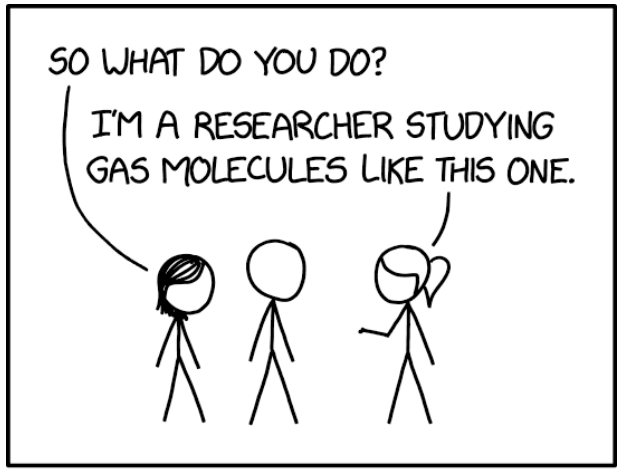
COMP 4446 / 5046
Lecture 1: Introduction & Representing Text

Jonathan K. Kummerfeld

Semester 1, 2025

[A lot of sentences undergo startling shifts in mood if you add "like this one" to the end, but high on the list is "I'm a neurologist studying dreams."]

Like This One



Source: https://xkcd.com/2879/

We recognise and pay respect to the Elders and communities – past, present, and emerging – of the lands that the University of Sydney's campuses stand on. For thousands of years they have shared and exchanged knowledges across innumerable generations for the benefit of all.

THE UNIVERSITY OF SYDNEY

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

Gadi - Grasstree
Gal - People

Gadigal - People of the Grasstree

https://www.botanicgardens.org.au/royal-botanic-garden-sydney/gadigal-country

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

What is this course about?

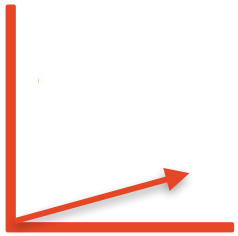
Natural Language Processing

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

What is this course about?



```
show_directions(from: current,
to: get_location(next class))
```

How can I get to the lecture theatre for my next class,

Natural Language Processing?

COURSE

¿Cómo puedo llegar a la sala de conferencias para mi próxima clase, Procesamiento del lenguaje natural?

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

What will I take away from this course?

The ability to:

build

evaluate

understand

integrate

NLP technology

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

What concepts are covered?

Data

Model based on counting

.
.
.

Large Language Models

Learning Methods

Under Construction

menti.com
4210 8267

Updated Lectures

Updated Workshops

New Projects

New Quizzes

menti.com
4210 8267

Motivation for all changes:
NLP is changing!
Student feedback!
Ideas for improvement!

My goal is for assignments, quizzes, and an exam that
everyone can pass
but also
hard work is reflected in your mark

Updated Lectures

menti.com
4210 8267

More content

Updated content

Bring in latest research where I can

Updated Workshops

menti.com
4210 8267

More interactive

No marks in-class

Pre-work required

New Projects

menti.com
4210 8267

For learning more than evaluation

A new, bigger project, possibly with an optional competition

New Quizzes

menti.com
4210 8267

Encourage revision through semester

Give a sense of certain exam question types

menti.com
4210 8267

Please do not post materials from this course online

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

How do the parts of the unit fit together?

	Learning	**Evaluating**
Concepts	Lectures Muddy Cards Assignments	Quizzes Exam
Skills	Workshops	

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

Who is Dr. Jonathan K. Kummerfeld?

Education and Career
- USyd, B Sc. Adv. (Hons) (Medal)
- Berkeley, PhD
- Michigan, Postdoc
- Harvard, Visiting Scholar
- USyd, Senior Lecturer

Teaching
- Sydney, Teaching Commendation for COMP 4446 / 5046 in 2023 and 2024
- Berkeley, Outstanding Graduate Student Instructor Award
- Guest talks all over the world - Cambridge, Oxford, Yale, etc

Research and Industry Experience
- Over 3,000 citations of my work
- Technical advisor to four startups
- Monash Scholar, ARC DECRA Fellow

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

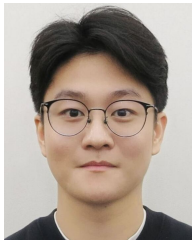
menti.com 4210 8267

Who are the other course staff?

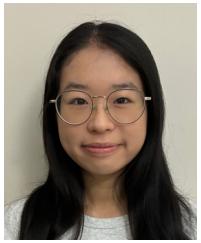
Chen Chen

James Douglas

Yidong Gan



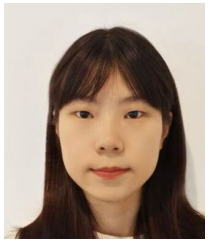
Qixuan (Cody) Hu



Ka Weng Pan

Ana Clarissa Miranda Pena

Ngoc Gia Hy Nguyen



Yongli Xiang

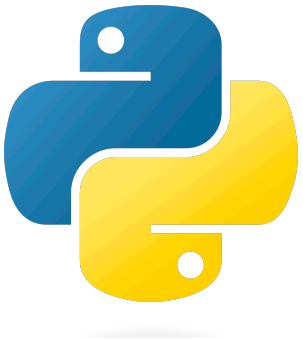
Henry Weld

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

What prerequisites / assumptions does the course have?



Python programming

$$\sum_{e^x} \frac{df}{dx}^{\%} log$$

Knowledge of maths and stats (high school)

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

How do lectures work?

5:05, Start
5:05 - 6:50, lecture with 1-3 short stretch breaks
6:50 - 6:55, muddy cards

Questions
- Ask and vote using Menti
- Answered after each break

Lectures are recorded

My advice:
- Write notes by hand
- No devices, except a phone for menti in breaks
- Sit where you don't see other devices

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

What is the assessment?

60% - exam

10% - assignment (group)

20% - assignments (individual)
5% each for 4 assignments

5% - quizzes
1.25% each, we take your top 4 (i.e., drop 1)

5% - muddy cards
0.5% each week, we take your top 10 (i.e., drop 3)

For the schedule, how to submit, etc, see Canvas

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

How does the exam work?

Must score at least 40% to pass

Main Exam - written
Replacement Exam - written
Second Replacement Exam - TBC, either written or oral

Covers everything in the course: lectures, labs, assignments

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

How do quizzes work?

Short, on paper, at the start of lecture

Will vary over semester in how they work

Goal: Spread revision over the semester

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

How do workshops work?

Pre-work (about 1 hour)

In-person activities (1 hour), mostly in pairs

No marks in the lab, but the content will be covered in quizzes

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

What are muddy cards?

At the end of each lecture you will a
- What was the **least clear** thing in
- Why did you give that response?

"I love the format of muddy card!"

Good examples:
"The derivation of the complexity of quicksort."
"Why do we use log probability rat

B
"
"

"The muddy card is really creative way of improving students' understanding."

"greatly helps me understand the course content"

Must be done between 6:45 and 7:05 on the lecture day

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

This year the muddy cards are part of a research study

Participants - All students in various large units of study across the university

Researchers - Thomas Elton (honours student) and me

Participation is optional

Participant Information Sheet

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

How do individual assignments work?

Smaller than a typical CS assignment

Released on Thursdays **(Assignment 1 out this week!)**

Due 8 working days later (e.g. out Thursday week 1, due Tuesday week 3).

Individual code writing, can discuss questions with others

No simple extensions because these are short release assignments. However, there are ‘slip days’:
- Can submit late with no penalty by using your slip days
- 5 slip days over semester
- Max 2 per assignment
- Automatically applied

My advice:

- Save these in case something happens. We will not grant more.

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

Where can I seek assistance if I have a disability?

You may not think of yourself as having a ‘disability’ but the definition under the Disability Discrimination Act (1992) is broad and includes temporary or chronic medical conditions, physical or sensory disabilities, psychological conditions and learning disabilities.

The types of disabilities USyd Disability Services see include:
Anxiety // Arthritis // Asthma // Autism // ADHD // Bipolar disorder // Broken bones // Cancer // Cerebral palsy // Chronic fatigue syndrome // Crohn’s disease // Cystic fibrosis // Depression Diabetes // Dyslexia // Epilepsy // Hearing impairment // Learning disability // Mobility impairment // Multiple sclerosis // Post-traumatic stress // Schizophrenia // Vision impairment
and much more.

Students needing assistance must register with Disability Services. It is advisable to do this as early as possible. Please contact us or review our website to find out more.

sydney.edu.au/disability - 02-8627-8422

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

Will plagiarism detectors be used?

Yes, we apply them to all assignments.

http://sydney.edu.au/elearning/student/EI/index.shtml:

“The University of Sydney is unequivocally opposed to, and intolerant of, plagiarism and academic dishonesty.

Academic dishonesty means seeking to obtain or obtaining academic advantage for oneself or for others (including in the assessment or publication of work) by dishonest or unfair means.

Plagiarism means presenting another person’s work as one’s own work by presenting, copying or reproducing it without appropriate acknowledgement of the source.”

Penalties for academic dishonesty or plagiarism can be severe

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

Can I use AI tools?

Yes, but…

(1) No AI in the exam (and remember, the exam will cover content from the assignments)
(2) Muddy card is less useful
(3) In the past, I had people come to office hours who had gotten confused by ChatGPT

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

How can I ask questions?

Admin
Email me with [COMP 4446 / 5046] in the subject:
jonathan.kummerfeld@sydney.edu.au

Course Content
1. Lecture (via menti)
2. Workshop - ask your tutor
3. Ed
4. Office Hours

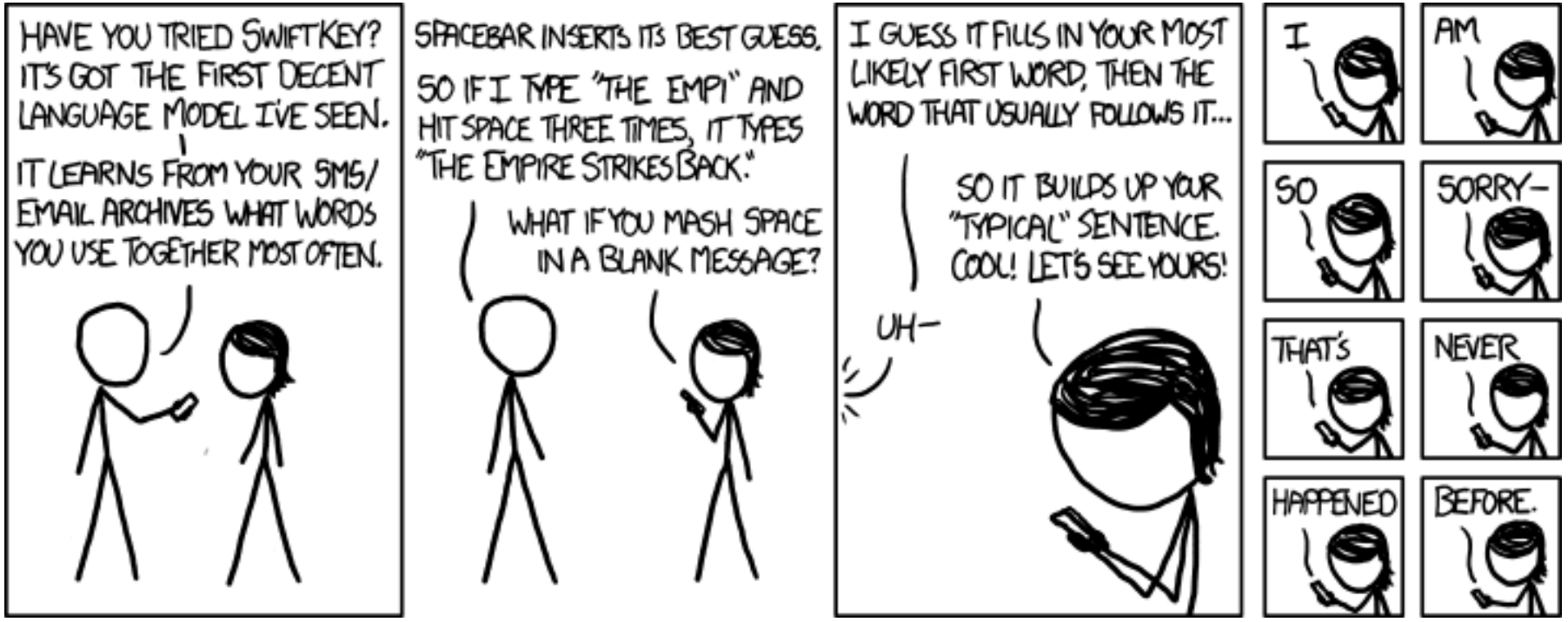
Check Canvas first!

Note:
- Do NOT send messages through Canvas or Direct Messages (DMs) in Ed
- We will (mostly) not respond on the weekend

3 minute Break - stretch and visit Menti

menti.com
4210 8267

Swiftkey



[Although the Markov chain-style text model is still rudimentary; it recently gave me "Massachusetts Institute of America". Although I have to admit it sounds prestigious.]

Source: https://xkcd.com/1068/

COMP 4446 / 5046
Lecture 1, 2025

menti.com 4210 8267

Representing Text

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

How we represent text determines what we can do

Computers work with bits, not language. How do we represent language with bits?

What should a word representation capture, to make it broadly useful?

- Morphology, e.g., does the word end in ’ed’?
- Animacy, e.g., is it a living thing?
- Number, e.g., is this a singular word?
- Concept similarity, e.g., ‘eat’ and ‘consume’ should be similar

…

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

Represent a word by its definition

Dictionary

right adj. located nearer the right hand esp. being on the right when facing the same direction as the observer.

left adj. located nearer to this side of the body than the right.

red n. the color of blood or a ruby.

blood n. the red liquid that circulates in the heart, arteries and veins of animals.

