Classic Word Representations & Vector Space Basics

Yu Meng
University of Virginia
yumeng5@virginia.edu

Sep 11, 2024

Join at
slido.com
#2339 801

Announcement

- Assignment 1 is due tonight 11:59pm!
- Assignment 2 is released (due 09/25 11:59pm)
- No instructor office hour today

Overview of Course Contents

- Week 1: Logistics & Overview
- Week 2: N-gram Language Models
- Week 3: Word Senses, Semantics & Classic Word Representations
- Week 4: Word Embeddings
- Week 5: Sequence Modeling and Transformers
- Week 6-7: Language Modeling with Transformers (Pretraining + Fine-tuning)
- Week 8: Large Language Models (LLMs) & In-context Learning
- Week 9-10: Knowledge in LLMs and Retrieval-Augmented Generation (RAG)
- Week 11: LLM Alignment
- Week 12: Language Agents
- Week 13: Recap + Future of NLP
- Week 15 (after Thanksgiving): Project Presentations

(Recap) Why Care About Word Semantics?

- Understanding word meanings helps us build better language models!
- Recall the example from N-gram lectures:

[BOS] The cat is on the mat [EOS]
[BOS] I have a cat and a mat [EOS]
[BOS] I like the cat [EOS]

$$p(\text{“cat”} \mid \text{“the”}) = \frac{2}{3}, \quad p(\text{“mat”} \mid \text{“the”}) = \frac{1}{3},$$

- Sparsity: many valid bigram counts are zero – count-based measures do not account for word semantics!
- If we know “cat” is semantically similar to “dog”, then $p(\text{“dog”} \mid \text{“the”}) \approx p(\text{“cat”} \mid \text{“the”})$

(Recap) Word Semantics & Relations in NLP

- **Synonyms**: words with similar meanings
  - "happy" & "joyful"
- **Antonyms**: words with opposite meanings
  - "hot" & "cold"
- **Hyponyms** & **hypernyms**: one word is a more specific instance of another
  - "rose" is a hyponym of "flower"
  - "flower" is a hypernym of "rose"
- **Polysemy**: A single word having multiple related meanings
  - "mouse" can mean small rodents or the device that controls a cursor
- **Lemma**: the base or canonical form of a word, from which other forms can be derived
- The study of these aspects of word meanings is called **lexical semantics** in linguistics

(Recap) Polysemy & Senses

Join at
slido.com
#2339 801

- **Polysemy**: a single word has multiple related meanings
  - "**Light**": "This bag is **light**" / "Turn on the **light**" / "She made a **light** comment"
- **Sense**: a particular meaning or interpretation of a word in a given context
- Word relations (e.g., synonyms, antonyms, hypernyms/hyponyms) are defined between word senses!
- **Word sense disambiguation (WSD):** determine which sense of a word is being used in a specific context
  - She went to the **bank** to deposit money
  - She lives by the river **bank**
- WSD can be challenging especially when the context is short/insufficient
  - Is the query "mouse info" looking for a pet or a tool?

(Recap) Word Similarity

- Most words may not have many perfect synonyms, but usually have lots of similar words
  - "cat" is not a synonym of "dog", but they are similar in meaning

vanish	disappear	9.8
belief	impression	5.95
muscle	bone	3.65
modest	flexible	0.98
hole	agreement	0.3

Word similarity (on a scale from 0 to 10) manually annotated by humans

- We'll introduce word embeddings to automatically learn word similarity next week!

Figure source: https://web.stanford.edu/~jurafsky/slp3/6.pdf

UNIVERSITY of VIRGINIA

Join at slido.com #2339 801

(Recap) Connotation

- Valence: the pleasantness of the stimulus
  - High: “happy” / “satisfied”; low: “unhappy” / “annoyed”
- Arousal: the intensity of emotion provoked by the stimulus
  - High: “excited”; low: “calm”
- Dominance: the degree of control exerted by the stimulus
  - High: “controlling”; low: “influenced”

	Valence	Arousal	Dominance
courageous	8.05	5.5	7.38
music	7.67	5.57	6.5
heartbreak	2.45	5.65	3.58
cub	6.71	3.95	4.24

Earliest work on representing words with multi-dimensional vectors!

