CS 333: Natural Language Processing

Fall 2025

Prof. Carolyn Anderson
Wellesley College

Reminders

- Extension for HW 4: **due Monday at 10pm**
- No quiz next week
- Midterm 1: Oct. 10th
- My next help hours: Thursday 4-5

New NLP stickers in my office — stop by! :)

Midterm 1

- In-class programming midterm
- Bring your own laptop to work on
- I will have starter code and documentation for you to download at the beginning, then ask you to turn off wifi.
- At the end, you will submit your code on Gradescope.
- May bring a 4x6in note card

Avoiding Harms in Classification

Harms in Sentiment Analysis

Kiritchenko and Mohammad (2018) found that most sentiment classifiers assign lower sentiment and more negative emotion to sentences with African American names in them.

This perpetuates negative stereotypes that associate African Americans with negative emotions

Harms in Toxicity Detection

Toxicity detection is the task of detecting hate speech, abuse, harassment, or other kinds of toxic language

But some toxicity classifiers incorrectly flag as being toxic sentences that are non-toxic but simply mention identities like blind people, women, or gay people.

This could lead to censorship of discussion about these groups.

Harms in Classification

Can be caused by:

- Problems in the training data; machine learning systems are known to amplify the biases in their training data.
- Problems in the human labels
- Problems in the resources used (like lexicons)
- Problems in model architecture (like what the model is trained to optimized)

Mitigation of these harms is an open research area

Word Meaning:
How is a raven like a writing desk?

What do words mean?

In the n-gram or text classification methods we've seen so far:

- Words are just strings (or indices w_i in a vocabulary list)
- That's not very satisfactory!

Desiderata

What should a theory of word meaning do for us?

HTTPS://WWW.YOUTUBE.COM/WATCH?V=NGQTMNSOV40&T=1333S

HTTPS://CONNECTING-WALL.NETLIFY.APP/

Relations between senses: Synonymy

Synonyms have the same meaning in some or all contexts.

- filbert / hazelnut
- couch / sofa
- big / large
- automobile / car
- vomit / throw up
- water / H_20

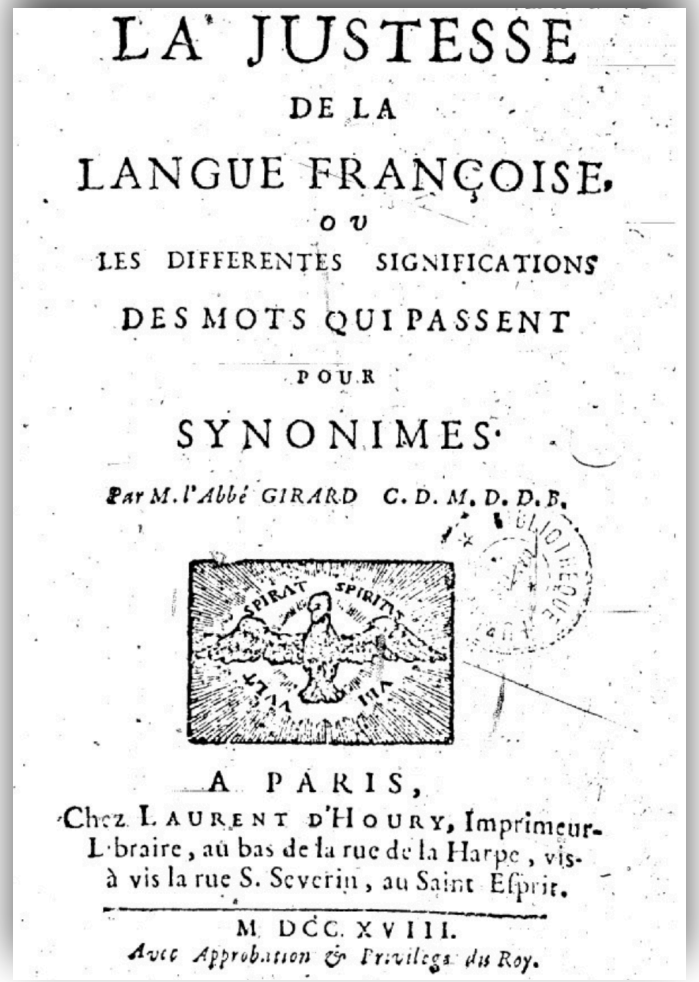
The Linguistic Principle of Contrast:

Difference in form → difference in meaning

Abbé Gabriel Girard (1718):

"je ne crois pas qu'il y ait de mot ſynonime dans aucune Langue. Je le dis par con-

[I do not believe that there is a synonymous word in any language]


LA JUSTESSE
DE LA
LANGUE FRANÇOISE,
OU
LES DIFFERENTES SIGNIFICATIONS
DES MOTS QUI PASSENT
POUR
SYNONIMES.
Par M. l'Abbé GIRARD C. D. M. D. D. B.
A PARIS,
Chez LAURENT D'HOURY, Imprimeur-Libraire, au bas de la rue de la Harpe, vis-à-vis la rue S. Severin, au Saint Eſprit.
M. DCC. XVIII.
Avec Approbation & Privilege du Roy.

Relation: Synonymy?

water/H20

"H20" in a surfing guide?

big/large

my big sister != my large sister

Relation: Similarity

Words with similar meanings. Not synonyms, but sharing some element of meaning:

car, bicycle

cow, horse

Relation: Word relatedness

Also called "word association"

Words can be related in any way, perhaps via a semantic frame or field

- coffee, tea: **similar**
- movie, popcorn: **related**, not similar

Ask humans how similar 2 words are

word1	word2	similarity (Class)	Dataset
vanish	disappear	9	9.8
behave	obey	7	7.3
belief	impression	4.5	5.99
muscle	bone	5	3.65
modest	flexible	2	0.98
hole	agreement	1	0.3

Simlex-999 dataset

Ask humans how similar 2 words are

word1	word2	similarity
vanish	disappear	9.8
behave	obey	7.3
belief	impression	5.95
muscle	bone	3.65
modest	flexible	0.98
hole	agreement	0.3

SimLex-999 dataset (Hill et al., 2015)

Desiderata

Concepts or word senses

- Have a complex many-to-many association with **words** (homonymy, multiple senses)

Have relations with each other

- Synonymy
- Antonymy
- Similarity
- Relatedness
- Connotation

Vector Semantics

Computational models of word meaning

Can we build representations of word meanings?

Most common approach: **vector semantics**

[1, 5, 6, 7, 0, -1, 3]

Key Insight # 1

"The meaning of a word is its use in the language" — Wittgenstein

Key Insight # 1

"The meaning of a word is its use in the language" — Wittgenstein

If A and B have almost identical environments we say that they are synonyms.

— Zellig Harris (1954)

What does recent English borrowing *ongchoi* mean?

Suppose you see these sentences:

- *Ong choi is delicious* ***sautéed with garlic****.*
- *Ong choi is superb* ***over rice***
- *Ong choi* ***leaves*** *with salty sauces*

And you've also seen these:

- *…spinach* ***sautéed with garlic over rice***
- *Chard stems and* ***leaves*** *are* ***delicious***
- *Collard greens and other* ***salty*** *leafy greens*

What does recent English borrowing *ongchoi* mean?

Suppose you see these sentences:

- *Ong choi is delicious* ***sautéed with garlic****.*
- *Ong choi is superb* ***over rice***
- *Ong choi* ***leaves*** *with salty sauces*

And you've also seen these:

- *…spinach* ***sautéed with garlic over rice***
- *Chard stems and* ***leaves*** *are* ***delicious***
- *Collard greens and other* ***salty*** *leafy greens*

Conclusion:

- Ongchoi is a leafy green like spinach, chard, or collard greens
- We could conclude this based on words like "leaves" and "delicious" and "sauteed"

Ongchoi: *Ipomoea aquatica "Water Spinach"*

空心菜
kangkong
rau muống
…

Yamaguchi, Wikimedia Commons, public domain

Key Insight #1: Defining meaning by linguistic distribution

Let's define the meaning of a word by its distribution in language use, meaning its **neighboring words** or grammatical environments.

Key Insight #2: Meaning as a point in multidimensional space

Each word is represented by a vector (not just "good" or "w45").

Similar words are "nearby in semantic space"

We build this space by seeing which words are nearby in text

How to represent word meaning numerically?

Idea: represent each word using a vector.

