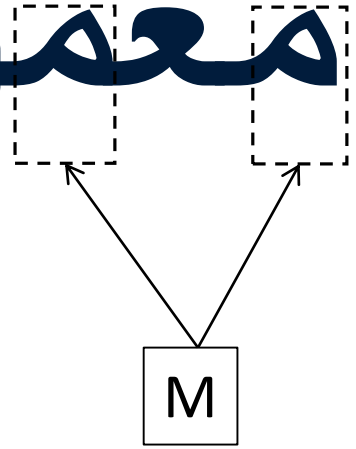
- Typos:
  - Brittany Spears -> Britney Spears
  - Catherine Hepburn -> Katharine Hepburn
  - Reciept -> receipt
- Variants in spelling:
  - Theater -> theatre

Who is this?

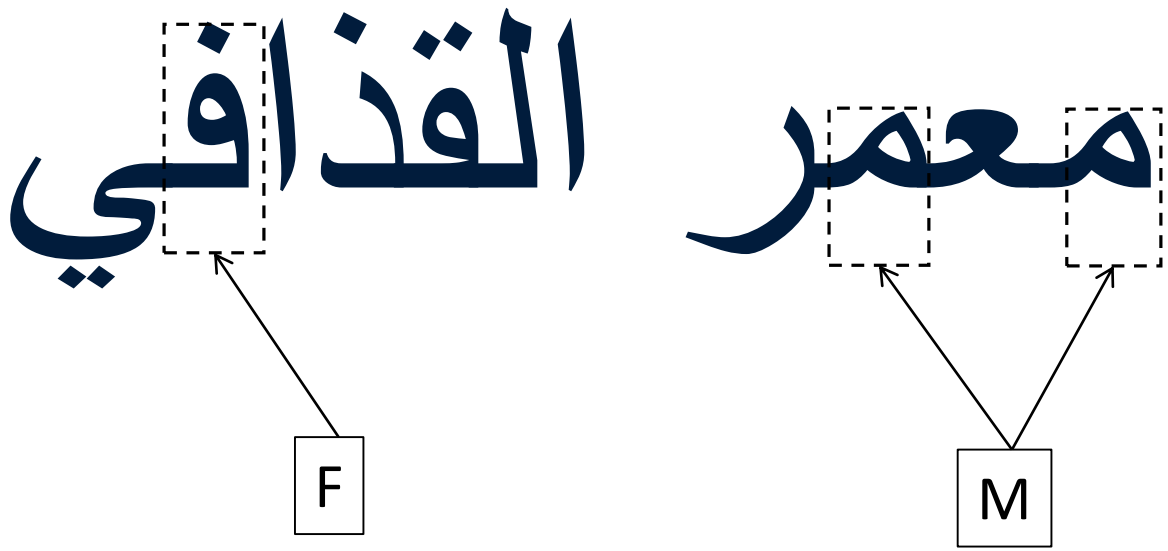
معمر القذافي

Hints

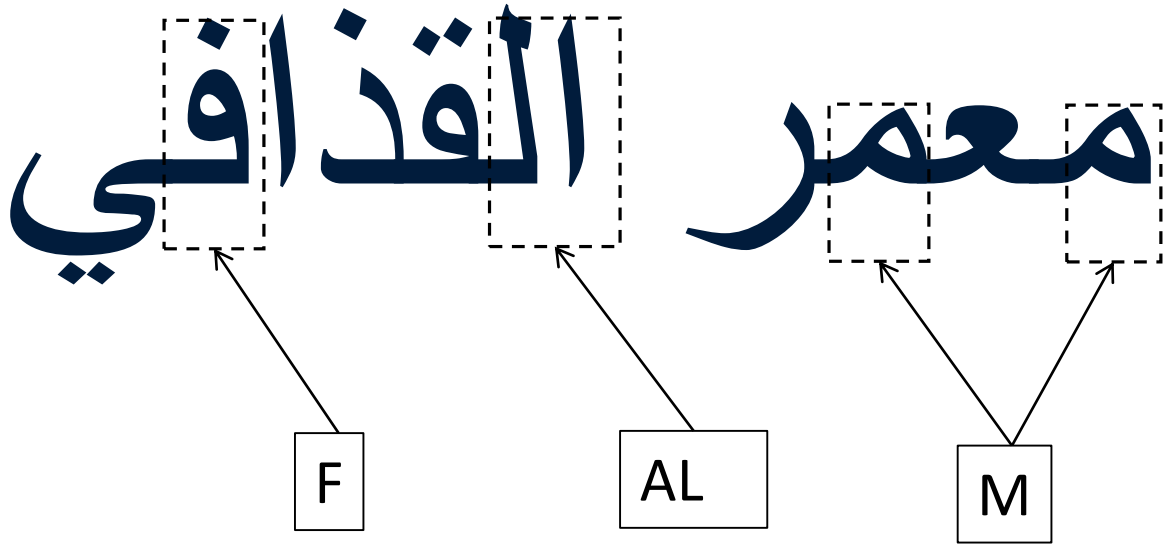
معمر القذافي



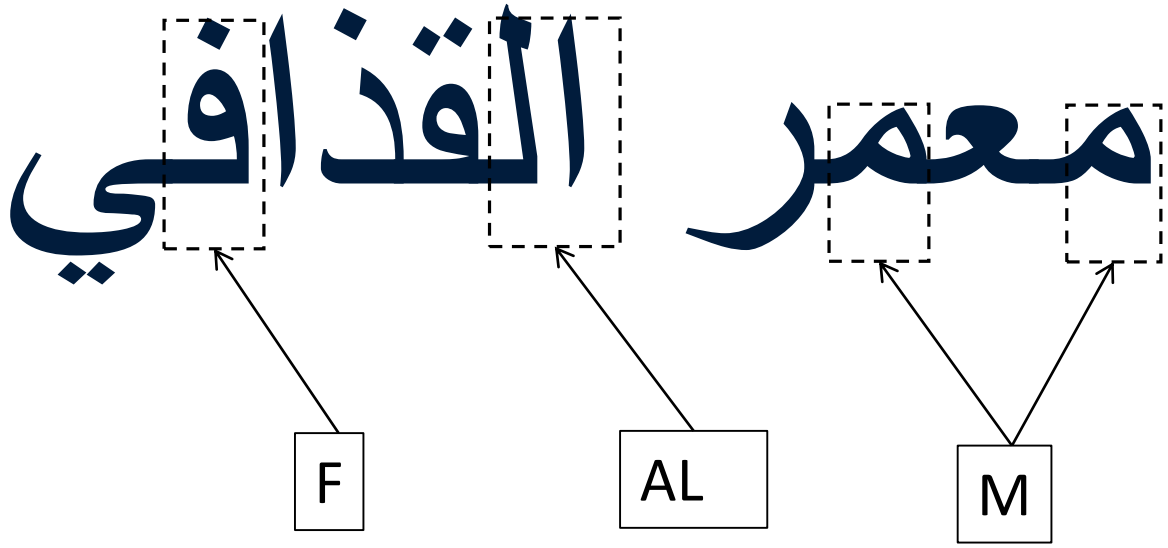
Hints



Hints



Hints



Muammar (al-)Gaddafi, or **Moamar Khadafi,** or …

Quiz

How many different transliterations can there be?

m u o a m mm a e r	el al El Al ø	Q G Gh K Kh a e u d dh ddh dhdh th zz a f ff i y

A lot!

m u o a m mm a e r		el al El Al ø		Q G Gh K Kh a e u d dh ddh dhdh th zz a f ff i y		
8	x	5	x	360	=	14,400

Edit Operations

- Insertion/deletion
  - behaviour - behavior
- Substitution
  - string - spring
- Multiple edits
  - sleep - slept

Levenshtein Method

- Based on dynamic programming
  - Insertions, deletions, and substitutions usually all have a cost of 1.

Example

		s	t	r	e	n	g	t	h
	0	1	2	3	4	5	6	7	8
t	1								
r	2								
e	3								
n	4								
d	5								

Recurrence relation

- Definitions
  - $s_1(i)$ – i^{th} character in string s_1
  - $s_2(j)$ – j^{th} character in string s_2
  - $D(i,j)$ – edit distance between a prefix of s_1 of length i and a prefix of s_2 of length j
  - $t(i,j)$ – cost of aligning the i^{th} character in string s_1 with the j^{th} character in string s_2

- Recursive dependencies

$$D(i,0)=i$$
$$D(0,j)=j$$
$$D(i,j)=\min[$$
$$\quad D(i-1,j)+1$$
$$\quad D(i,j-1)+1$$
$$\quad D(i-1,j-1)+t(i,j)$$
$$\quad]$$

- Simple edit distance:

$$t(i,j)=0 \ \textit{iff} \ s_1(i)=s_2(j)$$
$$t(i,j)=1, \textit{otherwise}$$

Example

		s	t	r	e	n	g	t	h
	0	1	2	3	4	5	6	7	8
t	1	1							
r	2								
e	3								
n	4								
d	5								

Example

		s	t	r	e	n	g	t	h
	0	1	2	3	4	5	6	7	8
t	1	1	1						
r	2								
e	3								
n	4								
d	5								

Example

		s	t	r	e	n	g	t	h
	0	1	2	3	4	5	6	7	8
t	1	1	1	2	3	4	5	6	7
r	2	2	2						
e	3								
n	4								
d	5								

Example

		s	t	r	e	n	g	t	h
	0	1	2	3	4	5	6	7	8
t	1	1	1	2	3	4	5	6	7
r	2	2	2						
e	3								
n	4								
d	5								

Example

		s	t	r	e	n	g	t	h
	0	1	2	3	4	5	6	7	8
t	1	1	1	2	3	4	5	6	7
r	2	2	2	1	2	3	4	5	6
e	3	3	3	2	1	2	3	4	5
n	4	4	4	3	2	1	2	3	4
d	5	5	5	4	3	2	2	3	4

Edit Transcript

		s	t	r	e	n	g	t	h
	0	1	2	3	4	5	6	7	8
t	1	1	1	2	3	4	5	6	7
r	2	2	2	1	2	3	4	5	6
e	3	3	3	2	1	2	3	4	5
n	4	4	4	3	2	1	2	3	4
d	5	5	5	4	3	2	2	3	4

Other Costs

- Damerau modification
  - Swaps of two adjacent characters also have a cost of 1
  - E.g., Lev(“cats”,”cast”) = 2, Dam(“cats”,”cast”) = 1

Quiz

- Some distance functions can be more specialized.
- Why do you think that the edit distances for these pairs are as follows?
  - Dist (“sit clown”,“sit down”) = 1
  - Dist (“qeather”,”weather”) = 1, *but* Dist (“leather”,”weather”) = 2

Quiz Answers

- Dist(“sit down”,”sit clown”) is lower in this example because we want to model the type of errors common with optical character recognition (OCR)
- Dist(“qeather”,”weather”) < Dist(“leather”,”weather”) because we want to model spelling errors introduced by “fat fingers” (clicking on an adjacent key on the keyboard)

Quiz: Guess the Language

```
AACCTGCGGAAGGATCATTACCGAGTGCGGGTCCTTTGGGCCCAACCTCCCATCCGTGTCTATTGTACCC
TGTTGCTTCGGCGGGCCCGCCGCTTGTCGGCCGCCGGGGGGGCGCCTCTGCCCCCCGGGCCCGTGCCCGC
CGGAGACCCCAACACGAACACTGTCTGAAAGCGTGCAGTCTGAGTTGATTGAATGCAATCAGTTAAAACT
TTCAACAATGGATCTCTTGGTTCCGGC
```

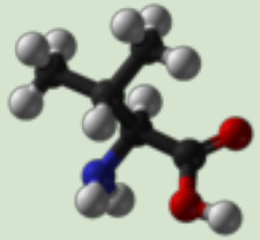
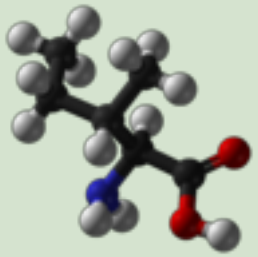
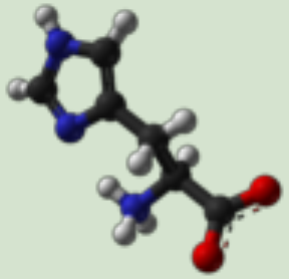
Quiz Answer

- This is a genetic sequence (nucleotides AGCT)

>U03518 Aspergillus awamori internal transcribed spacer 1 (ITS1)
AACCTGCGGAAGGATCATTACCGAGTGCGGGTCCTTTGGGCCCAACCTCCCATCCGTGTCTATTGTACCC
TGTTGCTTCGGCGGGCCCGCCGCTTGTCGGCCGCCGGGGGGGCGCCTCTGCCCCCCGGGCCCGTGCCCGC
CGGAGACCCCAACACGAACACTGTCTGAAAGCGTGCAGTCTGAGTTGATTGAATGCAATCAGTTAAAACT
TTCAACAATGGATCTCTTGGTTCCGGC

Other uses of Edit Distance

- In biology, similar methods are used for aligning non-textual sequences
  - Nucleotide sequences, e.g., GTTCGTGATGGAGCG, where A=adenine, C=cytosine, G=guanine, T=thymine, U=uracil, "-"=gap of any length, N=any one of ACGTU, etc.
  - Amino acid sequences, e.g., FMELSEDGIEMAGSTGVI, where A=alanine, C=cystine, D=aspartate, E=glutamate, F=phenylalanine, Q=glutamine, Z=either glutamate or glutamine, X="any", etc. The costs of alignment are determined empirically and reflect evolutionary divergence between protein sequences. For example, aligning V (valine) and I (isoleucine) is lower-cost than aligning V and H (histidine).

Valine	Isoleucine	Histidine
		

External URLs

- Levenshtein demo
  - http://www.let.rug.nl/~kleiweg/lev/
- Biological sequence alignment
  - http://www.bioinformatics.org/sms2/pairwise_align_dna.html
  - http://www.sequence-alignment.com/sequence-alignment-software.html
  - http://www.ebi.ac.uk/Tools/msa/clustalw2/
  - http://www.animalgenome.org/bioinfo/resources/manuals/seqformats

Text Similarity

- Motivation
  - People can express the same concept (or related concepts) in many different ways. For example, “the plane leaves at 12pm” vs “the flight departs at noon”
  - Text similarity is a key component of Natural Language Processing
- Uses in NLP
  - If the user is looking for information about cats, we may want the NLP system to return documents that mention kittens even if the word “cat” is not in them.
  - If the user is looking for information about “fruit dessert”, we want the NLP system to return documents about “peach tart” or “apple cobbler”.
  - A speech recognition system should be able to tell the difference between similar sounding words like the “Dulles” and “Dallas” airports.