Figure source: https://web.stanford.edu/~jurafsky/slp3/6.pdf

Agenda

- Introduction to Word Senses & Semantics
- Classic Word Representations
- Vector Space Model Basics

WordNet

- Word semantics is complex (multiple senses, various relations)!
- How did people represent word senses and relations in early NLP developments?
- **WordNet**: A manually curated large lexical database
- Three separate databases: one each for nouns, verbs and adjectives/adverbs
- Each database contains a set of lemmas, each one annotated with a set of senses
- Synset (synonym set): The set of near-synonyms for a sense
- Word relations (hypernym, hyponym, antonym) defined between synsets

WordNet: https://wordnet.princeton.edu/

WordNet Relations

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	$destruction^1 \Longleftrightarrow destroy^1$

Noun relations

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	$increase^1 \Longleftrightarrow decrease^1$

Verb relations

Figure source: https://web.stanford.edu/~jurafsky/slp3/G.pdf

UNIVERSITY of VIRGINIA

Join at
slido.com
#2339 801

WordNet as a Graph

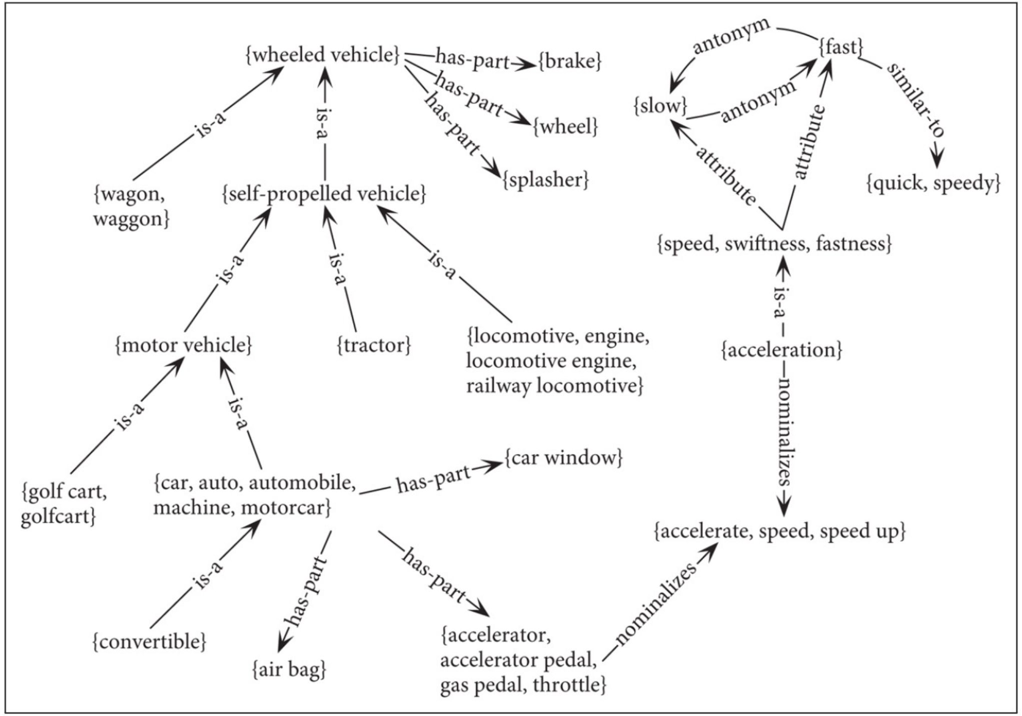


Figure source: https://academic.oup.com/edited-volume/42643/chapter/358151233

UNIVERSITY of VIRGINIA

Join at
slido.com
#2339 801

WordNet Demo

Category	Unique Strings
Noun	117798
Verb	11529
Adjective	22479
Adverb	4481

Figure source: https://lm-class.org/lectures/04%20-%20word%20embeddings.pdf

Word to search for: light [Search WordNet]

