DISTRIBUTIONAL SEMANTICS AND COMPOSITIONALITY

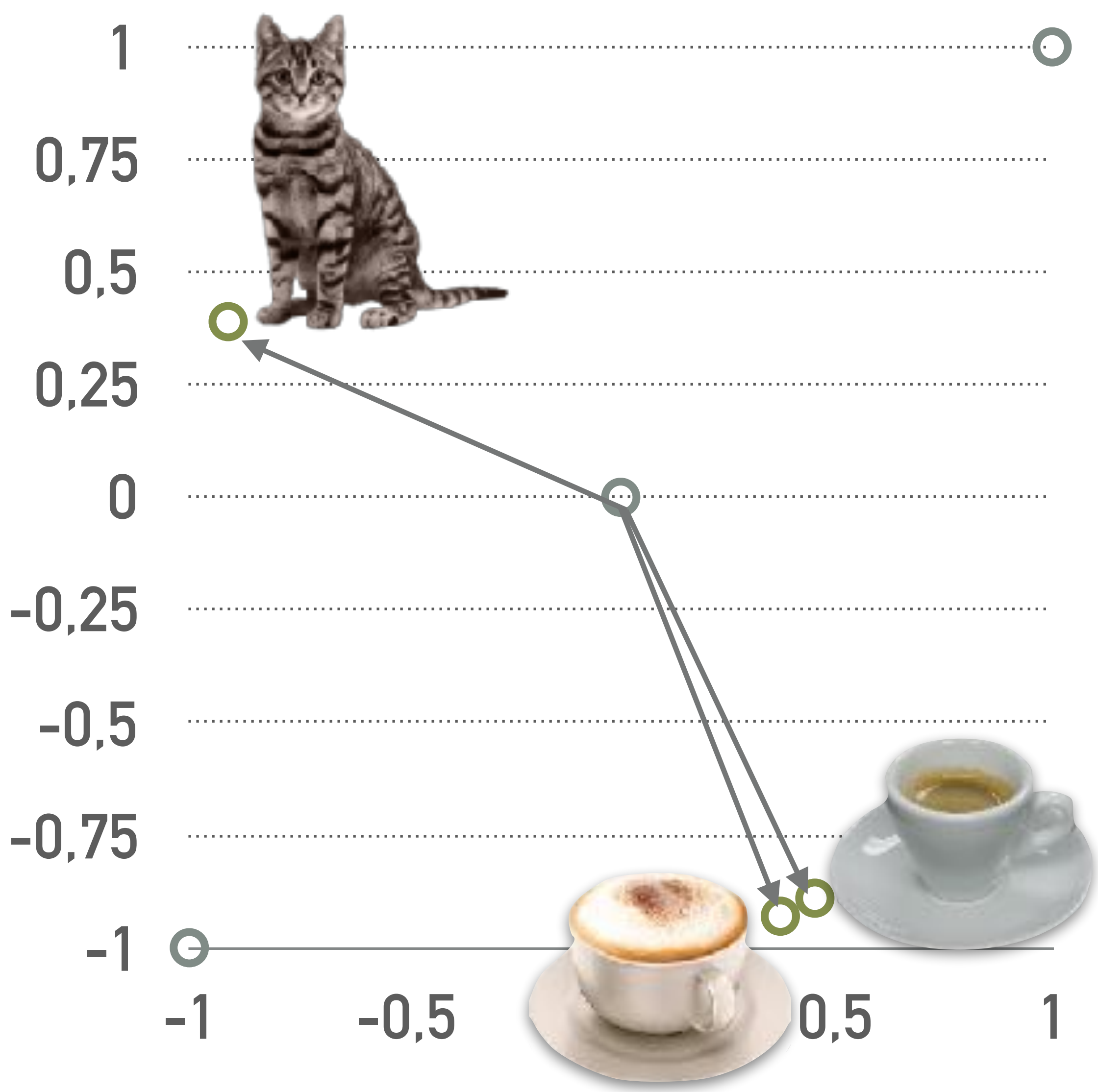
Corina Dima

April 23rd, 2019

COURSE LOGISTICS

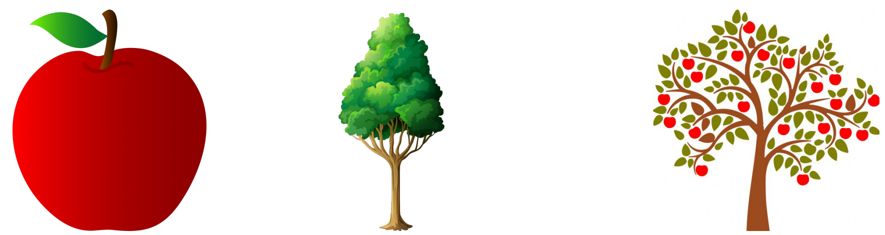
- Who?
  - Corina Dima
  - office: 1.05, Wilhelmstr. 19
  - email: corina.dima@uni-tuebingen.de
  - office hours: Tuesdays, 14-15 (please email me first)
- When?
  - Tuesdays, 8:30-10 (DS)
  - Thursdays, 8:30-10 (Comp)
- Where?
  - Room 1.13, Wilhelmstr. 19
- What?
  - Course webpage: https://dscomp2019.github.io/

DISTRIBUTIONAL SEMANTICS



- Word representations (word embeddings) based on distributional information are a key ingredient for state-of-the-art natural language processing applications.
- They represent similar words like ‘cappuccino’ and ‘espresso’ as similar vectors in vector space. Dissimilar vectors - like ‘cat’ - are far away.

COMPOSITIONALITY



Apfel + Baum → Apfelbaum

0.3	0.1	0.7	u	0.5	0.9	0.1	v	0.2	1.0	0.6	w

$$f\left(\begin{bmatrix}0.3 & 0.1 & 0.7\end{bmatrix} u,\ \begin{bmatrix}0.5 & 0.9 & 0.1\end{bmatrix} v\right) = \begin{bmatrix}?? & ?? & ??\end{bmatrix} p$$

What f makes p most similar to w?

wmask

$$p = g(W[u \odot u'; v \odot v''] + b)$$

where $p, u, u', v, v'', b \in \mathbb{R}^n$; $W \in \mathbb{R}^{n \times 2n}$; $g = tanh$

multimatrix

$$p = Wg(W_1[u; v] + b_1; W_2[u; v] + b_2; \ldots; W_k[u; v] + b_k) + b$$

where $p, u, v, b, b_i \in \mathbb{R}^n$; $W_i \in \mathbb{R}^{n \times 2n}$; $W \in \mathbb{R}^{n \times kn}$; $g = relu$

- Composition models for distributional semantics extend the vector spaces by learning how to create representations for complex words (e.g. ‘apple tree’) and phrases (e.g. ‘black car’) from the representations of individual words.
- The course will cover several approaches for creating and composing distributional word representations.