Jurafsky and Martin, Appendix G

COMP 4446 / 5046
Lecture 1, 2025

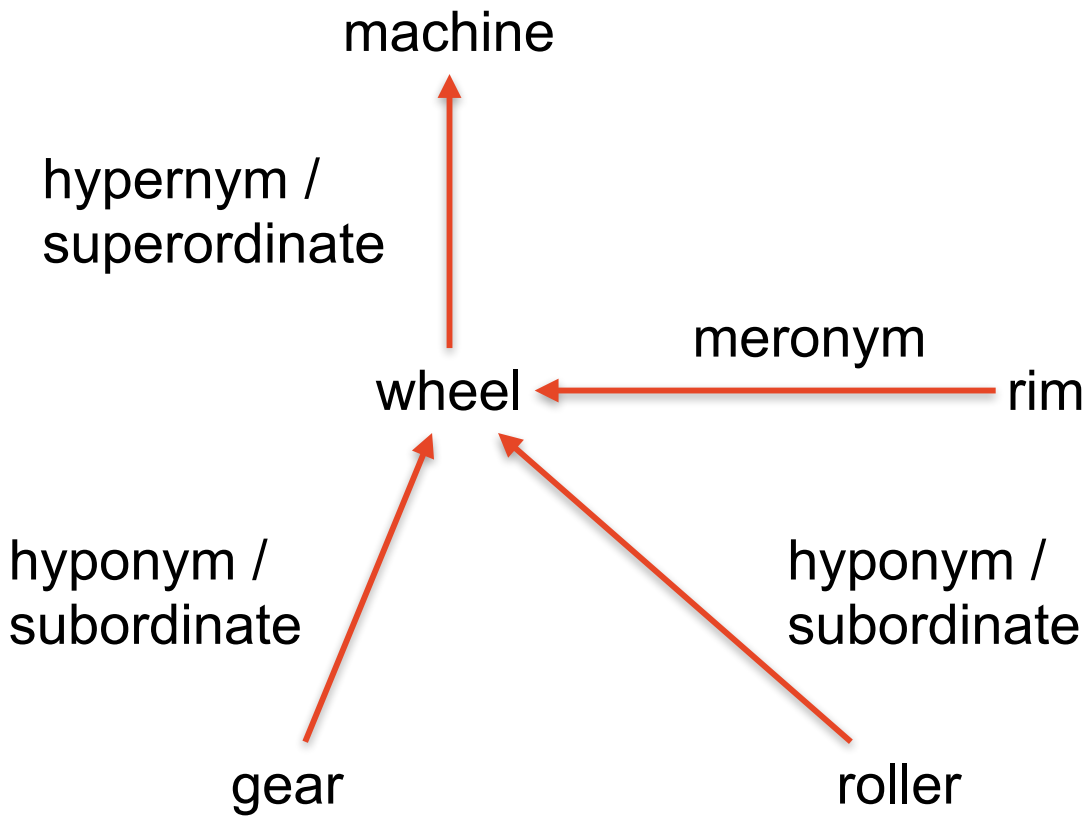
Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

Represent a word by its relationships with other words

WordNet

a simple machine consisting of a circular frame …



http://wordnetweb.princeton.edu/perl/webwn

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

Represent a word by its relationships with other words

WordNet

Synonym set for ‘run’

scat, run, scarper, turn tail, lam, run away, hightail it, bunk, head for the hills, take to the woods, escape, fly the coop, break away

English:
- 117,798 nouns
- 11,529 verbs
- 22,479 adjectives
- 4,481 adverbs.

Also in 200+ other languages! But… smaller

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

Represent words using numbers

What data structure for storage?

Vector ('one-hot')

Index 27 = 1 to indicate 'hall'

Sparse Vector
[27]

Set
{27}

We also need a dictionary that maps from words to their numbers
OR
We can use a hash function (and live with some collisions)

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

Represent words using numbers

What data structure for storage?

Vector ('one-hot')

Hot code 00 ... 010 ...

Sparse Vector

```
[(27, 3)]
```

↳ 3 times

Map

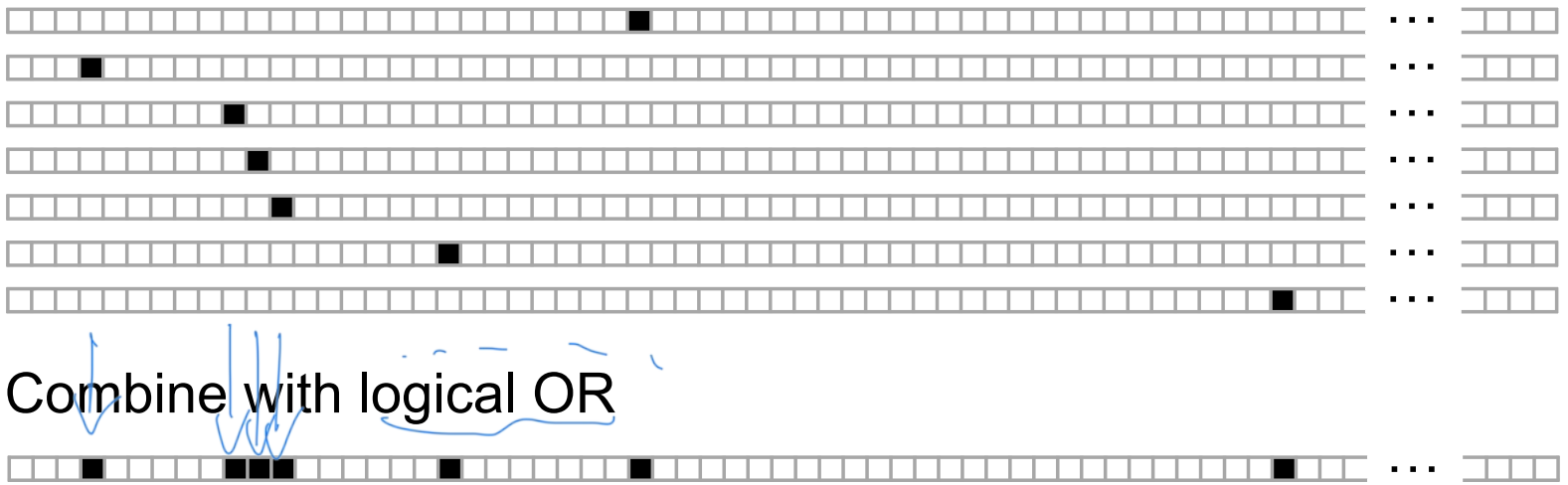
```
{27: 3}
```

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

Represent **documents** using numbers



Combine with logical OR

We are losing word frequencies!

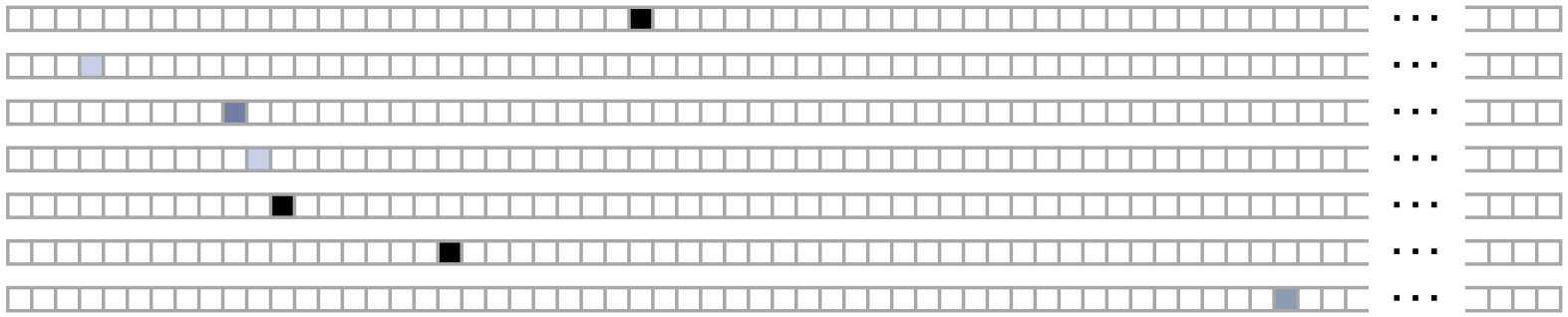
Sometimes called a
Bag of Words

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

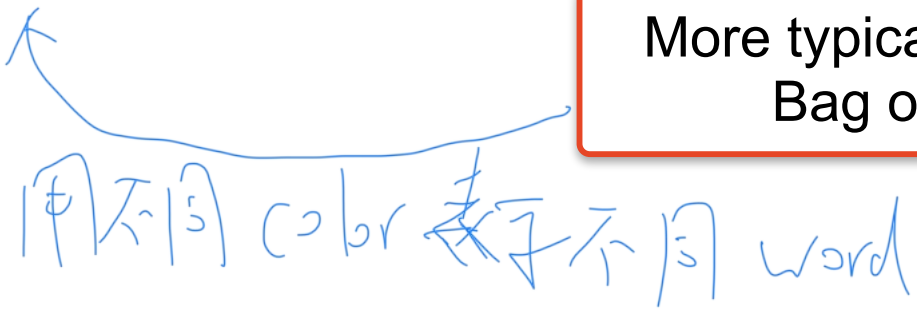
menti.com 4210 8267

Represent **documents** using numbers



Added together:

More typical meaning of Bag of Words



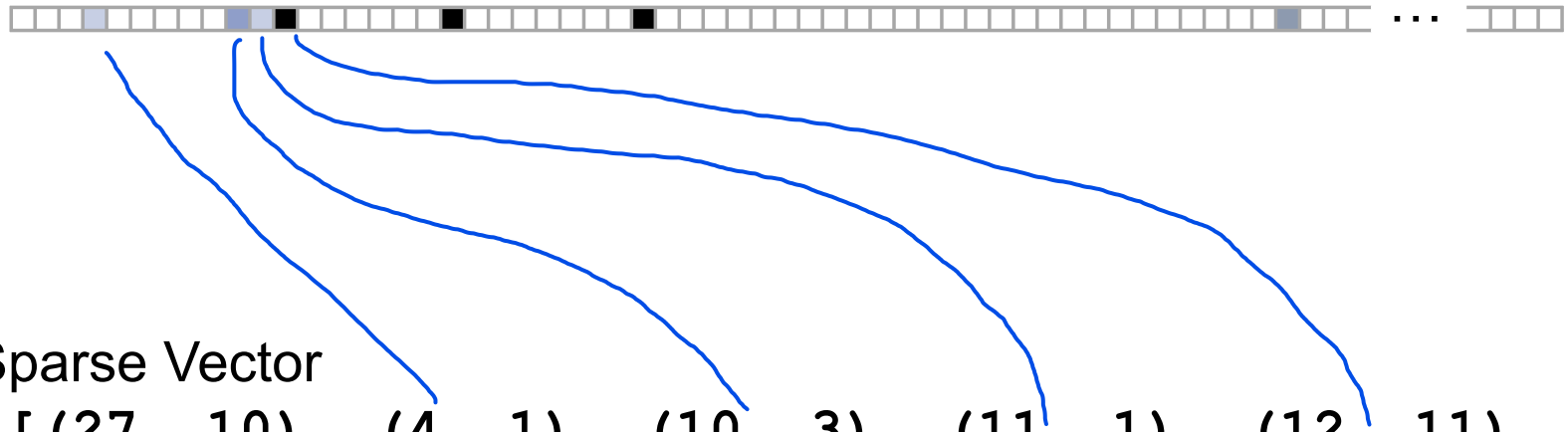
COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

Represent **documents** using numbers

Vector



Sparse Vector

```
[(27, 10), (4, 1), (10, 3), (11, 1), (12, 11),
(19, 12), (54, 6)]
```

Map

```
{27:10, 4:1, 10:3, 11:1, 12:11, 19:12, 54:6}
```

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

What are the tradeoffs in performance between these storage methods?

	Size	Time to compare
Vector	|Vocab|	|Vocab|
Sparse Vector	|Unique tokens|	Depends
Map	Depends	|Unique tokens|

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

Represent **documents** using numbers

This will be a common word, e.g., ‘the’, which is not so informative!

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

If you have a document collection, adjust counts based on it

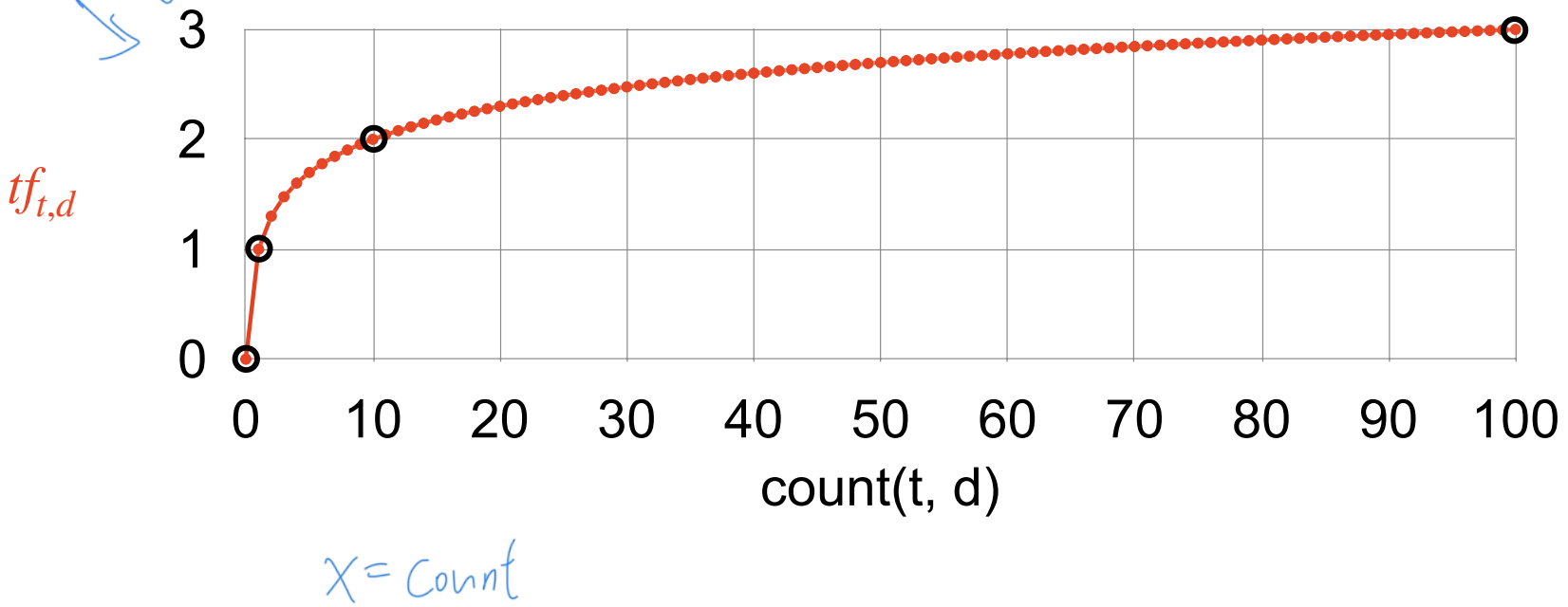
count(t, d) = how often token t occurs in document d
N = number of documents
df(t) = how many documents token t occurs in