Types of Text Similarity

- Many types of text similarity exist:
  - Morphological similarity (e.g., respect-respectful)
  - Spelling similarity (e.g., theater-theatre)
  - Homophony (e.g., raise-raze-rays)
  - **Synonymy (e.g., talkative-chatty**), including across languages
  - **Semantic similarity (e.g., cat-tabby)**
  - Sentence similarity (e.g., paraphrases)
  - Document similarity (e.g., two news stories on the same event)

Notes

- Similarity vs. relatedness
  - car, bicycle: similar
  - car, gasoline: related, not similar
- Semantic field
  - doctor, nurse, hospital, syringe, medication, scalpel

Human Judgments of Similarity

tiger	cat	7.35
tiger	tiger	10.00
book	paper	7.46
computer	keyboard	7.62
computer	internet	7.58
plane	car	5.77
train	car	6.31
telephone	communication	7.50
television	radio	6.77
media	radio	7.42
drug	abuse	6.85
bread	butter	6.19
cucumber	potato	5.92

[Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin, "Placing Search in Context: The Concept Revisited", ACM Transactions on Information Systems, 20(1):116-131, January 2002]

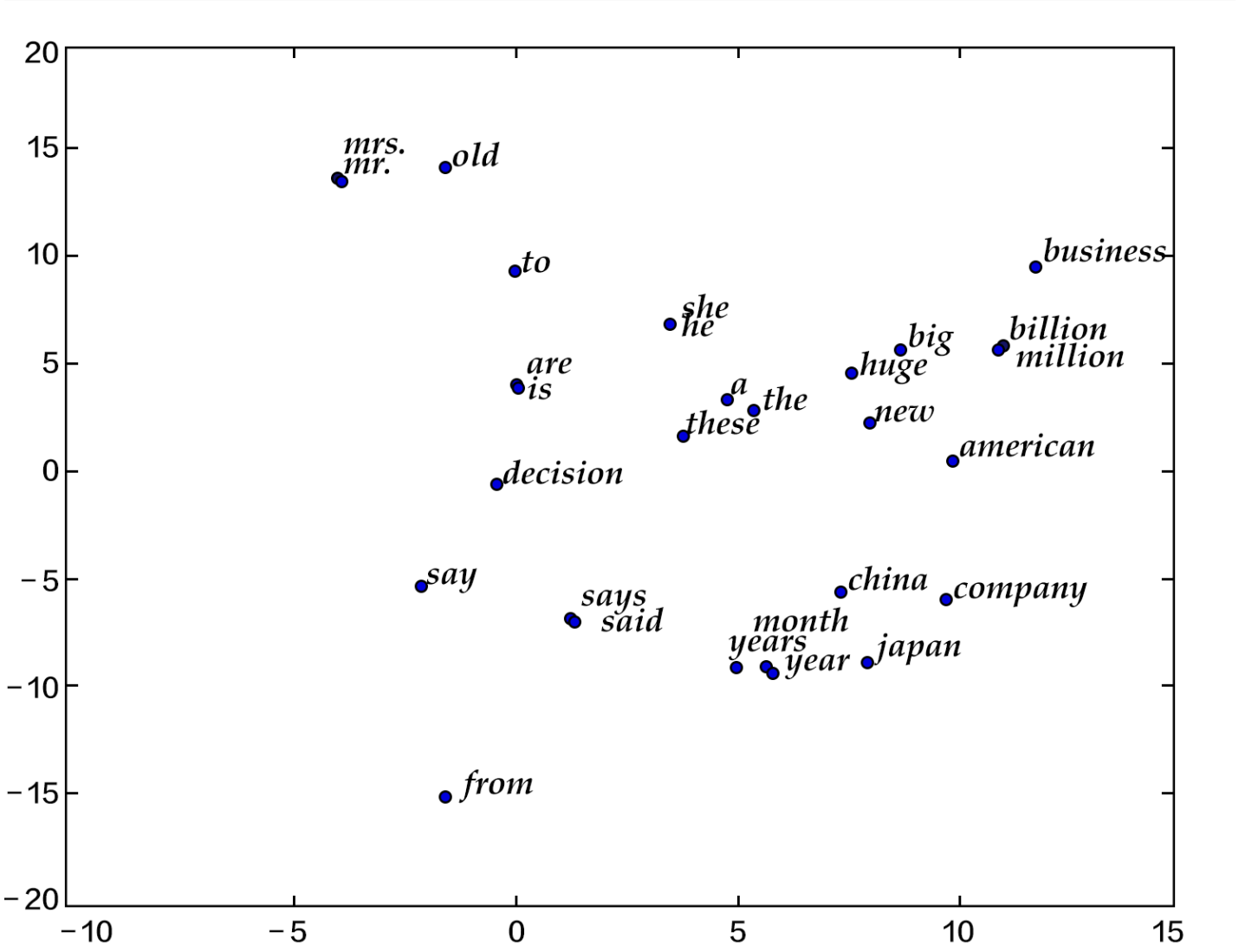
http://wordvectors.org/suite.php

Automatic Similarity Computation

spain	0.679
belgium	0.666
netherlands	0.652
italy	0.633
switzerland	0.622
luxembourg	0.610
portugal	0.577
russia	0.572
germany	0.563
catalonia	0.534

- Words most similar to “France”
- Computed using word2vec
  - [Mikolov et al. 2013]

Two-dimensional Representations



Natural Language Processing

314.

Semantic Similarity

Synonyms and paraphrases

- Example: post-close market announcements

The S&P 500	climbed	6.93, or 0.56 percent, to 1,243.72,	its best close	since June 12, 2001.
The Nasdaq	gained	12.22, or 0.56 percent, to 2,198.44	for its best showing	since June 8, 2001.
The DJIA	rose	68.46, or 0.64 percent, to 10,705.55,	its highest level	since March 15.

Synonyms

- Different words (and also word compounds) can have similar meanings.
  - *tepid* and *lukewarm* have very similar meanings and can be substituted for one another (*tepid water* vs. *lukewarm water).*
- True synonyms are actually relatively rare.
  - even though *big* and *large* are often thought of as synonyms, consider the difference between *Big Leagues* and *Large Leagues*. ☺
- The verbs *sweat* and *perspire* are also near synonyms.
  - However, they differ in their frequency of use and the type of text in which they are likely to appear.

Polysemy


SWEARING
Two electric windmills are standing in a field.
One turns to the other and asks "What kind of music do you like?"
The second one replies "I'm a huge metal fan".

Polysemy

- Polysemy is the property of words to have multiple senses.
- For example, the noun *book* can refer to the following:
  - A literary work (e.g., “Anna Karenina”)
  - A stack of pages (e.g., a notebook)
  - A record of business transactions (think “bookkeeper”)
  - A record of bets (think “bookmaker”)
  - A list of buy and sell orders in a financial market

Polysemy

- The same word can also have multiple parts of speech, each with its own set of senses. For example, the word *book*, as a verb can mean “make a reservation for” or “occupy”.
- The different senses of the same word don’t have to be equally frequent.
- Some of the senses may overlap (e.g., the first two senses of *book* on the previous slide). That’s partially why different dictionaries list different sets of word senses for the same word.
  - “My favorite books are Anna Karenina and my father’s checkbook” ☺
- Some words can be highly polysemous (e.g., the verb “get” has at least 35 different meanings, according to Wordnet).

Other Semantic Relations

- Antonymy (near opposites)
  - *raise-lower*
- Hypernymy
  - a *deer* is a hypernym for *elk*
- Hyponymy (the inverse of hypernymy)
- Membership Meronymy:
  - a *flock* includes *sheep* (or *birds)*
- Part Meronymy:
  - a *table* has *legs*

Synsets

- Semantic relations hold between word senses, not between words.
- Examples:
  - the antonym of *hot* can be either *mild* or *cold* (or *unattractive)* depending on the specific sense of *hot*.
  - the immediate hypernym of *bar* can be one of the following, among others: *room*, *musical notation*, *obstruction*, *profession*, depending on the sense of *bar*.
- The term *synset* is used to group together all synonyms of the same word. If a word is polysemous, it may be associated with multiple synsets.

Text Similarity

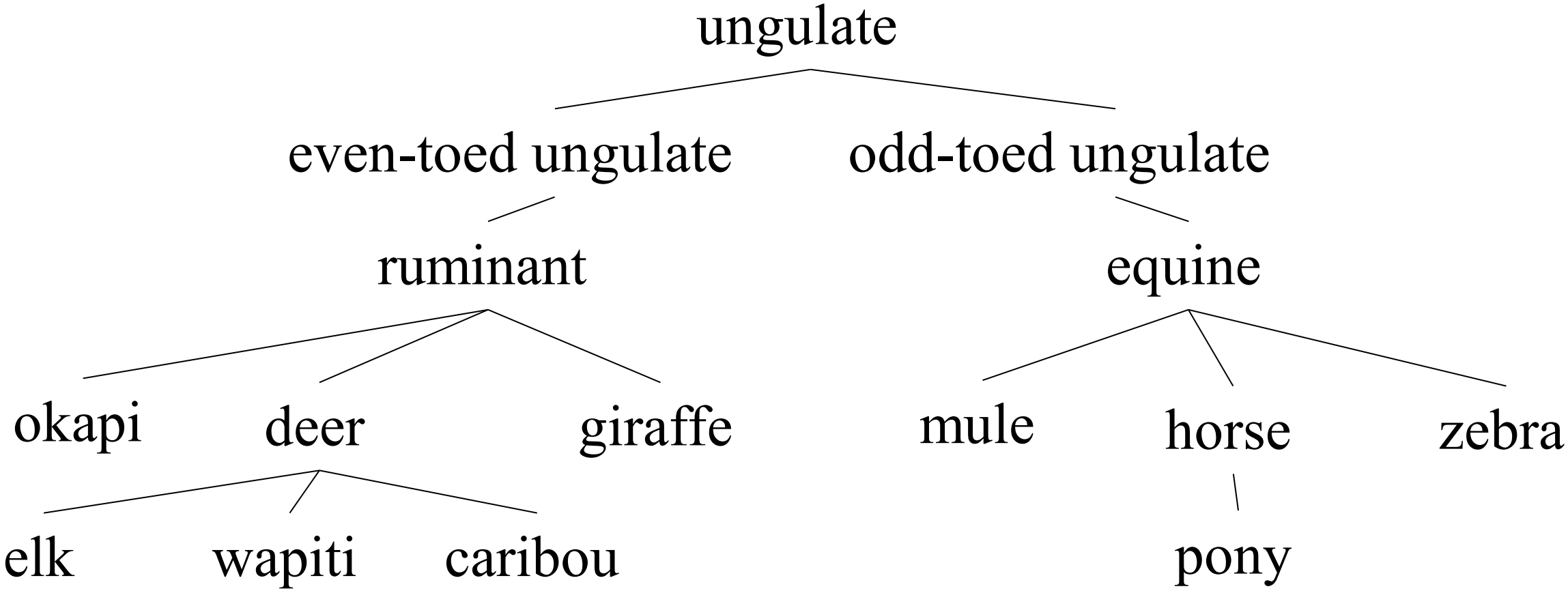
315.

Wordnet

Wordnet

- Wordnet is a project run by George Miller (1920-2012) and Christiane Fellbaum at Princeton University.
- It includes a database of words (mainly nouns and verbs but also adjectives and adverbs) and semantic relations between them.
- The main relation is hypernymy, so the overall structure of the database is more tree-like (see next slide).
- References:
  - George A. Miller (1995). WordNet: A Lexical Database for English. Communications of the ACM Vol. 38, No. 11: 39-41.
  - Christiane Fellbaum (1998, ed.) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Sample Noun Taxonomy



Wordnet Example (1/6)

The **noun** bar has 11 senses
1. barroom, bar, saloon, ginmill, taproom -- (a room where alcoholic drinks are served over a counter)
2. bar -- (a counter where you can purchase food or drink)
3. bar -- (a rigid piece of metal)
4. measure, bar -- (notation for a repeating pattern of musical beats; written followed by a vertical bar)
5. bar -- (usually metal placed in windows to prevent escape)
6. prevention, bar -- (the act of preventing)
7. bar -- (a unit of pressure equal to a million dynes per square centimeter)
8. bar -- (a submerged (or partly submerged) ridge in a river or along a shore)
9. legal profession, bar, legal community -- (the body of individuals qualified to practice law)
10. cake, bar -- (a block of soap or wax)
11. bar -- ((law) a railing that encloses the part of the courtroom where the the judges and lawyers sit and the case is tried)

The **verb** bar has 4 senses
1. bar, debar, exclude -- (prevent from entering; keep out; "He was barred from membership in the club")
2. barricade, block, blockade, block off, block up, bar -- (render unsuitable for passage; "block the way"; "barricade the streets")
3. banish, relegate, bar -- (expel, as if by official decree; "he was banished from his own country")
4. bar -- (secure with, or as if with, bars; "He barred the door")

Wordnet Example (2/6)

Sense 1
barroom, bar, saloon, ginmill, taproom
=> room
=> area
=> structure, construction
=> artifact, artefact
=> object, physical object
=> entity, something

Sense 2
bar
=> counter
=> table
=> furniture, piece of furniture, article of furniture
=> furnishings
=> instrumentality, instrumentation
=> artifact, artefact
=> object, physical object
=> entity, something

Wordnet Example (3/6)

Sense 3
bar
=> implement
=> instrumentality, instrumentation
=> artifact, artefact
=> object, physical object
=> entity, something

Sense 4
measure, bar
=> musical notation
=> notation, notational system
=> writing, symbolic representation
=> written communication, written language
=> communication
=> social relation
=> relation
=> abstraction

Wordnet Example (4/6)

Sense 5
bar
=> obstruction, impediment, impedimenta
=> structure, construction
=> artifact, artefact
=> object, physical object
=> entity, something

Sense 6
prevention, bar
=> hindrance, interference, interfering
=> act, human action, human activity

Sense 7
bar
=> pressure unit
=> unit of measurement, unit
=> definite quantity
=> measure, quantity, amount, quantum
=> abstraction

Wordnet Example (5/6)

Sense 8
bar
=> ridge
=> natural elevation, elevation
=> geological formation, geology, formation
=> natural object
=> object, physical object
=> entity, something
=> barrier
=> mechanism
=> natural object
=> object, physical object
=> entity, something

Wordnet Example (6/6)

Sense 9
legal profession, bar, legal community
=> profession, community
=> occupation, vocation, occupational group
=> body
=> gathering, assemblage
=> social group
=> group, grouping

Sense 10
cake, bar
=> block
=> artifact, artefact
=> object, physical object
=> entity, something

Top-Level Categories

Noun

GROUP	1469	*place*
PERSON	1202	*people*
ARTIFACT	971	*car*
COGNITION	771	*way*
FOOD	766	*food*
ACT	700	*service*
LOCATION	638	*area*
TIME	530	*day*
EVENT	431	*experience*
COMMUNIC.*	417	*review*
POSSESSION	339	*price*
ATTRIBUTE	205	*quality*
QUANTITY	102	*amount*
ANIMAL	88	*dog*
BODY	87	*hair*
STATE	56	*pain*
NATURAL OBJ.	54	*flower*
RELATION	35	*portion*
SUBSTANCE	34	*oil*
FEELING	34	*discomfort*
PROCESS	28	*process*
MOTIVE	25	*reason*
PHENOMENON	23	*result*
SHAPE	6	*square*
PLANT	5	*tree*
OTHER	2	*stuff*

Verb

STATIVE	2922	*is*
COGNITION	1093	*know*
COMMUNIC.*	974	*recommend*
SOCIAL	944	*use*
MOTION	602	*go*
POSSESSION	309	*pay*
CHANGE	274	*fix*
EMOTION	249	*love*
PERCEPTION	143	*see*
CONSUMPTION	93	*have*
BODY	82	*get...done*
CREATION	64	*cook*
CONTACT	46	*put*
COMPETITION	11	*win*
WEATHER	0	—

[Example from J&M, based on Schneider and Smith 2013]

