Welcome back to CTASS!

Introductions

A quick reminder of our goals here

Building an interdisciplinary community

Sharing knowledge and building a common vocabulary

Providing support

Expanding the acceptance and legitimacy of text analysis methods in the social sciences

Upcoming events

Grad Student lightning presentations

Sign up here if you're interested: https://bit.ly/ctasslightning

Jen Pan faculty talk (1/24)

CASM: A Deep-Learning Approach for Identifying Collective Action Events with Text and Image Data from Social Media

RSVP here: https://bit.ly/ctassjenpan

Colloquium on topic models (2/7)

Twitter API Workshop (3/6)

Interested in joining the CTASS leadership?

Email us:

kkanopka@stanford.edu

avanloon@stanford.edu

sheridan.a.stewart@stanford.edu

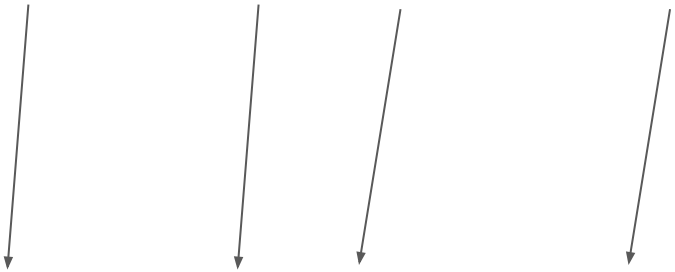
Word Embeddings:

What are they and what role do they play in social science?

The orthodontist was smiling, happy to see her friend.

The dentist was grinning, glad to see her buddy.

The orthodontist was smiling, happy to see her friend.



The dentist was grinning, glad to see her buddy.

Distributed Representations

One way to uniquely identify three grad students:

Austin	1	0	...	0
Sheridan	0	1	...	0
...	...	...	...	...
Klint	0	0	...	1

What happens if we want to add another grad student to this database?

What are some alternatives to this kind of representation?

Distributed Representations

What about this?

	D_1	D_2	...	D_K
Austin	1	0.25	...	1
Sheridan	3	0.4	...	0
...	...	...	...	...
Klint	2	0.25	...	1

If *N* is the number of grad students and *K* is the number of dimensions, could we build our representation such that *D* < *K*? How about D << K?

What are some additional benefits to storing data in this way?

Distributed Representations of Language

Our goal is to build a distributed representation of linguistic items (let’s say for now we’re only interested in words)

Deep learning will build our (dense) distributed representation as long as we can provide a metric of how similar any two words should be

So what is a metric we can use for how similar two words are?

Distributional Hypothesis

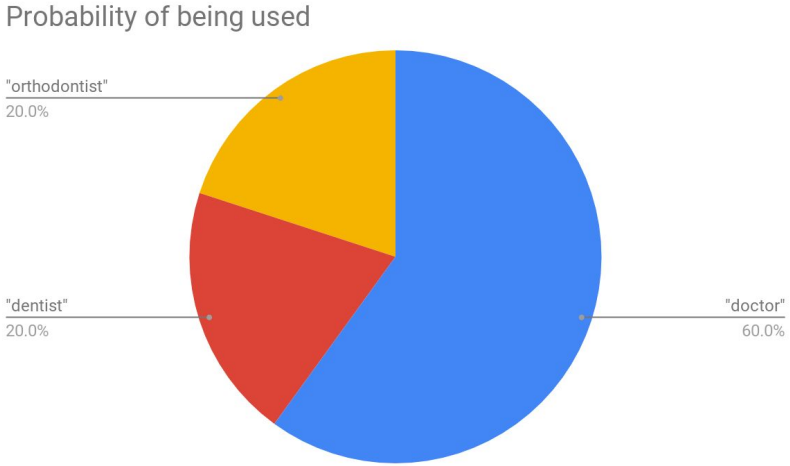
The distributional hypothesis: the more similar two words are, the more they will tend to occur in similar contexts…

“...you shall know a word by the company it keeps” (J.R. Firth)

This tells us that the meaning of the word is actually encoded in (or just is) the context in which the word is found

Consider the following contexts:

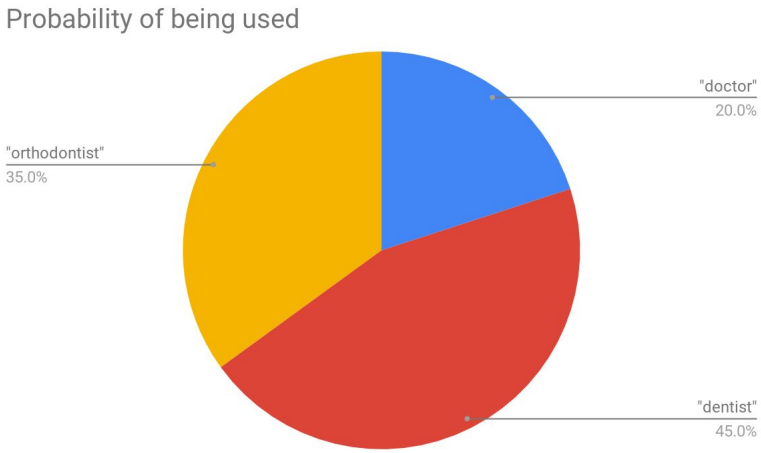
The ___ will see you now.



Consider the following contexts:

The ___ will see you now.

The ___ examined her teeth.

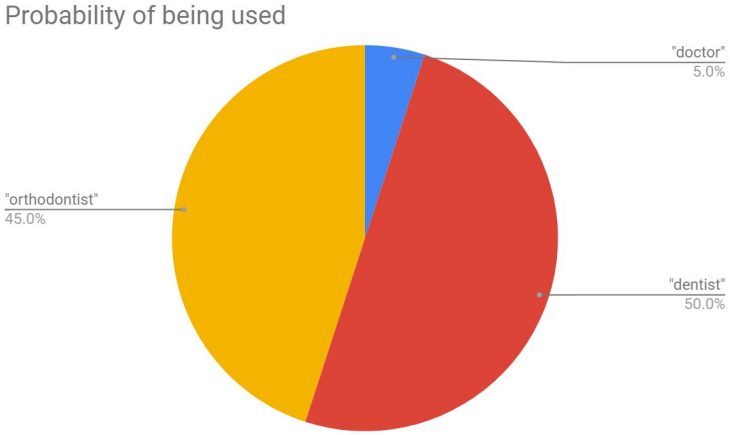


Consider the following contexts:

The ___ will see you now.

The ___ examined her teeth.

Four out of five ___ recommend Colgate toothpaste



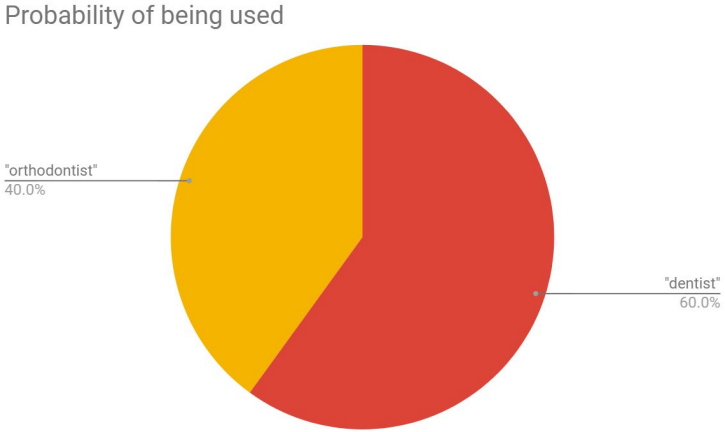
Consider the following contexts:

The ___ will see you now.

The ___ examined her teeth.

Four out of five ___ recommend Colgate toothpaste

John had a cavity, so he went to see his ____.