COURSE PREREQUISITES

- Prerequisites
  - linear algebra (matrix-vector multiplications, dot product, Hadamard product, vector norm, unit vectors, cosine similarity, cosine distance, matrix decomposition, orthogonal and diagonal matrices, tensor, scalar)
  - programming (Java III), computational linguistics (Statistical NLP) - ISCL-BA-08 or equivalent; programming in Python (+numpy, Tensorflow/PyTorch) for the project
  - machine learning (regression, classification, optimization objective, dropout, recurrent neural networks, autoencoders, convolutions)

GRADING

- For 6 CP
  - Active participation in class (30%)
  - Presenting a paper (70%)
- For 9 CP
  - Active participation in class (30%)
  - Doing a project (paper(s)-related) & writing a paper (70%)
- **Strict deadline** for the project: end of lecture time (27.07.2019)
- Both presentations and projects are **individual**

REGISTRATION

- **register using your GitHub account until 29.04.2019**
- Info
  - Last name(s)
  - First name(s)
  - Email address
  - Native language(s)
  - Other natural languages
  - Programming languages
  - Student ID (Matrikelnr.)
  - Degree program, semester (e.g. ISCL BA, 5th semester)
  - Chosen variant of the course: 6CP/9CP

EXAMPLE PROJECTS (1)

- Implement a PMI-based tool for the automatic discovery of English noun-noun compounds in a corpus. The tool should be able to discover both two-part as well as multi-part compounds.
  - References:
    - Church & Hanks (1990) - *Word Association Norms, Mutual Information and Lexicography*
    - Mikolov et al. (2013) - *Distributed Representations of Words and Phrases and their Compositionality*

EXAMPLE PROJECTS (2)

- Implement a recursive composition model that uses subword representations.
  - E.g. ’Apfelbaum’ ~ ‘Apfe’, ‘pfel’, ‘felb’, ‘elba’, ‘lbau’, ‘baum’
- recursively compose each two ngrams, each time replacing the two composed ngrams with the composed representation
- References:
  - Piotr Bojanowski, Edouard Grave, Armand Joulin, Tomas Mikolov. 2017. *Enriching Word Vectors with Subword Information*.
  - Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher Manning, Andrew Ng, Christopher Potts. 2013. *Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank*.

NEXT WEEK

- Tuesday, 30.04 (DS)
  - (word2vec paper) Tomas Mikolov, Kai Chen, Greg Corrado, Jefferey Dean. 2013. *Efficient Estimation of Word Representations in Vector Space* (Corina)
- Thursday, 2.05 (COMP)
  - Jeff Mitchell and Mirella Lapata. 2010. *Composition in Distributional Models of Semantics* (Corina)

IN TWO WEEKS

- Tuesday, 7.05 (DS)
  - Kenneth Church and Patrick Hanks. 1990. *Word Association Norms, Mutual Information and Lexicography* (?)
- Thursday, 9.05 (COMP)
  - Emiliano Guevara. 2010. *A Regression Model of Adjective-Noun Compositionality in Distributional Semantics* (?)
  - Marco Baroni and Roberto Zamparelli. 2010. *Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space* (?)

HOW TO WRITE A RESEARCH PAPER

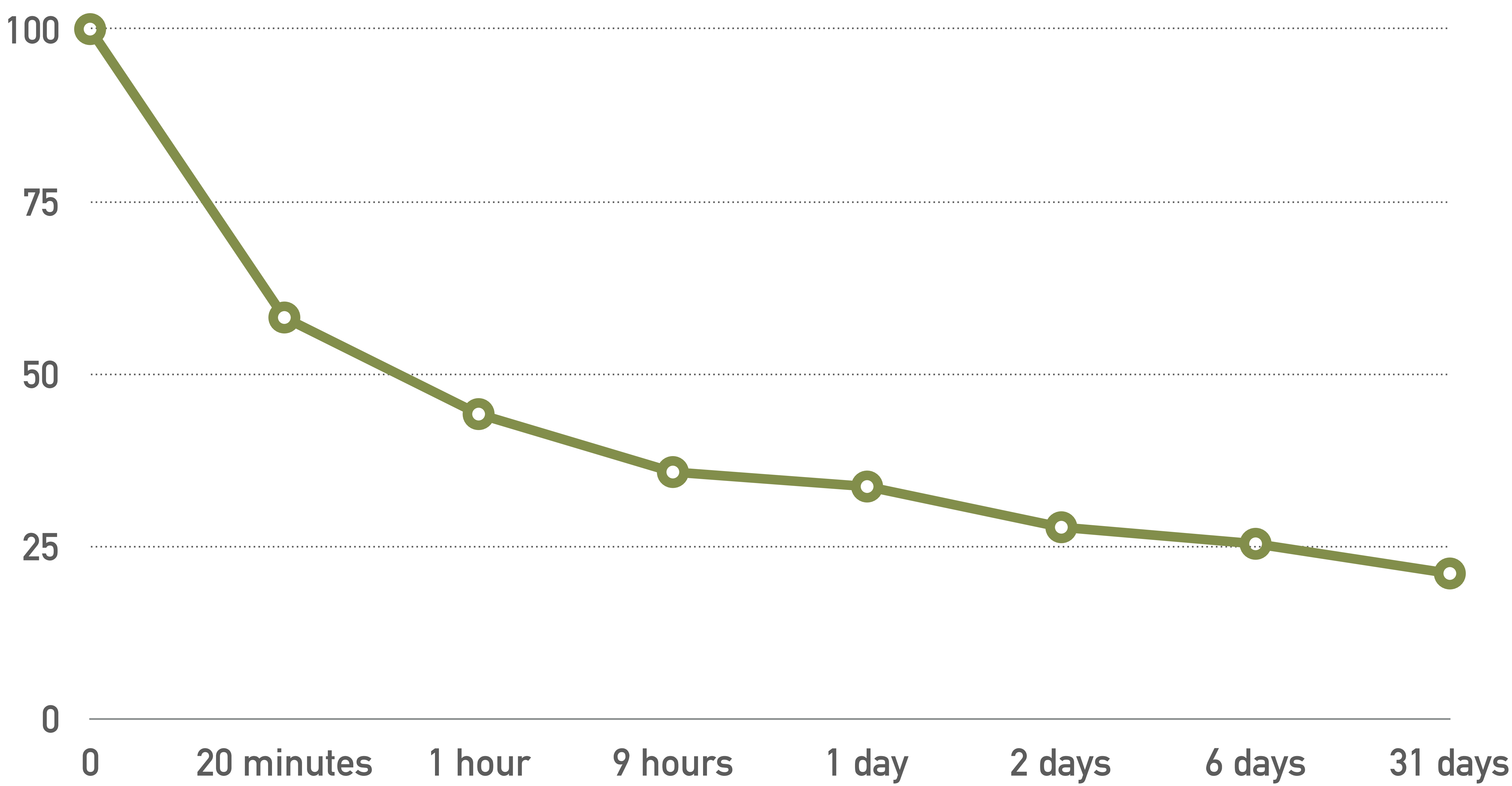
- Jason Eisner's blog post *Write the paper first* (https://www.cs.jhu.edu/~jason/advice/write-the-paper-first.html)
    - "Writing is the best use of limited time"
    - "If you run out of time, it is better to have a great story with incomplete experiments than a sloppy draft with complete experiments"
    - "Writing is a form of thinking and planning. Writing is therefore part of the research process—just as it is part of the software engineering process. When you write a research paper, or when you document code, you are not just explaining the work to other people: you are thinking it through for yourself."