Term Frequency

$$tf_{t,d} = \begin{cases} 1 + \log_{10} \text{count(t, d)} & \text{if count(t, d)} > 0 \\ 0 & \text{otherwise} \end{cases}$$

rescaler

Frequency = y



X = Count

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

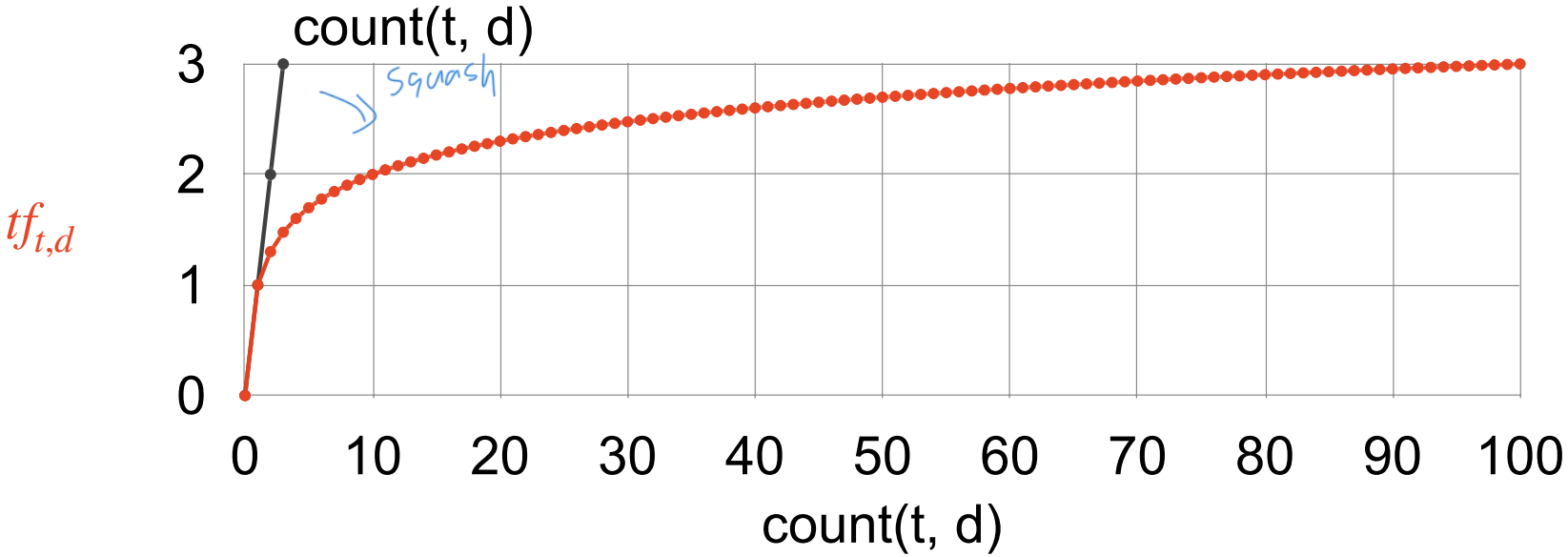
menti.com 4210 8267

If you have a document collection, adjust counts based on it

count(t, d) = how often token t occurs in document d
N = number of documents
df(t) = how many documents token t occurs in

Term Frequency

$$tf_{t,d} = \begin{cases} 1 + \log_{10} \text{count(t, d)} & \text{if count(t, d)} > 0 \\ 0 & \text{otherwise} \end{cases}$$



COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

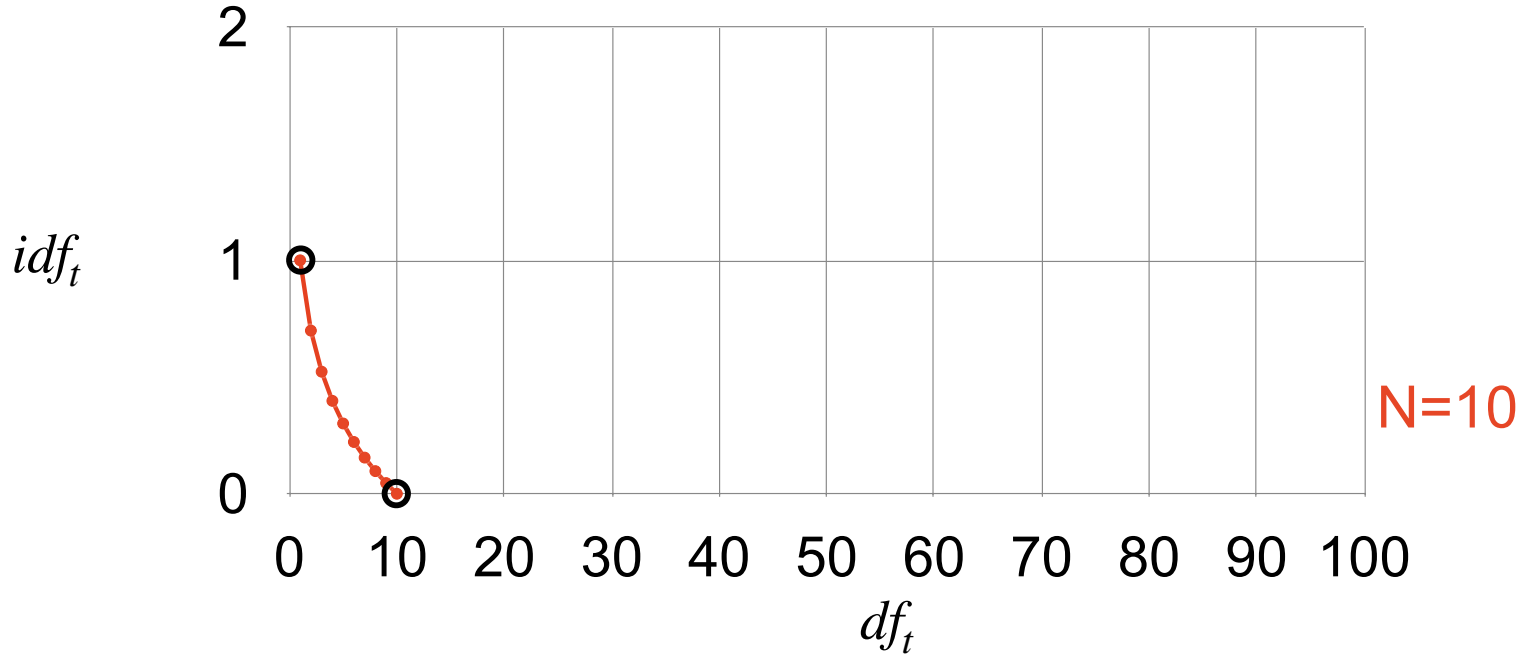
menti.com 4210 8267

If you have a document collection, adjust counts based on it

count(t, d) = how often token t occurs in document d
N = number of documents
df(t) = how many documents token t occurs in

Inverse Document Frequency

$$idf_t = \log_{10}\left(\frac{N}{df_t}\right)$$



COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

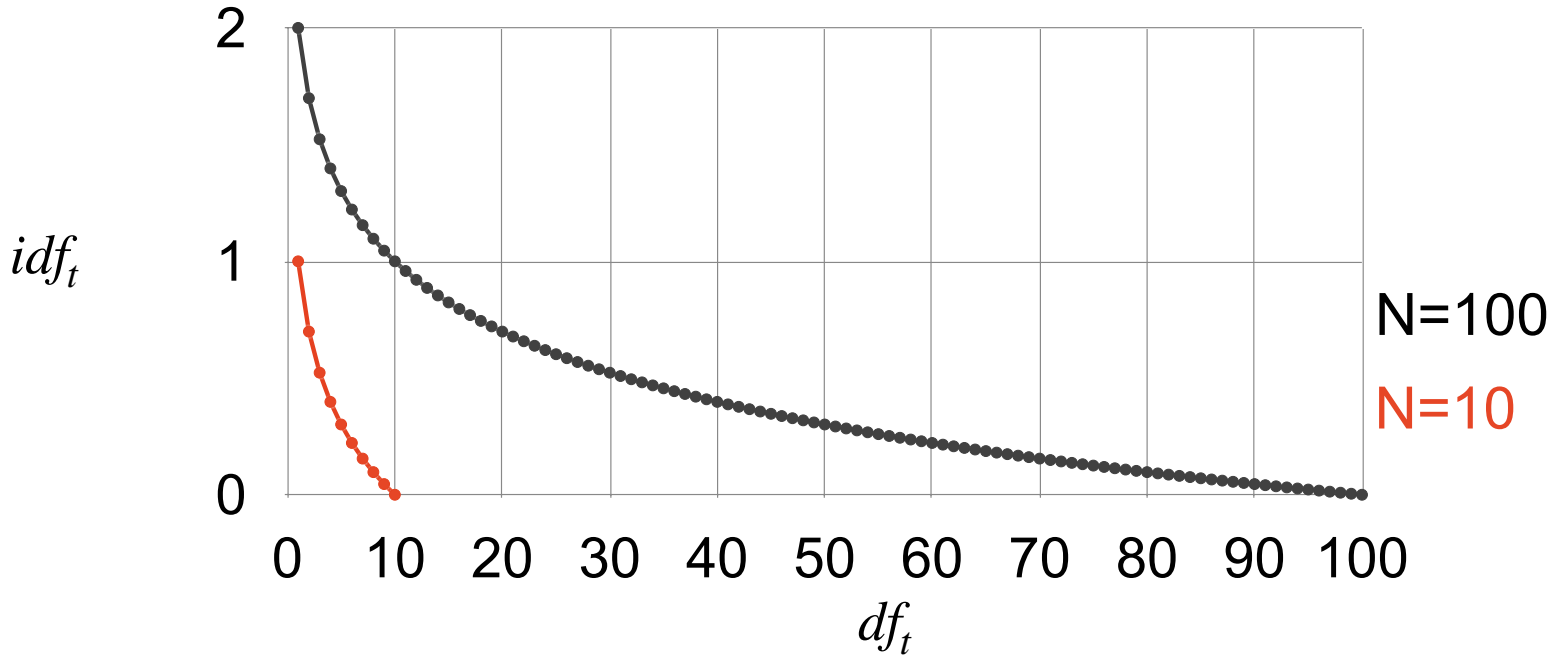
menti.com 4210 8267

If you have a document collection, adjust counts based on it

count(t, d) = how often token t occurs in document d
N = number of documents
df(t) = how many documents token t occurs in

Inverse Document Frequency

$$idf_t = \log_{10}\left(\frac{N}{df_t}\right)$$



COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

If you have a document collection, adjust counts based on it

count(t, d) = how often token t occurs in document d
N = number of documents
df(t) = how many documents token t occurs in

Term Frequency ①

$$tf_{t,d} = \begin{cases} 1 + \log_{10} \text{count(t, d)} & \text{if count(t, d)} > 0 \\ 0 & \text{otherwise} \end{cases}$$

Inverse Document Frequency ②

$$idf_t = \log_{10}\left(\frac{N}{df_t}\right)$$

TF-IDF ③

$$w_t = tf_{t,d} \times idf_t$$

3 ways. understand these

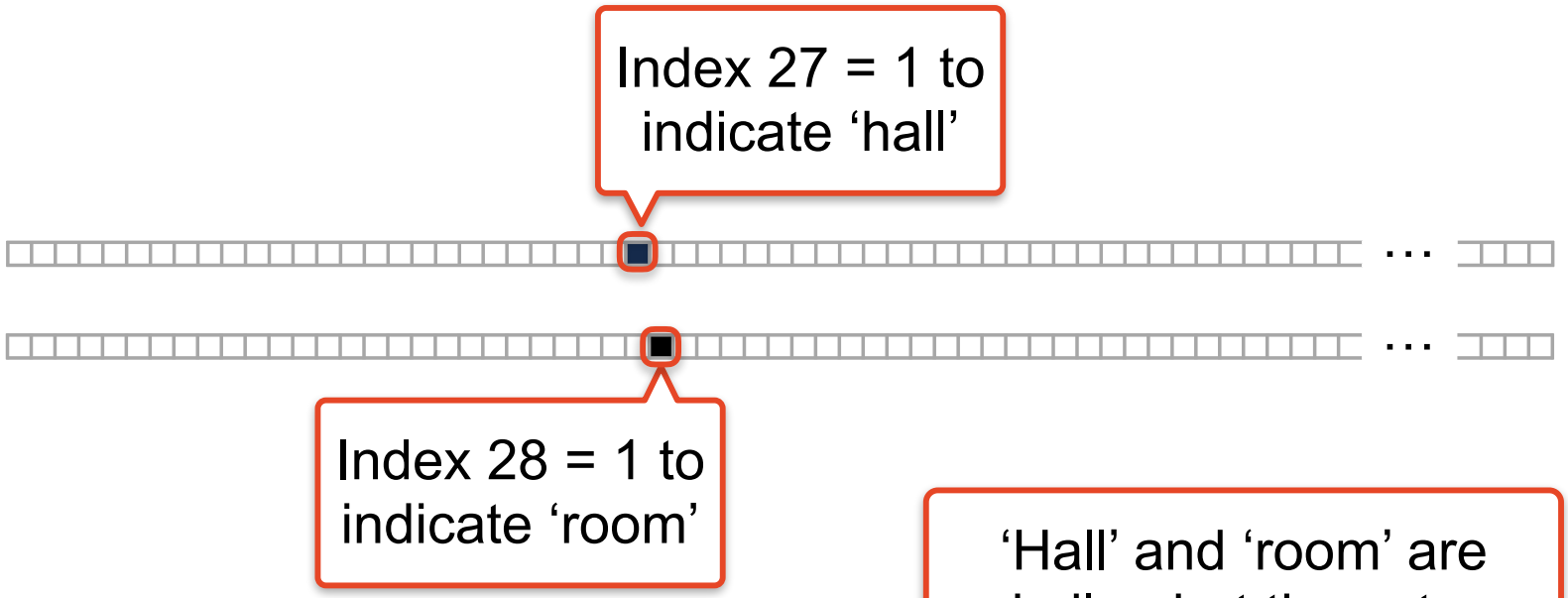
One of many variants of TF-IDF!

