Word Embeddings

CS 490A, Fall 2021

Applications of Natural Language Processing
https://people.cs.umass.edu/~brenocon/cs490a_f21

Brendan O'Connor & Laure Thompson

College of Information & Computer Sciences
University of Massachusetts Amherst

Administrivia

- HW3 due this Friday, 10/29
- Project Proposal Feedback coming soon!
- Keep a lookout for Project Proposal Meeting sign-ups!
- Midterm Review: 11/4
- In-Class Midterm: 11/9

What is a *pawpaw*?

I. Look it up in a dictionary

https://www.merriam-webster.com/

https://www.oed.com/

https://en.wiktionary.org/

pawpaw noun

Save Word

paw·paw

variants: *or less commonly* **papaw**

Definition of *pawpaw*

1 \ pə-ˈpȯ \ **:** PAPAYA

2 \ ˈpä-(ˌ)pȯ, ˈpȯ- \ **:** a North American tree (*Asimina triloba*) of the custard-apple family with purple flowers and an edible green-skinned fruit

also **:** its fruit

Lemma

pawpaw noun

Save Word

paw·paw

variants: *or less commonly* **papaw**

Word Senses

Definition of *pawpaw*

1 \ pə-ˈpȯ \ : PAPAYA

2 \ ˈpä-(ˌ)pȯ, ˈpȯ- \ : a North American tree (*Asimina triloba*) of the custard-apple family with purple flowers and an edible green-skinned fruit

also : its fruit

Definition

II. Look it at how its used

"**<u>Pawpaw</u>**, Most Neglected American Fruit." — NYTimes <u>1922</u>

"**<u>Pawpaw</u>** Recommended by U.S. Food Experts, Along With Persimmon, as War Nutrition" — NYTimes <u>1942</u>

"The **<u>pawpaw</u>** is also pollinated by flies and other insects rather than by honeybees…"— NYTimes <u>2020</u>

"Many people also cook with ripe **<u>pawpaws</u>**, making bread, beer, ice cream, or this **<u>pawpaw</u>** pudding…" — NYTimes <u>2020</u>

II. Look it at how its used

"*Pawpaw*, Most Neglected **American Fruit**." — NYTimes 1922

"*Pawpaw* Recommended by U.S. Food Experts, Along With **Persimmon**, as War **Nutrition**" — NYTimes 1942

"The *pawpaw* is also **pollinated** by **flies** and other insects rather than by honeybees…"— NYTimes 2020

"Many people also **cook** with **ripe** *pawpaws*, making **bread**, **beer**, **ice cream**, or this *pawpaw* **pudding**…" — NYTimes 2020

Word Relations

Synonyms

- couch / sofa
- oculist / eye-doctor
- car / automobile
- water / H_2O
- draft / draught

Antonyms

- yes / no
- dark / light
- hot / cold
- up / down
- clip / clip

Word Relations

Similarity

- cat / dog
- cardiologist / pulmonologist
- car / bus
- sheep / goat
- glass / mug

Relatedness

- coffee / cup
- waiter / menu
- farm / cow
- house / roof
- theater / actor

Quantifying Similarity

Ask humans how ***similar*** two words are on a scale of 1-10

Word 1	**Word 2**	**SimLex-999**
area	region	9.47
horse	mare	8.33
water	ice	6.7
hill	cliff	4.28
absence	presence	0.4
princess	island	0.3

Hill et al. 2015

Task Design is Difficult

Similarity and **Relatedness** do not capture the same relations

Word 1	Word 2	SimLex-999	WordSim-353
coast	shore	8.83	9.10
clothes	closet	3.27	8.00

Hill et al. 2015; Finkelstein et al. 2002

…but what about computers?

Word Embeddings

Represent each word as a **vector**

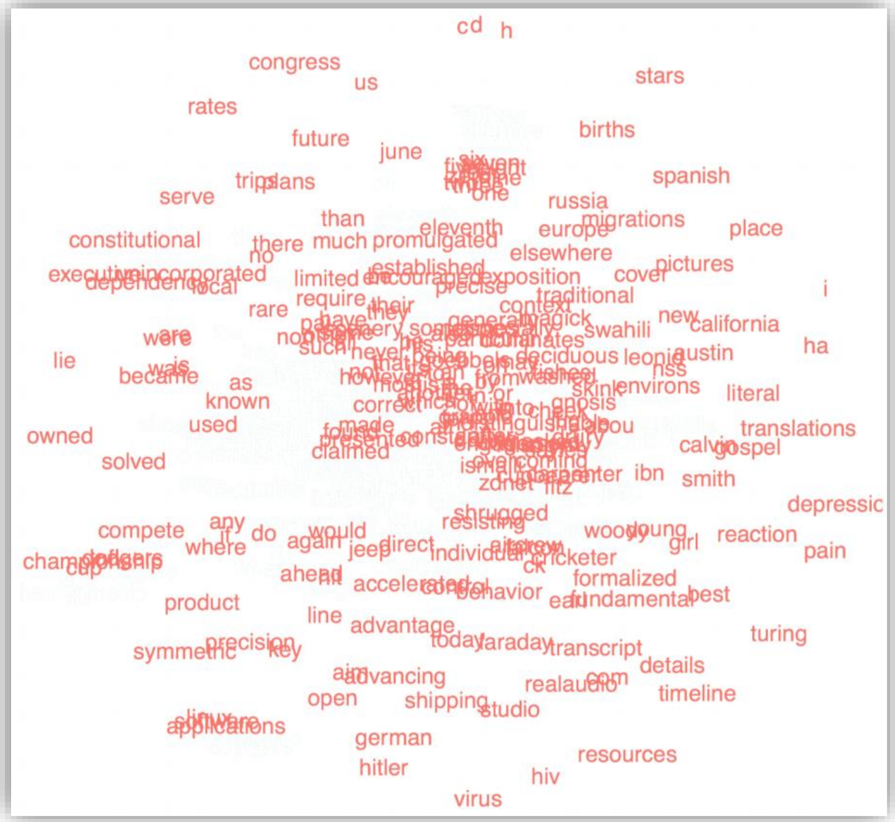
On Vectors:

- A **vector** is a list of numbers
- A **vector** can also be considered a **point** in a *k*-dimensional space

Capturing Word Similarity

Operationalize word similarity by computationally **comparing** vectors

Distance reflects semantic relationships



Closer vectors represent more similar words

More distant vectors represent less similar words pawpaw

Top | Latest | People | Photos | Videos

Scott BaroooOooOOooooOooo @sbarolo · Jul 7

#Fruitbracket Match #36:
PAPAYA vs. KIWIFRUIT

Papaya	30.6%
Kiwifruit	**69.4%**

376 votes · Final results

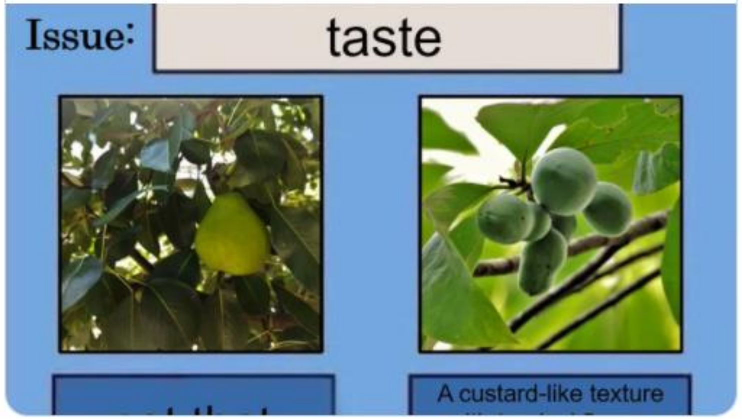
1 4

Show this thread

Scott BaroooOooOOooooOooo @sbarolo · Jul 5

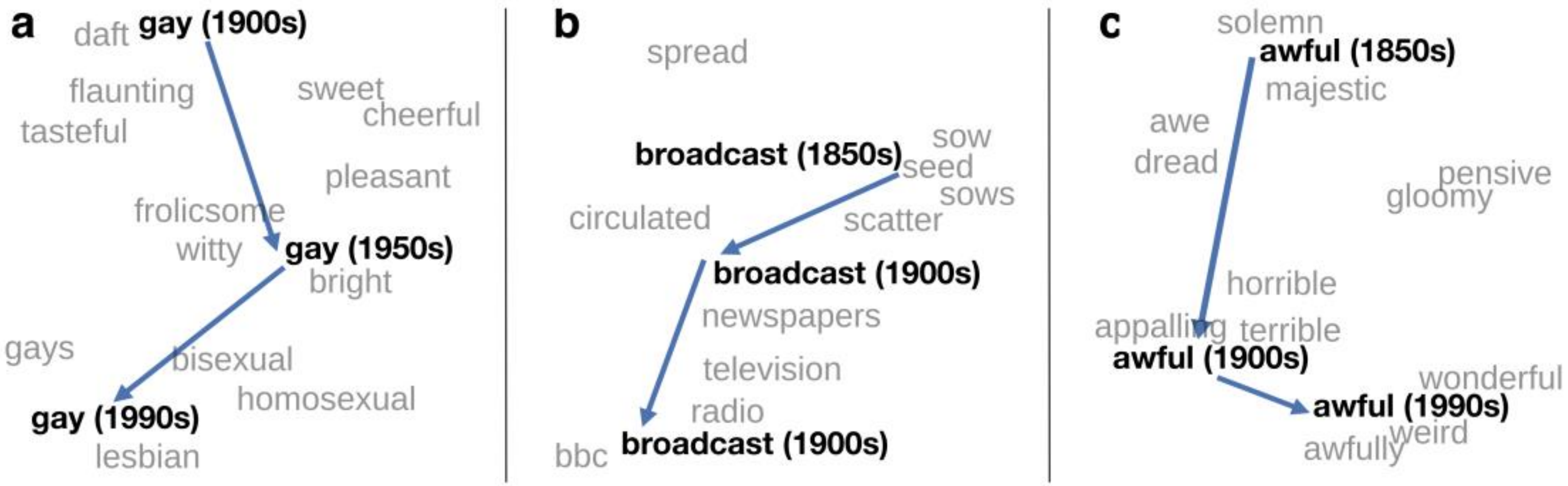
Big **Pawpaw** is coming out swinging

> **Jeff L'épouvantail** @letourjeff · Jul 5
> twitter.com/sbarolo/status...