Consider the following contexts:

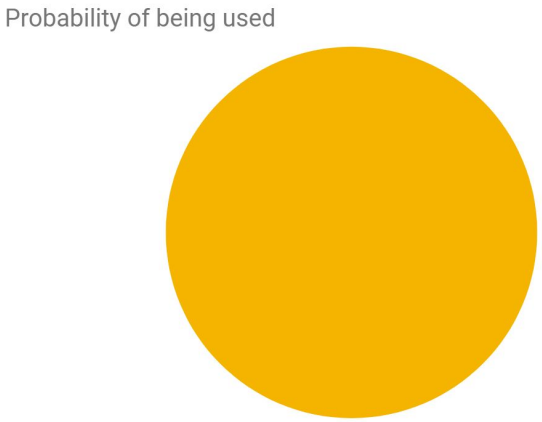
The ___ will see you now.

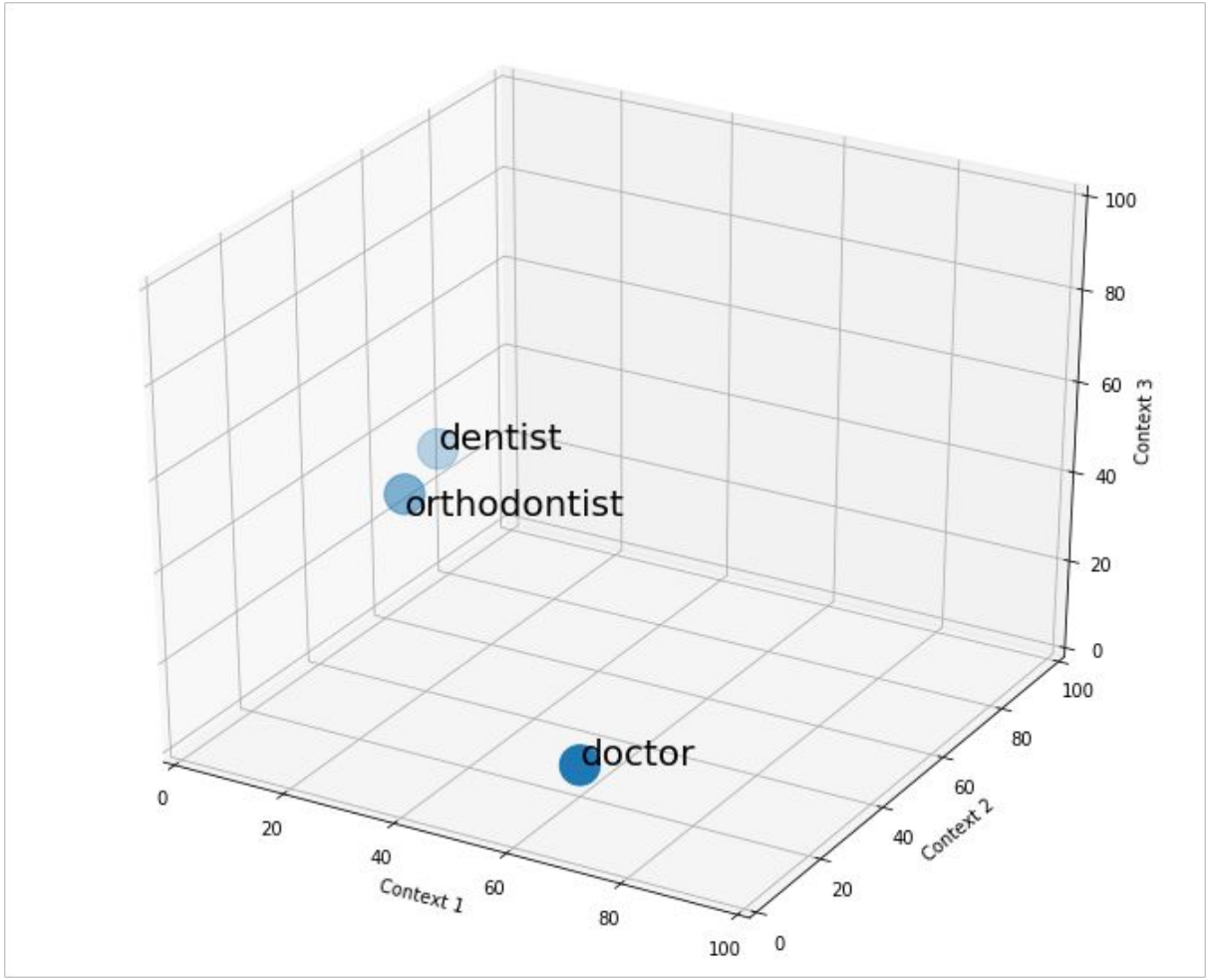
The ___ examined her teeth.

