RICE SOCIAL SCIENCES
Department of Linguistics

vector semantics

LING 430 Computational Linguistics Fall 2026
lecture 11a
Nov 3 2026

Siyu Liang
siyu.liang@rice.edu

last time

- parsing
  - constituency
  - context-free grammars
  - CKY algorithm

agenda

- word meaning
- word vectors
- cosine similarity
- tf-idf

word meaning

word meaning

- what does *ongchoi* mean?
  - *Ongchoi is delicious sauteed with garlic.*
  - *Ongchoi is superb over rice.*
  - *…ongchoi leaves with salty sauces…*

word meaning

- suppose you have seen these before
  - *…spinach sauteed with garlic over rice…*
  - *…chard stems and leaves are delicious…*
  - *…collard greens and other salty leafy greens…*
- so *ongchoi* is probably a leafy green like spinach, chard and collard greens

word meaning

- *Ipomoea aquatica*, water spinach
  - 蕹菜 *ongchoi* in Cantonese
  - 空心菜 *kōngxīncài* in Mandarin
  - *kangkong* in Tagalog and Malay
  - *rau muống* in Vietnamese
- what did we do?
  - we didn’t look it up
  - we looked at the words around it to guess the meaning



Yamaguchi, Wikimedia Commons

word meaning

a model of meaning

- so far a word has been a string of letters, or an index in a vocabulary list
- what would we want a model of word meaning to tell us?
  - the most similar word to *Android*
  - *lemon* is to *orange* as *apple* is to …?
  - which one is different: *lemon, orange, apple, cucumber*
  - that *mouse* the rodent and *mouse* the device are two senses of one lemma

word meaning

similarity vs. relatedness

- *cat* is not a synonym of *dog*, but the two are **similar**
  - we can ask people how similar, on a scale from 0 to 10
- *coffee* and *cup* are not similar at all, but they are **related**

pair	rating
vanish, disappear	9.8
belief, impression	5.95
muscle, bone	3.65
hole, agreement	0.3

word meaning

- “the meaning of a word is its use in the language”
  - Ludwig Wittgenstein, *Philosophical Investigations* §43, 1953
- but how do we capture the use?

word meaning

the distributional hypothesis

- “you shall know a word by the company it keeps” - J. R. Firth, 1957
- two words with almost identical environments have almost identical meanings - Zellig Harris, 1954
- this is the **distributional hypothesis**
  - we just tested it with *ongchoi*
- the idea is from the 1950s
  - but corpora and computational systems are based on it

word meaning

- from now on: computing with meaning representations instead of string representations

荃者所以在魚，得魚而忘荃

nets are for fish;

once you get the fish, you can forget the net

言者所以在意，得意而忘言

words are for meaning;

once you get the meaning, you can forget the words

庄子 Zhuangzi, chapter 26

word vectors

word vectors

- we would like to compare word meanings
  - e.g. how different/similar are *cherry*, *strawberry*, *digital*, *information*
- for each word, look at the four words on either side of it
  - the context window: 8

context	word	context
is traditionally followed by	**cherry**	pie, a traditional dessert
often mixed, such as	**strawberry**	rhubarb pie. Apple pie
computer peripherals and personal	**digital**	assistants. These devices usually
a computer. This includes	**information**	available on the internet

word vectors

- how often do *a*, *computer* and *pie* occur in each window?
- we can create a **word-word matrix**
  - $|V|$ rows and $|V|$ columns
  - *cherry* is now a vector [1, 0, 1]

	a	***computer***	***pie***
cherry	1	0	1
strawberry	0	0	2
digital	0	1	0
information	1	1	0

word vectors

- we could expand this using the entire Wikipedia
  - picking six of the $|V|$ dimensions
  - *cherry* and *strawberry* share *pie* and *sugar*
  - *digital* and *information* share *computer* and *data*

	…	***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	…
…	…	…	…	…	…	…	…	…

word vectors

- a real row has tens of thousands of dimensions, but most of them are 0
  - the vector is **sparse**
  - we'll talk about how to improve it on Thursday