Display Options: (Select option to change) [Change]

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

Noun

- S: (n) **light**, visible light, visible radiation ((physics) electromagnetic radiation that can produce a visual sensation) *"the light was filtered through a soft glass window"*
  - *direct hyponym* / *full hyponym*
  - *domain category*
  - *direct hypernym* / *inherited hypernym* / *sister term*
  - *part holonym*
  - *derivationally related form*
- S: (n) **light**, light source (any device serving as a source of illumination) *"he stopped the car and turned off the lights"*
- S: (n) **light** (a particular perspective or aspect of a situation) *"although he saw it in a different light, he still did not understand"*
- S: (n) luminosity, brightness, brightness level, luminance, luminousness, **light** (the quality of being luminous; emitting or reflecting light) *"its luminosity is measured relative to that of our sun"*
- S: (n) **light** (an illuminated area) *"he stepped into the light"*
  - *direct hypernym* / *inherited hypernym* / *sister term*
  - *derivationally related form*
- S: (n) **light**, illumination (a condition of spiritual awareness; divine illumination) *"follow God's light"*
- S: (n) **light**, lightness (the visual effect of illumination on objects or scenes as created in pictures) *"he could paint the lightest light and the darkest dark"*
- S: (n) **light** (a person regarded very fondly) *"the light of my life"*
- S: (n) **light**, lighting (having abundant light or illumination) *"they played as long as it was light"; "as long as the lighting was good"*
- S: (n) **light** (mental understanding as an enlightening experience) *"he finally saw the light"; "can you shed light on this problem?"*
- S: (n) sparkle, twinkle, spark, **light** (merriment expressed by a brightness or gleam or animation of countenance) *"he had a sparkle in his eye"; "there's a perpetual twinkle in his eyes"*
- S: (n) **light** (public awareness) *"it brought the scandal to light"*
- S: (n) Inner Light, **Light**, Light Within, Christ Within (a divine presence

WordNet web browser: http://wordnetweb.princeton.edu/perl/webwn

WordNet for Word Sense Disambiguation

- All words WSD task: map all input words (nouns/verbs/adjectives/adverbs) to WordNet senses
- Strong baseline: map to the first sense in WordNet (most frequent)
- Modern approaches: sequence modeling architectures (later lectures!)

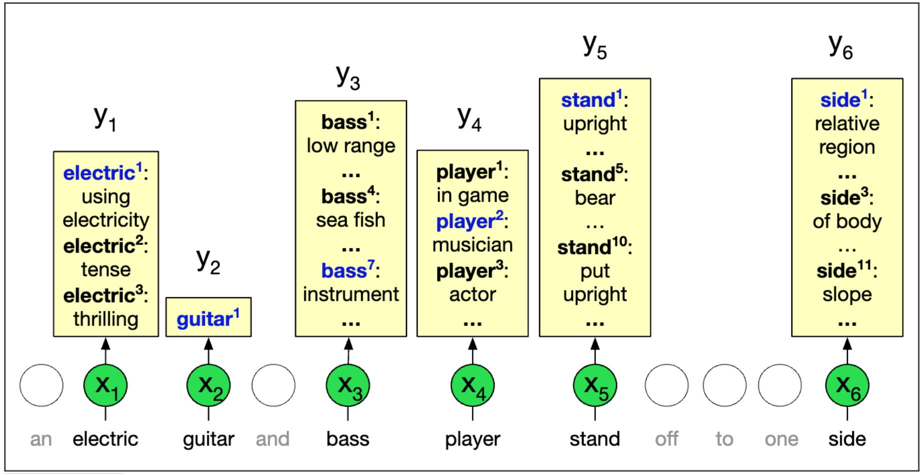


Figure source: https://web.stanford.edu/~jurafsky/slp3/G.pdf

WordNet Limitations

- Require significant efforts to construct and maintain/update
  - Hard to keep up with rapidly evolving language usage
- Limited coverage of domain-specific terms & low-resource language
  - No coverage of specialized, domain-specific terms (e.g., medical, legal, or technical)
- Only support individual words and their meanings
  - Do not account for idiomatic expressions, phrasal verbs, or collocations

A more automatic, scalable, and contextualized word semantic learning approach is needed!