These vectors are called "embeddings" because they are **embedded** into a space.

Every modern NLP algorithm uses embeddings as the representation of word meaning.

Word vectors provide a fine-grained model of meaning for similarity.

Why vectors?

Consider sentiment analysis:

With **words**, a feature is a word identity

- Feature 5: 'The previous word was "terrible"'
- requires **exact same word** to be in training and test

With **embeddings**:

- Feature is a word vector
- 'The previous word was vector [35,22,17…]
- Now in the test set we might see a similar vector [34,21,14]
- We can generalize to **similar but unseen** words!!!

Vector Representations

We'll discuss 2 kinds of embeddings:

tf-idf

- Information Retrieval workhorse!
- A common baseline model
- **Sparse** vectors
- Words are represented by (a simple function of) the **counts** of nearby words

Word2vec

- **Dense** vectors
- Representation is created by training a classifier to **predict** whether a word is likely to appear nearby
- Later we'll discuss extensions called **contextual embeddings**

Words and Vectors

Term-document matrix

Each document is represented by a vector of words:

	Emma	Persuasion	Sense & Sensibility
admiral	0	69	0
dance	49	11	21
admire	31	14	18
horse	40	15	24

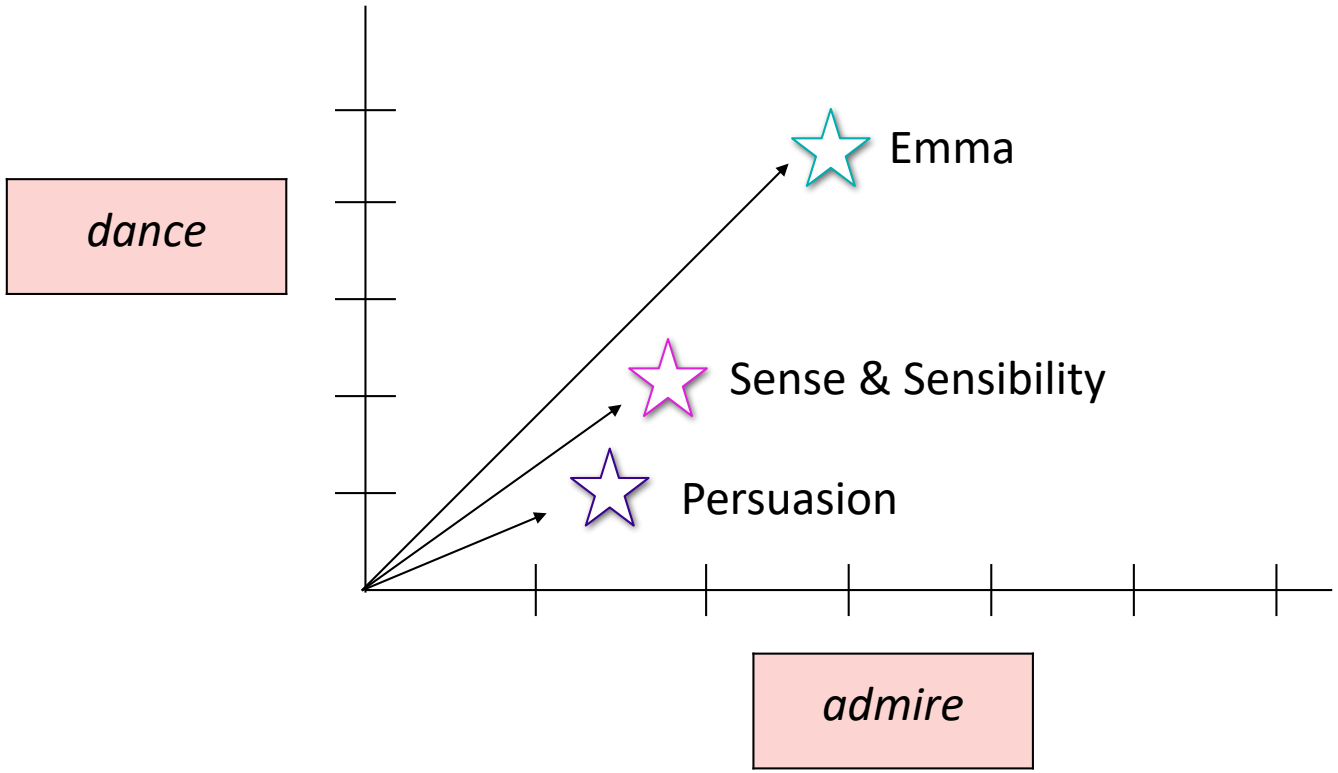
"admiral" is the kind of word that shows up in Persuasion