	…	*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	…
…	…	…	…	…	…	…	…	…

cosine similarity

cosine similarity

the dot product

- how similar are two vectors v and w?
  - we want one number
- multiply dimension by dimension and add up

$$v \cdot w = \sum_i v_i w_i = v_1 w_1 + v_2 w_2 + \cdots + v_N w_N$$

- high when both vectors have large values in the same dimensions
- 0 when they have their zeros in different places

cosine similarity

vector length

- a frequent word has more neighbors, so it gets a longer vector
  - and a bigger dot product with everything
- we want similarity regardless of size, so we divide the lengths out
- the length of a vector

$$|v| = \sqrt{\left(\sum_i {v_i}^2\right)}$$

cosine similarity

- the **dot product** divided by the two **lengths** is the cosine of the angle between the vectors

$$\text{cosine}(v, w) = \frac{v \cdot w}{|v||w|}$$

- 1 for vectors pointing the same way
- 0 for vectors at right angles
- −1 for opposite directions
- counts are never negative, so for us the cosine is between 0 and 1

cosine similarity

1. the dot product

- is *digital* closer to *information* or to *cherry*?

$$\text{digital} \cdot \text{information} = 5{\times}5 + 1683{\times}3982 + 1670{\times}3325 = 12254481$$

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

cosine similarity

2. the lengths

$$|\text{digital}| = \sqrt{(5^2 + 1683^2 + 1670^2)} = \sqrt{5621414} = 2370.95$$
$$|\text{information}| = \sqrt{(5^2 + 3982^2 + 3325^2)} = \sqrt{26911974} = 5187.68$$

- *information* is the more frequent word, so its vector is longer, and the dot product alone would reward that

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

cosine similarity

3. divide

$$\cos(\text{digital}, \text{information}) = \frac{12254481}{2370.95 \times 5187.68} = \frac{12254481}{12299728} = .996$$

- for *cherry*

  $\text{digital} \cdot \text{cherry} = 5 \times 442 + 1683 \times 8 + 1670 \times 2 = 19014$

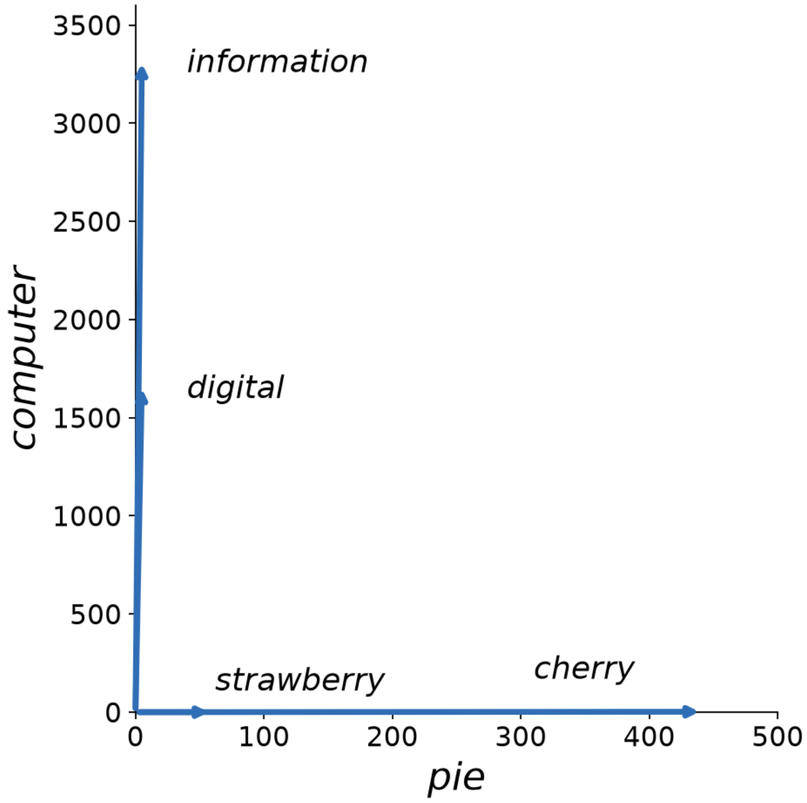
  $|\text{cherry}| = \sqrt{(442^2 + 8^2 + 2^2)} = \sqrt{195432} = 442.08$

  $$\cos(\text{digital}, \text{cherry}) = \frac{19014}{2370.95 \times 442.08} = .018$$