Agenda

- Introduction to Word Senses & Semantics
- Classic Word Representations
- Vector Space Model Basics

UNIVERSITY of VIRGINIA

Join at
slido.com
#2339 801

Motivation: Representing Texts with Vectors

- Word similarity computation is important for understanding semantics

Word similarity (on a scale from 0 to 10) manually annotated by humans

vanish	disappear	9.8
belief	impression	5.95
muscle	bone	3.65
modest	flexible	0.98
hole	agreement	0.3

Word semantics can be multi-faceted

	Valence	Arousal	Dominance
courageous	8.05	5.5	7.38
music	7.67	5.57	6.5
heartbreak	2.45	5.65	3.58
cub	6.71	3.95	4.24

- How to represent words numerically? Using multi-dimensional vectors!

Figure source: https://web.stanford.edu/~jurafsky/slp3/6.pdf

Vector Semantics

- Represent a word as a point in a multi-dimensional semantic space
- A desirable vector semantic space: words with similar meanings are nearby in space

2D visualization of a desirable high-dimensional vector semantic space

Figure source: https://web.stanford.edu/~jurafsky/slp3/6.pdf

Vector Space Basics

- Vector notation: an N-dimensional vector $\boldsymbol{v} = [v_1, v_2, \ldots, v_N] \in \mathbb{R}^N$
- Vector dot product/inner product:

$$\text{dot product}(\boldsymbol{v}, \boldsymbol{w}) = \boldsymbol{v} \cdot \boldsymbol{w} = v_1 w_1 + v_2 w_2 + \cdots + v_n w_n = \sum_{i=1}^{N} v_i w_i$$

- Vector length/norm:

$$|\boldsymbol{v}| = \sqrt{\boldsymbol{v} \cdot \boldsymbol{v}} = \sqrt{\sum_{i=1}^{N} v_i^2}$$

Other (less commonly-used) vector norms: Manhattan norm, *p*-norm, infinity norm…

- Cosine similarity between vectors:

$$\cos(\boldsymbol{v}, \boldsymbol{w}) = \frac{\boldsymbol{v} \cdot \boldsymbol{w}}{|\boldsymbol{v}||\boldsymbol{w}|} = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2}\sqrt{\sum_{i=1}^{N} w_i^2}}$$

Vector Space Basics: Example

- Consider two 4-dimensional vectors $\boldsymbol{v} = [1, 0, 1, 0] \in \mathbb{R}^4 \quad \boldsymbol{w} = [0, 1, 1, 0] \in \mathbb{R}^4$
- Vector dot product/inner product:

$$\boldsymbol{v} \cdot \boldsymbol{w} = \sum_{i=1}^{N} v_i w_i = 1$$

- Vector length/norm:

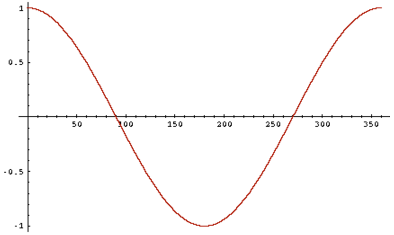
$$|\boldsymbol{v}| = \sqrt{\sum_{i=1}^{N} v_i^2} = \sqrt{2} \quad |\boldsymbol{w}| = \sqrt{\sum_{i=1}^{N} w_i^2} = \sqrt{2}$$

- Cosine similarity between vectors:

$$\cos(\boldsymbol{v}, \boldsymbol{w}) = \frac{\boldsymbol{v} \cdot \boldsymbol{w}}{|\boldsymbol{v}||\boldsymbol{w}|} = \frac{1}{2}$$

Vector Similarity

- Cosine similarity is the most commonly used metric for similarity measurement
  - Symmetric: $\cos(\boldsymbol{v}, \boldsymbol{w}) = \cos(\boldsymbol{w}, \boldsymbol{v})$
  - Not influenced by vector length
  - Has a normalized range: [-1, 1]
  - Intuitive geometric interpretation