2 24

Study word use over time



Hamilton et al. 2016

One–Hot Vectors

Each word is represented by a vector with a 1 in the word's index in the vocabulary and 0's elsewhere

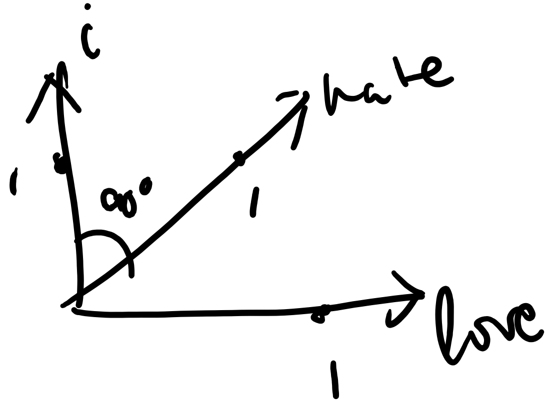
Term	Vector
i	<1, 0, 0, 0, 0, 0>
hate	<0, 1, 0, 0, 0, 0>
love	<0, 0, 1, 0, 0, 0>
the	<0, 0, 0, 1, 0, 0>
movie	<0, 0, 0, 0, 1, 0>
film	<0, 0, 0, 0, 0, 1>

Sparse!

Q: What are some issues with these representations?

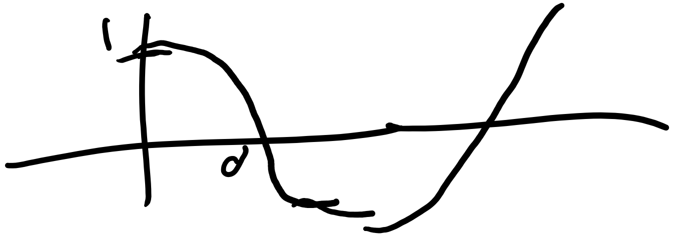
Q: What are some issues with these representations?

1. Vocabularies are large!



2. They're all equidistant

cosine(90°) = 0



Distributional Semantics

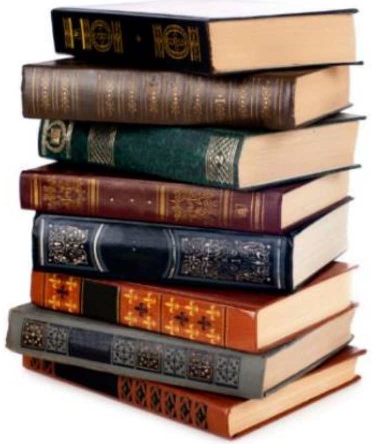
“You shall know a word by the company it keeps!” — Firth (1957)

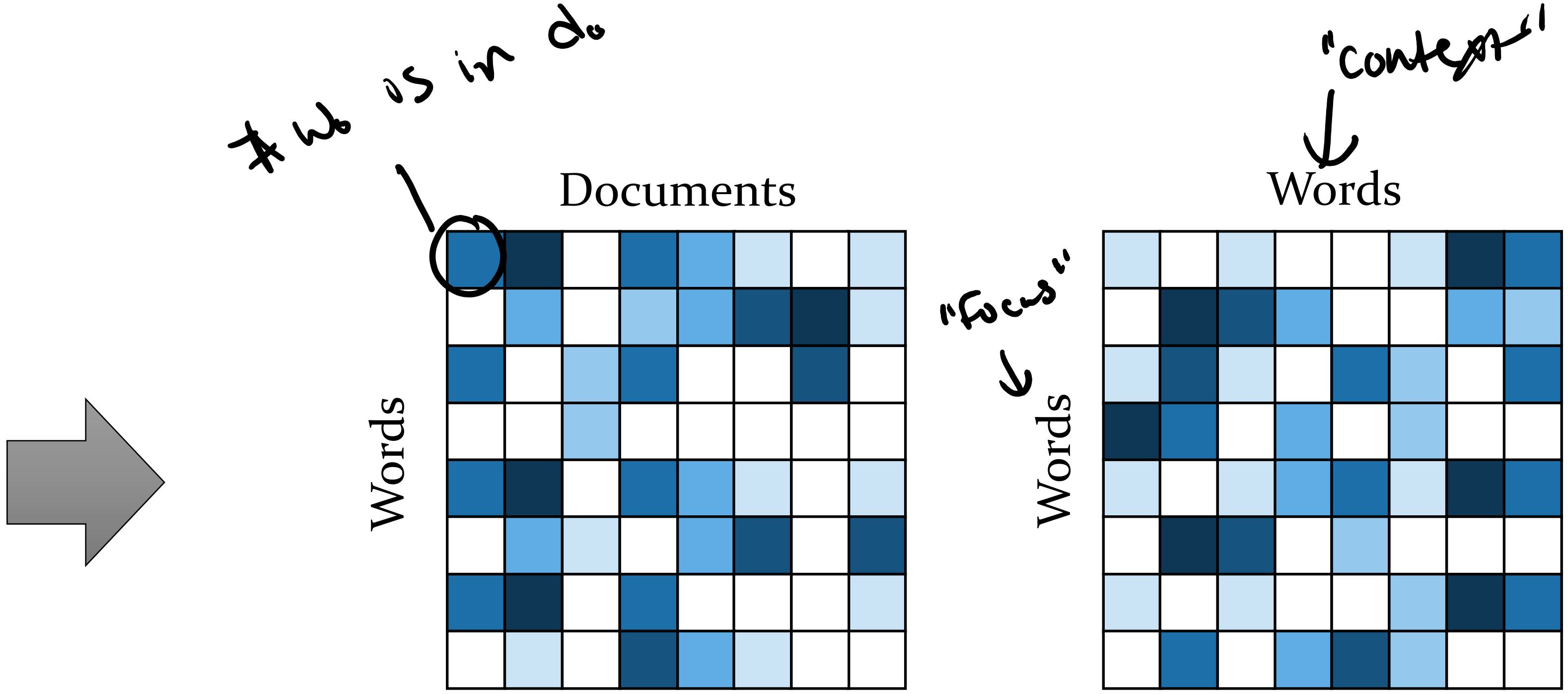
Intuitions: Harris (1954)