- *digital* is more similar to *information* than to *cherry*

cosine similarity

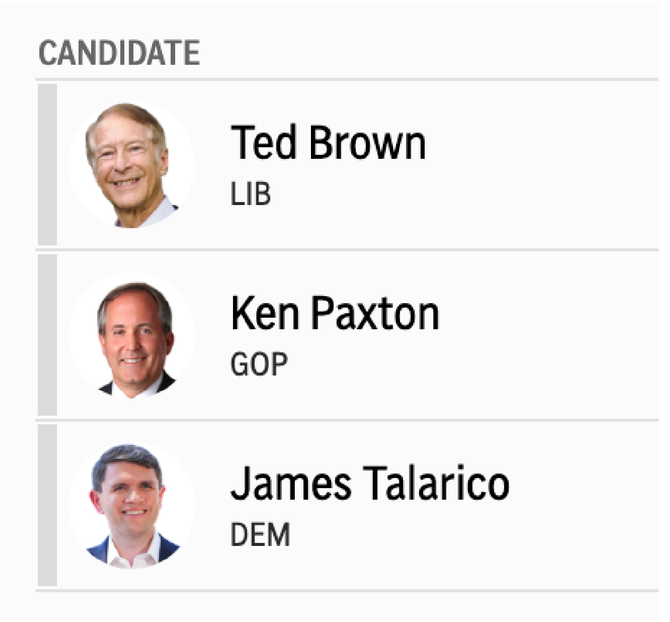
- we can visualize cosine similarity in planes
- picking 2 dimensions for a 2D plane
  - more dimensions can be used too
- a small angle means a large cosine
  - more similar



activity 1

- suppose you've never heard of Ted Brown, but you want to know whether its **distributional semantics** is closer to Ken Paxton or Jamaes Talarico
- you looked at all the texts these names occur next to
- use cosine similarity to calculate if Ted Brown closer to Ken Paxton or to James Talarico
- disclaimer: the counts are made up for class!

	taxes	***border***	***schools***
Ted Brown	6	1	2
Ken Paxton	4	7	1
James Talarico	1	1	7



activity 1

debrief

$$\cos(\text{Brown}, \text{Paxton}) = \frac{6\times4 + 1\times7 + 2\times1}{\sqrt{41}\times\sqrt{66}} = \frac{33}{52.02} = .634$$

$$\cos(\text{Brown}, \text{Talarico}) = \frac{6\times1 + 1\times1 + 2\times7}{\sqrt{41}\times\sqrt{51}} = \frac{21}{45.73} = .459$$

tf-idf

tf-idf

- we have built vectors for words; what about documents?
  - a document can be a vector too: its bag of words
  - the dimensions are the words, the values are their counts
- so the columns are now documents, and a row is a word's count in each document
  - a **term-document matrix**

tf-idf

- if we have bags of words for four Shakespeare plays
  - *As You Like It*, *Twelfth Night*, *Julius Caesar*, *Henry V*
- which is which depending on the word use?

	?	?	?	?
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

tf-idf

- common words like *good* don't say a lot
- but uncommon words do
  - the comedies have the fools and the wit, the other two the battles

	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

tf-idf

term frequency

- how often does the word occur in this document? count(t, d)
- but a word that appears 100 times is not 100 times as important, so we take the log

$$\mathrm{tf}_{t,d} = 1 + \log_{10}\mathrm{count}(t, d)$$

- and tf = 0 when the count is 0, since there is no log of 0
  - 1 occurrence: tf = 1
  - 10 occurrences: tf = 2
  - 100: tf = 3
  - …

tf-idf

document frequency

- how many documents does the word occur in? df_t
- a word found in a few documents helps us tell them apart; a word found everywhere does not
- so we weight by the **inverse document frequency** (Spärck Jones 1972), with N documents in the collection

$$\text{idf}_t = \log_{10} \frac{N}{\text{df}_t}$$

- the fewer documents a word is in, the higher its idf
- a word in every document gets 0

tf-idf

- what does idf look like over all 37 plays?
- *Romeo* is in one play: high idf
- *good* is in all 37: low idf

word	df	idf
Romeo	1	1.57
salad	2	1.27
Falstaff	4	0.967
battle	21	0.246
wit	34	0.037
fool	36	0.012
good	37	0