HOW TO READ A RESEARCH PAPER

- Jason Eisner's blog post *How to Read a Technical Paper* (https://www.cs.jhu.edu/~jason/advice/how-to-read-a-paper.html)
  - multi-pass reading (skim first, more thorough second pass)
  - write as you read (low-level notes, high-level notes)
  - start early!

- Michael Nielsen's blog post *Augmenting Long-Term Memory* (http://augmentingcognition.com/ltm.html)
  - Using Anki to thoroughly read research papers (++remember)

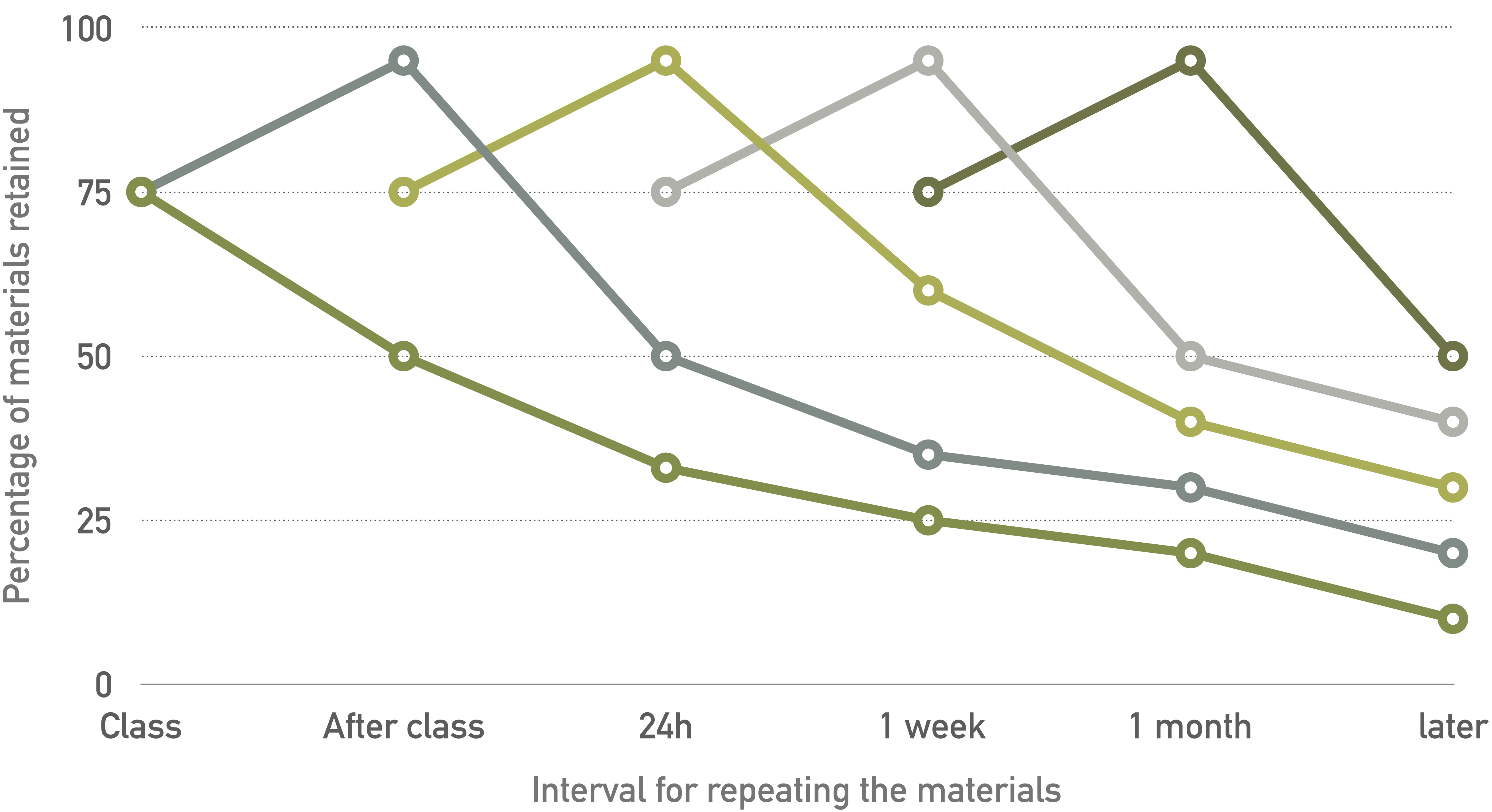
EBBINGHAUS'S FORGETTING CURVE



LEARNING HOW TO LEARN

- Barbara Oakely & Terrence Sejnowski's *Learning how to learn* course on Coursera (https://www.coursera.org/learn/learning-how-to-learn)
- Main points:
  - learning doesn't happen overnight - you need several passes through some material to really understand it
  - re-reading/highlighting materials can give you the illusion of learning - avoid it by practicing active recall (testing yourself)
  - spaced repetition can help you learn & remember forever-ish

THE EFFECTS OF SPACED REPETITION ON THE FORGETTING CURVE



HELPFUL POINTERS

- Khan Academy's Linear Algebra course (https://www.khanacademy.org/math/linear-algebra)
- Dan Jurafsky and James H. Martin. *Speech and Language Processing*. 3rd edition draft (https://web.stanford.edu/~jurafsky/slp3/), esp. Ch. 6, Vector Semantics

INTRO TO DISTRIBUTIONAL SEMANTICS

- What does a word mean?

cappuccino | ˌkapʊˈtʃiːnəʊ |

noun (plural **cappuccinos**)

a type of coffee made with espresso and milk that has been frothed up with pressurized steam.

ORIGIN

from Italian, literally **'Capuchin'**, because its colour resembles that of a Capuchin's habit.