Term-document matrix

Each document is represented by a vector of words:

	Emma	**Persuasion**	**Sense & Sensibility**
admiral	0	69	0
dance	49	11	21
admire	31	14	18
horse	40	15	24

Visualizing document vectors



Vectors are the basis of information retrieval

	Emma	Persuasion	Sense & Sensibility	Paradise Lost
admiral	0	69	0	
dance	49	11	21	
admire	31	14	18	
horse	40	15	24	

Vectors are the basis of information retrieval

	Emma	Persuasion	Sense & Sensibility	Paradise Lost
admiral	0	69	0	0
dance	49	11	21	27
admire	31	14	18	11
horse	40	15	24	5

Vectors are similar for the Austen novels, but Milton is different: fewer horses.

Idea for word meaning: Words can be vectors too!!!

	Emma	Persuasion	Sense & Sensibility	Paradise Lost
admiral	0	69	0	0
dance	49	11	21	27
admire	31	14	18	11
horse	40	15	24	5

horse = [40 15 24 5]

Idea for word meaning: Words can be vectors too!!!

	Emma	Persuasion	Sense & Sensibility	Paradise Lost
admiral	0	69	0	0
dance	49	11	21	27
admire	31	14	18	11
horse	40	15	24	5

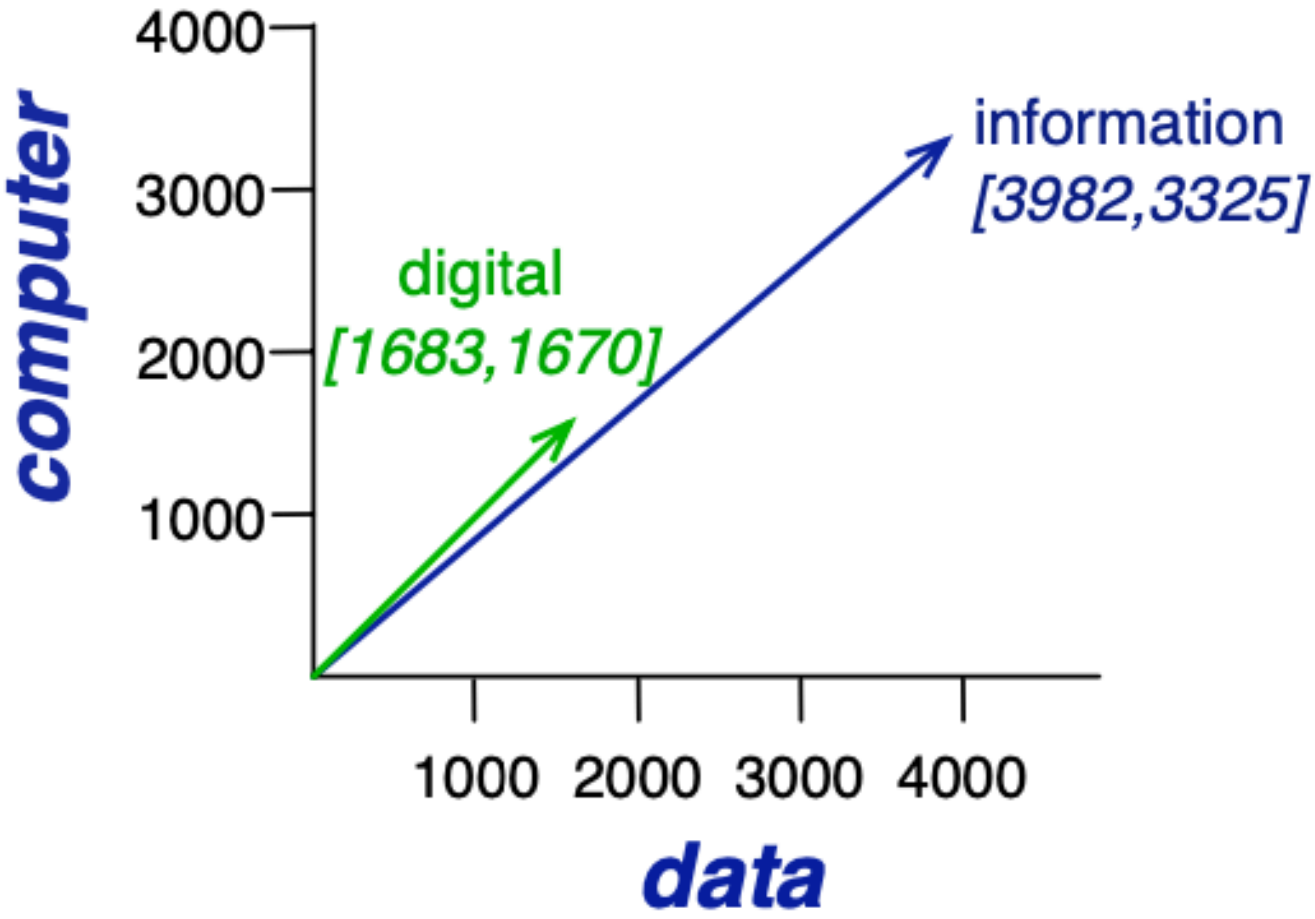
admiral is "the kind of word that occurs in Persuasion"