Noun Relations in WordNet

Relation	Also Called	Definition	Example
Hypernym	Superordinate	From concepts to superordinates	*breakfast*^{1} $\rightarrow$ *meal*^{1}
Hyponym	Subordinate	From concepts to subtypes	*meal*^{1} $\rightarrow$ *lunch*^{1}
Instance Hypernym	Instance	From instances to their concepts	*Austen*^{1} $\rightarrow$ *author*^{1}
Instance Hyponym	Has-Instance	From concepts to their instances	*composer*^{1} $\rightarrow$ *Bach*^{1}
Part Meronym	Has-Part	From wholes to parts	*table*^{2} $\rightarrow$ *leg*^{3}
Part Holonym	Part-Of	From parts to wholes	*course*^{7} $\rightarrow$ *meal*^{1}
Antonym		Semantic opposition between lemmas	*leader*^{1} $\Longleftrightarrow$ *follower*^{1}
Derivation		Lemmas w/same morphological root	*destruction*^{1} $\Longleftrightarrow$ *destroy*^{1}

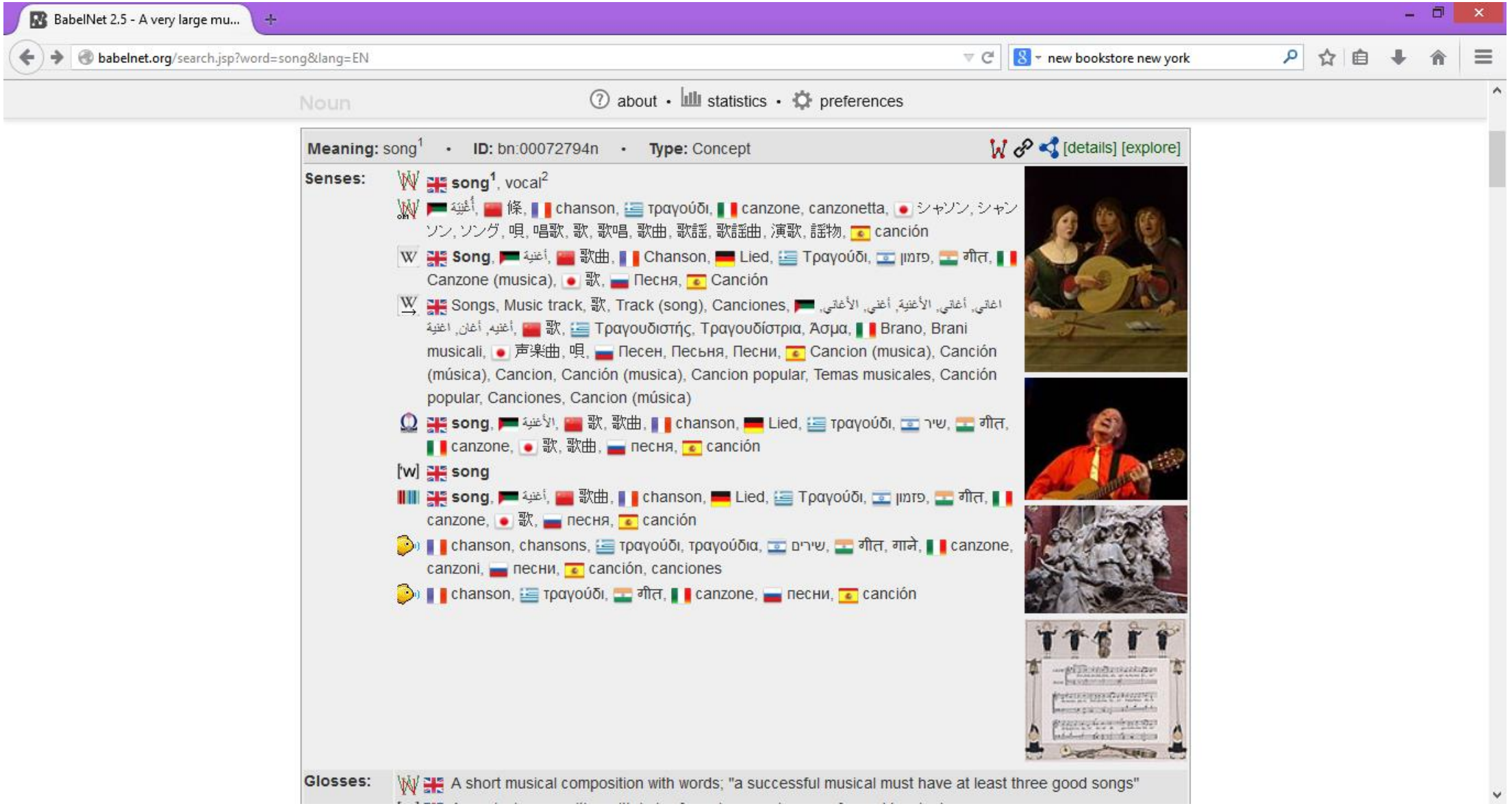
Figure 19.3 Some of the noun relations in WordNet.

Verb Relations in WordNet

Relation	Definition	Example
Hypernym	From events to superordinate events	$fly^9 \rightarrow travel^5$
Troponym	From events to subordinate event	$walk^1 \rightarrow stroll^1$
Entails	From verbs (events) to the verbs (events) they entail	$snore^1 \rightarrow sleep^1$
Antonym	Semantic opposition between lemmas	$increase^1 \Longleftrightarrow decrease^1$

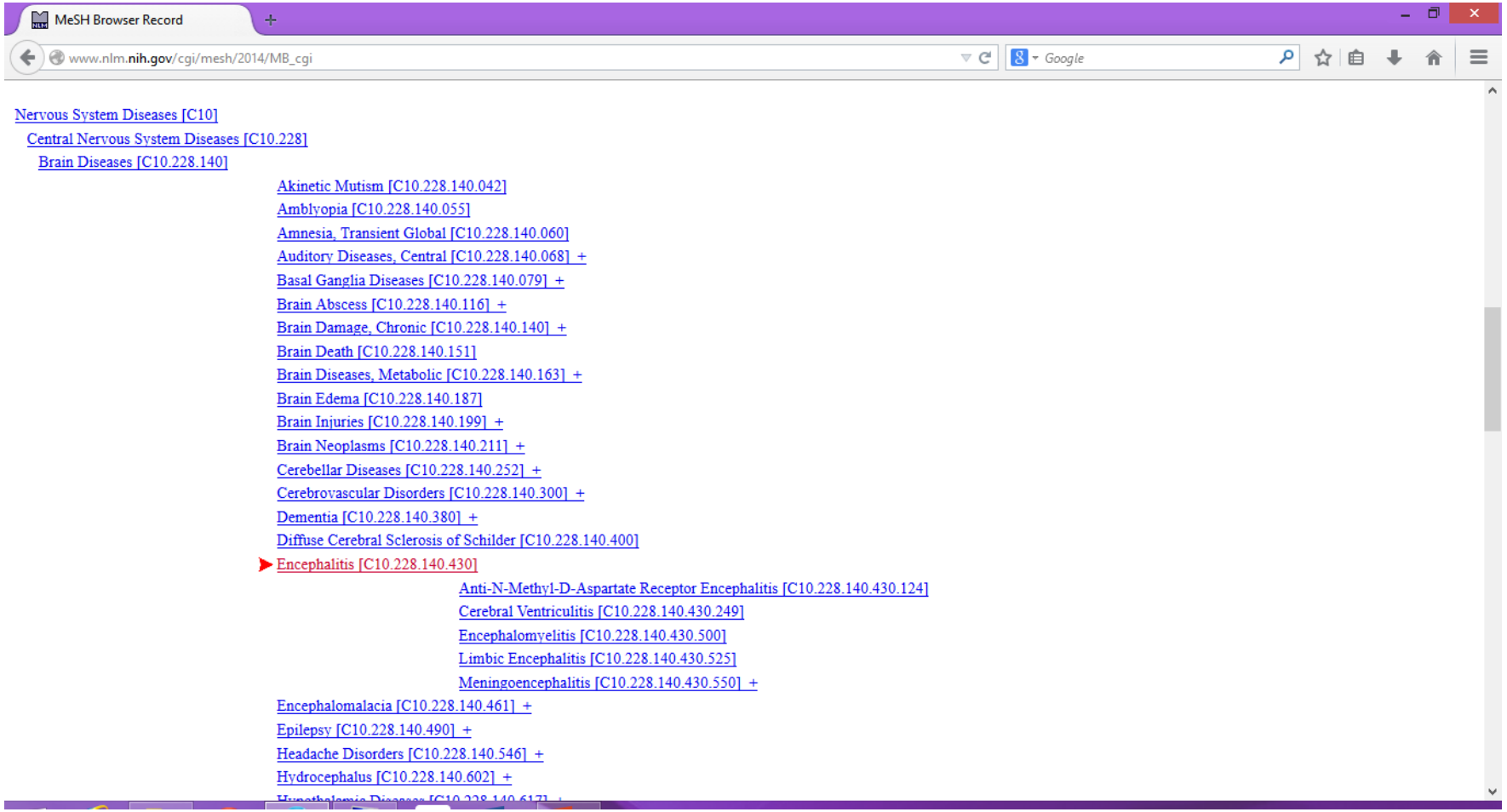
Figure 19.4 Some verb relations in WordNet.

BabelNet Example



MeSH

Medical Subject Headings



http://www.nlm.nih.gov/mesh/MBrowser.html

External Links

- EuroWordNet
  - http://www.illc.uva.nl/EuroWordNet/
- Open Thesaurus
  - http://www.openthesaurus.de/
- Freebase
  - http://www.freebase.com
- DBPedia
  - http://www.dbpedia.org
- BabelNet
  - http://babelnet.org
- Various thesauri
  - https://sites.google.com/site/openrogets/

Text Similarity

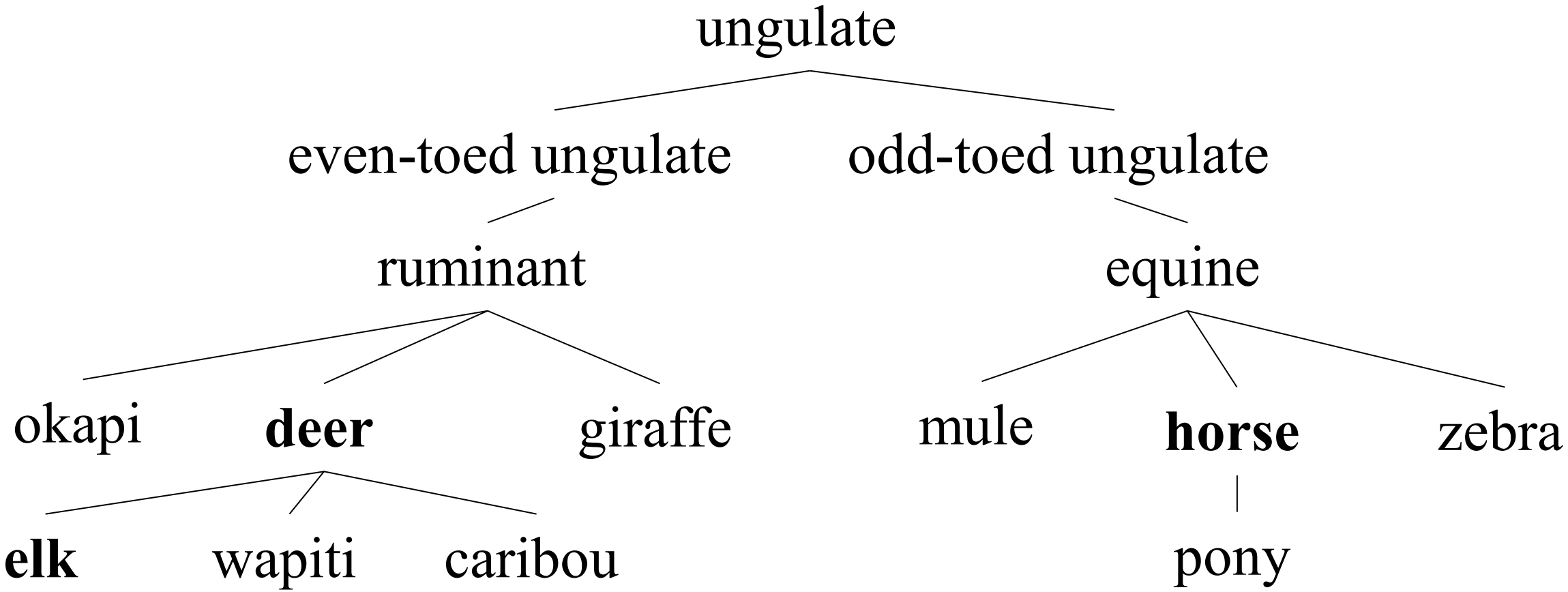
316.

Thesaurus-based Word Similarity Methods

Quiz Answer

- Which pair of words exhibits the greatest similarity?
  - **1. Deer-elk**
  - 2. Deer-horse
  - 3. Deer-mouse
  - 4. Deer-roof
- Why?

Remember Wordnet



Path Similarity

- Version 1
  - Sim (v,w) = - pathlength (v,w)
- Version 2
  - Sim (v,w) = - log pathlength (v,w)

Problems with this Approach

- There may be no tree for the specific domain or language
- A specific word (e.g., a term or a proper noun) may not be in any tree
- IS-A (hypernym) edges are not all equally apart in similarity space

Path similarity between two words

- Version 3 (Philip Resnik)

  Sim (v,w) = - log *P*(LCS(v,w))

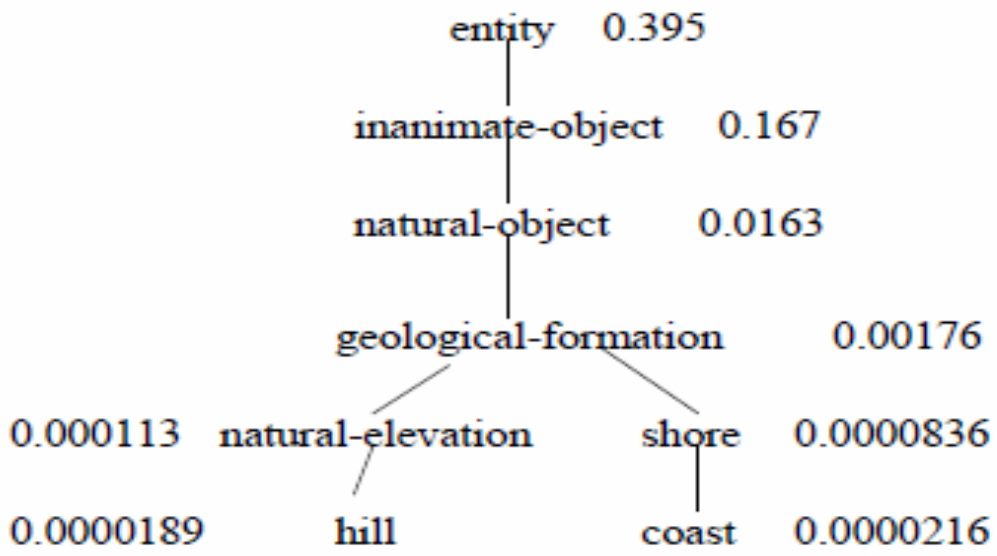
  where LCS = lowest common subsumer, e.g.,

  ungulate for deer and horse

  deer for deer and elk

Information content

- Version 4 (Dekang Lin)
  - Wordnet augmented with probabilities (Lin 1998)
  - IC(c) = -log P(c)
  - Sim (v,w) = 2 x log *P*(LCS(v,w)) / (log *P*(v) + log *P*(w))



$$\text{sim}(\text{Hill, Coast}) = \frac{2 \times \log P(\text{Geological-Formation})}{\log P(\text{Hill}) + \log P(\text{Coast})}$$

$$= 0.59$$

Text Similarity

321.