“If A and B have almost identical environments except chiefly sentences which contain both, we say they are synonyms: *oculist* and *eye-doctor*.”

petrol , gas

Build vectors based on context





Q: What are some issues with these representations?

Q: What are some issues with these representations?

- Still size of vocab!
- These are sparse

Trouble with raw frequency

Words occur at different frequencies irrespective of context

So, raw frequency does not necessarily correspond to significant, informative use.

peach , fruit → relationship

the , fruit → not so meaningful

the , peach

Move away from raw frequency

Term-Document Matrix

Apply tf-idf weighting

$$= \frac{term\ frequency}{document\ frequency}$$

tf : raw count

"freq" $\frac{\text{raw count}}{\text{total}}$

Word-Context Matrix

Use PPMI (Positive Pointwise Mutual Information)

$$= \max(PMI(w_a, w_b), 0)$$

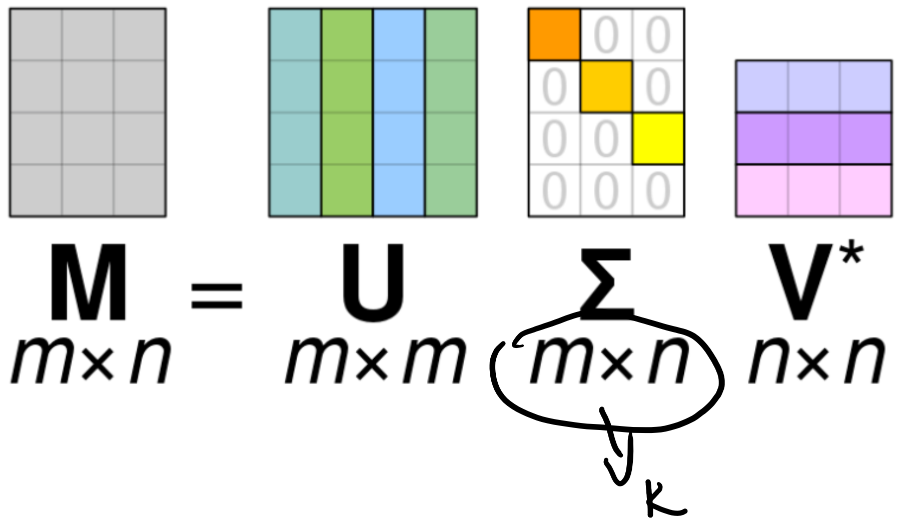
$$= \max\left(\log_2 \frac{P(w_a, w_b)}{P(w_a)P(w_b)}, 0\right)$$