INTRO TO DISTRIBUTIONAL SEMANTICS

- How can the meaning of a word be represented on a computer?
- One-hot vectors
  - each word is represented by a 1 in a particular dimension of the vector, with the other elements of the vector being 0
  - local representation: no interaction between the different dimensions

[1, 0, 0]

[0, 1, 0]

[0, 0, 1]

INTRO TO DISTRIBUTIONAL SEMANTICS

- Local representations, problem 1: word similarity does not correspond to vector similarity
  - ‘cappuccino’ and ‘espresso’ are just as similar/dissimilar as ‘cappuccino’ and ‘cat’
  - one-hot vectors are orthogonal to each other

INTRO TO DISTRIBUTIONAL SEMANTICS

- measure cosine similarity in vector space

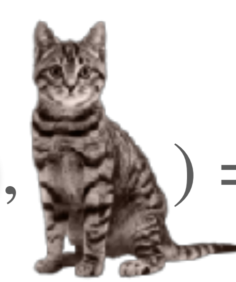
$$\cos(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\|_2 \|\mathbf{v}\|_2} = \frac{\sum_{i=1}^{n} \mathbf{u}_i \mathbf{v}_i}{\sqrt{\sum_{i=1}^{n} \mathbf{u}_i^2}\sqrt{\sum_{i=1}^{n} \mathbf{v}_i^2}}$$

 *[1, 0, 0]*

 *[0, 1, 0]*

 *[0, 0, 1]*

cos( , ) $= \frac{1 \cdot 0 + 0 \cdot 1 + 0 \cdot 0}{\sqrt{1^2 + 0^2 + 0^2}\sqrt{0^2 + 1^2 + 0^2}} = \frac{0}{1} = 0$

cos( , ) $= \frac{1 \cdot 0 + 0 \cdot 0 + 0 \cdot 1}{\sqrt{1^2 + 0^2 + 0^2}\sqrt{0^2 + 0^2 + 1^2}} = \frac{0}{1} = 0$

cosine of 0 means angle of 90° between the vectors

➛ *orthogonal vectors*

INTRO TO DISTRIBUTIONAL SEMANTICS

- Local representations, problem 2: representing new words

[1, 0, 0]

[0, 1, 0]

[0, 0, 1]

[?, ?, ?]

[1, 0, 0, 0]

[0, 1, 0, 0]

[0, 0, 1, 0]

[0, 0, 0, 1]

- representing a new word involves expanding the vector, since the existing components are already “used up”

INTRO TO DISTRIBUTIONAL SEMANTICS

- Solution: distributed representations (Hinton, McClelland and Rumelhart, 1986)
- meaning is distributed over the different dimensions of the vector
- each word is represented by a configuration over the components of the vector representations
- each component contributes to the representation of every word in the vocabulary

[0.37, -0.93]

[0.45, -0.89]

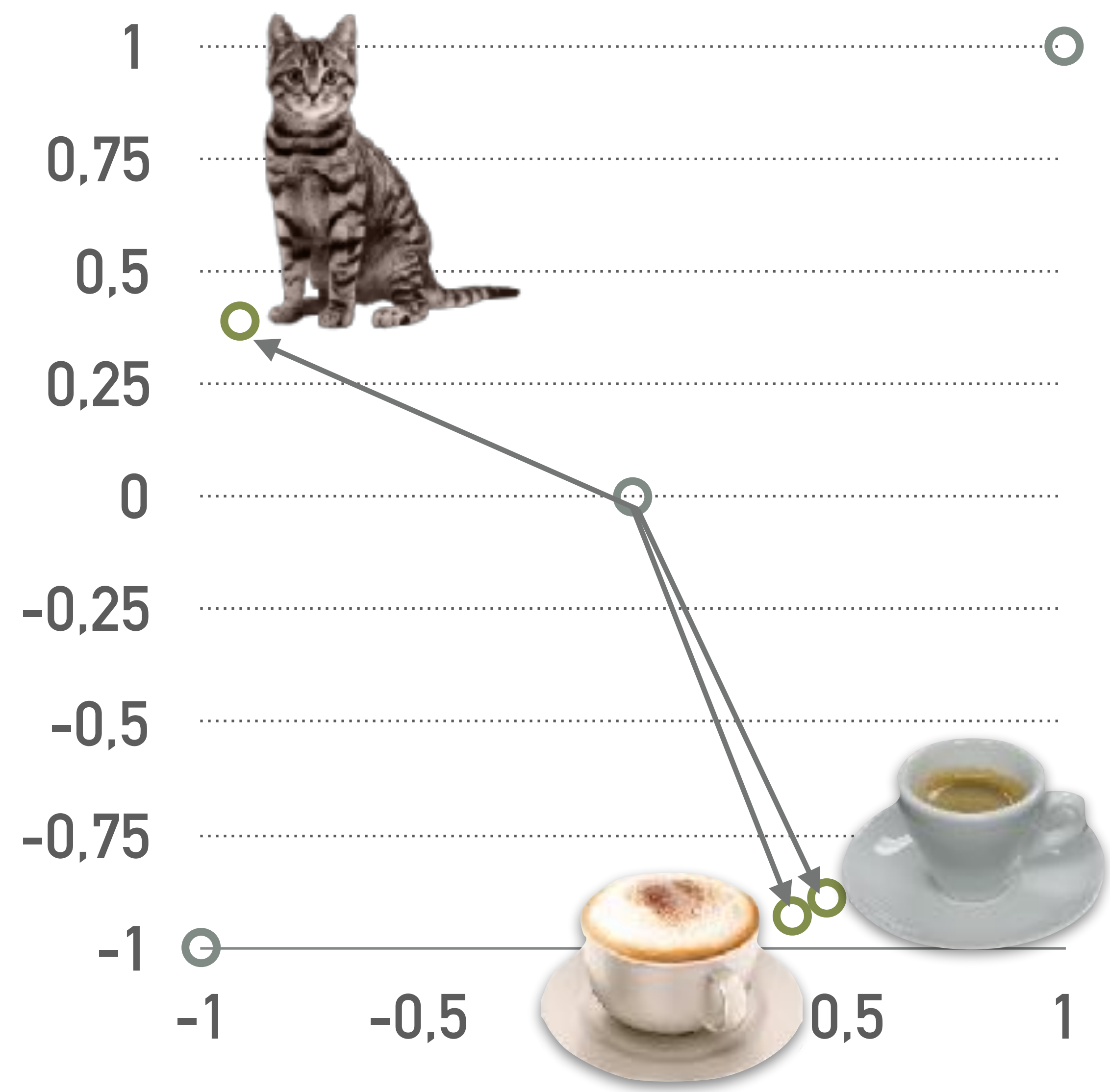
[-0.92, 0.39]

INTRO TO DISTRIBUTIONAL SEMANTICS

[0.37, -0.93]

[0.45, -0.89]

[-0.92, 0.39]

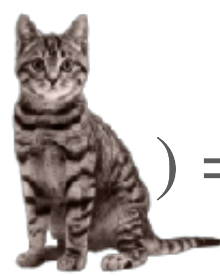


INTRO TO DISTRIBUTIONAL SEMANTICS

- Distributed representations solve problem 1: similar words can have similar vectors

cos( , )

$$= \frac{0.37 \cdot 0.45 + (-0.93) \cdot (-0.89)}{\sqrt{0.37^2 + (-0.93)^2}\sqrt{0.45^2 + (-0.89)^2}} \approx 0.9965$$