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

How well does this capture similarity?

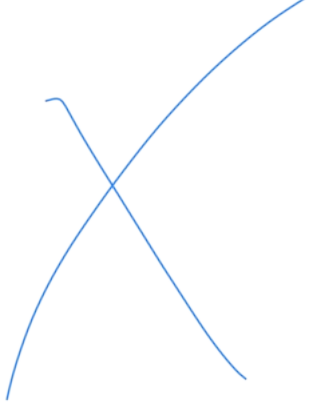


COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

Represent a word by the contexts it appears in



“the meaning of words lies in their use”
Ludwig Wittgenstein

“You shall know a word by the company it keeps!”
J. R. Firth

“language can be described in terms of a distributional structure”
Zellig Harris

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

Distributional similarity

A **book** is a medium for recording information in the form of writing or images.

A paperback **book** is one with a thick paper or paperboard cover.

Shrek! is a fantasy comedy picture **book** published in 1990.

The **book** is recognized as a classic in children's literature and is one of the best-selling books of all time

related to "book"

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

Distributional similarity

Context words:

drove
door
luxury
speed
crashed
insurance

car
automobile

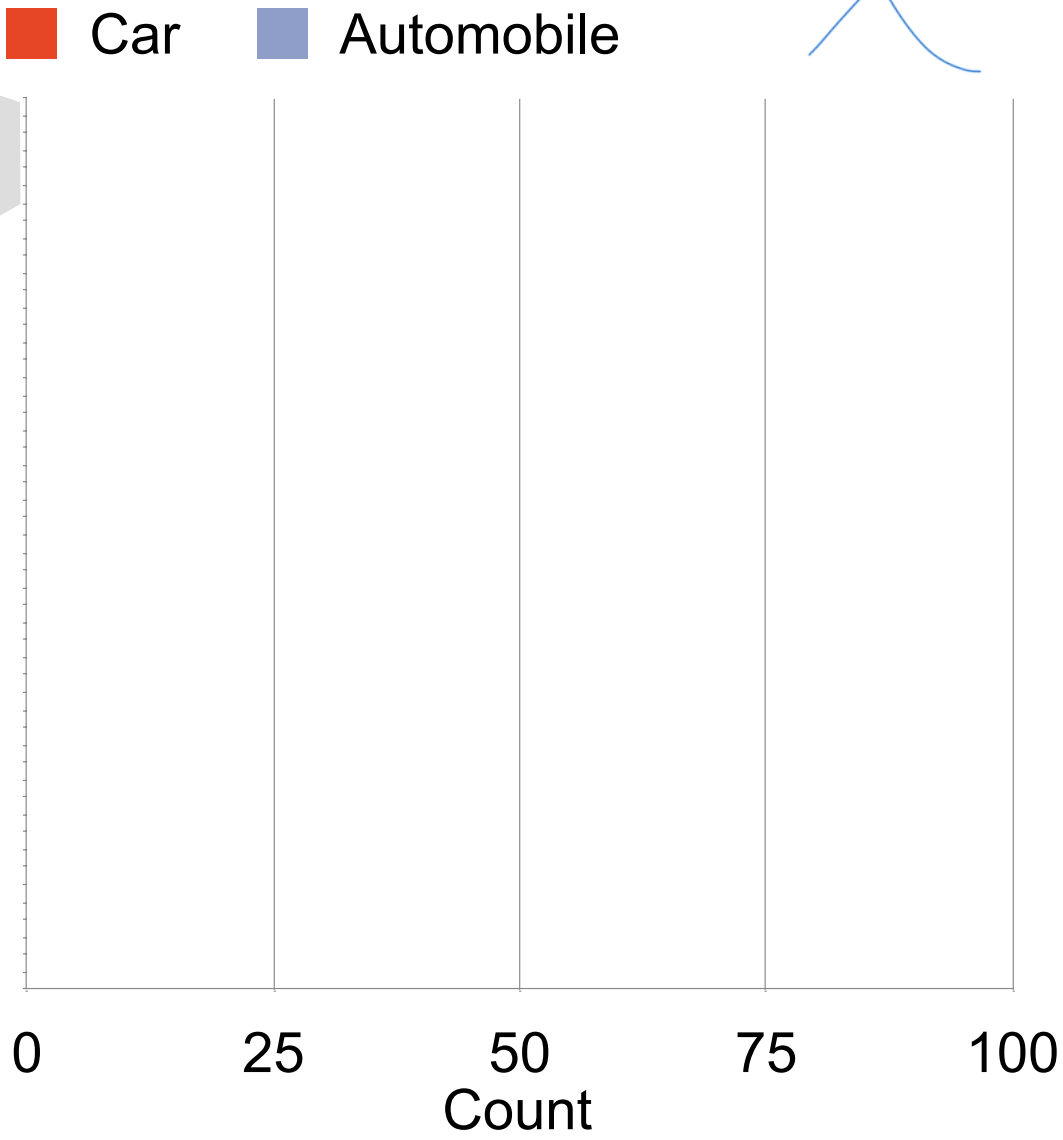
COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

Distributional similarity

drove
door
luxury
speed
crashed
insurance



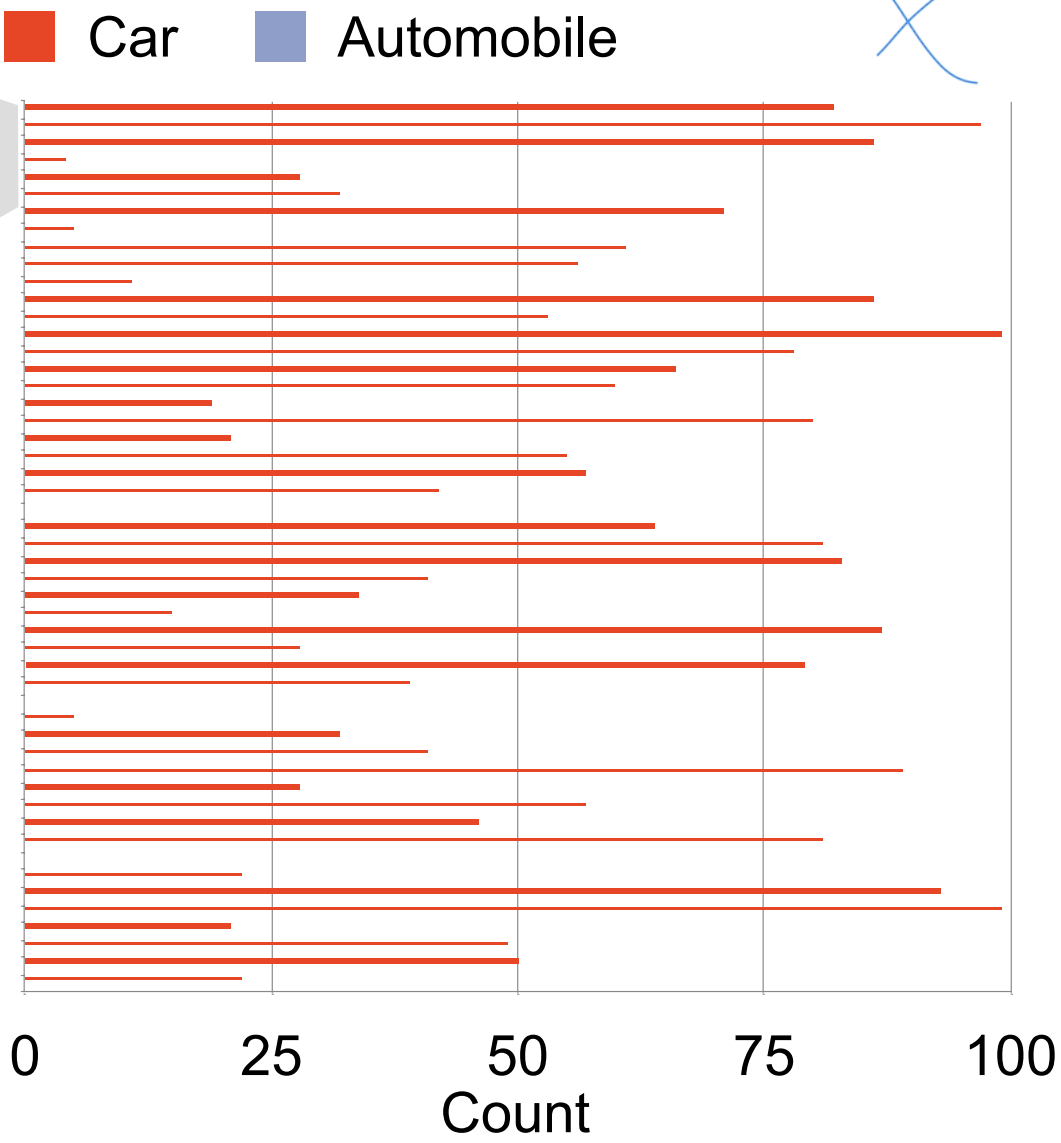
COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

Distributional similarity

drove
door
luxury
speed
crashed
insurance



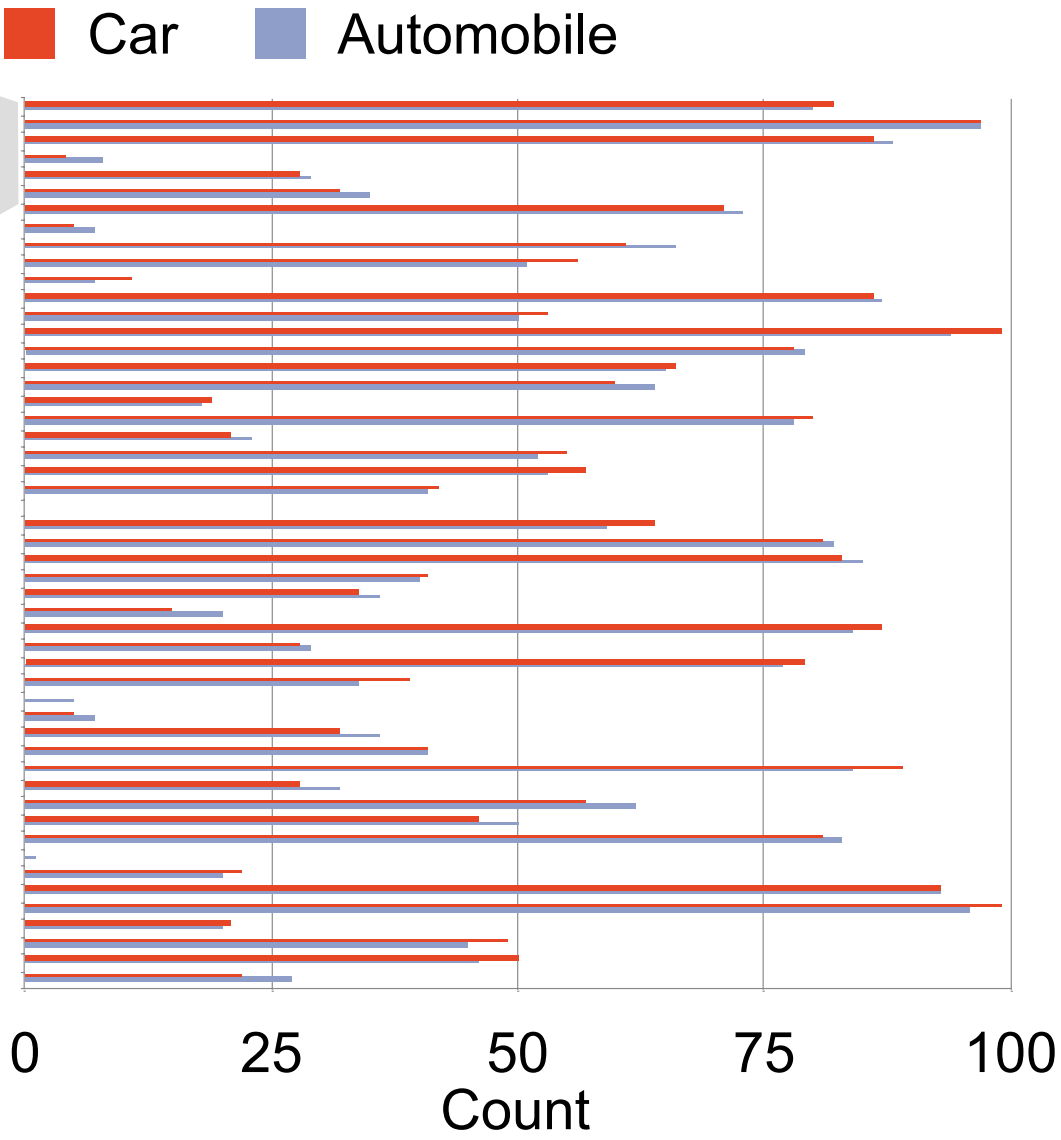
COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