Move to smaller, dense embeddings

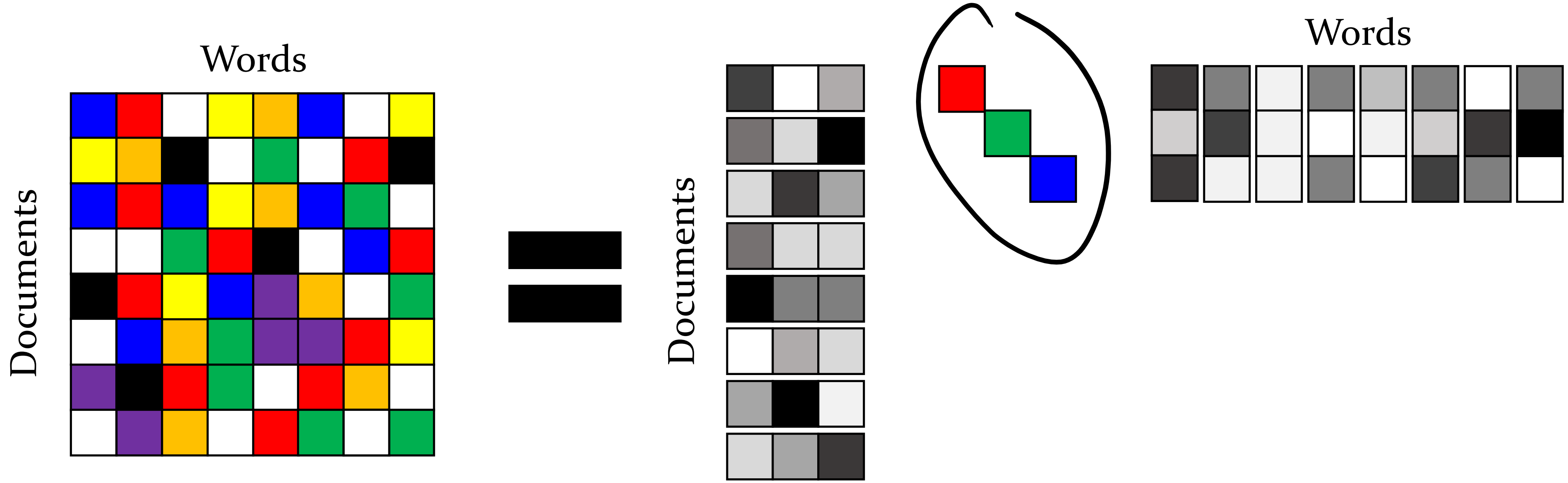
Use **matrix factorization** to build a more compact representation

Matrix factorization *decomposes* a matrix into the product of several (smaller) matrices

E.g., Singular Value Decomposition (SVD)



Latent Semantic Analysis (LSA)