The Vector Space Model

Vectors, Matrices, and Tensors

- $X = \langle x_1, x_2, \ldots, x_n \rangle$: a vector of n dimensions.
  - $x_1, \ldots, x_n$ can take either binary values {0, 1}, or real values
- Vectors and matrices provide a natural way to represent the occurrence of words in a document/query.
  - In text analysis, n is usually the size of the vocabulary, so each dimension corresponds to a unique word
  - X can be used to represent a document, or a query, or …
  - So x_i indicates either "whether the i-th word in the vocabulary appears" (binary value), or "how many times does the i-th word appear" (real value).
- The entire collection is thus represented as a matrix.
  - Next slide

Example of Document Vectors

- Doc 1= “information retrieval”
- Doc 2 = “computer information retrieval”
- Doc 3 = “computer retrieval”

- Vocabulary: information, retrieval, computer
- Doc 1 = <1, 1, 0>
- Doc 2 = <1, 1, 1>
- Doc 3 = <0, 1, 1>

information, retrieval, computer

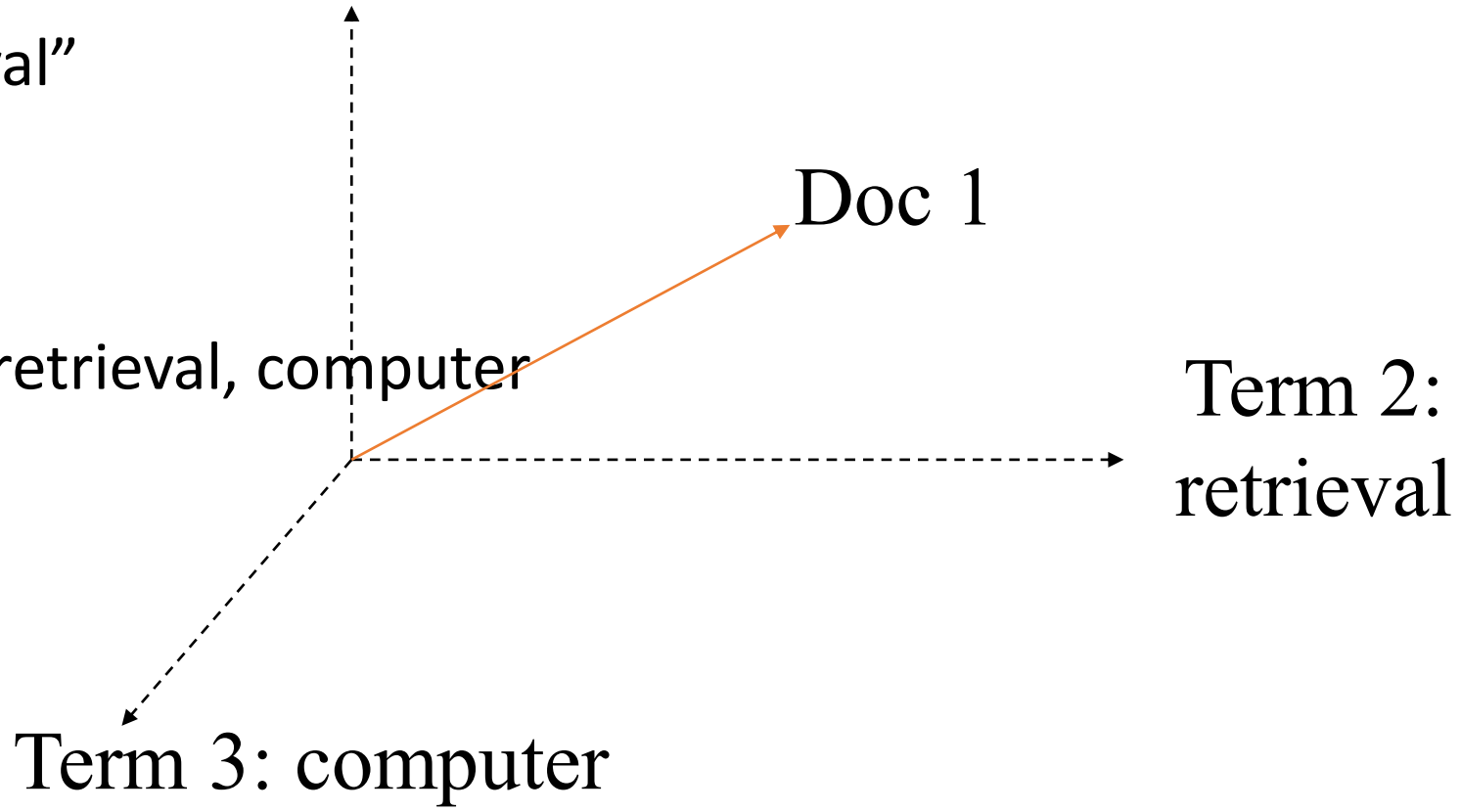
$$D = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$

Question: Doc 4 = “retrieval information retrieval” ?

Documents in a Vector Space

- Doc 1= “information retrieval”
- Doc 2 = “computer information retrieval”
- Doc 3 = “computer retrieval”

- Vocabulary: information, retrieval, computer
- Doc 1 = <1, 1, 0>

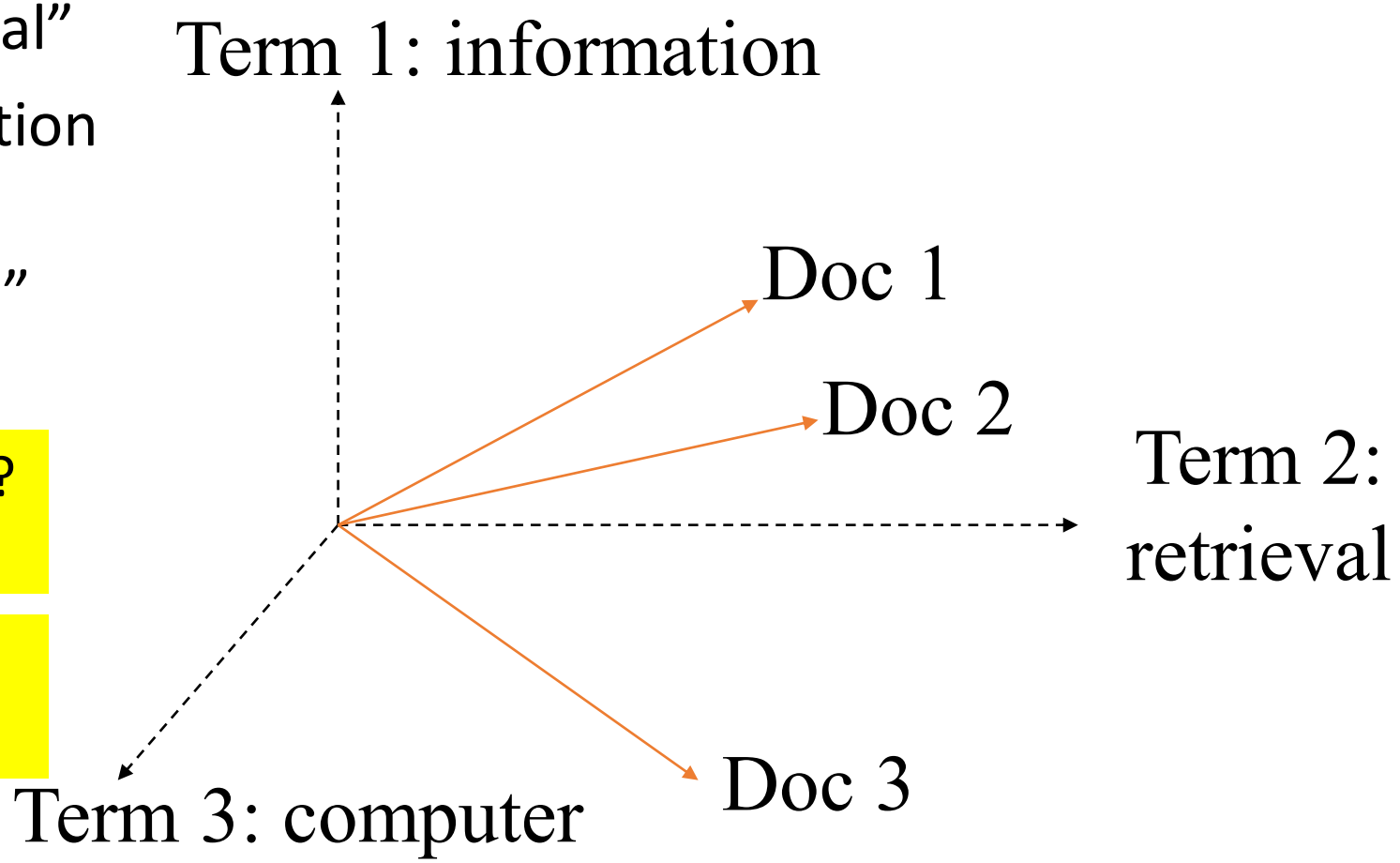


Relevance as Vector Similarity

- Doc 1= “information retrieval”
- Doc 2 = “computer information retrieval”
- Doc 3 = “computer retrieval”

Which document is closer to Doc 1? Doc 2 or Doc 3?

What if we have a query “retrieval”?



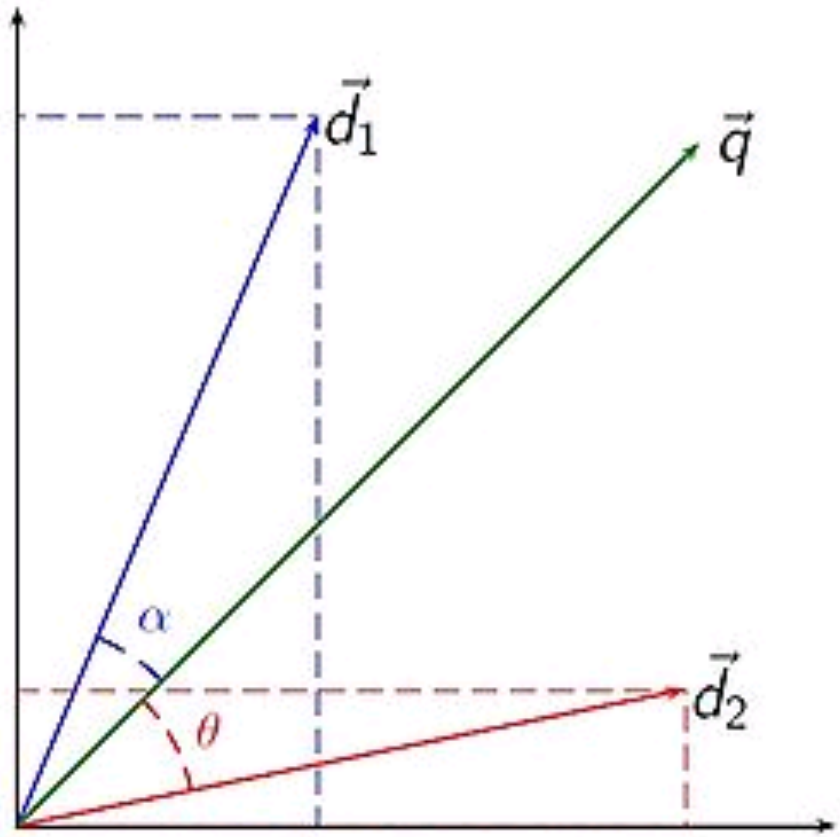
Distance/Similarity Calculation

- The similarity/relevance of two vectors can be calculated based on distance/similarity measures
- S: X, Y → (0, 1)
- X: $<x_1, x_2, \ldots, x_n>$
- Y: $<y_1, y_2, \ldots, y_n>$

- S(X, Y) = ?
  - The more dimensions in common, the larger the similarity
  - What about real values?
  - Normalization needed.

Document Similarity

- Used in information retrieval to determine which document (d_1 or d_2) is more similar to a given query q
- Documents and queries are represented in the same space
- Angle (or cosine) is a proxy for similarity between two vectors



Similarity Measures

- The Jaccard similarity (Similarity of Two Sets)

$$Jaccard(X, Y) = \frac{|X \cap Y|}{|X \cup Y|}$$

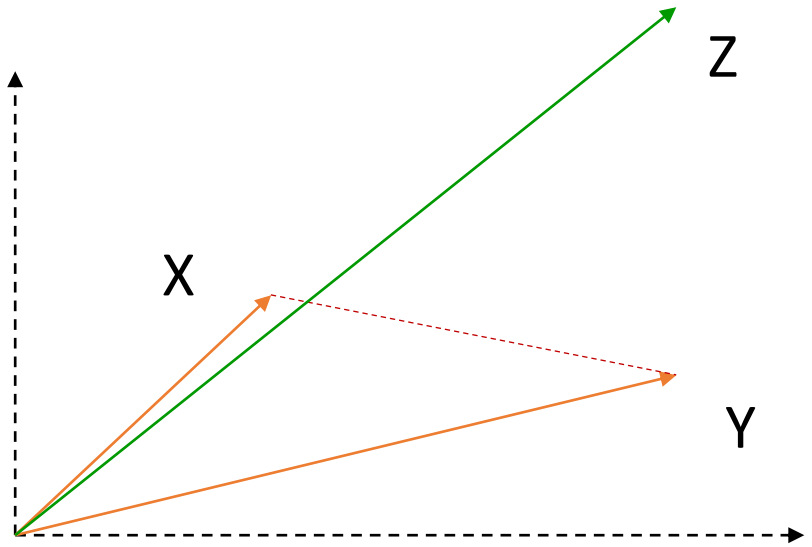
- D1 = “information retrieval class”
- D2 = “information retrieval algorithm”
- D3 = “processing information”
- What’s the Jaccard similarity of S(D1, D2)? S(D1, D3)?

- What about D3 = “information of information retrieval”

Similarity Measures

- Euclidean Distance – distance of two points

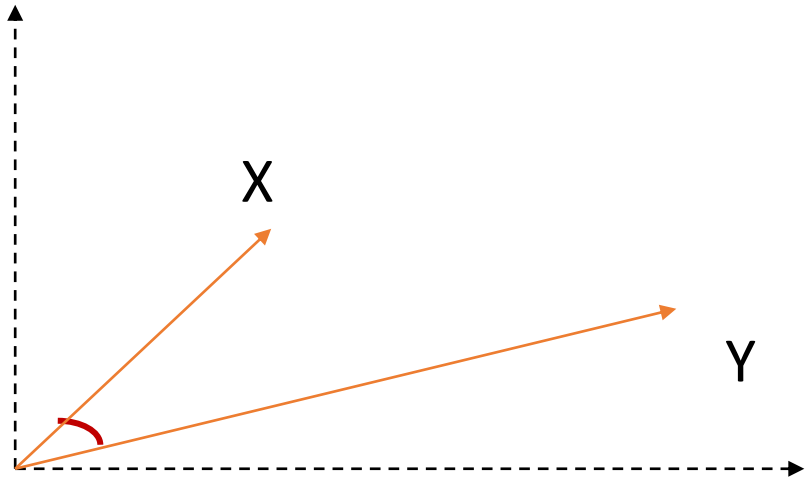
$$D(X,Y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$



Similarity Measures (Cont.)