Cosine function values under different angles

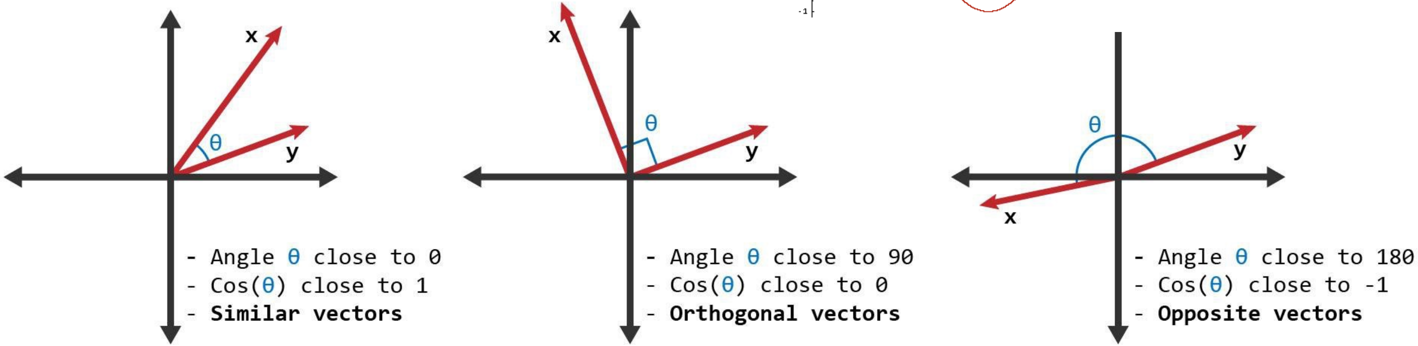


Figure source: https://www.learndatasci.com/glossary/cosine-similarity/

How to Represent Words as Vectors?

- Given a vocabulary $\mathcal{V} = \{\text{good}, \text{feel}, \text{I}, \text{sad}, \text{cats}, \text{have}\}$
- Most straightforward way to represent words as vectors: use their indices
- One-hot vector: only one high value (1) and the remaining values are low (0)
- Each word is identified by a unique dimension

$$
\begin{aligned}
\boldsymbol{v}_{\text{good}} &= [1, 0, 0, 0, 0, 0] \\
\boldsymbol{v}_{\text{feel}} &= [0, 1, 0, 0, 0, 0] \\
\boldsymbol{v}_{\text{I}} &= [0, 0, 1, 0, 0, 0] \\
\boldsymbol{v}_{\text{sad}} &= [0, 0, 0, 1, 0, 0] \\
\boldsymbol{v}_{\text{cats}} &= [0, 0, 0, 0, 1, 0] \\
\boldsymbol{v}_{\text{have}} &= [0, 0, 0, 0, 0, 1]
\end{aligned}
$$

Represent Sequences by Word Occurrences

- Consider the mini-corpus with three documents

$$d_1 = \text{“I feel good”}$$
$$d_2 = \text{“I feel sad”}$$
$$d_3 = \text{“I have cats”}$$

$$\boldsymbol{v}_{\text{good}} = [1, 0, 0, 0, 0, 0]$$
$$\boldsymbol{v}_{\text{feel}} = [0, 1, 0, 0, 0, 0]$$
$$\boldsymbol{v}_{\text{I}} = [0, 0, 1, 0, 0, 0]$$
$$\boldsymbol{v}_{\text{sad}} = [0, 0, 0, 1, 0, 0]$$
$$\boldsymbol{v}_{\text{cats}} = [0, 0, 0, 0, 1, 0]$$
$$\boldsymbol{v}_{\text{have}} = [0, 0, 0, 0, 0, 1]$$

- Straightforward way of representing documents: look at which words are present

$$\boldsymbol{v}_{d_1} = [1, 1, 1, 0, 0, 0]$$
$$\boldsymbol{v}_{d_2} = [0, 1, 1, 1, 0, 0]$$
$$\boldsymbol{v}_{d_3} = [0, 0, 1, 0, 1, 1]$$