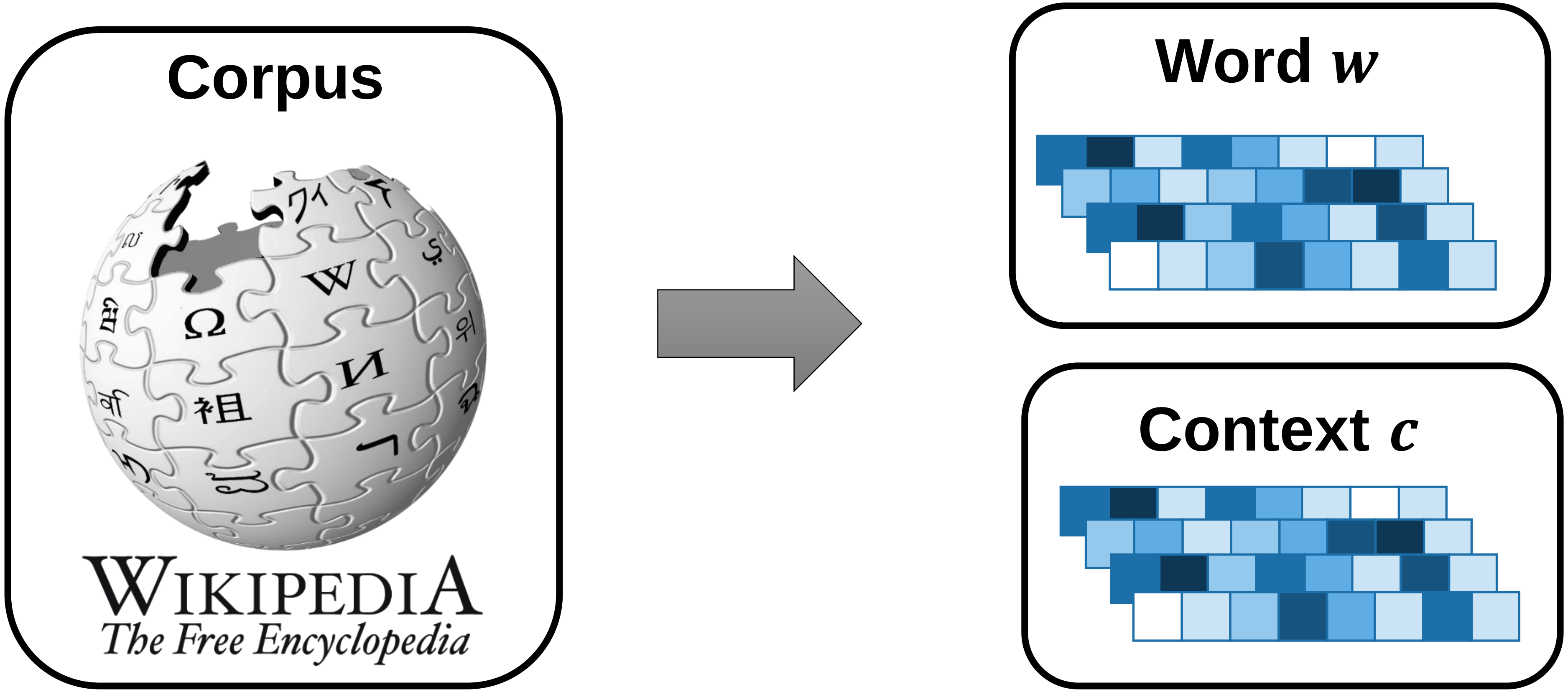
Deerwester et al. 1990

Newer, neural models also use matrix factorization

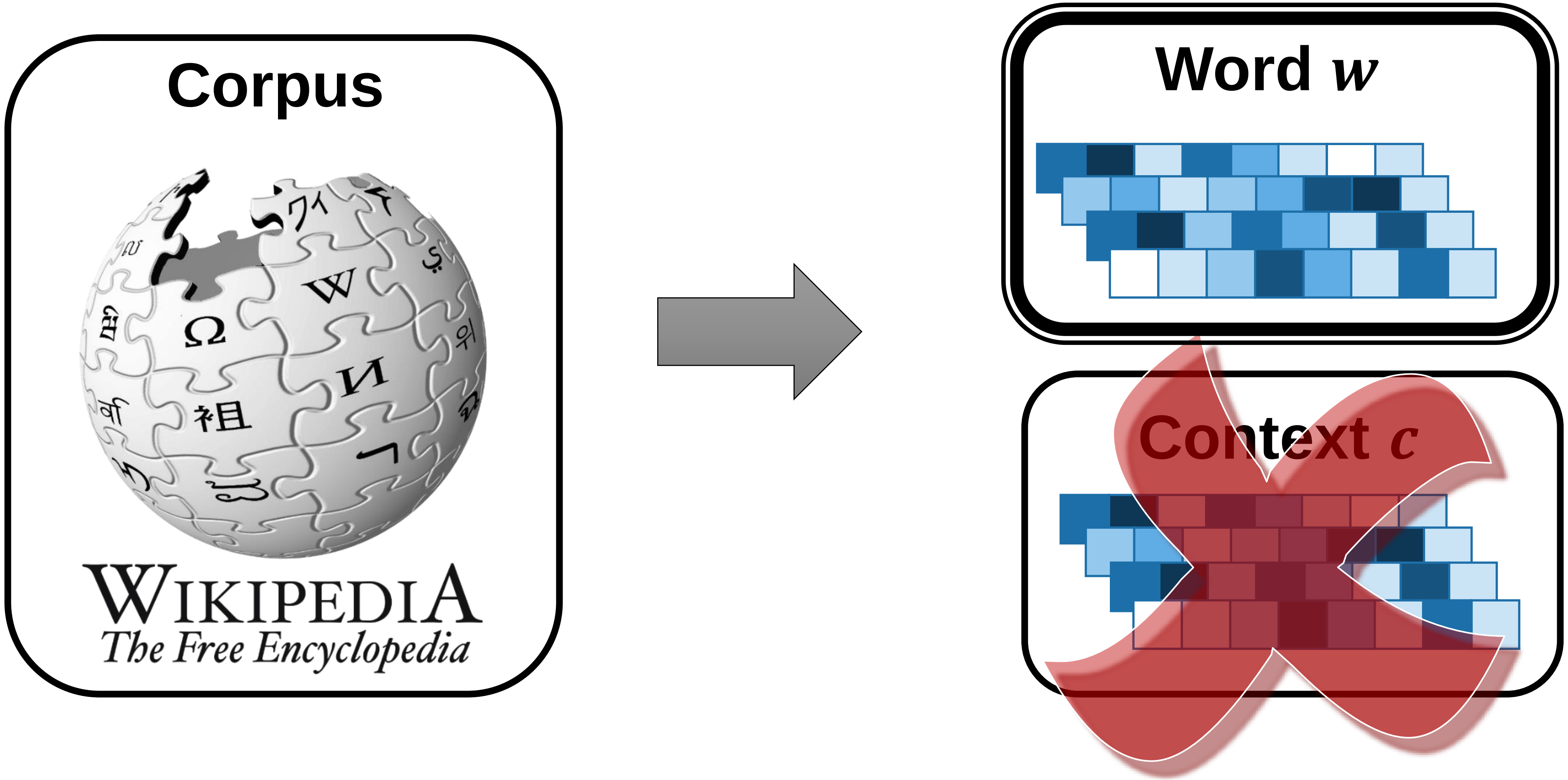
E.g., GloVE and SGNS

word2vec

Neural Word Embeddings



Neural Word Embeddings



Skip-Gram with Negative Sampling (SGNS)

The brown fox jumps over the lazy dog.

<u>SG</u>NS: Skip-Gram Model

The brown fox **jumps** over the lazy dog.

<u>**SG**</u>NS: Skip-Gram Model

The brown fox **jumps** over the lazy dog.

Context Window Size = 2

SGNS: Skip-Gram Model

The brown fox **jumps** over the lazy dog.