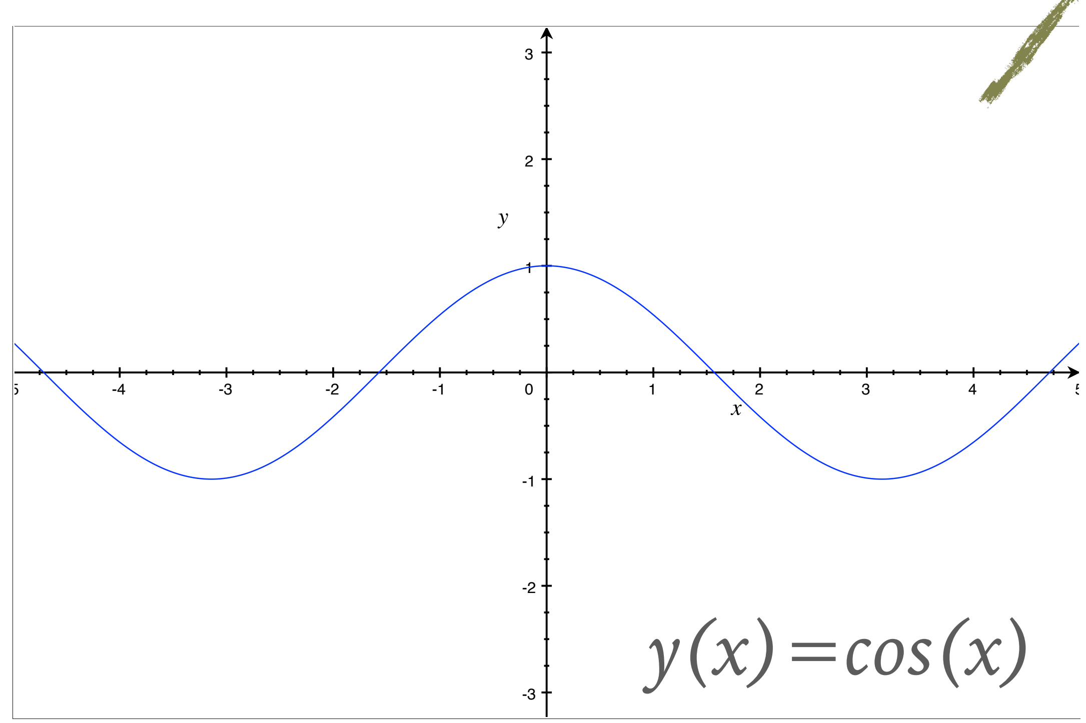
cos( , )

$$= \frac{0.37 \cdot (-0.92) + (-0.93) \cdot 0.39}{\sqrt{0.37^2 + (-0.93)^2}\sqrt{(-0.92)^2 + 0.39^2}} \approx -0.7071$$

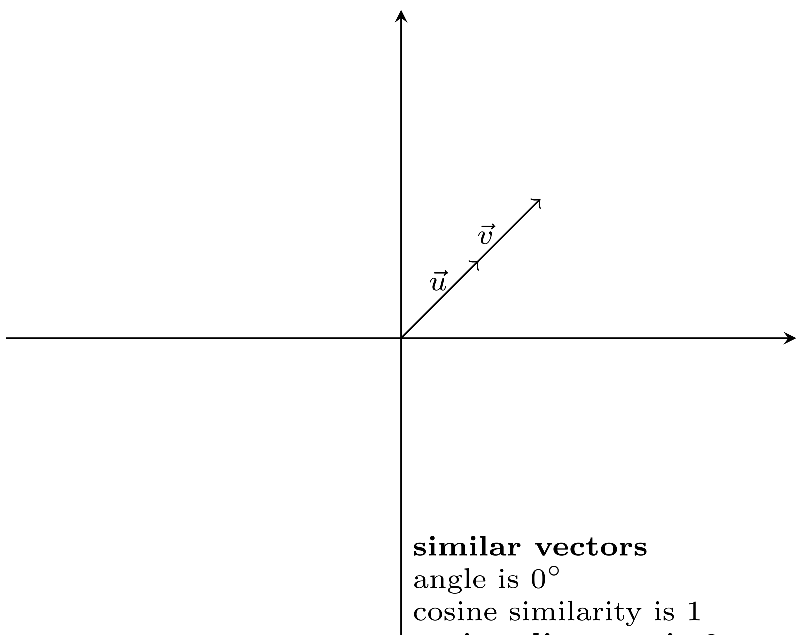
[0.37, -0.93]

[0.45, -0.89]

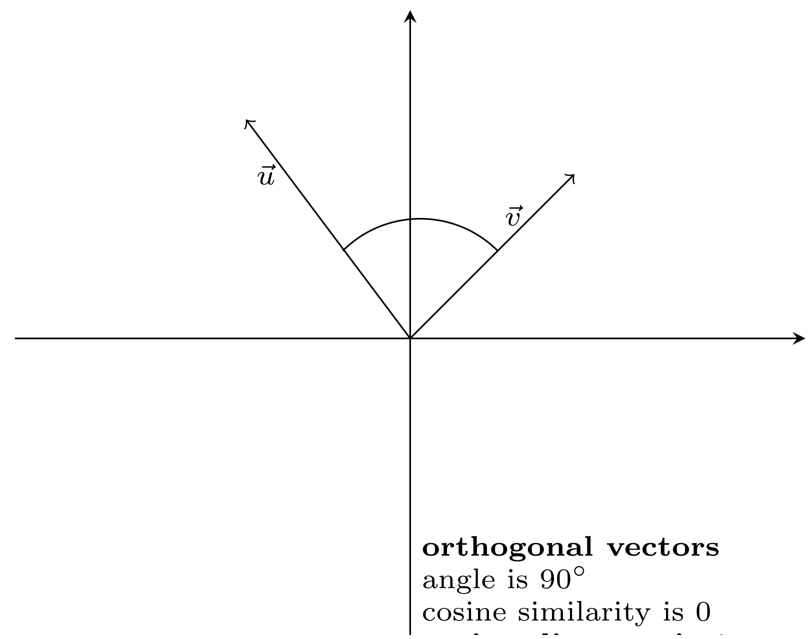
[-0.92, 0.39]