Document vector similarity

$$\cos(\boldsymbol{v}_{d_1}, \boldsymbol{v}_{d_2}) = \frac{2}{3}$$
$$\cos(\boldsymbol{v}_{d_1}, \boldsymbol{v}_{d_3}) = \frac{1}{3}$$
$$\cos(\boldsymbol{v}_{d_2}, \boldsymbol{v}_{d_3}) = \frac{1}{3}$$

Term-Document Matrix

- With larger text collections, word frequencies in documents entail rich information
- Consider the four plays by Shakespeare and obtain the word frequency statistics
- Look at 4 manually-picked words: “battle” “good” “fool” “wit”

	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

There are many more words!

- Document vector representation with word frequencies:

$$\boldsymbol{v}_{d_1} = [1, 114, 36, 20] \quad \boldsymbol{v}_{d_2} = [0, 80, 58, 15] \quad \boldsymbol{v}_{d_3} = [7, 62, 1, 2] \quad \boldsymbol{v}_{d_4} = [13, 89, 4, 3]$$

Figure source: https://web.stanford.edu/~jurafsky/slp3/6.pdf

Document Similarity

- Document vector representation with word frequencies:

$$\boldsymbol{v}_{d_1} = [1, 114, \boxed{36, 20}] \quad \boldsymbol{v}_{d_2} = [0, 80, \boxed{58, 15}] \quad \boldsymbol{v}_{d_3} = [7, 62, 1, 2] \quad \boldsymbol{v}_{d_4} = [13, 89, 4, 3]$$

	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

- "fool" and "wit" occur much more frequently in d_1 and d_2 than d_3 and d_4
- d_1 and d_2 are comedies $\cos(\boldsymbol{v}_{d_1}, \boldsymbol{v}_{d_2}) = 0.95 \quad \cos(\boldsymbol{v}_{d_2}, \boldsymbol{v}_{d_3}) = 0.81$
- Word frequencies in documents do reflect the semantic similarity between documents!

Figure source: https://web.stanford.edu/~jurafsky/slp3/6.pdf

Words Represented with Documents

- "Battle": "the kind of word that occurs in Julius Caesar and Henry V (history plays)"
- "Fool": "the kind of word that occurs in comedies"

	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

- Represent words using their co-occurrence counts with documents:

$$\boldsymbol{v}_{\text{battle}} = [1, 0, 7, 13]$$
$$\boldsymbol{v}_{\text{good}} = [114, 80, 62, 89]$$
$$\boldsymbol{v}_{\text{fool}} = [36, 58, 1, 4]$$
$$\boldsymbol{v}_{\text{wit}} = [20, 15, 2, 3]$$

UNIVERSITY *of* VIRGINIA

Join at
slido.com
#2339 801

Words Represented with Documents

	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

$$\boldsymbol{v}_{\text{battle}} = [1, 0, 7, 13]$$
$$\boldsymbol{v}_{\text{good}} = [114, 80, 62, 89]$$
$$\boldsymbol{v}_{\text{fool}} = [36, 58, 1, 4]$$
$$\boldsymbol{v}_{\text{wit}} = [20, 15, 2, 3]$$

$$\cos(\boldsymbol{v}_{\text{fool}}, \boldsymbol{v}_{\text{wit}}) = 0.93$$
$$\cos(\boldsymbol{v}_{\text{fool}}, \boldsymbol{v}_{\text{battle}}) = 0.09$$

Previously:

$$\boldsymbol{v}_{\text{battle}} = [1, 0, 0, 0]$$
$$\boldsymbol{v}_{\text{good}} = [0, 1, 0, 0]$$
$$\boldsymbol{v}_{\text{fool}} = [0, 0, 1, 0]$$
$$\boldsymbol{v}_{\text{wit}} = [0, 0, 0, 1]$$

$$\cos(\boldsymbol{v}_{\text{fool}}, \boldsymbol{v}_{\text{wit}}) = 0$$
$$\cos(\boldsymbol{v}_{\text{fool}}, \boldsymbol{v}_{\text{battle}}) = 0$$

Document co-occurrence statistics provide coarse-grained contexts

Fine-Grained Contexts: Word-Word Matrix

Instead of using documents as contexts for words, we can also use words as contexts

4 words to the left	center word	4 words to the right
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

Figure source: https://web.stanford.edu/~jurafsky/slp3/6.pdf

Fine-Grained Contexts: Word-Word Matrix

Count how many times words occur in a ±4 word window around the center word

context word

center word

	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	...