Context Window Size = 2

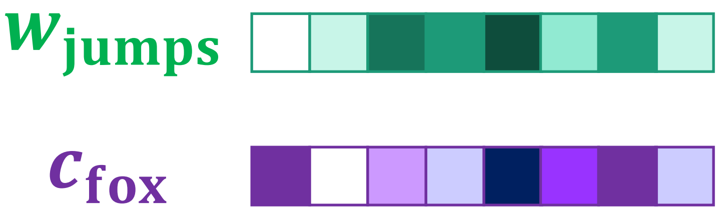
jumps → { brown, fox, over, the }

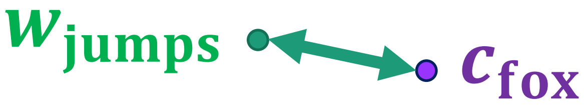
6

CBOW context $\xrightarrow{\text{predict}}$ word

SG**<u>NS</u>**: Negative Sampling

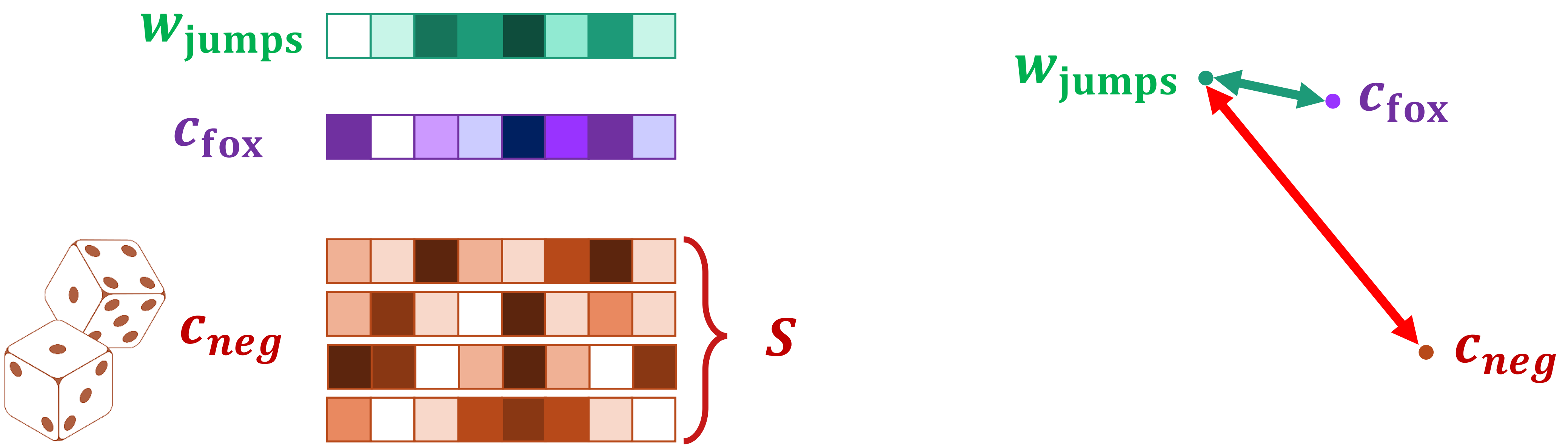
Co-occurrence **jumps** , **fox**:





SG<u>NS</u>: Negative Sampling

Co-occurrence **jumps** , **fox**:

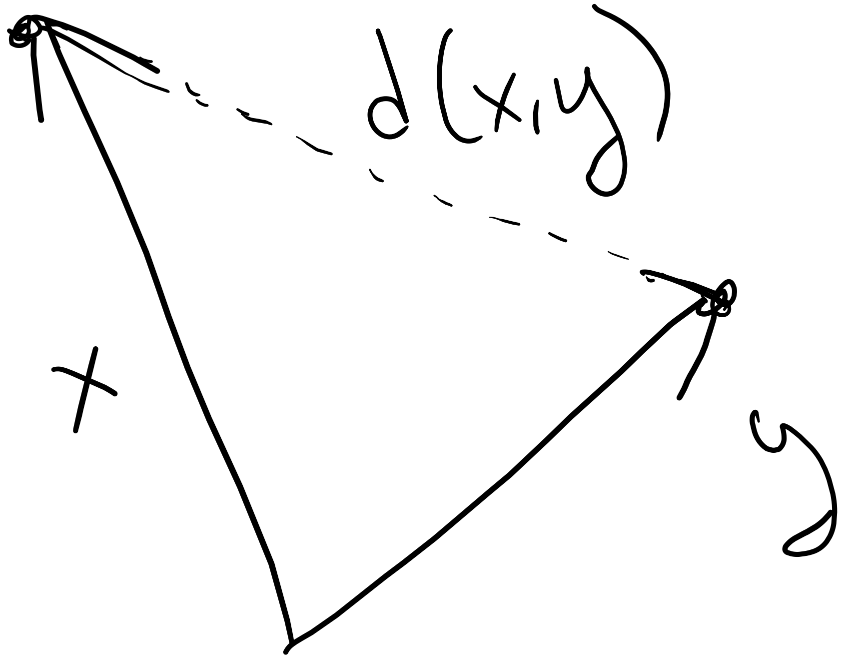


How do we compare vectors?