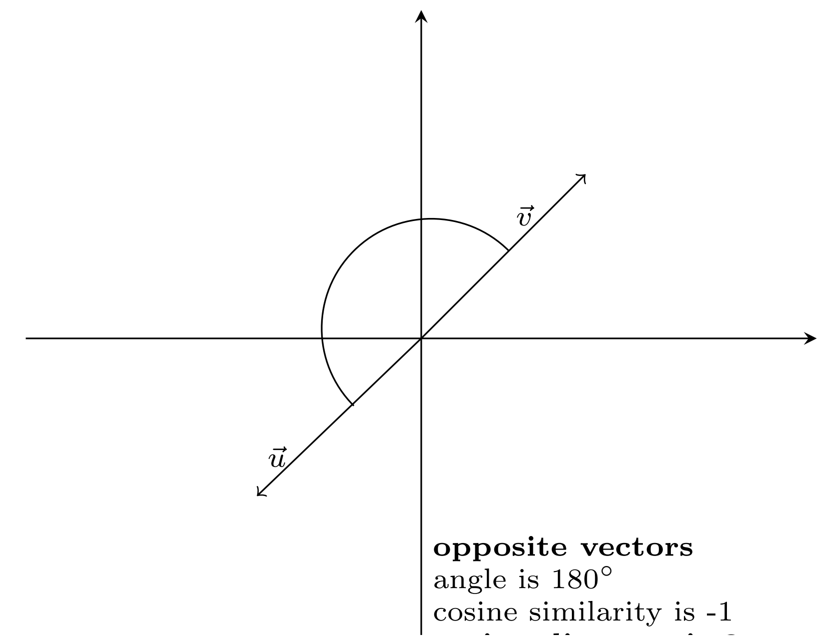
INTRO TO DISTRIBUTIONAL SEMANTICS



similar vectors
angle is 0°
cosine similarity is 1



orthogonal vectors
angle is 90°
cosine similarity is 0



opposite vectors
angle is 180°
cosine similarity is -1

INTRO TO DISTRIBUTIONAL SEMANTICS

- Distributed representations solve problem 2: new words can be added to the vector space without changing the dimensions of the vectors

[0.37, -0.93]

[0.45, -0.89]

[-0.92, 0.39]

[0.32, -0.95]

INTRO TO DISTRIBUTIONAL SEMANTICS

- What information can be used to create the (local/distributed) word representations?
- Distributional semantics
  - Harris (1954): “Meaning as a function of distribution”
  - Firth (1957): “You shall know a word by the company it keeps!”

“

If we consider *oculist* and *eye-doctor*[17] we find that, as our corpus of actually-occurring utterances grows, these two occur in almost the same environments, except for such sentences as *An oculist is just an eye-doctor under a fancier name*, or *I told him Burns was an oculist, but since he didn't know the professional titles, he didn't realize that he could go to him to have his eyes examined*. If we ask informants for any words that may occupy the same place as *oculist* in sentences like the above (i.e. have these same environments), we will not in general obtain *eye-doctor;* but in almost any other sentence we would. In contrast, there are many sentence environments in which *oculist* occurs but *lawyer* does not: e.g. *I've had my eyes examined by the same oculist for twenty years*, or *Oculists often have their prescription blanks printed for them by opticians*.

- Zelling S. Harris (1954)

“

The *placing* of a *text* as a constituent in a context of situation contributes to the statement of meaning since situations are set up to recognize *use.* As Wittgenstein says, ‘the meaning of words lies in their use.’ [4] The day to day practice of playing language games recognizes customs and rules. It follows that a text in such established usage may contain sentences such as ‘Don’t be such an ass!’, ‘You silly ass!’, ‘What an ass he is!’ In these examples, the word *ass* is in familiar and habitual company, commonly collocated with *you silly—, he is a silly—, don’t be such an—.* You shall know a word by the company it keeps! One of the meanings of *ass* is its habitual collocation with such other words as those above quoted.[5] Though Wittgenstein was dealing with another problem, he also recognizes the plain face-value, the physiognomy of words. They look at us! [6] ‘The sentence is composed of the words and that is enough.’

-J.R. Firth (1957)

“

We found a cute, hairy wampimuk sleeping behind the tree.

Lazaridou et. al, 2014

“

We found a cute, hairy wampimuk sleeping behind the tree.

Lazaridou et. al, 2014

INTRO TO DISTRIBUTIONAL SEMANTICS

36	MAG	the waitress in a neat black and white uniform . My	cappuccino	came with the correct amount of froth , sprinkled with chocolate
37	NEWS	ice cream in whimsical flavors like white pistachio and	cappuccino	chocolate crunch . The pair soon owned a string of stores ,
38	FIC	's an expensive wedding cake of a store , adorned with	cappuccino	colored carpeting and a headless mannequin wearing a two
39	MAG	from halfway around the world . Everyone agrees that the best	cappuccino	comes from the caf run by the French soldiers , but it
40	ACAD	added to create a chocolaty beverage . # Like today 's	cappuccino	connoisseurs , elite Maya and Aztec cherished foam atop their
41	FIC	. # I did n't ask the natural next question .	Cappuccino	cream mustache on her upper lip , she volunteered . # "
42	NEWS	you 're rewarded with a lush , velvety custard . The	cappuccino	creme brulee ($ 5) served in a coffee cup had
43	NEWS	for a custard mousse affair) . Served in an oversized	cappuccino	cup , the rich and velvety custard was topped with a twirl
44	SPOK	OK. Ms-RAY : She serves it in little espresso cups or	cappuccino	cups . So here 's all you do . Here , you
45	MAG	be gratis . COFFEE . With no more office pot ,	cappuccino	doses run $4 a day -- that 's up to $1,460 a
46	MAG	Scottie 's after school as the waiter slammed two mugs of	cappuccino	down in front of Meghan and me . She had been after
47	FIC	blackmail me into doing a job . # Tommy brakes and	cappuccino	flies . Hawk half-heartedly tries to lick up with his fingers .
48	NEWS	On this early morn , it 's all for one and	cappuccino	for all . Everyone except Malkovich lights up strong European
49	MAG	bag . To avoid feeling deprived , sip a frothy skim-milk	cappuccino	for dessert . The second approach involves selecting a
50	MAG	satisfying drinks-like root-beer floats for kids or an iced	cappuccino	for grown-ups-and you 're dishing the kind of bliss that your
51	FIC	have to . " # Lyric and I get iced decaf	cappuccinos	from the store next to the condo . We get an Orangina
52	SPOK	subtle movement , his body temperature began to rise . And	Cappuccino	gambled again , using a cooling machine to lower Everett 's body
53	NEWS	Institute) (pg . B2) 1427 # From a	cappuccino	grande at Starbucks to a plain cup of black no-sugar at a
54	SPOK	difference in people 's lives . BLAKE : Right . A	cappuccino	here in New York , three or four bucks . Save the
55	MAG	made my next change : I gave up the sugary convenience-store	cappuccinos	I 'd been drinking several times a day for lower-calorie vanilla
56	NEWS	away from sides of pan , 25-30 minutes . 71915 #	Cappuccino	Icebox Cookies | # You can keep a roll of this dough in
57	MAG	cozy fireplaces and live Andean music . We stop for a	cappuccino	in a corner bar that advertises 15 types of coffee . But
58	NEWS	Norte , his city 's main newspaper , while sipping a	cappuccino	in anticipation of a shopping spree . " Settling down in McAllen
59	SPOK	guess , makers -- oh , look , oh , just	cappuccino	in general with their faces in the foam . GIFFORD : Oh
60	MAG	their A-list friends in trendy clubs , they prefer sipping soy	cappuccinos	in local cafes . Moder is an outdoorsy guy who enjoys running
61	SPOK	theaters adapt and they now have multi-screen and now they have	cappuccino	in some movie theaters . So they adapt , and I think

https://www.wordandphrase.info/, made by Mark Davies, BYU
Corpus of Contemporary American English (COCA)