Counts derived from the Wikipedia corpus

Figure source: https://web.stanford.edu/~jurafsky/slp3/6.pdf

Word Similarity Based on Word Co-occurrence

- Word-word matrix with ±4 word window

	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	...

- “digital” and “information” both co-occur with “computer” and “data” frequently
- “cherry” and “strawberry” both co-occur with “pie” and “sugar” frequently
- Word co-occurrence statistics reflect word semantic similarity!
- Issues? Sparsity!

Is Raw Frequency A Good Representation?

- On the one hand, high frequency can imply semantic similarity
- On the other hand, there are words with universally high frequencies

	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

- Can we reweight the raw frequencies so that distinctively high frequency terms are highlighted?

Term Frequency (TF)

- A word appearing 100 times in a document doesn't make it 100 times more likely to be relevant to the meaning of the document
- Instead of using the raw counts, we squash the counts with log scale

$$\mathrm{TF}(w,d) = \begin{cases} 1 + \log_{10} \mathrm{count}(w,d) & \mathrm{count}(w,d) > 0 \\ 0 & \text{otherwise} \end{cases}$$

Document Frequency (DF)

- Motivation: Give a higher weight to words that occur only in a few documents
  - Terms that are limited to a few documents are more discriminative
  - Terms that occur frequently across the entire collection aren't as helpful
- Document frequency (DF): count how many documents a word occurs in

$$\mathrm{DF}(w) = \sum_{i=1}^{N} \mathbb{1}(w \in d_i)$$

Evaluates to 1 if w occurs in d_i
otherwise evaluates to 0

- DF is NOT defined to be the total count of a word across all documents (collection frequency)!

	Collection Frequency	Document Frequency
Romeo	113	1
action	113	31

Figure source: https://web.stanford.edu/~jurafsky/slp3/6.pdf

Inverse Document Frequency (IDF)

- We want to emphasize discriminative words (with low DF)
- Inverse document frequency (IDF): total number of documents (N) divided by DF, in log scale

$$\mathrm{IDF}(w) = \log_{10}\left(\frac{N}{\mathrm{DF}(w)}\right)$$

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

DF & IDF statistics in the Shakespeare corpus

Figure source: https://web.stanford.edu/~jurafsky/slp3/6.pdf

TF-IDF Weighting

The TF-IDF weighted value characterizes the "salience" of a term in a document

$$\text{TF-IDF}(w, d) = \text{TF}(w, d) \times \text{IDF}(w)$$

TF-IDF weighted

	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	0.0012	0.0019
wit	0.085	0.081	0.048	0.054

$$\cos(\boldsymbol{v}_{d_2}, \boldsymbol{v}_{d_3}) = 0.10 \quad \cos(\boldsymbol{v}_{d_3}, \boldsymbol{v}_{d_4}) = 0.99$$

Raw counts

	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

$$\cos(\boldsymbol{v}_{d_2}, \boldsymbol{v}_{d_3}) = 0.81 \quad \cos(\boldsymbol{v}_{d_3}, \boldsymbol{v}_{d_4}) = 0.99$$

How to Define Documents?

- The concrete definition of documents is usually open to different design choices
  - Wikipedia article/page
  - Shakespeare play
  - Book chapter/section
  - Paragraph/sentence
  - …
- Larger documents provide broader context; smaller ones provide focused insights
- Depends on the analysis need: interested in global trends across documents (e.g., news articles) vs. more local patterns (e.g., specific sections of a legal document)?

Thank You!

Yu Meng
University of Virginia
yumeng5@virginia.edu