tf-idf

- **tf-idf** product of the term frequency and inverse document frequency

$$\text{tf-idf}(t, d) = \text{tf}_{t,d} \cdot \text{idf}_t$$

- e.g. for battle
  - *battle* occurs 13 times in *Henry V*

$$\text{tf} = 1 + \log_{10} 13 = 1 + 1.114 = 2.114$$

  - *battle* occurs in 21 of the 37 plays

$$\text{idf} = \log_{10} \frac{37}{21} = 0.246$$

  - multiply

$$\text{tf-idf} = 2.114 \times 0.246 = 0.520$$

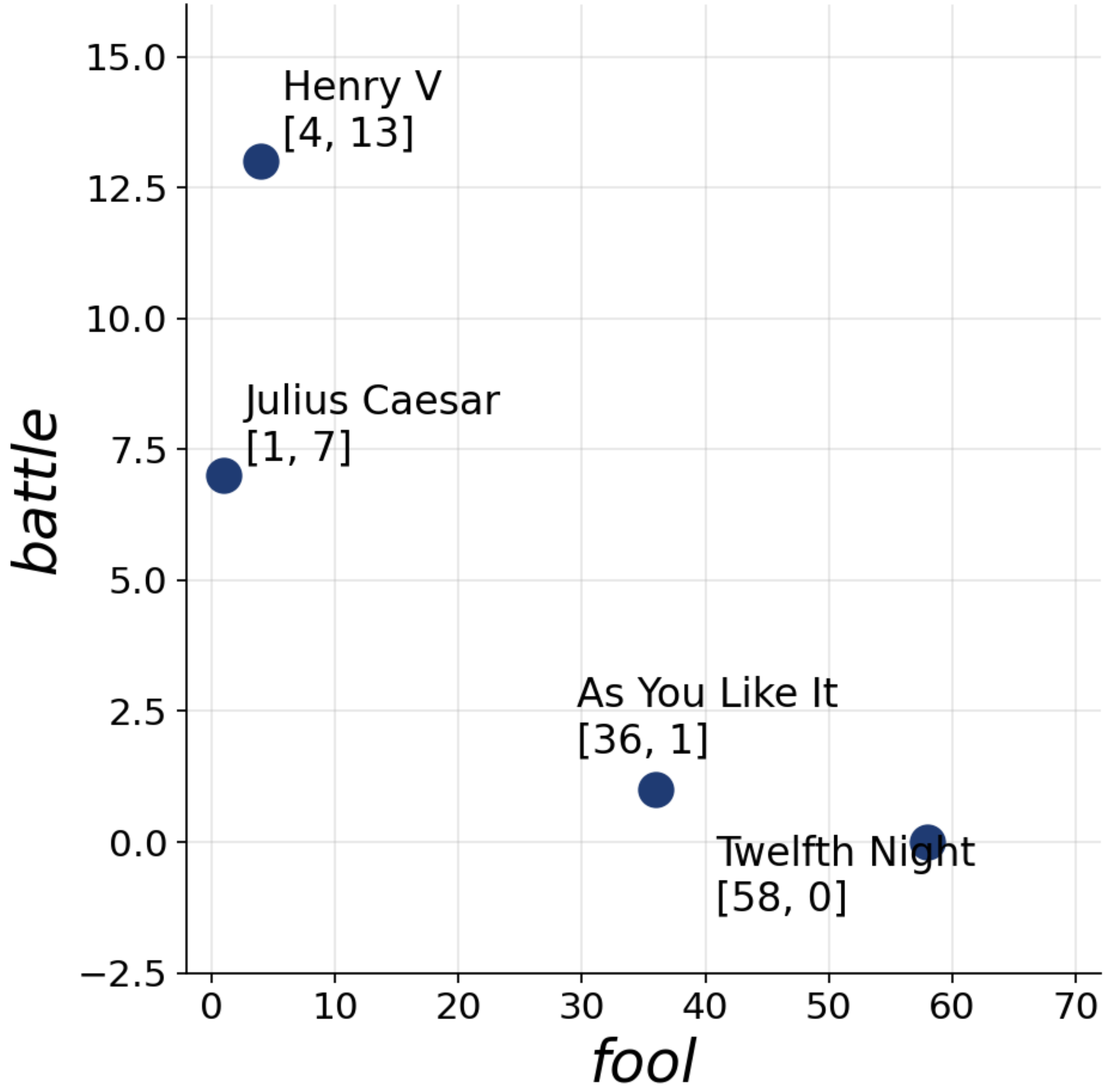
tf-idf

- plugging tf-idf in the table

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	0.246	0	0.454	0.520
good	0	0	0	0
fool	0.030	0.033	0.012	0.019
wit	0.085	0.080	0.048	0.054

tf-idf

- similar vector representation in 2D



activity 2

- which play is this line from?
  - “better a witty fool than a foolish wit”
- calculate the line’s tf-idf vector and cosine with the play you think it’s from and one other
  - the line’s vector → the line’s tf-idf → the cosine similarity with a play’s tf-idf

	As You Like It	**Twelfth Night**	**Julius Caesar**	**Henry V**
battle	0.246	0	0.454	0.520
good	0	0	0	0
fool	0.030	0.033	0.012	0.019
wit	0.085	0.080	0.048	0.054

next time

- Thursday, Nov 5
  - word embeddings
- HW 6 to be released Thursday