Distributional similarity

drove
door
luxury
speed
crashed
insurance



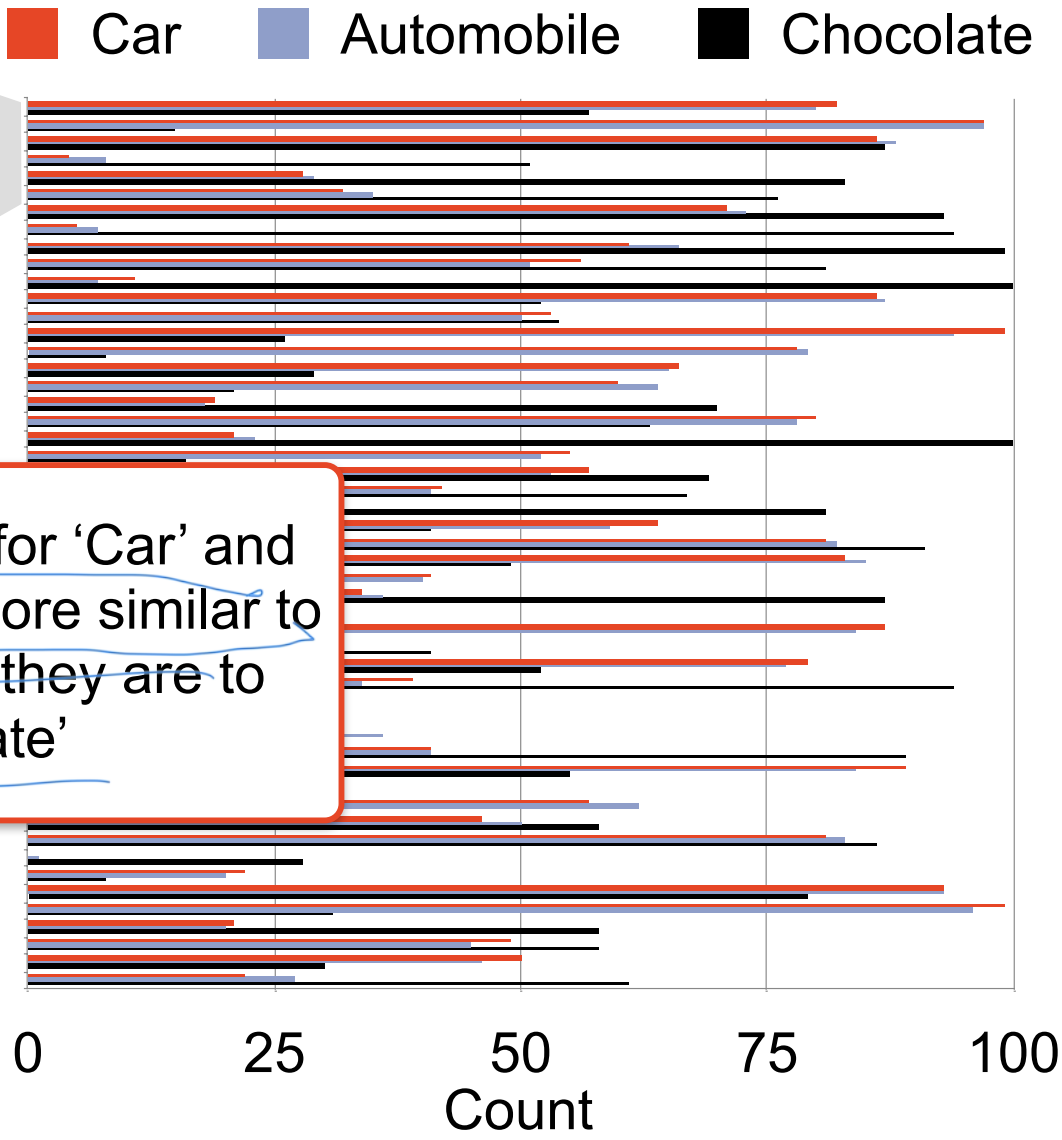
COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

Distributional similarity

drove
door
luxury
speed
crashed
insurance



The distributions for ‘Car’ and ‘Automobile’ are more similar to each other than they are to ‘Chocolate’

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

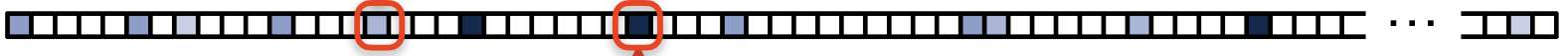
Idea 1: Learn a representation by counting

Method 1
learn the relationship between words

1. Get a lot of data
2. For each word *w*, count how many times every other word appears with *w* in a sentence
3. Put those counts in a vector

Count for being in the same sentence as 'drove'

Word Representation



Count for being in the same sentence as 'the'

These vectors are big and sparse…

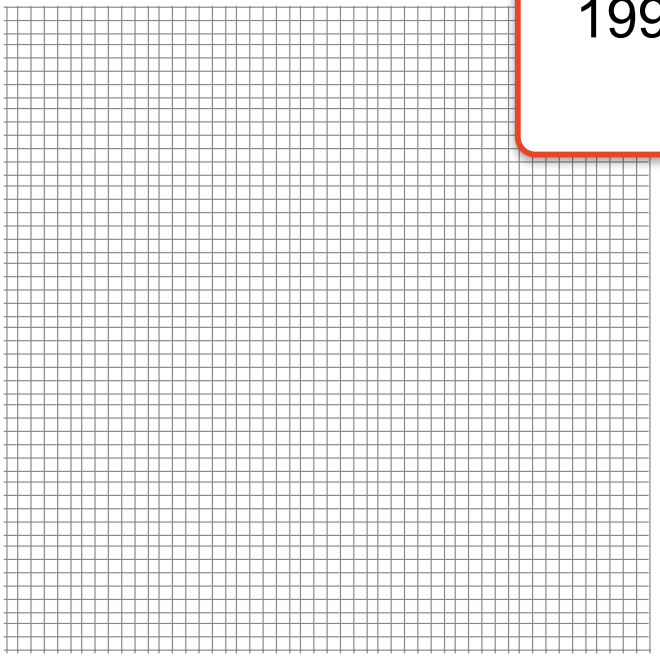
COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

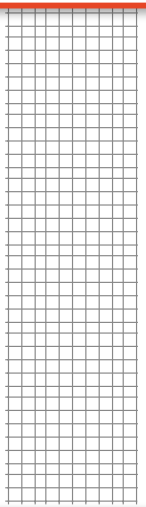
menti.com 4210 8267

Idea 2: Use dimensionality reduction to get a smaller vector

This idea was first introduced in 1993 (Schütze, ACL), but did not take off until later



Singular Value Decomposition



These vectors are small and dense

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

Can you guess the next word?

Today is Monday — World state

1, 2, 3, 4 — Numerical patterns

We are at The University of Sydney — Speaker context

Qantas bought a 787 from Boeing — World knowledge

I like games. Portal is a game. Therefore, I like Portal. — Logic

Winter's a good time to stay in and cuddle
But put me in summer and I'll be a happy snowman! puddle

Pop Culture — Rhyme

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

Alternative Idea: Make up a task and learn vectors with it

Data

Raw text:
the key to an excellent hot chocolate is using enough chocolate

Task Input:
the key to an excellent hot [ ] ? is using enough chocolate

Task Output:
chocolate

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

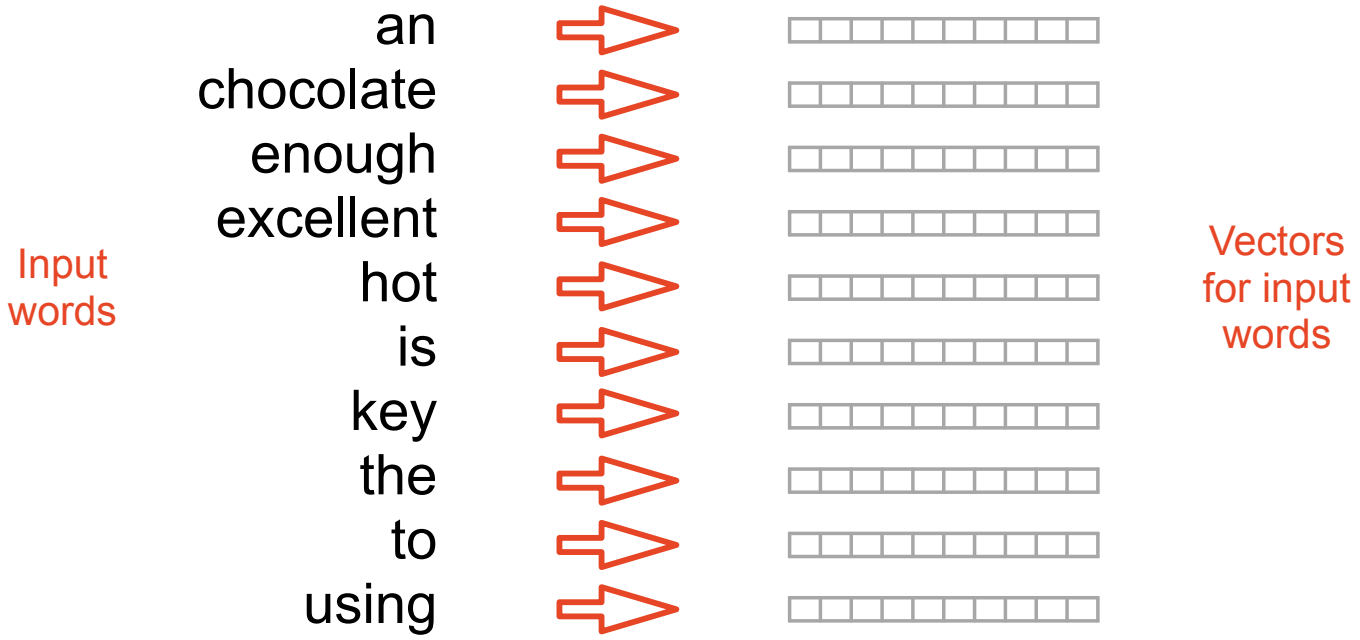
menti.com 4210 8267

Predict next word

Alternative Idea: Make up a task and learn vectors with it

Model

(1) Look up vectors for context words



COMP 4446 / 5046
Lecture 1, 2025

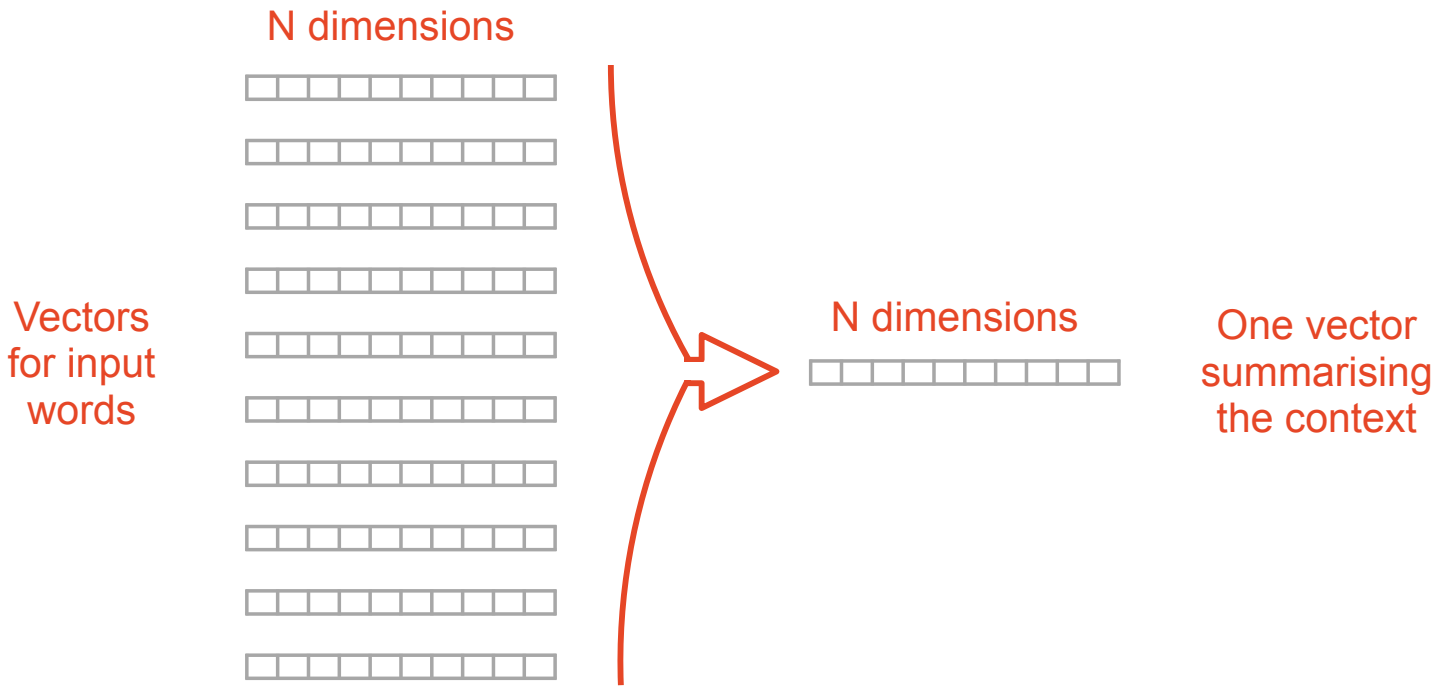
Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

Alternative Idea: Make up a task and learn vectors with it

Model

(2) Calculate mean



COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

Alternative Idea: Make up a task and learn vectors with it

Model

(3) Matrix Multiply

One vector summarising the context

N dimensions

Weight matrix

| vocabulary | dimensions

N dimensions

Scores for each possible word to guess

| vocabulary |

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

Alternative Idea: Make up a task and learn vectors with it

Model

(4) Normalise the vector

Scores for each possible word to guess

| vocabulary |

$$\frac{e^{x_i}}{\sum_{x_j} e^{x_j}}$$

Probabilities for each possible word to guess

| vocabulary |

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

Alternative Idea: Make up a task and learn vectors with it

Model

(4) Calculate loss

Probabilities for each possible word to guess

$\hat{y}$

| vocabulary |

$$-\sum_{y_i} y_i \, log(\hat{y}_i)$$

True answer

y

| vocabulary |

loss

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

Alternative Idea: Make up a task and learn vectors with it

Model

(5) Update weights

Future lecture!