- Cosine similarity: similarity of two vectors, normalized

$$\cos(X,Y) = \frac{x_1y_1 + x_2y_2 + \ldots + x_ny_n}{\sqrt{x_1^2 + \ldots + x_n^2} \times \sqrt{y_1^2 + \ldots + y_n^2}} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \times \sqrt{\sum_{i=1}^{n} y_i^2}}$$



Which one do you think is suitable for retrieval?
Jaccard? Euclidean? Cosine?

Example

- What is the cosine similarity between:
  - D= “cat,dog,dog” = <1,2,0>
  - Q= “cat,dog,mouse,mouse” = <1,1,2>
- Answer:

$$\sigma(D,Q) = \frac{1\times1+2\times1+0\times2}{\sqrt{1^2+2^2+0^2}\sqrt{1^2+1^2+2^2}} = \frac{3}{\sqrt{5}\sqrt{6}} \approx 0.55$$

- In comparison:

$$\sigma(D,D) = \frac{1\times1+2\times2+0\times0}{\sqrt{1^2+2^2+0^2}\sqrt{1^2+2^2+0^2}} = \frac{5}{\sqrt{5}\sqrt{5}} = 1$$

Quiz

- Given three documents

  D_1 = <1,3>

  D_2 = <10,30>

  D_3 = <3,1>

- Compute the cosine scores

  $\sigma(D_1,D_2)$

  $\sigma(D_1,D_3)$

- What do the numbers tell you?

Answers to the Quiz

$\sigma(D_1,D_2) = 1$

one of the two documents is a scaled version of the other

$\sigma(D_1,D_3) = 0.6$

swapping the two dimensions results in a lower similarity

Quiz

- What is the range of values that the cosine scores can take?

Answer to the Quiz

- Mathematically, the cosine function has a range of [-1,1]
- However, when the two vectors are both in the first quadrant (since all word counts are non-negative), the range is [0,1]
- For word embeddings, the range is [-1,1] (since the values don’t have to be non-negative)

Term-Document Matrix

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	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.

Term-Document Matrix

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

Figure 6.3 The term-document matrix for four words in four Shakespeare plays. The red boxes show that each document is represented as a column vector of length four.

Representing Documents as Vectors

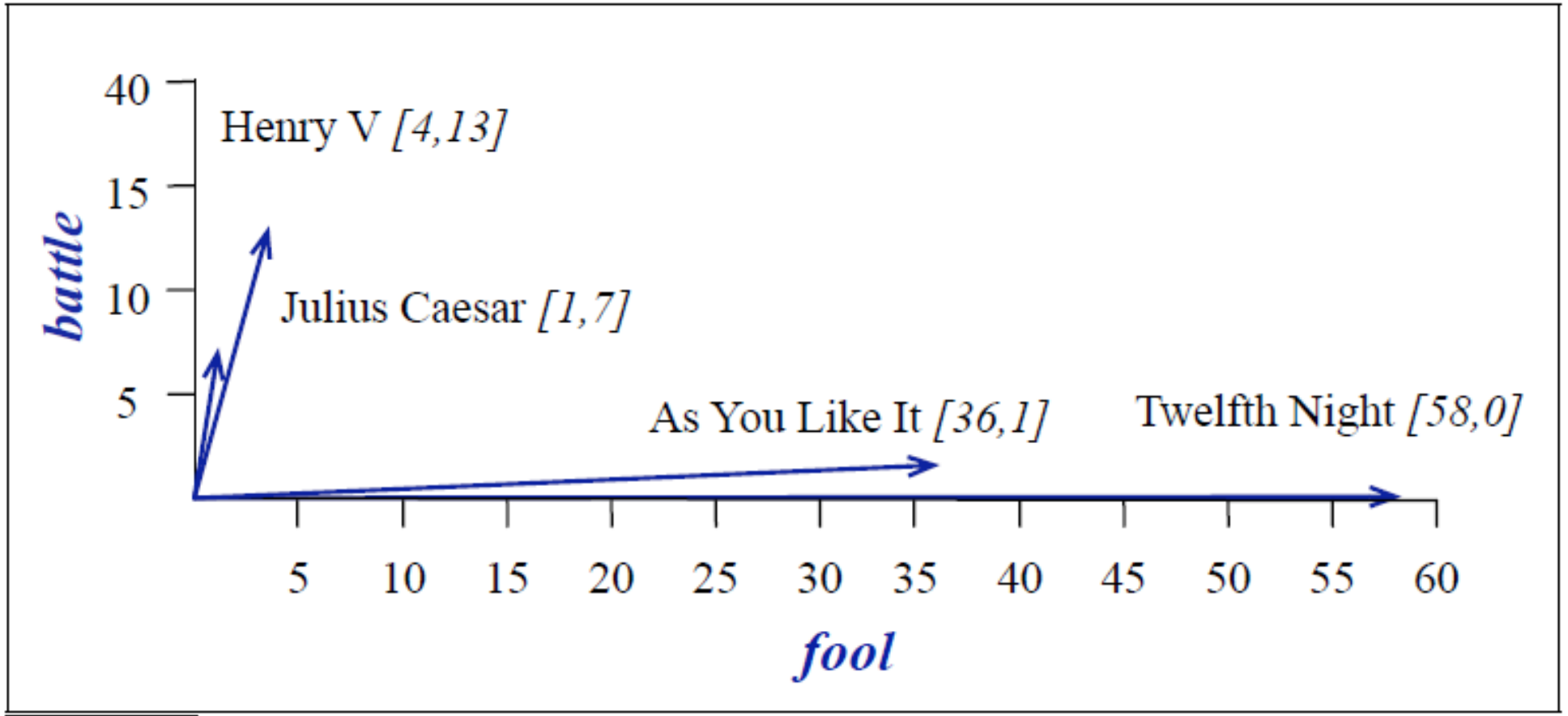


Figure 6.4 A spatial visualization of the document vectors for the four Shakespeare play documents, showing just two of the dimensions, corresponding to the words *battle* and *fool*. The comedies have high values for the *fool* dimension and low values for the *battle* dimension.

Example (Cont'd)

	aardvark	...	computer	data	result	pie	sugar	...
cherry	0	...	2	8	9	442	25	
strawberry	0	...	0	0	1	60	19	
digital	0	...	1670	1683	85	5	4	
information	0	...	3325	3982	378	5	13	

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

Example (Cont'd)

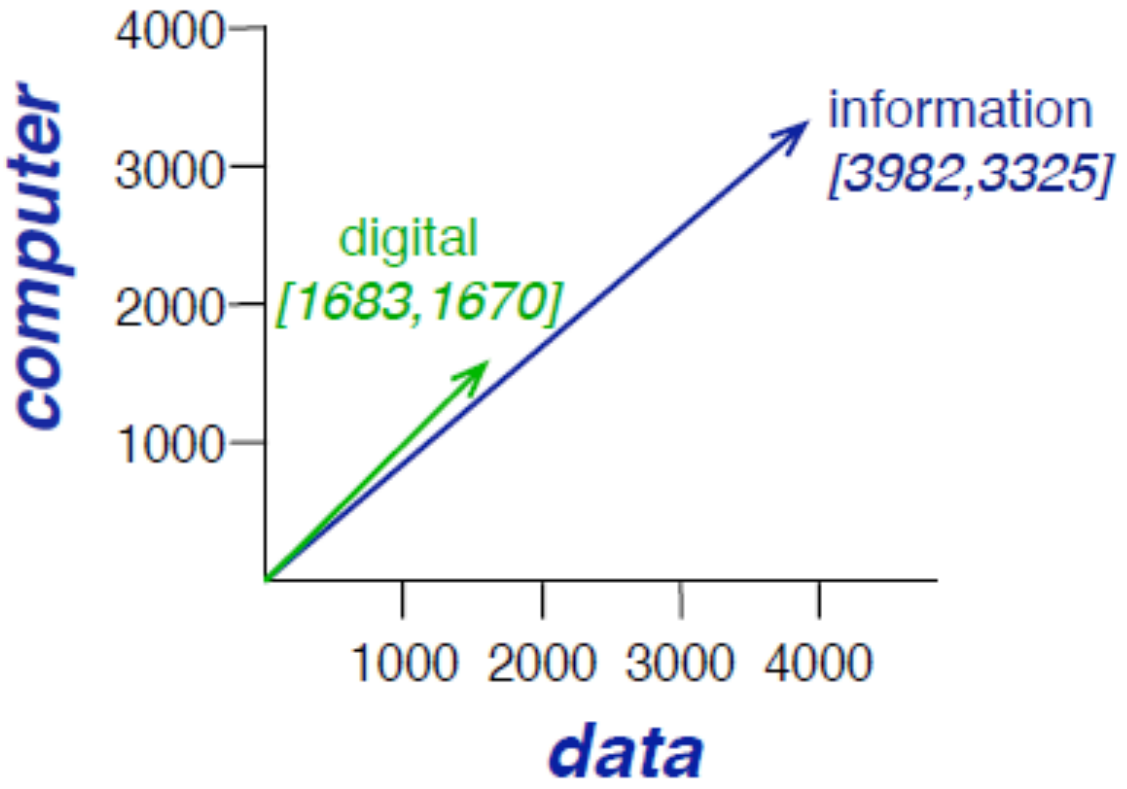


Figure 6.6 A spatial visualization of word vectors for *digital* and *information*, showing just two of the dimensions, corresponding to the words *data* and *computer*.

Example (Cont'd)

	pie	data	computer
cherry	442	8	2
digital	5	1683	1670
information	5	3982	3325

$$\cos(\text{cherry}, \text{information}) = \frac{442*5+8*3982+2*3325}{\sqrt{442^2+8^2+2^2}\sqrt{5^2+3982^2+3325^2}} = .017$$

$$\cos(\text{digital}, \text{information}) = \frac{5*5+1683*3982+1670*3325}{\sqrt{5^2+1683^2+1670^2}\sqrt{5^2+3982^2+3325^2}} = .996$$

Example (Cont'd)

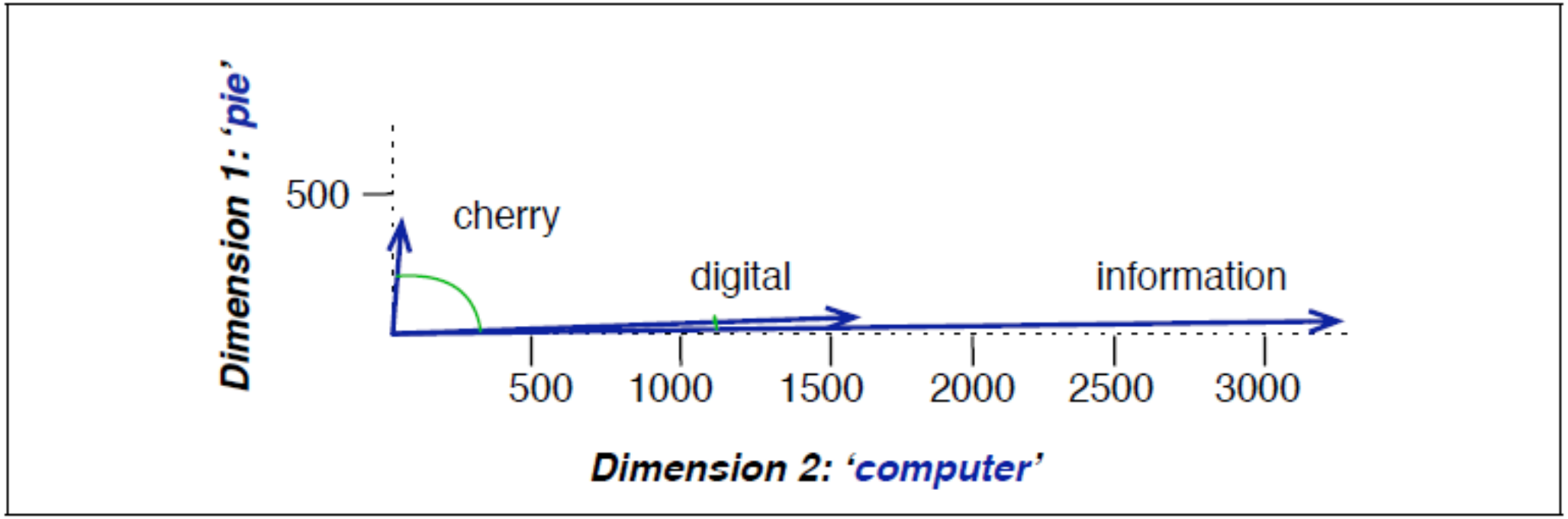


Figure 6.7 A (rough) graphical demonstration of cosine similarity, showing vectors for three words (*cherry*, *digital*, and *information*) in the two dimensional space defined by counts of the words *computer* and *pie* nearby. Note that the angle between *digital* and *information* is smaller than the angle between *cherry* 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°); the cosine of all other angles is less than 1.

Term Frequency and Inverse Document Frequency

	Collection Frequency	Document Frequency
Romeo	113	1
action	113	31

$$\text{idf}_t = \log_{10}\left(\frac{N}{\text{df}_t}\right)$$

Word	df	idf
Romeo	1	1.57
salad	2	1.27
Falstaff	4	0.967
forest	12	0.489
battle	21	0.246
wit	34	0.037
fool	36	0.012
good	37	0
sweet	37	0

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	0.074	0	0.22	0.28
good	0	0	0	0
fool	0.019	0.021	0.0036	0.0083
wit	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. For example the 0.049 value for *wit* in *As You Like It* is the product of $\text{tf} = \log_{10}(20+1) = 1.322$ and $\text{idf} = .037$. 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*.

Review: Mutual Information

$$I(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)}$$

$$\text{PMI}(w,c) = \log_2 \frac{P(w,c)}{P(w)P(c)}$$

$$\text{PPMI}(w,c) = \max(\log_2 \frac{P(w,c)}{P(w)P(c)}, 0)$$

Mutual Information: Example

$$p_{ij} = \frac{f_{ij}}{\sum_{i=1}^{W}\sum_{j=1}^{C} f_{ij}} \quad p_{i*} = \frac{\sum_{j=1}^{C} f_{ij}}{\sum_{i=1}^{W}\sum_{j=1}^{C} f_{ij}} \quad p_{*j} = \frac{\sum_{i=1}^{W} f_{ij}}{\sum_{i=1}^{W}\sum_{j=1}^{C} f_{ij}}$$

$$\text{PPMI}_{ij} = \max(\log_2 \frac{p_{ij}}{p_{i*}p_{*j}}, 0)$$

	computer	data	result	pie	sugar	count(w)
cherry	2	8	9	442	25	486
strawberry	0	0	1	60	19	80
digital	1670	1683	85	5	4	3447
information	3325	3982	378	5	13	7703
count(context)	4997	5673	473	512	61	11716

Figure 6.9 Co-occurrence counts for four words in 5 contexts in the Wikipedia corpus, together with the marginals, pretending for the purpose of this calculation that no other words/contexts matter

Example (Cont'd)

$$P(\text{w=information,c=data}) = \frac{3982}{11716} = .3399$$

$$P(\text{w=information}) = \frac{7703}{11716} = .6575$$

$$P(\text{c=data}) = \frac{5673}{11716} = .4842$$

$$\text{ppmi(information,data)} = \log 2(.3399/(.6575 * .4842)) = .0944$$

Example (Cont'd)

	p(w,context)					p(w)
	computer	**data**	**result**	**pie**	**sugar**	**p(w)**
cherry	0.0002	0.0007	0.0008	0.0377	0.0021	0.0415
strawberry	0.0000	0.0000	0.0001	0.0051	0.0016	0.0068
digital	0.1425	0.1436	0.0073	0.0004	0.0003	0.2942
information	0.2838	0.3399	0.0323	0.0004	0.0011	0.6575
p(context)	0.4265	0.4842	0.0404	0.0437	0.0052	

Figure 6.10 Replacing the counts in Fig. 6.5 with joint probabilities, showing the marginals around the outside.