- Similarity measurements
  - Larger values → similar vectors → similar words
  - Smaller values → dissimilar vectors → dissimilar words

- Distance / dissimilarity measurements
  - Note: distance metric requires triangle inequality
  - Larger values → dissimilar vectors → dissimilar words
  - Smaller values → similar vectors → similar words

Euclidean Distance

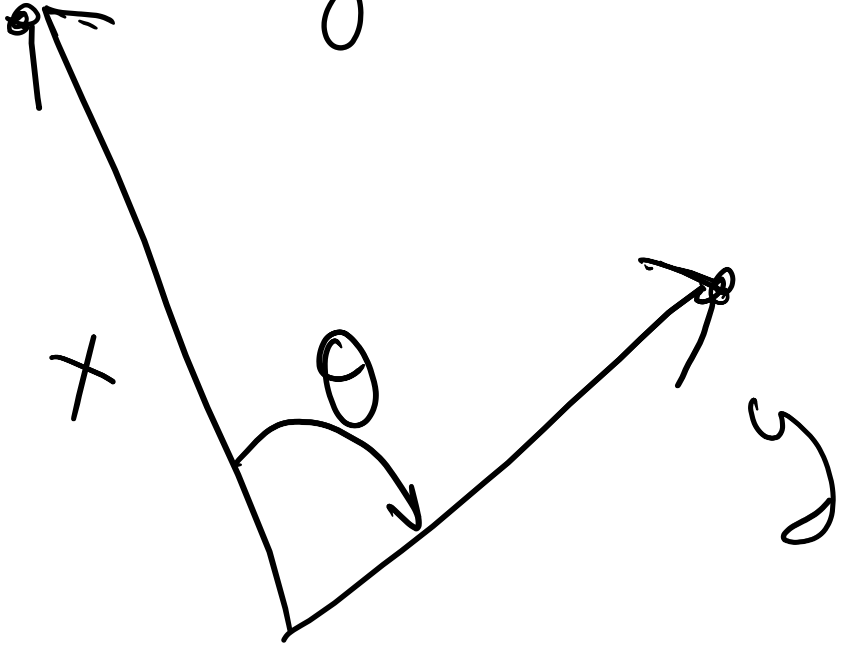


$$d(x, y) = \sqrt{\sum_i (x_i - y_i)^2}$$

Issue: Vector length depends on frequency. More frequent words will have longer vectors.

Cosine Similarity

$s(x,y) = \cos\theta$

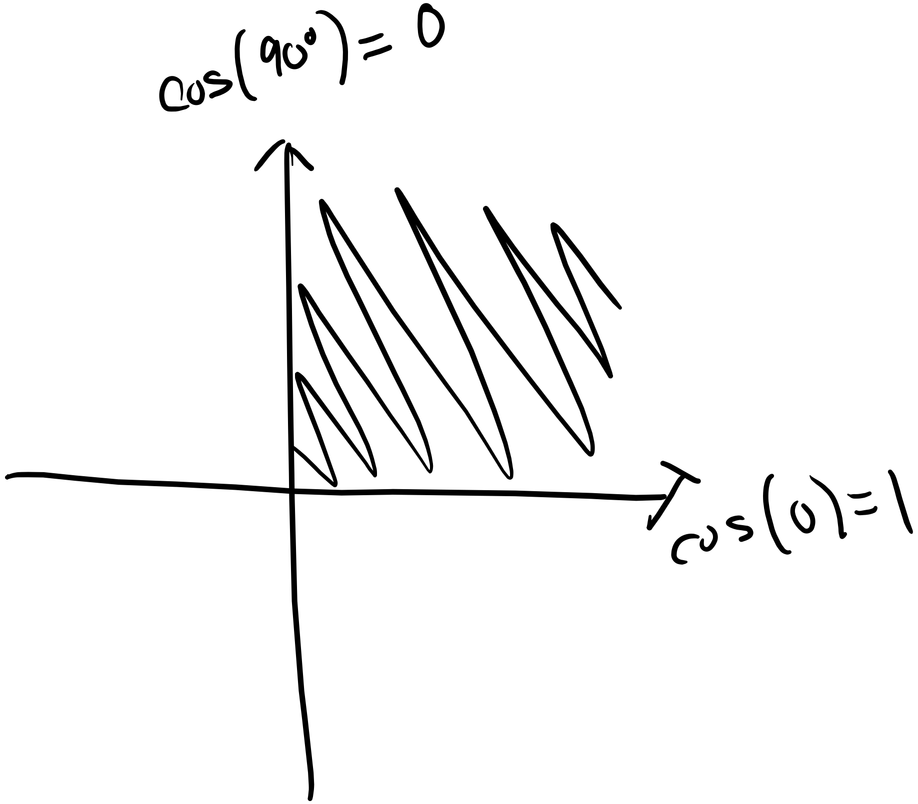


$$s(x, y) = \frac{x \cdot y}{|x||y|}$$

Only depends on vector angle

Range: [-1, 1]

Non-negative vectors & cosine similarity



If all vectors have non-negative values, then their cosine similarity will be between 0 and 1