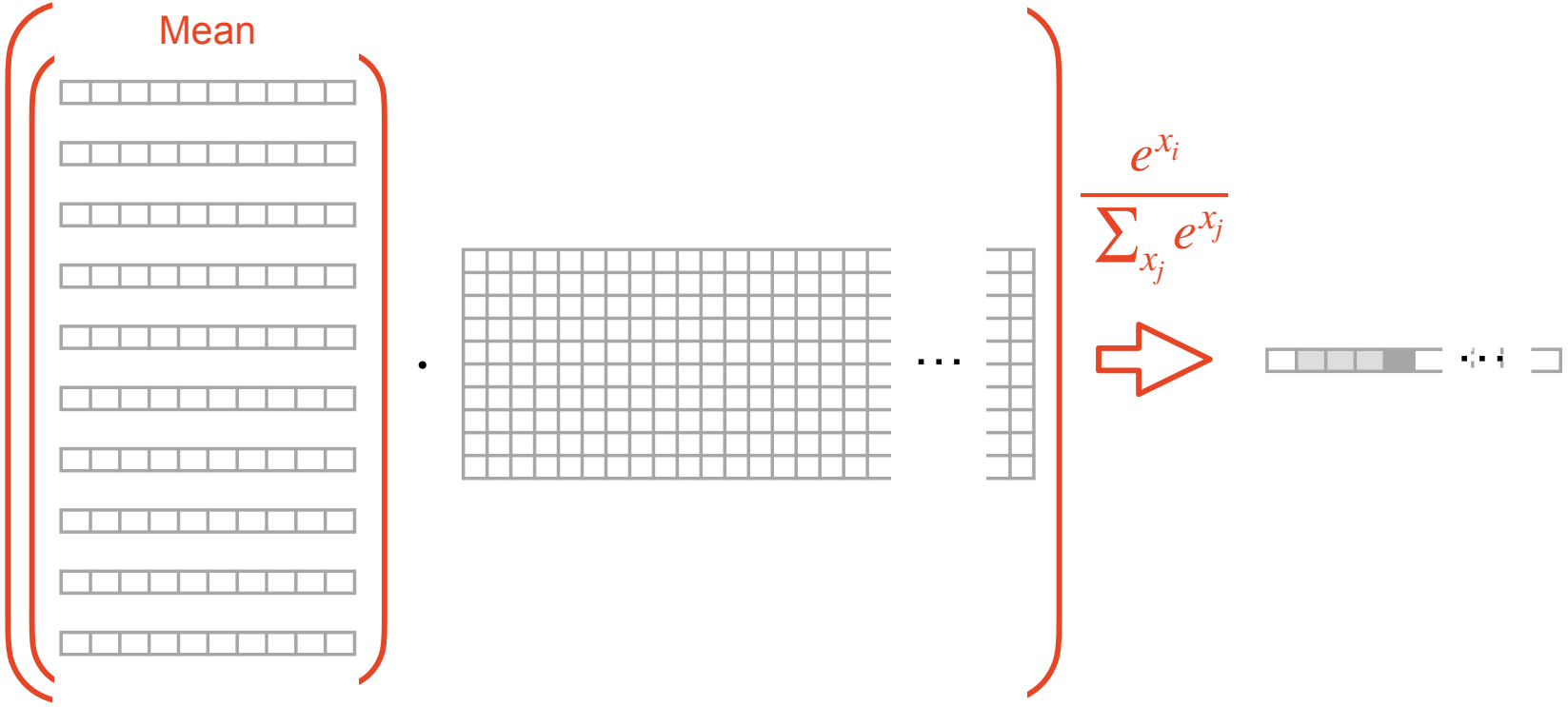
COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

Alternative Idea: Make up a task and learn vectors with it

Model



$$x = \frac{\sum_{\text{words}} \text{word vectors}}{|\text{words}|} \cdot \text{weight matrix}$$

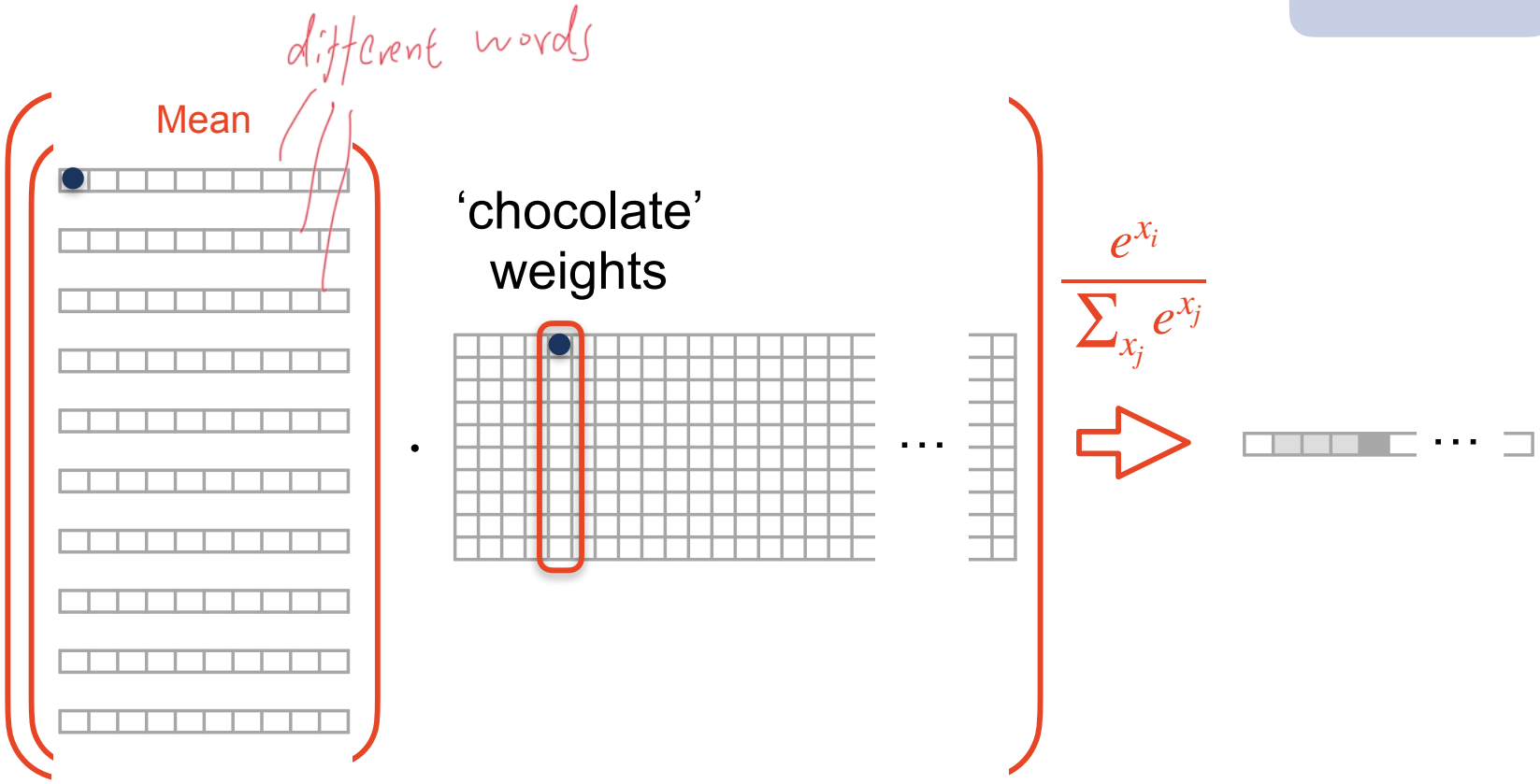
COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

Alternative Idea: Make up a task and learn vectors with it

Model



$$x = \frac{\sum_{\text{words}} \text{word vectors}}{|\text{words}|} \cdot \text{weight matrix}$$

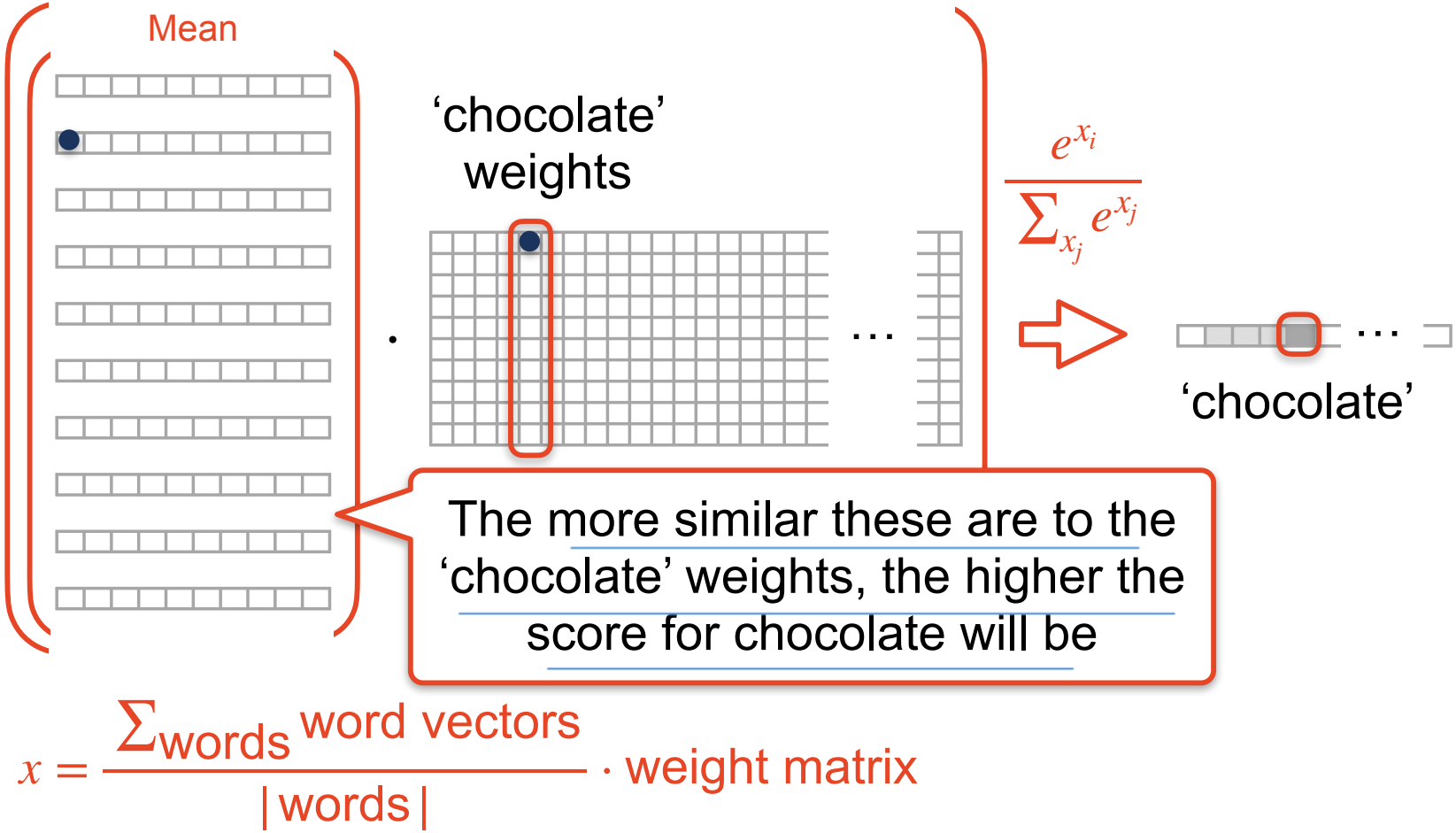
COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

Alternative Idea: Make up a task and learn vectors with it

Model



$$x = \frac{\sum_{\text{words}} \text{word vectors}}{|\text{words}|} \cdot \text{weight matrix}$$

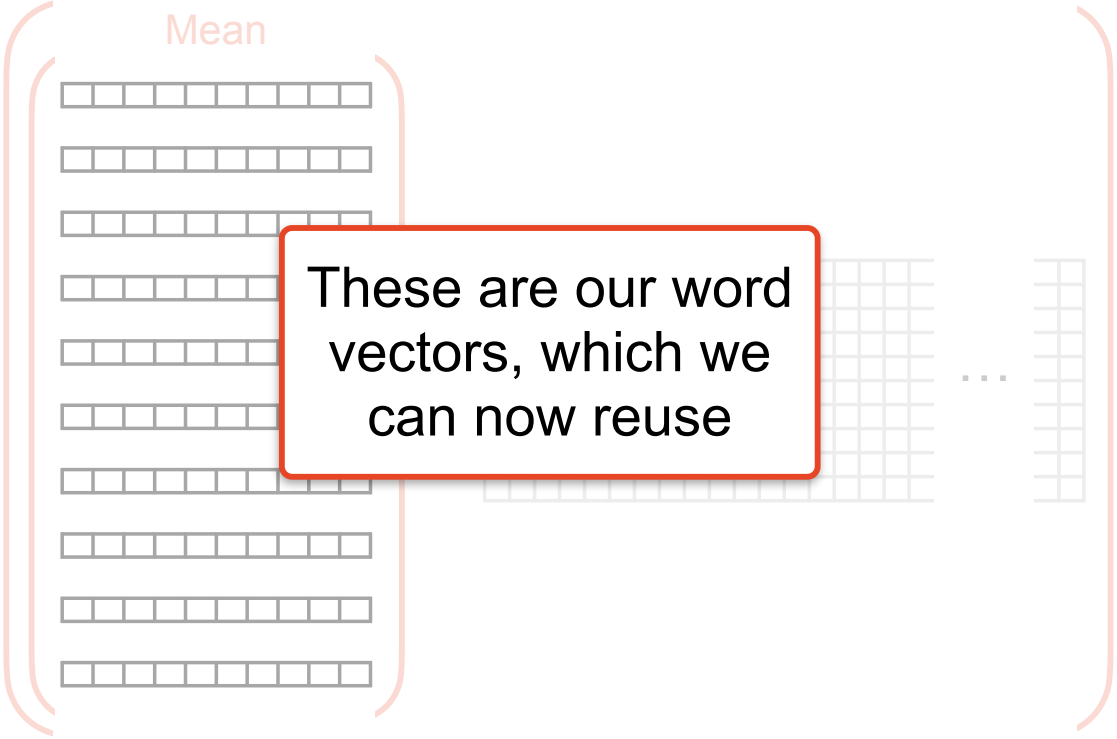
COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