Example (Cont'd)

	computer	data	result	pie	sugar
cherry	0	0	0	4.38	3.30
strawberry	0	0	0	4.10	5.51
digital	0.18	0.01	0	0	0
information	0.02	0.09	0.28	0	0

Figure 6.11 The PPMI matrix showing the association between words and context words, computed from the counts in Fig. 6.10. Note that most of the 0 PPMI values are ones that had a negative PMI; for example PMI(*cherry,computer*) = -6.7, meaning that *cherry* and *computer* co-occur on Wikipedia less often than we would expect by chance, and with PPMI we replace negative values by zero.

Text Similarity

322.

Vector Semantics

What does “acerola” mean?

- *acerola* is a significant source of vitamin C.
- the pulp of the *acerola* is very soft
- *acerola* are now found growing in most sub-tropical regions of the world.
- *acerola* can be eaten fresh or used to make jams or jellies.

Distributional similarity

- Two words that appear in similar contexts are likely to be semantically related, e.g.,
  - schedule a test ***drive*** and investigate **Honda**'s financing options
  - **Volkswagen** debuted a new version of its front-wheel-***drive*** Golf
  - the **Jeep** reminded me of a recent ***drive***
  - Our test ***drive*** took place at the wheel of loaded **Ford** EL model
- “You will know a word by the company that it keeps.” (J.R. Firth 1957)

Basic Ideas

- Represent words as vectors
  - For example, based on nearby words
- Similar words (synonyms) should have similar representations
- Different senses of the same word should have different representations
- Relations should be preserved
  - For example, “cat”-”kitten” should be similar to “dog”-”puppy”

Context Features

- The context features can be any of the following:
  - The word before the target word
  - The word after the target word
  - Any word within *n* words of the target word
  - Any word within a specific syntactic relationship with the target word (e.g., the head of the dependency or the subject of the sentence)
  - Any word within the same sentence
  - Any word within the same document

Example

- S1: schedule a test ***drive*** and investigate **Honda**'s financing options
- S2: **Volkswagen** debuted a new version of its front-wheel-***drive*** Golf
- S3: the **Jeep** reminded me of a recent ***drive***
- S4: Our test ***drive*** took place at the wheel of loaded **Ford** EL model

	schedule	test	drive	version	front	recent	model
Honda	1	1	**1**				
Vokswagen			**1**	1	1		
Jeep			**1**			1	
Ford		1	**1**				1

t-SNE Projection

Figure 6.1 A two-dimensional (t-SNE) projection of embeddings for some words and phrases, showing that words with similar meanings are nearby in space. The original 60-dimensional embeddings were trained for sentiment analysis. Simplified from Li et al. (2015).

Text Similarity

341.

Dimensionality Reduction

Issues with Vector Similarity

- Polysemy (sim < cos)
  - bar, bank, jaguar, hot
- Synonymy (sim > cos)
  - building/edifice, large/big, spicy/hot
- Relatedness (people are really good at figuring this)
  - doctor/patient/nurse/treatment

Semantic Matching

Query = “natural language processing”

Document 1 = “linguistics semantics viterbi learning”

Document 2 = “welcome to new haven”

- Which one should we rank higher?
- Query vocabulary & doc vocabulary mismatch!
- If only we can represent documents/queries as concepts!
- That’s where dimensionality reduction helps

Semantic Concepts

	election	vote	president	tomato	salad
NEWS1	4	4	4	0	0
NEWS2	3	3	3	0	0
NEWS3	1	1	1	0	0
NEWS4	5	5	5	0	0
RECIPE1	0	0	0	1	1
RECIPE2	0	0	0	4	4
RECIPE3	0	0	0	1	1

Semantic Concepts

	election	vote	president	tomato	salad
NEWS1	4	4	4	0	0
NEWS2	3	3	3	0	0
NEWS3	1	1	1	0	0
NEWS4	5	5	5	0	0
RECIPE1	0	0	0	1	1
RECIPE2	0	0	0	4	4
RECIPE3	0	0	0	1	1

Concept Space = Dimension Reduction

- Number of concepts (K) is smaller than the number of words (N) or number of documents (M).
- If we represent a document as a N-dimensional vector; and the corpus as an M*N matrix…
  - The goal is to reduce the dimensionality from N to K.
  - But how can we do that?

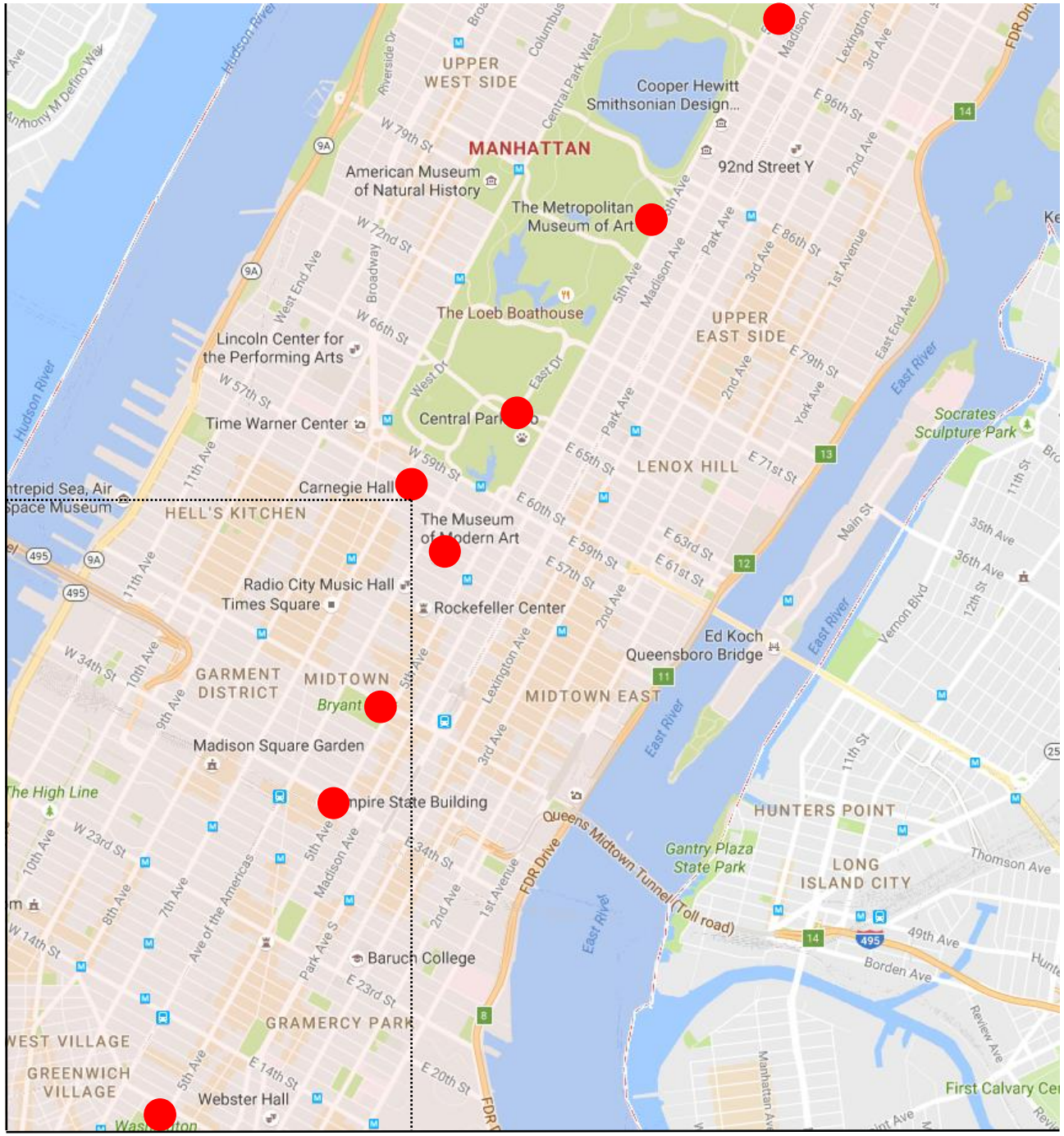
TOEFL Synonyms and SAT Analogies

- Word similarity vs. analogies

Stem:		**levied**
Choices:	**(a)**	**imposed**
	(b)	**believed**
	(c)	**requested**
	(d)	**correlated**
Solution:	**(a)**	**imposed**

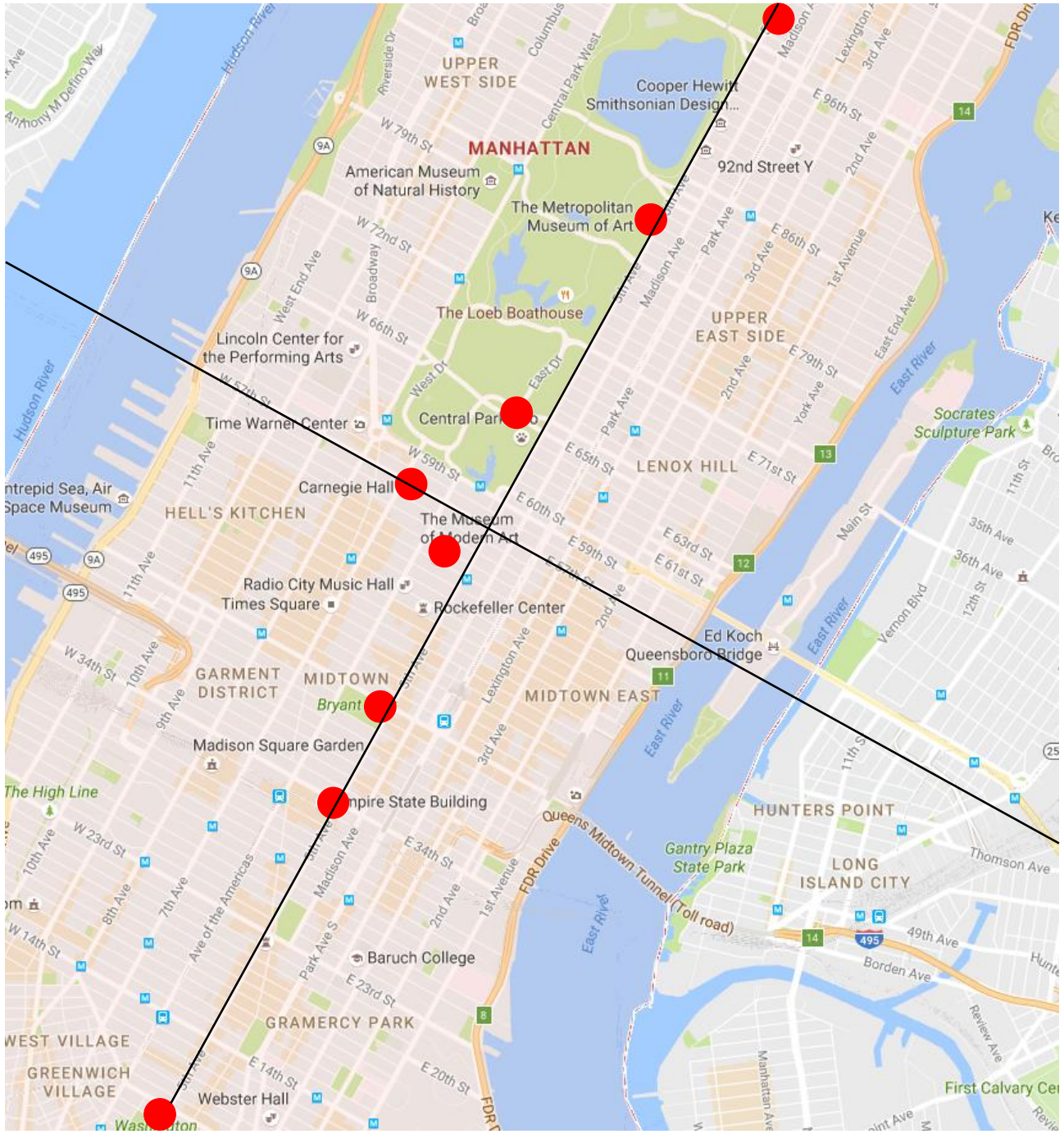
Stem:	**mason:stone**
Choices:	**(a) teacher:chalk**
	(b) carpenter:wood
	(c) soldier:gun
	(d) photograph:camera
	(e) book:word
Solution:	**(b) carpenter:wood**

Example from Peter Turney


UPPER WEST SIDE
MANHATTAN
Cooper Hewitt Smithsonian Design...
92nd Street Y
American Museum of Natural History
The Metropolitan Museum of Art
The Loeb Boathouse
UPPER EAST SIDE
Lincoln Center for the Performing Arts
Time Warner Center
Central Park
Carnegie Hall
LENOX HILL
Socrates Sculpture Park
HELL'S KITCHEN
The Museum of Modern Art
Radio City Music Hall
Times Square
Rockefeller Center
Ed Koch Queensboro Bridge
GARMENT DISTRICT
MIDTOWN
MIDTOWN EAST
Bryant
Madison Square Garden
The High Line
Empire State Building
HUNTERS POINT
Gantry Plaza State Park
LONG ISLAND CITY
Queens Midtown Tunnel (Toll road)
FDR Drive
East River
Baruch College
GRAMERCY PARK
WEST VILLAGE
GREENWICH VILLAGE
Webster Hall
First Calvary Ce
Hudson River
East River


UPPER
WEST SIDE
MANHATTAN
Cooper Hewitt
Smithsonian Design...
92nd Street Y
American Museum
of Natural History
The Metropolitan
Museum of Art
The Loeb Boathouse
UPPER
EAST SIDE
Lincoln Center for
the Performing Arts
Time Warner Center
Central Park
Carnegie Hall
The Museum
of Modern Art
LENOX HILL
HELL'S KITCHEN
Intrepid Sea, Air
Space Museum
Radio City Music Hall
Times Square
Rockefeller Center
Ed Koch
Queensboro Bridge
GARMENT
DISTRICT
MIDTOWN
MIDTOWN EAST
Bryant
Madison Square Garden
The High Line
Empire State Building
HUNTERS POINT
Gantry Plaza
State Park
LONG
ISLAND CITY
Queens Midtown Tunnel (Toll road)
FDR Drive
Baruch College
GRAMERCY PARK
WEST VILLAGE
GREENWICH
VILLAGE
Webster Hall
Socrates
Sculpture Park
Hudson River
East River
Broadway
Park Ave
Madison Ave
1st Avenue
2nd Ave
3rd Ave
Lexington Ave
West End Ave
East End Ave
York Ave
Main St
Vernon Blvd
Borden Ave
49th Ave
Thomson Ave
First Calvary Ce