INTRO TO DISTRIBUTIONAL SEMANTICS

CAPPUCCINO *n* (RANK 17250, FREQ 595)

	SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC
CLICK BAR TO LIMIT					
STORED	21	56	61	57	7
MORE	**56**	**200**	**180**	**144**	**15**

adj iced, double, orthopedic, frothy, hot, instant, steaming, tall, excellent, fat-free **noun** cup, machine, espresso, latte, bar, coffee, sip, cafe, maker, decaf **verb** sip, drink, serve, order, buy, sell, finish, enjoy

INTRO TO DISTRIBUTIONAL SEMANTICS

co-occurence matrix

context words

target words

	iced	(to) drink	owner	in
cappuccino	6	2	0	3
espresso	1	1	0	4
cat	0	1	4	3
latte	6	5	0	4
leaf	0	0	0	5

co-occurence

INTRO TO DISTRIBUTIONAL SEMANTICS

- *the pointwise mutual information (PMI) between a target word **t** and a context word **c** is defined as*

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

INTRO TO DISTRIBUTIONAL SEMANTICS

- *the pointwise mutual information (PMI) between a target word **t** and a context word **c** is defined as*

*how often are **t** and **c** are observed **together***

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

INTRO TO DISTRIBUTIONAL SEMANTICS

- *the pointwise mutual information (PMI) between a target word **t** and a context word **c** is defined as*

*how often are **t** and **c** are observed **together***

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

*how often would we **expect** **t** and **c** to co-occur (assuming each occurs independently)*

INTRO TO DISTRIBUTIONAL SEMANTICS

- *the pointwise mutual information (PMI) between a target word **t** and a context word **c** is defined as*

*how often are **t** and **c** are observed **together***

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

the ratio is an estimate of how much more the two words co-occur than is expected by chance

*how often would we **expect** **t** and **c** to co-occur (assuming each occurs independently)*

INTRO TO DISTRIBUTIONAL SEMANTICS

- the PMI for '*Humpty Dumpty*' is 22.5
- the pair (*Humpty, Dumpty*) occurs 6,000,000 ($\sim 2^{22.5}$) times more than one would expect from the frequencies of *Humpty* and *Dumpty* - from Brown et al. (1992)
- order matters!
- PMI(*Humpty, Dumpty*) $\neq$ PMI(*Dumpty, Humpty*)
- positive point wise mutual information (PPMI) is used

INTRO TO DISTRIBUTIONAL SEMANTICS

	iced	(to) drink	owner	p(t)
cappuccino	6	2	0	8
espresso	1	1	0	2
cat	0	1	4	5
p(c)	7	4	4	15

$$P(t = cappuccino, c = iced) = \frac{6}{15} = 0.4$$

$$P(t = cappuccino) = \frac{8}{15} = 0.53 \qquad P(c = iced) = \frac{7}{15} = 0.47$$

$$PMI(t = cappuccino, c = iced) = \log_2 \frac{0.4}{0.53 * 0.47} = \log_2 1.6 = 0.68$$

INTRO TO DISTRIBUTIONAL SEMANTICS

- vocabularies contain typically 10,000-1,000,000 words
- sparse vectors (most components are 0) - most words will co-occur with a small subset of other words in the vocabulary
- use dimensionality reduction techniques to transform high-dimensional, sparse representations to low-dimensional, dense representations

INTRO TO DISTRIBUTIONAL SEMANTICS

- singular value decomposition (SVD)

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\top}$$

  - where $\mathbf{A} \in \mathbb{R}^{mxn}$
  - $\mathbf{U} \in \mathbb{R}^{mxn}$ is a matrix with orthogonal columns
  - $\mathbf{\Sigma} \in \mathbb{R}^{nxn}$ is a diagonal matrix of singular values; the singular values are, by convention, ordered from the largest to the smallest
  - $\mathbf{V}^{\top} \in \mathbb{R}^{nxn}$ is an orthogonal matrix ($\mathbf{V}^{-1} = \mathbf{V}^{\top}$)
- by taking only the top *k* singular values, $k \ll n$, SVD obtains an approximation of A, A_k, such that the distance between the matrices (the 2-norm, $\|A - A_k\|_2$) is minimized

INTRO TO DISTRIBUTIONAL SEMANTICS

- where does the dimensionality reduction come from?
- singular value decomposition separates any matrix into simple pieces
- m=30,000; n = 10,000; k = 300
- size of initial A: 30,000 x 10,000 = 300,000,000 numbers

$$\mathbf{A_k} = \mathbf{U_k \Sigma_k V_k}^{\top}$$

INTRO TO DISTRIBUTIONAL SEMANTICS

$$\underset{A}{\begin{bmatrix} x & x & x \\ x & x & x \\ x & x & x \\ x & x & x \\ x & x & x \end{bmatrix}} = \underset{U}{\begin{bmatrix} x & x & x \\ x & x & x \\ x & x & x \\ x & x & x \\ x & x & x \end{bmatrix}} \underset{\Sigma}{\begin{bmatrix} x & o & o \\ o & x & o \\ o & o & x \end{bmatrix}} \underset{V^T}{\begin{bmatrix} x & x & x \\ x & x & x \\ x & x & x \end{bmatrix}}$$

$n = 5, \; m = 3$

INTRO TO DISTRIBUTIONAL SEMANTICS

$$\overset{A}{\begin{bmatrix} x & x & x \\ x & x & x \\ x & x & x \\ x & x & x \\ x & x & x \end{bmatrix}} = \overset{u_1 \quad U}{\begin{bmatrix} \boxed{\begin{matrix} x \\ x \\ x \\ x \\ x \end{matrix}} & \begin{matrix} x \\ x \\ x \\ x \\ x \end{matrix} & \begin{matrix} x \\ x \\ x \\ x \\ x \end{matrix} \end{bmatrix}} \overset{\sigma_1 \quad \Sigma}{\begin{bmatrix} x & 0 & 0 \\ 0 & x & 0 \\ 0 & 0 & x \end{bmatrix}} \overset{V^T}{\begin{bmatrix} \boxed{x \quad x \quad x} \\ x \quad x \quad x \\ x \quad x \quad x \end{bmatrix}} \; v_1^T$$

$m = 5, \; n = 3$

$$A = \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T + \sigma_3 u_3 v_3^T$$