Alternative Idea: Make up a task and learn vectors with it

Model



$$x = \frac{\sum_{\text{words}} \text{word vectors}}{|\text{words}|} \cdot \text{weight matrix}$$

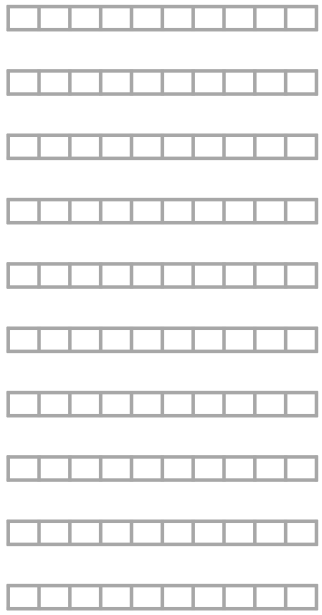
COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

Alternative Idea: Make up a task and learn vectors with it

Model



Raw text:
the key to an excellent hot chocolate is using enough chocolate

We can get lots of raw text very easily

Task Input:
the key to an excellent hot ? is using enough chocolate

Creating the inputs and outputs is also easy

Task Output:
chocolate

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

A few extra notes / details

Model

"Continuous Bag of Words", from word2vec

- 'Continuous' because the embeddings / vectors have a floating point value, rather than the integers in a Bag of Words
- Used only the 4 words before and 4 words after as context
- Also proposed the reverse task: given a word, predict the words in its context (Skip-gram), which tends to work better, but is less intuitive to explain

"Efficient Estimation of Word Representations in Vector Space", Mikolov, et al., (2013)
"Distributed representations of words and phrases and their compositionality", Mikolov, et al., (2013)

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

A few extra notes / details

Model

Details impact what is learned

- Small context (+/- 2 words), more syntactically similar words
- Larger context (+/- 5 words), more semantically similar words

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

Does it capture the phenomena we mentioned?

Model

We asked:
What should go into it, to make it broadly useful?
- Morphology, e.g., does the word end in 'ed'?
- Animacy, e.g., is it a living thing?
- Number, e.g., is this a singul
…

Negative here means inanimate

Option 1: An axis records it

Option 2: A direction records it (generalisation of 1)

Word Embedding

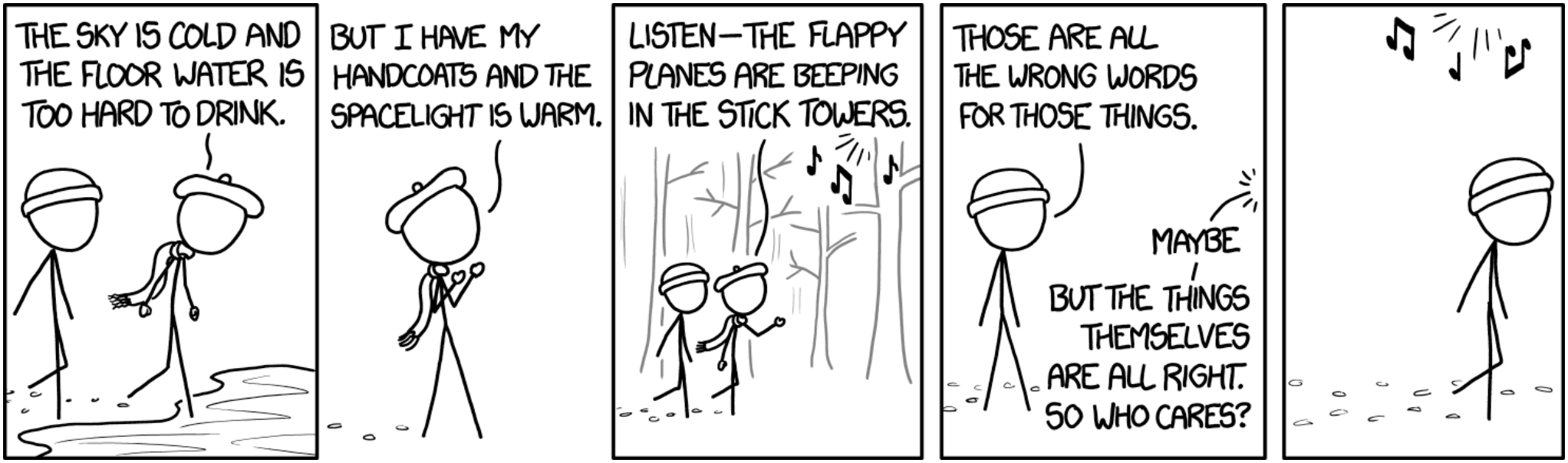
Overlap with 'animacy dimension'

Option 3: A region records it

3 minute Break - stretch and visit Menti

menti.com
4210 8267

Winter



[Stay warm, little flappers, and find lots of plant eggs!]

Source: https://xkcd.com/1322/

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

Evaluation

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

Comparing vectors

How to compare vector

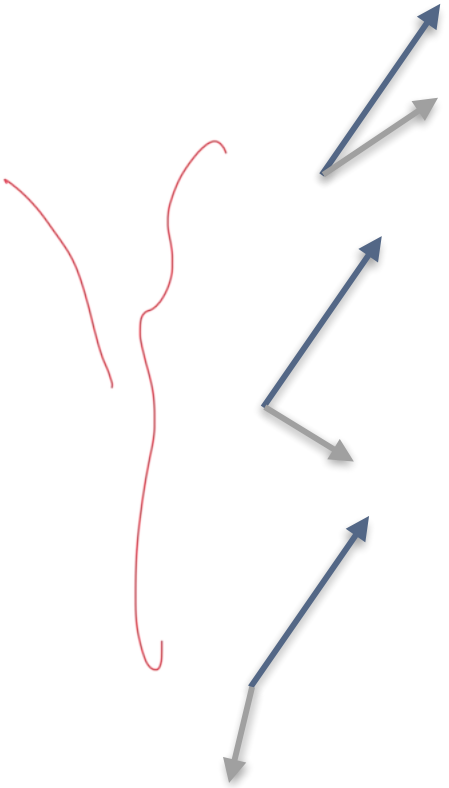
Idea 1: Dot product
- Favours long vectors

$$u \cdot v$$

Idea 2: Cosine similarity
- Widely used

更常用

$$\frac{u \cdot v}{|u||v|}$$



High similarity

Zero similarity

Negative similarity

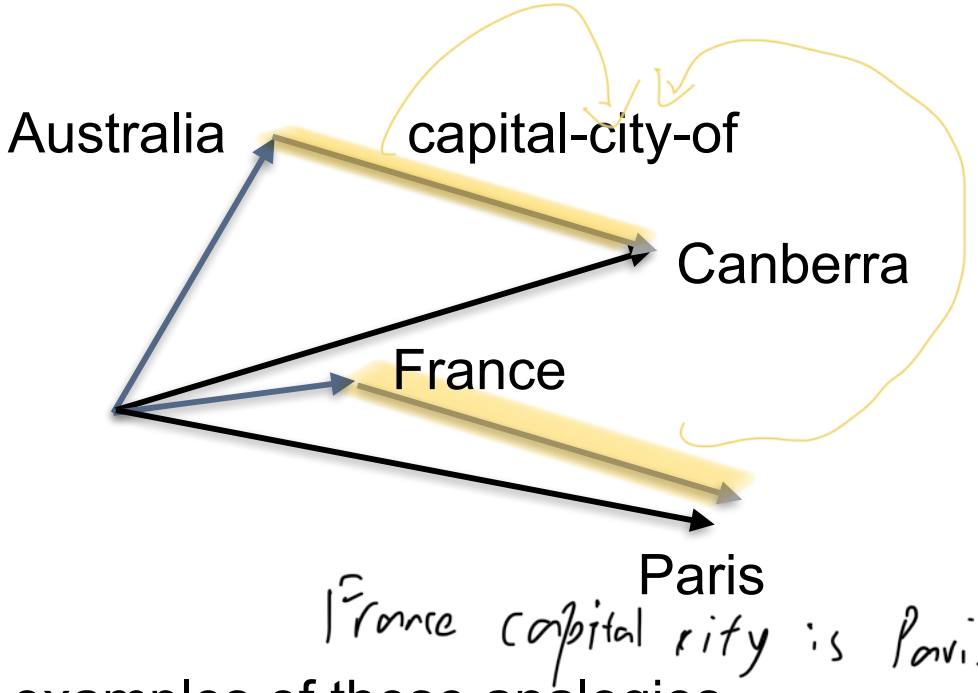
COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

Word analogy task

Intrinsic metric



Manually create examples of these analogies.
See how many the model word embedding gets right.

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

What does this learn?

Calculate [$word_1$ - $word_2$ + $word_3$] and find the word with a vector closest to the result that is not one of the input words

Italy capital city is Rome

Paris - France + Italy = Rome

Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza

"Efficient Estimation of Word Representations in Vector Space", Mikolov, et al., (2013)

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

What does this learn?

Caveats
- Mainly works for frequent words
- Only some relations
- Small distances in the vector space

Still cool!

And more research in progress…

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

"Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings"

Bolukbasi, et al., NeurIPS 2016

bias Resources

Word embeddings trained on news text

man - woman ≈ computer programmer - homemaker

father - mother ≈ doctor - nurse

Developed 'de-biasing' methods that adjust word embeddings to avoid specific patterns

COMP 4446 / 5046
Lecture 1, 2025

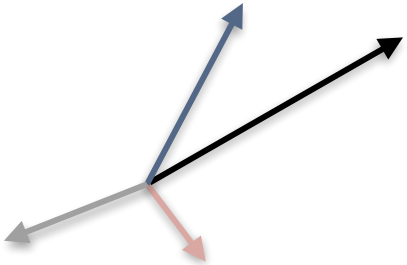
Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

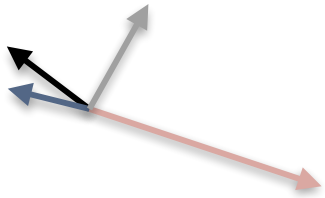
These representations can be quite variable

Keep the same:
Data, model, learning method, inference method
Vary the random initial values of the weights

Measure:
For each word, see how many of its nearest neighbours are the same



Embeddings v1



Embeddings v2

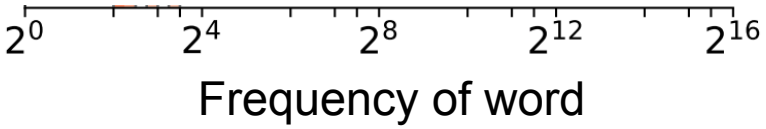
"Factors Influencing the Surprising Instability of Word Embeddings", Wendlandt, Kummerfeld, and Mihalcea (2018)

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

These representations can be quite variable



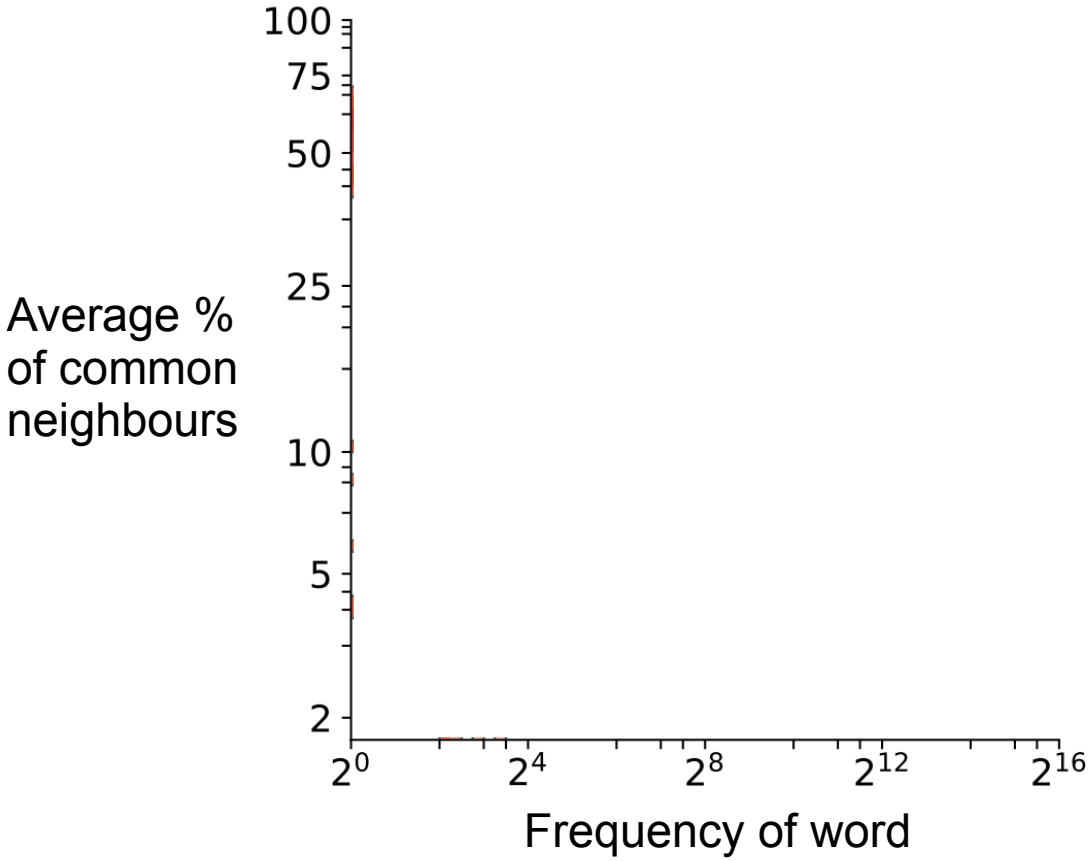
"Factors Influencing the Surprising Instability of Word Embeddings",
Wendlandt, Kummerfeld, and Mihalcea (2018)