29 degrees


UPPER
WEST SIDE
MANHATTAN
Cooper Hewitt
Smithsonian Design...
92nd Street Y
American Museum
of Natural History
The Metropolitan
Museum of Art
The Loeb Boathouse
Lincoln Center for
the Performing Arts
UPPER
EAST SIDE
Time Warner Center
Central Park
Socrates
Sculpture Park
LENOX HILL
Carnegie Hall
HELL'S KITCHEN
The Museum
of Modern Art
Radio City Music Hall
Times Square
Rockefeller Center
Ed Koch
Queensboro Bridge
GARMENT
DISTRICT
MIDTOWN
MIDTOWN EAST
Bryant
Madison Square Garden
The High Line
Empire State Building
HUNTERS POINT
Gantry Plaza
State Park
LONG
ISLAND CITY
Queens Midtown Tunnel (Toll road)
FDR Drive
Baruch College
GRAMERCY PARK
WEST VILLAGE
GREENWICH
VILLAGE
Webster Hall
Hudson River
East River

29 degrees

Vectors and Matrices

- A matrix is an m x n table of objects (in our case, numbers)
- Each row (or column) is a vector.
- Matrices of compatible dimensions can be multiplied together.
- What is the result of the multiplication below?

$$\begin{bmatrix} 1 & 2 & 4 \\ 2 & 5 & 7 \\ 4 & 9 & 14 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \\ -1 \end{bmatrix} = \begin{bmatrix} \quad \\ \quad \\ \quad \end{bmatrix}$$

Answer to the Quiz

$$\begin{bmatrix} 1 & 2 & 4 \\ 2 & 5 & 7 \\ 4 & 9 & 14 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \\ -1 \end{bmatrix} = \begin{bmatrix} 1 \times 2 + 2 \times 1 + 4 \times (-1) \\ 2 \times 2 + 5 \times 1 + 7 \times (-1) \\ 4 \times 2 + 9 \times 1 + 14 \times (-1) \end{bmatrix} = \begin{bmatrix} 0 \\ 2 \\ 3 \end{bmatrix}$$

Eigenvectors and Eigenvalues

- An eigenvector is an implicit “direction” for a matrix A

$$Av = \lambda v$$

  - v (the eigenvector) is non-zero
  - λ (the eigenvalue) can be any complex number, in principle.
- Computing eigenvalues:

$$\det(A - \lambda I) = 0$$

Eigenvectors and Eigenvalues

Example:

$$A=\begin{pmatrix}-1 & 3\\ 2 & 0\end{pmatrix} \qquad A-\lambda I=\begin{pmatrix}-1-\lambda & 3\\ 2 & -\lambda\end{pmatrix}$$

$\det(A-\lambda I) = (-1-\lambda)*(-\lambda)-3*2=0$

Then: $\lambda+\lambda^2-6=0$; $\lambda_1=2$; $\lambda_2=-3$

For $\lambda_1=2$:

$$\begin{pmatrix}-3 & 3\\ 2 & -2\end{pmatrix}\begin{pmatrix}v_1\\ v_2\end{pmatrix}=0$$

Solutions: $v_1=v_2$

Matrix decomposition

- If Σ is a square matrix, it can be decomposed into $U\Lambda U^{-1}$, where

  U = matrix of eigenvectors

  Λ = diagonal matrix of eigenvalues

$\Sigma U = U\Lambda$

$U^{-1}\Sigma U = \Lambda$

$\Sigma = U\Lambda U^{-1}$

Example

$$S=\begin{pmatrix}2 & 1\\ 1 & 2\end{pmatrix}, \lambda_1=1, \lambda_2=3$$

$$U=\begin{pmatrix}1 & 1\\ -1 & 1\end{pmatrix}$$

$$U^{-1}=\begin{pmatrix}1/2 & -1/2\\ 1/2 & 1/2\end{pmatrix}$$

$$S=U\Lambda U^{-1}=\begin{pmatrix}1 & 1\\ -1 & 1\end{pmatrix}\begin{pmatrix}1 & 0\\ 0 & 3\end{pmatrix}\begin{pmatrix}1/2 & -1/2\\ 1/2 & 1/2\end{pmatrix}$$

SVD: Singular Value Decomposition

- $A=U\Sigma V^T$
  - U is the matrix of orthogonal eigenvectors of AA^T
  - V is the matrix of orthogonal eigenvectors of A^TA (co-variance matrix)
  - The components of Σ are the eigenvalues of A^TA
- Properties
  - This decomposition exists for all matrices and is unique
  - U, V are column orthonormal
  - $U^T U = I$; $V^T V = I$
  - Σ is diagonal and sorted by absolute value of the singular values (large to small)
  - Each column (row) of Σ corresponds to a principal component
  - If A has 5 columns and 3 rows, then U will be 5x5 and V will be 3x3

Example (Berry and Browne)

terms

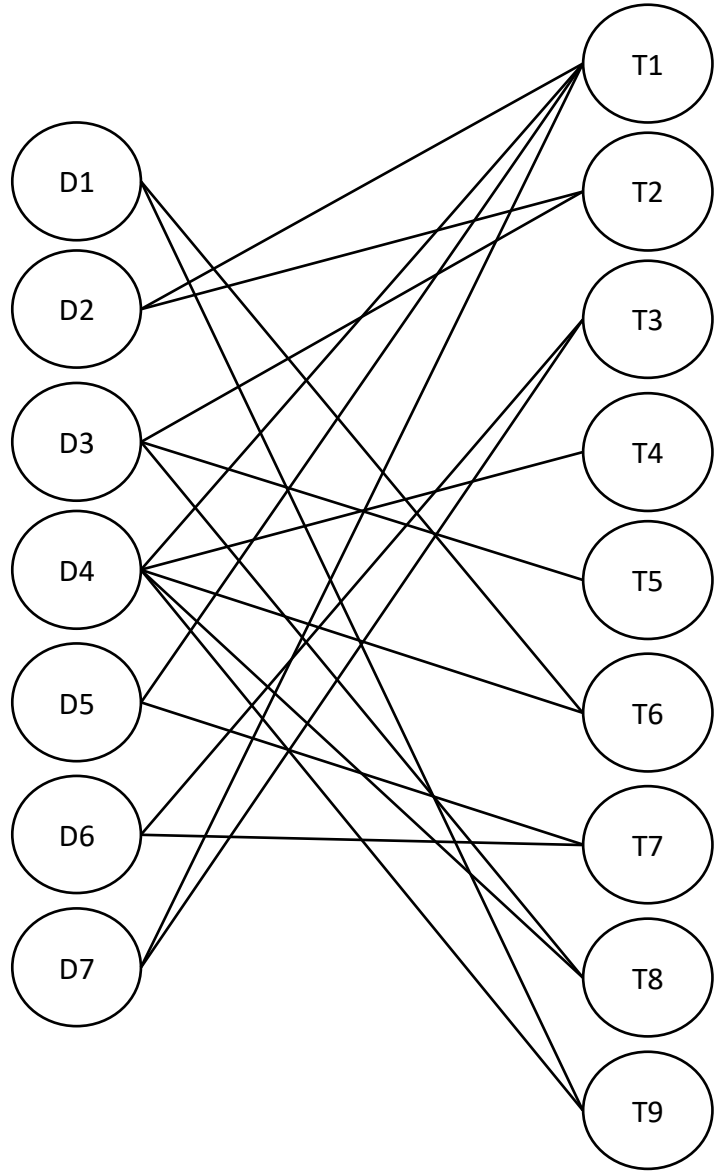
T1: baby
T2: child
T3: guide
T4: health
T5: home
T6: infant
T7: proofing
T8: safety
T9: toddler

documents

D1: infant & toddler first aid
D2: babies & children's room (for your home)
D3: child safety at home
D4: your baby's health and safety: from infant to toddler
D5: baby proofing basics
D6: your guide to easy rust proofing
D7: beanie babies collector's guide

Example

D1: T6, T9
D2: T1, T2
D3: T2, T5, T8
D4: T1, T4, T6, T8, T9
D5: T1, T7
D6: T3, T7
D7: T1, T3



Document-Term Matrix

raw

$$A=\begin{bmatrix} 0 & 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

normalized

$$A^{(n)}=\begin{bmatrix} 0 & 0.58 & 0 & 0.45 & 0.71 & 0 & 0.71 \\ 0 & 0.58 & 0.58 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.71 & 0.71 \\ 0 & 0 & 0 & 0.45 & 0 & 0 & 0 \\ 0 & 0.58 & 0.58 & 0 & 0 & 0 & 0 \\ 0.71 & 0 & 0 & 0.45 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.71 & 0.71 & 0 \\ 0 & 0 & 0.58 & 0.45 & 0 & 0 & 0 \\ 0.71 & 0 & 0 & 0.45 & 0 & 0 & 0 \end{bmatrix}$$

Dimensionality Reduction

- Low rank matrix approximation

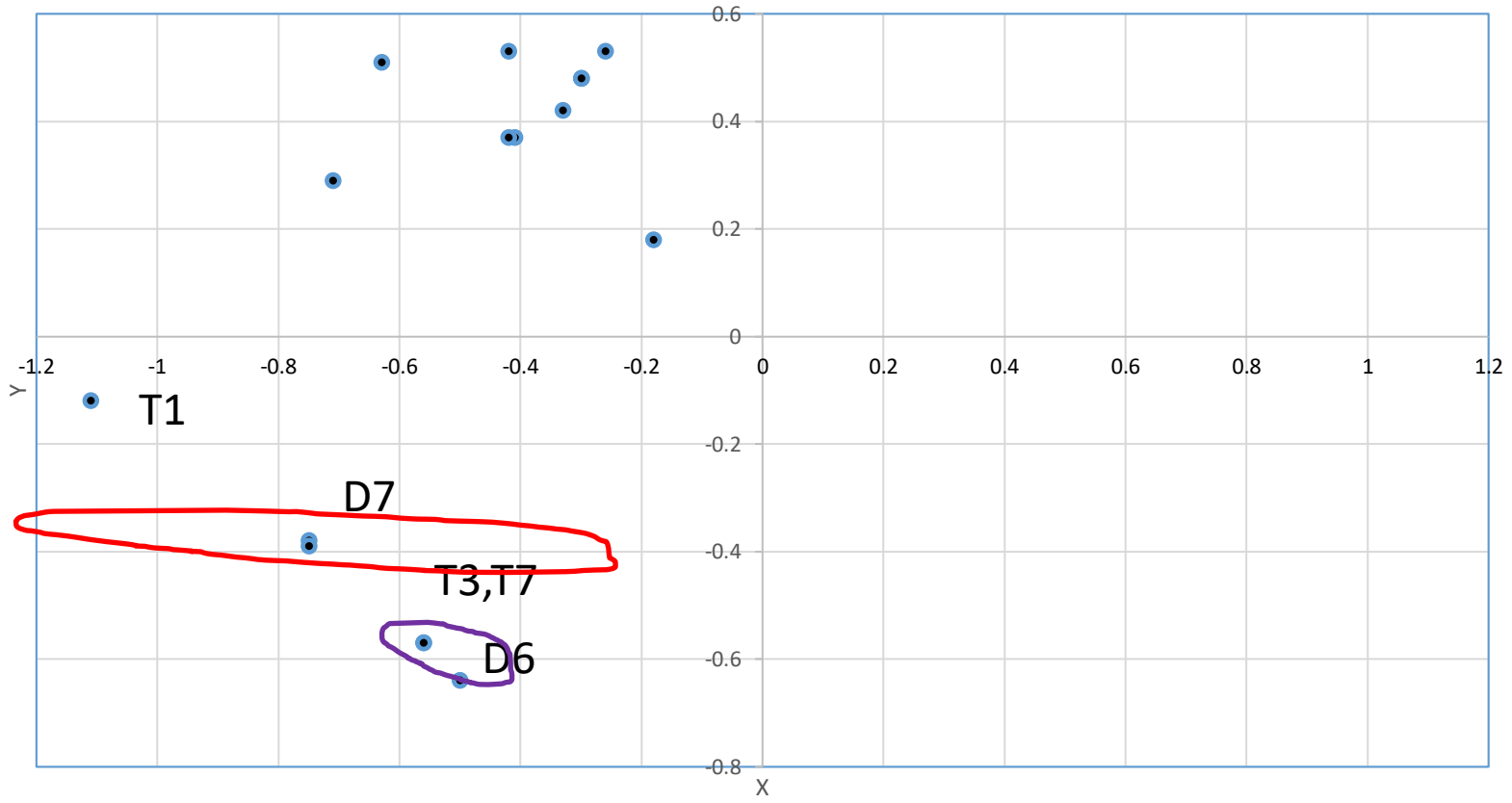
  $A_{[m*n]} = U_{[m*m]}\Sigma_{[m*n]}V^T_{[n*n]}$

- Σ is a diagonal matrix of eigenvalues
- If we only keep the largest *r* eigenvalues

  $A \approx U_{[m*r]}\Sigma_{[r*r]}V^T_{[n*r]}$

D1	D2	D3	D4	D5	D6	D7
-0.26	-0.71	-0.42	-0.62	-0.74	-0.50	-0.74
0.53	0.29	0.54	0.51	-0.38	-0.64	-0.38

T1	T2	T3	T4	T5	T6	T7	T8	T9
-1.10	-0.41	-0.56	-0.18	-0.41	-0.30	-0.56	-0.33	-0.30
-0.12	0.38	-0.57	0.18	0.38	0.47	-0.57	0.42	0.47



Semantic Concepts

	election	vote	president	tomato	salad
NEWS1	4	4	4	0	0
NEWS2	3	3	3	0	0
NEWS3	1	1	1	0	0
NEWS4	5	5	5	0	0
RECIPE1	0	0	0	1	1
RECIPE2	0	0	0	4	4
RECIPE3	0	0	0	1	1

Adding Noise

	election	vote	president	tomato	salad
NEWS1	4	4	4	0	0
NEWS2	3	**0**	3	0	**2**
NEWS3	1	1	1	**1**	0
NEWS4	5	5	**3**	0	0
RECIPE1	0	0	0	1	1
RECIPE2	0	**1**	0	4	4
RECIPE3	0	0	0	**0**	1

Quiz

- Let A be a document x term matrix.
- What is A*A’?
- What about A’*A?

Interpretation of SVD

- Best direction to project on
  - The principal eigenvector is the dimension that explains most of the variance
- Finding hidden concepts
  - Mapping documents, terms to a lower-dimensional space
- Turning the matrix into block-diagonal form
  - (same as finding bi-partite cores)
- In the NLP/IR literature, SVD is called LSA (LSI)
  - Latent Semantic Analysis (Indexing)
- Keep as many dimensions as necessary to explain 80-90% of the data (energy)
  - In practice, use 300 dimensions or so

Problem #5 (20 points). In a series of experiments run in Carnegie Mellon University (Pittsburgh, USA) in 2010, volunteers were first shown some English words, while activity was being registered in different locations of their brains. Then the volunteers were asked to think of some other words from a preselected list of 60 words, while the researchers were measuring their brain activity again. Using the obtained data, the researchers were able to determine the words the volunteers were thinking of quite successfully.

Below you can find some data on the activity levels for four brain locations depending on which word the volunteers were thinking of.