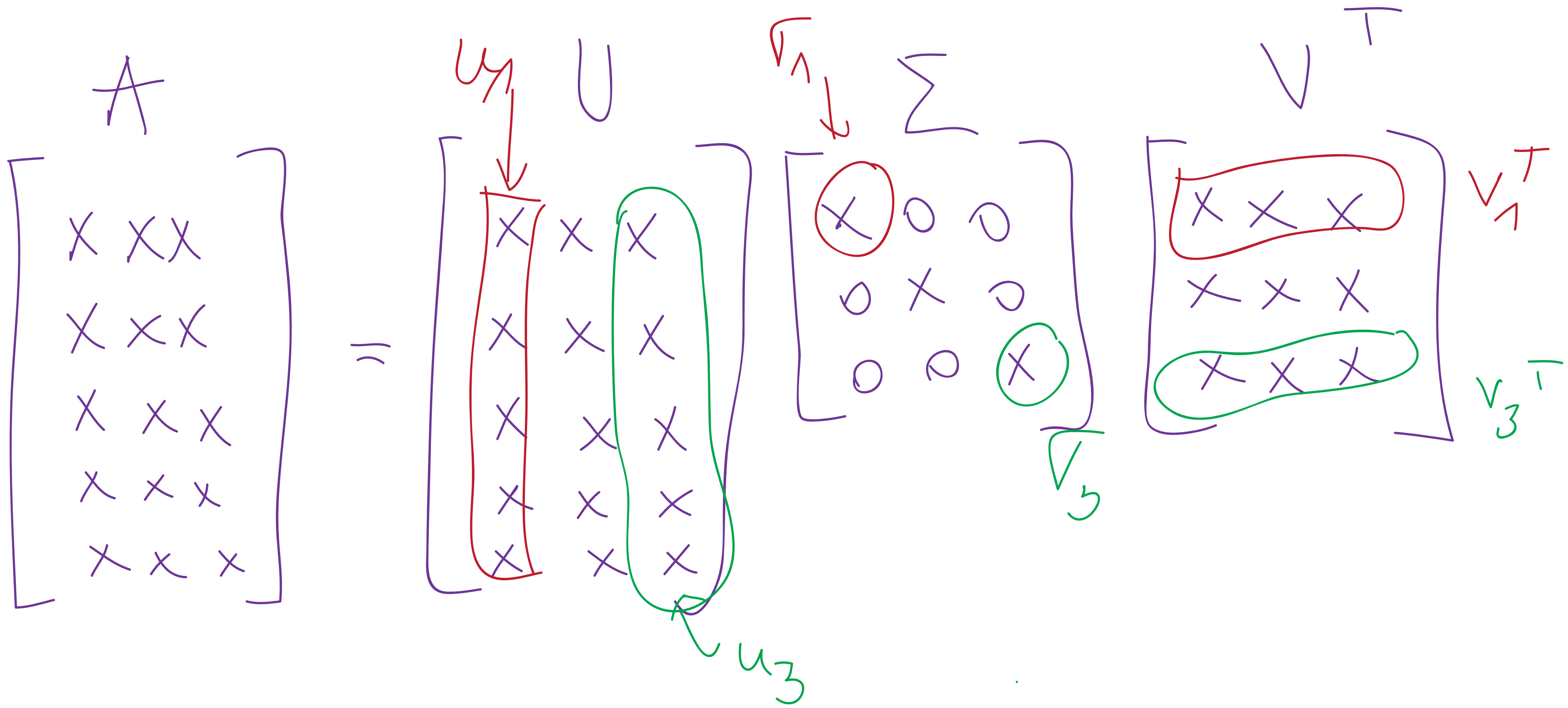
INTRO TO DISTRIBUTIONAL SEMANTICS

$$\underset{6 \times 1}{\overset{w}{\begin{bmatrix}1\\1\\1\\1\\1\\1\end{bmatrix}}} \underset{1 \times 6}{\overset{v^T}{\begin{bmatrix} a & a & b & b & c & c \end{bmatrix}}} = \underset{6 \times 6}{\begin{bmatrix} a & a & b & b & c & c \\ a & a & b & b & c & c \\ a & a & b & b & c & c \\ a & a & b & b & c & c \\ a & a & b & b & c & c \\ a & a & b & b & c & c \end{bmatrix}}$$

rows

columns

INTRO TO DISTRIBUTIONAL SEMANTICS



$m = 5, \; n = 3$

$$A = \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T + \sigma_3 u_3 v_3^T$$

$$A \approx \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T$$

INTRO TO DISTRIBUTIONAL SEMANTICS

- m=30,000; n = 10,000; k = 300
- size of initial A: 30,000 x 10,000 = 300,000,000 numbers

$$\mathbf{A_k} = \mathbf{U_k \Sigma_k V_k}^\top$$

- size of A_k: 30,000 x 300 ($\mathbf{U}$) + 300 ($\mathbf{\Sigma}$) + 300 x 10,000 ($\mathbf{V}^\top$) = 9,000,000 + 300 + 3,000,000 = 12,000,300 numbers
- words can now be represented as reduced-dimensionality vectors

$$\mathbf{W}^{SVD_p} = \mathbf{U}_k \cdot \Sigma_k^p$$

$$\mathbf{W} \in \mathbb{R}^{m \times k}$$

in our example $\mathbf{W} \in \mathbb{R}^{30{,}000 \times 300}$

$$p = 0, \mathbf{W}^{SVD} = \mathbf{U}_k$$

$$p = \frac{1}{2}, \mathbf{W}^{SVD} = \mathbf{U}_k \cdot \sqrt{\Sigma_k}$$

$$p = 1, \mathbf{W}^{SVD} = \mathbf{U}_k \cdot \Sigma_k$$

INTRO TO DISTRIBUTIONAL SEMANTICS

- after dimensionality reduction, a particular vector component no longer has an associated "meaning"
- the information is "spread" over the dimensions
- more difficult to interpret individual vector components

REFERENCES

- J.R. Firth. 1957. *A synopsis of linguistic theory 1930-55*. In Studies in Linguistic Analysis (special volume of the Philological Society), 1-32. Oxford.
- Zelling S. Harris. 1954. *Distributional Structure*. Word, 10:2-3, 146-162, DOI: 10.1080/00437956.1954.11659520
- G.E. Hinton, J.L. McClelland, D.E. Rumelhart. 1986. *Distributed Representations*. In Parallel Distributed Processing, Volume 1: Foundations. Editors: David E. Rumelhart, James L. McClelland and the PDP Research Group.
- Peter Brown, Peter deSouza, Robert Mercer, Vincent Della Pietra, Jenifer Lai. 1992. *Class-based n-gram Models of Natural Language*.
- A. Lazaridou, E. Bruni, M. Baroni. 2014. *Is this a wampimuk? Cross-modal mapping between distributional semantics and the visual world*. ACL 2014.

Image credits

Creative Commons 4.0 BY-NC:
http://pngimg.com/download/49645
http://pngimg.com/download/50514
http://pngimg.com/download/27425
https://commons.wikimedia.org/wiki/File:Espresso_BW_1.jpg

CC BY-SA 2.5
https://commons.wikimedia.org/wiki/Category:Latte_macchiato?uselang=de#/media/File:Latte_macchiato_with_coffee_beans.jpg

Public domain
https://de.wikipedia.org/wiki/Datei:Humpty_Dumpty_1_-_WW_Denslow_-_Project_Gutenberg_etext_18546.jpg