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

These representations can be quite variable



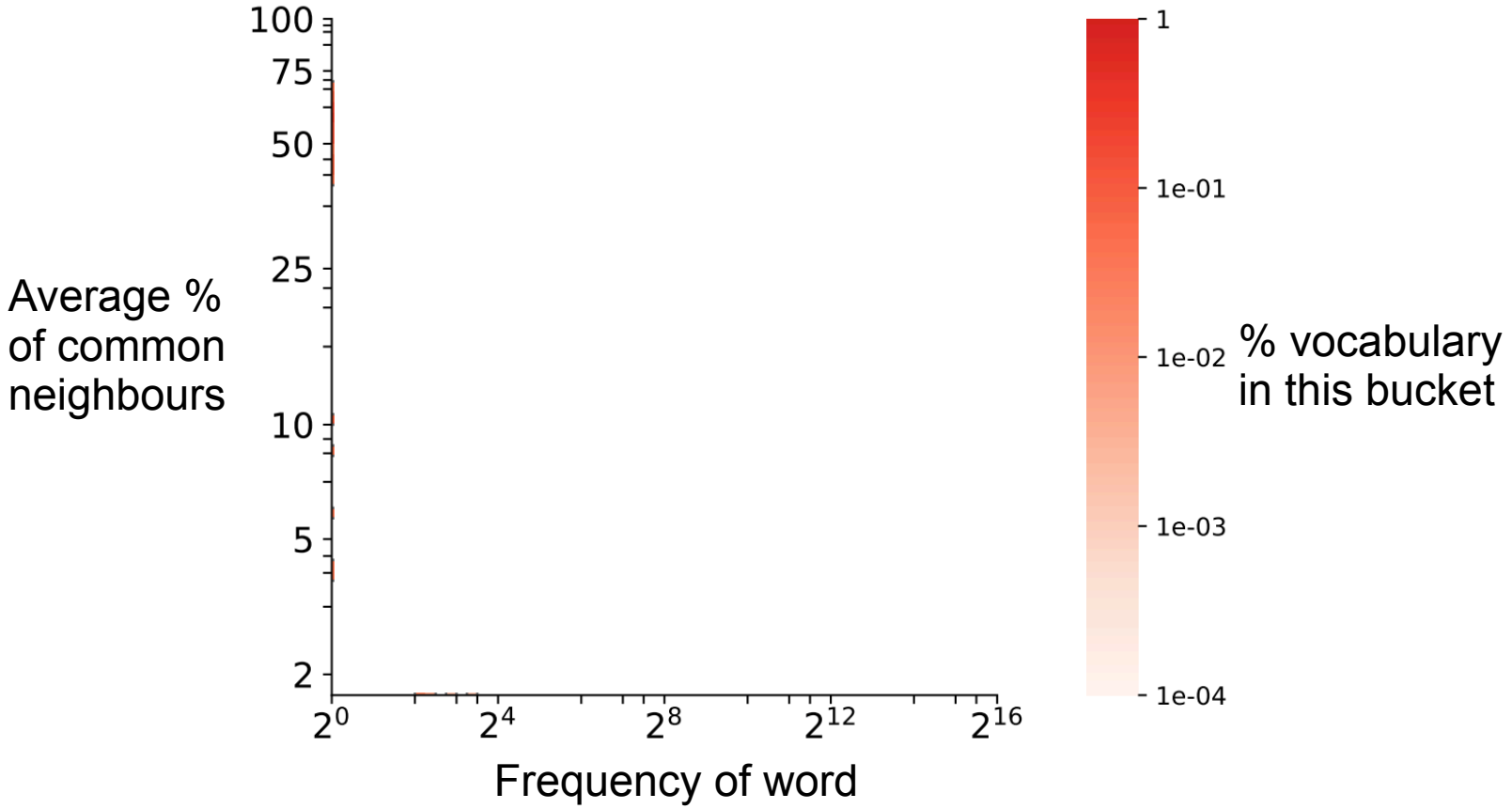
"Factors Influencing the Surprising Instability of Word Embeddings", Wendlandt, Kummerfeld, and Mihalcea (2018)

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

These representations can be quite variable



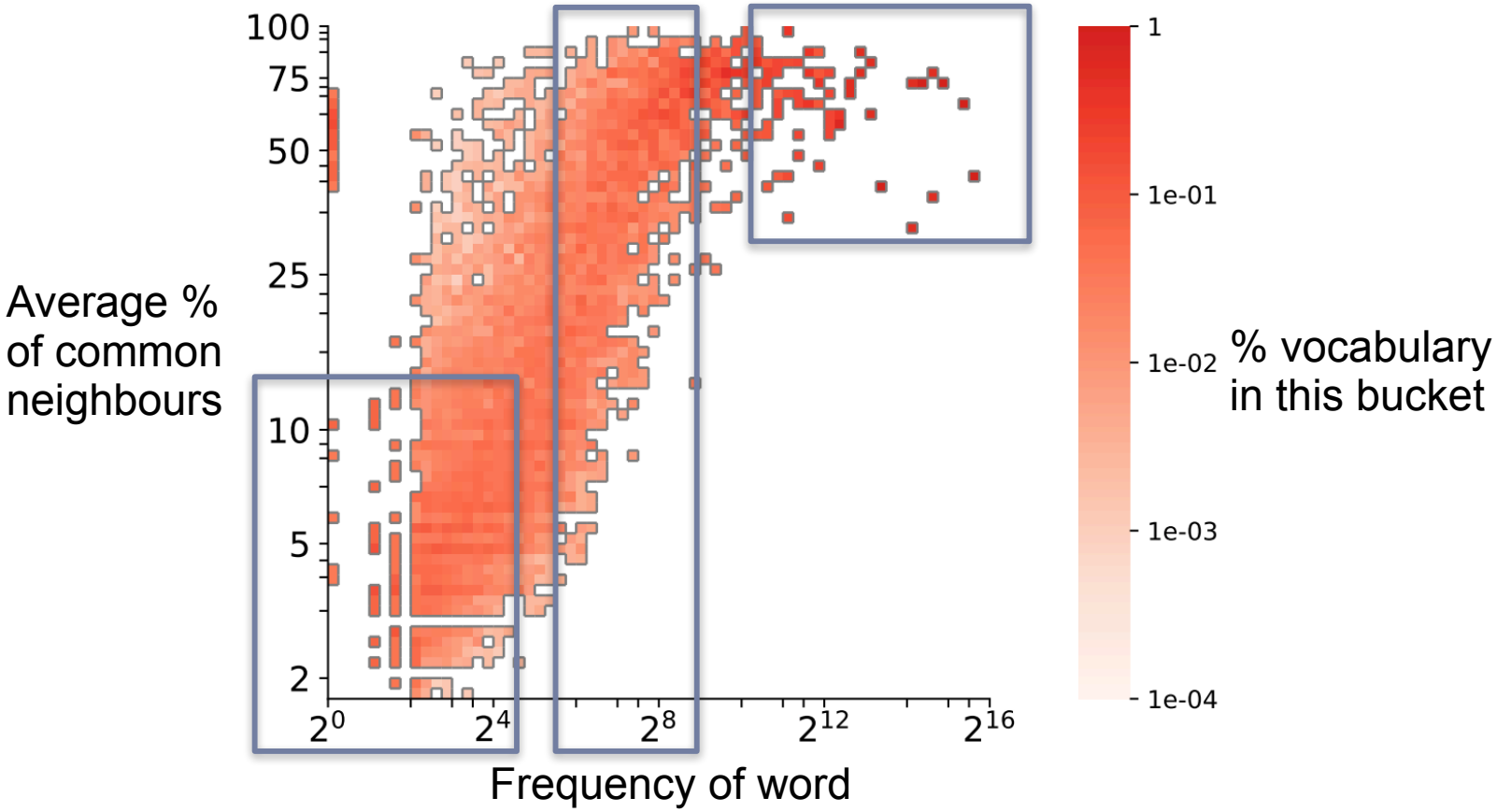
"Factors Influencing the Surprising Instability of Word Embeddings", Wendlandt, Kummerfeld, and Mihalcea (2018)

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

These representations can be quite variable



"Factors Influencing the Surprising Instability of Word Embeddings", Wendlandt, Kummerfeld, and Mihalcea (2018)

高 频词的 common neighbor 小
低 多

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

Workshop Preview

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

Pre-Workshop:

Unit policy quiz

Practise code submission

In Workshop:

Collaborative exploration of word embeddings

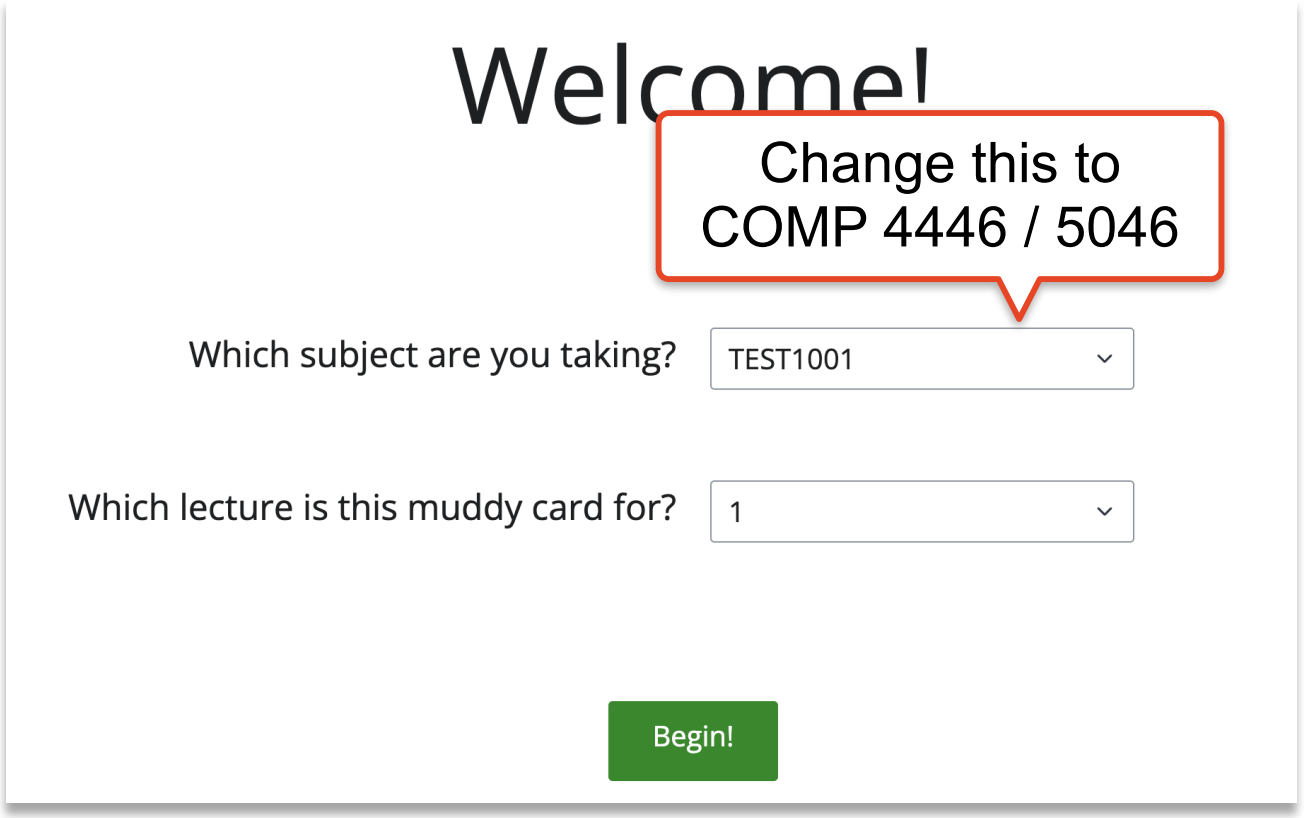
Group discussion of embedding bias found

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

Muddy Card



COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

Muddy Card

Enter your SID:

Ensure you enter this correctly, that is how you get credit!

uddy Card (TEST1001)

What was least clear to you in this lecture?

- Please write the ONE most confusing part of the lecture.
- Don't write anything other than what was confusing.
- Be specific.
- Bad examples: *"Photosynthesis"*, *"Merge sort"*
- Good examples: *"Why sunlight is needed in photosynthesis"*, *"How merge sort is considered a divide+conquer algorithm"*

I wrote this muddy card response because I...

- (•) do not understand this.
- ( ) think I understand this but want to check.
- ( ) would like to learn more about this.
- ( ) just wanted/needed to do the muddy cuddy.
- ( ) [some other reason].

Do you consent to your *class* **and** *lectu*

ecorded and used in a publicly available data set and subsequent research publication(s)?
Your SID number will NOT be stored ar
data set. If you have any questions about this resarch project,please read the student participant information sheet here.

- (•) Yes - I consent.
- ( ) No - I DO NOT consent

This choice will have no impact on your marks

Return

Submit

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

Muddy Card

Would an answer to any of the following also answer your question? You can choose **zero or more** options. When finished, press "submit" to continue.

- [ ] Why sunlight is needed in photosynthesis
- [ ] Why sunlight is needed in photosynthesis
- [ ] Why is sunlight needed in photosynthesis?
- [ ] What is sunlight needed in photosynthesis?

Submit

COMP 4446 / 5046
Lecture 1, 2025

Introduction
Representing Text
Evaluation
Workshop Preview

menti.com 4210 8267

Muddy Card

Open now, closes at 7:05pm

https://saipll.shinyapps.io/student-interface/

If you do not wish to participate in the study, use the Ed form instead

Go to Ed → Lessons → Muddy Cards Lecture 1