Word	Translation	Location A	Location B	Location C	Location D
airplane	airplane	high	low	low	high
apartment	apartment	high	low	low	high
arm	arm	low	high	low	low
corn	corn	low	low	high	low
cup	cup	low	low	high	low
igloo	igloo	high	low	low	low
key	key	high	high	low	low
lettuce	lettuce	low	low	high	high
screwdriver	screwdriver	low	high	low	high

The same information is given below on six more words the volunteers were thinking of: ***bed*** 'bed', ***butterfly*** 'butterfly', ***cat*** 'cat', ***cow*** 'cow', ***refrigerator*** 'refrigerator', ***spoon*** 'spoon'.

Word	Location A	Location B	Location C	Location D
1	low	low	high	high
2	low	low	high	low
3	high	low	low	low
4	low	low	low	high
5	low	high	high	low
6	low	low	low	low

Determine the correct correspondences.

—*Boris Iomdin*

Problem #5. Location A is activated by the idea of shelter. Location B is activated by the idea of manipulation. Location C is activated by the idea of eating. Location D is activated by long words. The researchers claim that the first three factors have high ecological validity (i. e., the results of the experiment conform to the data on human behaviour in real life) and survival value, and that Location D is responsible for a low-level visual representation of the printed word.

Word	Translation	Location A (shelter)	Location B (manipulation)	Location C (eating)	Location D (long words)
refrigerator	'refrigerator'	low	low	high	high
cow	'cow'	low	low	high	low
bed	'bed'	high	low	low	low
butterfly	'butterfly'	low	low	low	high
spoon	'spoon'	low	high	high	low
cat	'cat'	low	low	low	low

fMRI example

- fMRI
  - functional MRI (magnetic resonance imaging)
  - Used to measure activity in different parts of the brain when exposed to various stimuli
- Factor analysis
- Paper
  - Just, M. A., Cherkassky, V. L., Aryal, S., & Mitchell, T. M. (2010). A neurosemantic theory of concrete noun representation based on the underlying brain codes. PLoS ONE, 5, e8622

Table 1. 60 stimulus words grouped into 12 semantic categories.

Category	Exemplar 1	Exemplar 2	Exemplar 3	Exemplar 4	Exemplar 5
body parts	leg	arm	eye	foot	hand
furniture	chair	table	bed	desk	dresser
vehicles	car	airplane	train	truck	bicycle
animals	horse	dog	bear	cow	cat
kitchen utensils	glass	knife	bottle	cup	spoon
tools	chisel	hammer	screwdriver	pliers	saw
buildings	apartment	barn	house	church	igloo
building parts	window	door	chimney	closet	arch
clothing	coat	dress	shirt	skirt	pants
insects	fly	ant	bee	butterfly	beetle
vegetables	lettuce	tomato	carrot	corn	celery
man-made objects	refrigerator	key	telephone	watch	bell

doi:10.1371/journal.pone.0008622.t001

[Just et al. 2010]

Table 2. Ten words with highest factor scores (in descending order) for each of the 4 factors.

Shelter	*Manipulation*	*Eating*	*Word length*
apartment	pliers	carrot	butterfly
church	saw	lettuce	screwdriver
train	screwdriver	tomato	telephone
house	hammer	celery	refrigerator
airplane	key	cow	bicycle
key	knife	saw	apartment
truck	bicycle	corn	dresser
door	chisel	bee	lettuce
car	spoon	glass	chimney
closet	arm	cup	airplane

doi:10.1371/journal.pone.0008622.t002

[Just et al. 2010]

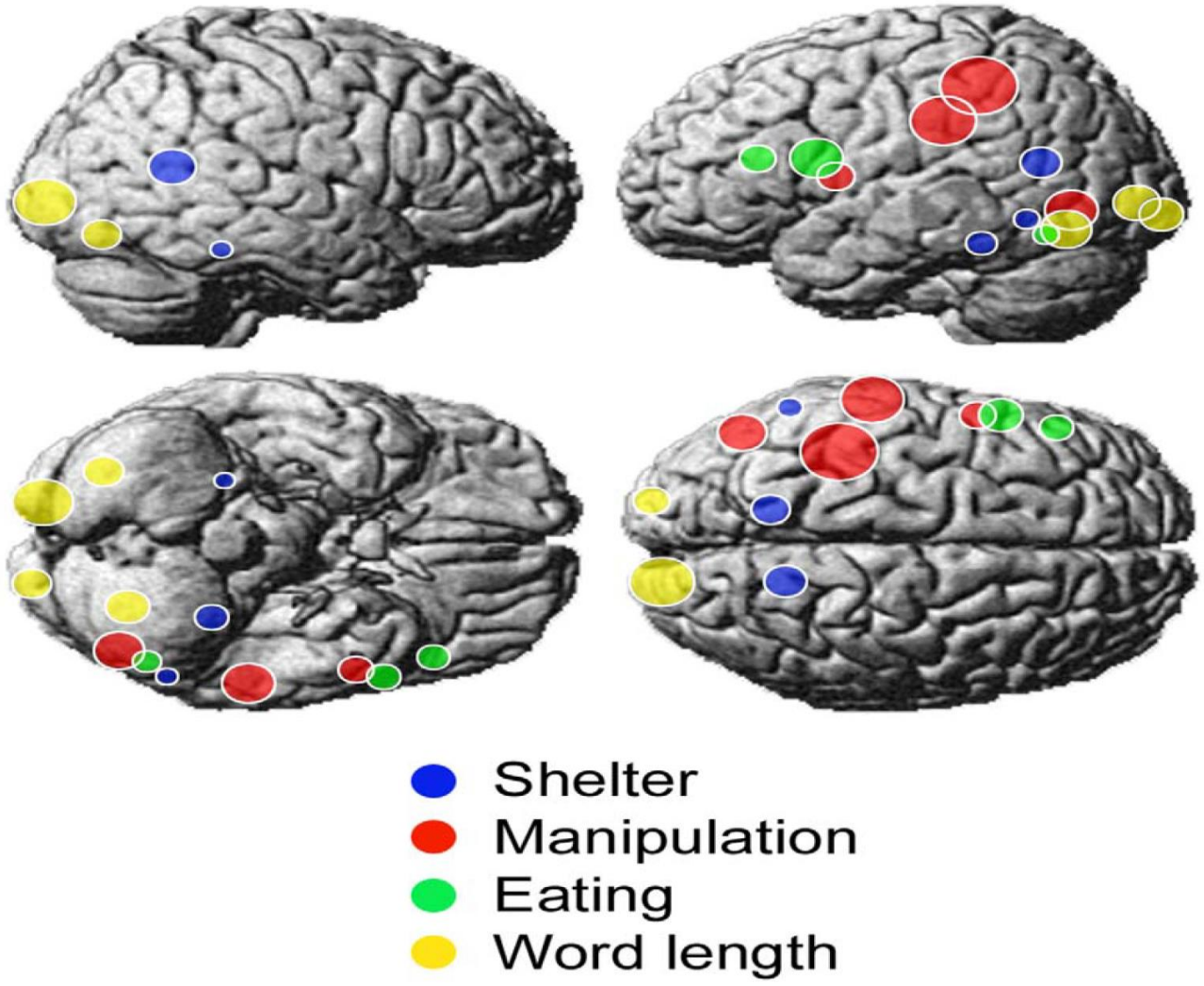


Figure 1. Locations of the voxel clusters (spheres) associated with the four factors. The spheres (shown as surface projections) are centered at the cluster centroid, with a radius equal to the mean radial dispersion of the cluster voxels.
doi:10.1371/journal.pone.0008622.g001

[Just et al. 2010]

Table 3. Locations (MNI centroid coordinates) and sizes of the voxel clusters associated with the four factors.

Factor	Cluster location	x	y	z	No. of voxels	Radius (mm)
shelter	L Fusiform Gyrus/Parahippocampal Gyrus (PPA)	−32	−42	−18	26	6
	R Fusiform Gyrus/Parahippocampal Gyrus (PPA)	26	−38	−20	6	4
	L Precuneus	−12	−60	16	40	8
	R Precuneus	16	−54	14	36	8
	L Inf Temporal Gyrus	−56	−56	−8	12	4
manipulation	L Supramarginal Gyrus	−60	−30	34	51	10
	L Postcentral/Supramarginal Gyri	−38	−40	48	21	12
	L Precentral Gyrus	−54	4	10	18	6
	L Inf Temporal Gyrus	−46	−70	−4	34	8
eating	L Inf Frontal Gyrus	−54	10	18	26	8
	L Mid/Inf Frontal Gyri	−48	28	18	10	6
	L Inf Temporal Gyrus	−52	−62	−14	7	4
word length	L Occipital Pole	−18	−98	−6	24	6
	R Occipital Pole	16	−94	0	47	10
	L Lingual/Fusiform Gyri	−28	−68	−12	20	8
	R Lingual/Fusiform Gyri	30	−76	−14	14	6

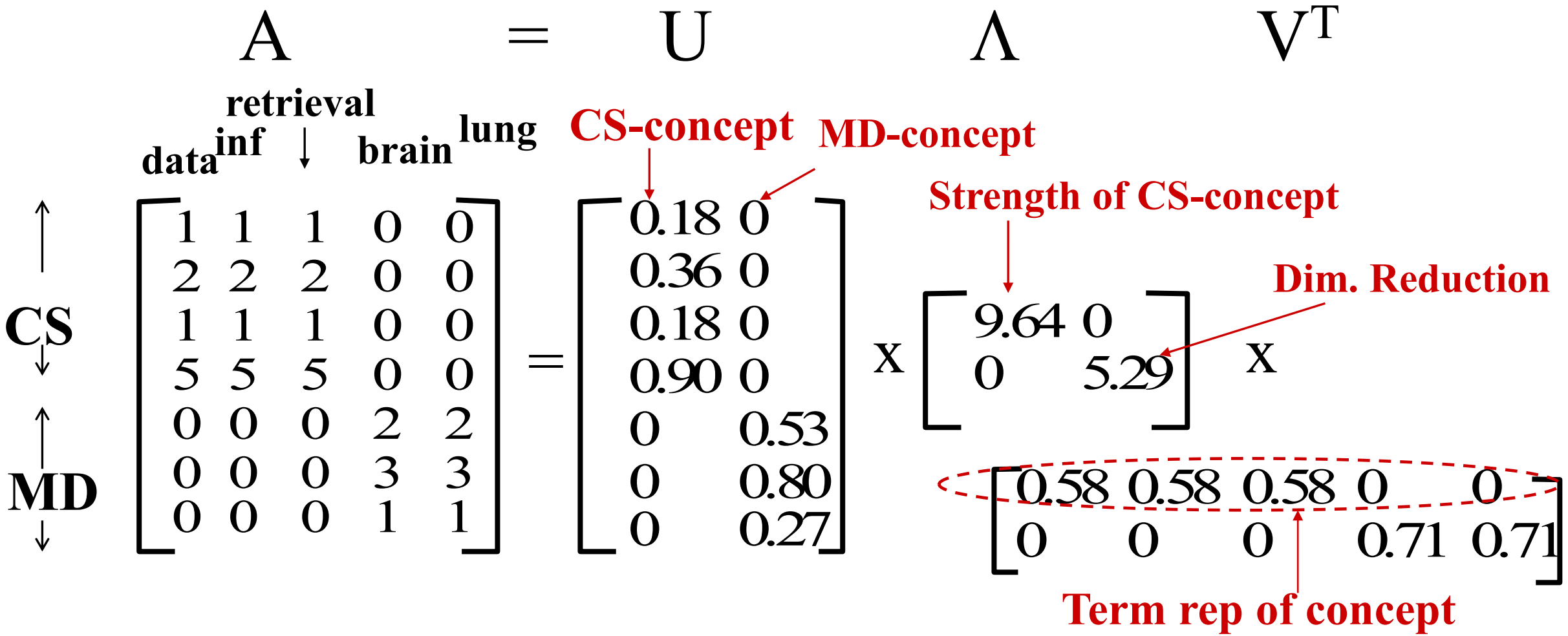
doi:10.1371/journal.pone.0008622.t003

[Just et al. 2010]

External pointers

- http://lsa.colorado.edu
- http://www.cs.utk.edu/~lsi

Example of LSI

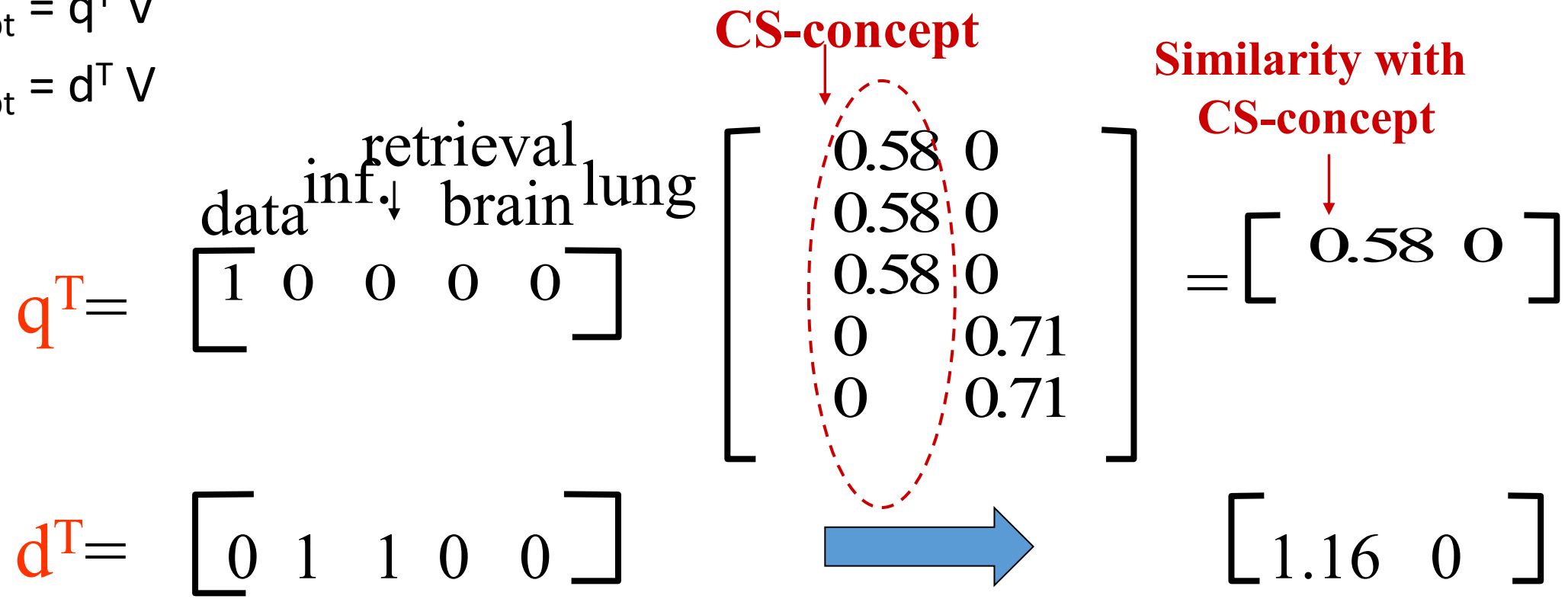


[Example modified from Christos Faloutsos]

Mapping Queries and Docs to the Same Space

$q^T_{concept} = q^T V$

$d^T_{concept} = d^T V$



[Example modified from Christos Faloutsos]