Word-word matrix (or "term-context matrix")

Two words are similar in meaning if their context vectors are similar:

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

	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	...


4000
3000
2000
1000
computer
digital
[1683,1670]
information
[3982,3325]
1000
2000
3000
4000
data

Computing Word Similarity

Computing word similarity

The dot product between two vectors is a scalar:

$$\text{dot}(v, w) = v \cdot w$$

$$= \sum_{i=1}^{N} v_i w_i$$

$$= v_1 w_1 + v_2 w_2 + \ldots + v_N w_N$$

The dot product is high when two vectors have large values in the same dimensions

Problem with raw dot-product

Dot product is higher if a vector is longer (has high values in many dimensions).

Vector length:

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

Frequent words have long vectors

Solution: normalize by vector length

Vector length: $$|\mathbf{v}| = \sqrt{\sum_{i=1}^{N} v_i^2}$$

Normalized dot product: $$\frac{\mathbf{v} \cdot \mathbf{w}}{|\mathbf{v}|\,|\mathbf{w}|} = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2}\sqrt{\sum_{i=1}^{N} w_i^2}}$$

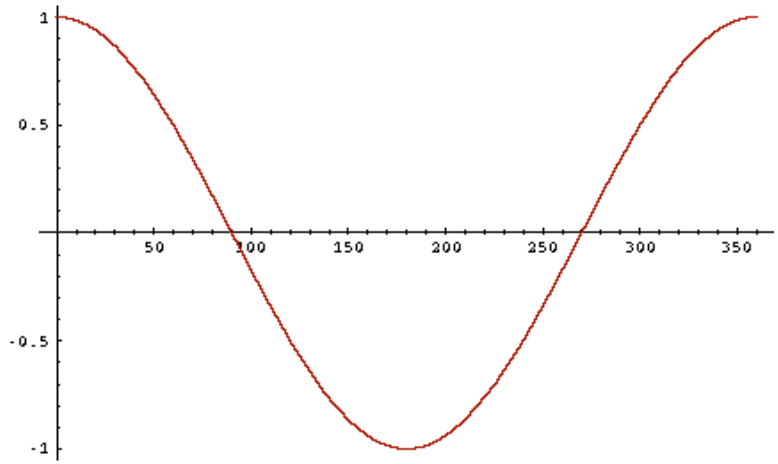
Surprise! This is the cosine of the angle between the vectors.

Cosine as a similarity metric

-1: vectors point in opposite directions

+1: vectors point in same directions

0: vectors are orthogonal



Since raw frequency values are non-negative, the cosine for term-term matrix vectors ranges from 0–1.

(Later we will see other vector representations that have similarities from -1 to +1).

Cosine examples

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

$\cos(\text{cherry}, \text{information}) = 0.018$

$\cos(\text{digital}, \text{information}) = 0.996$

Visualizing cosine similarity