Four out of five ___ recommend Colgate toothpaste

John had a cavity, so he went to see his ____.

Austin needed to have his teeth straightened, and he found out that ___, unlike dentists, specialize in such matters.




dentist
orthodontist
doctor
Context 1
Context 2
Context 3
0
20
40
60
80
100

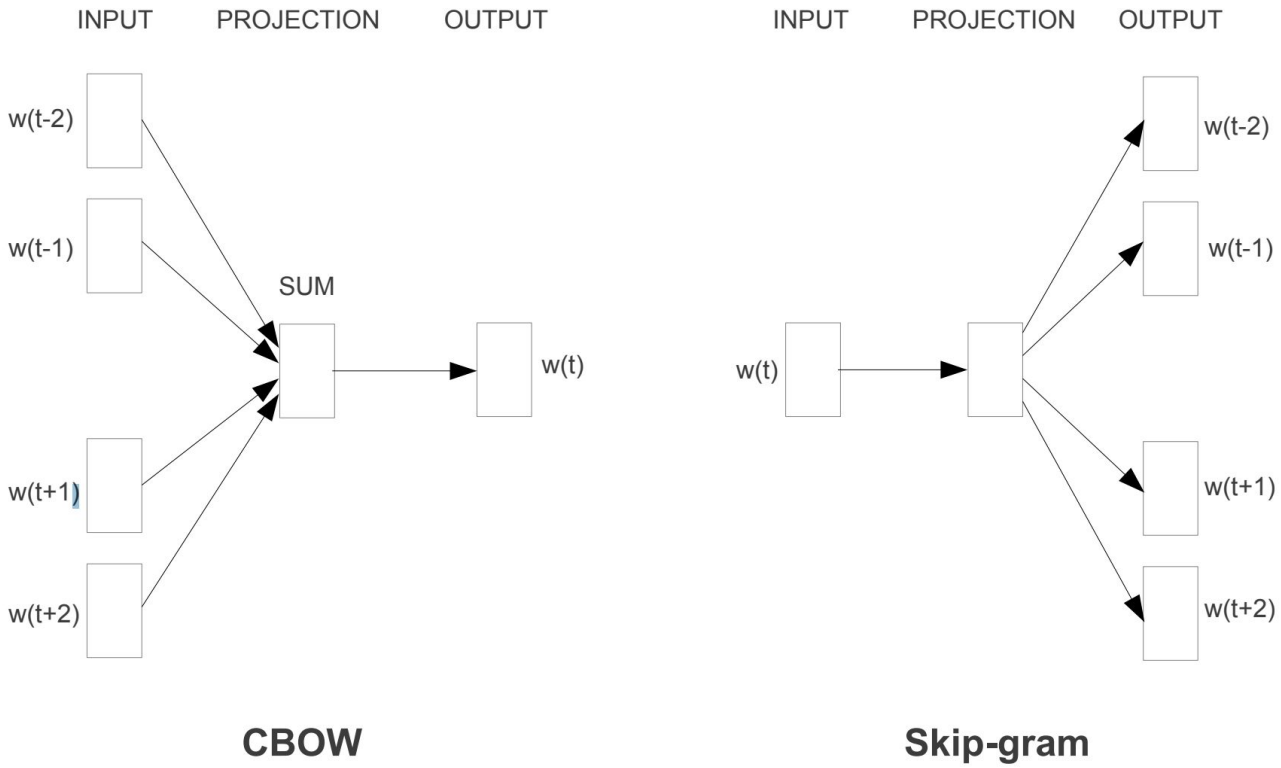


Figure 1: New model architectures. The CBOW architecture predicts the current word based on the context, and the Skip-gram predicts surrounding words given the current word.

____ of a feather flock together.

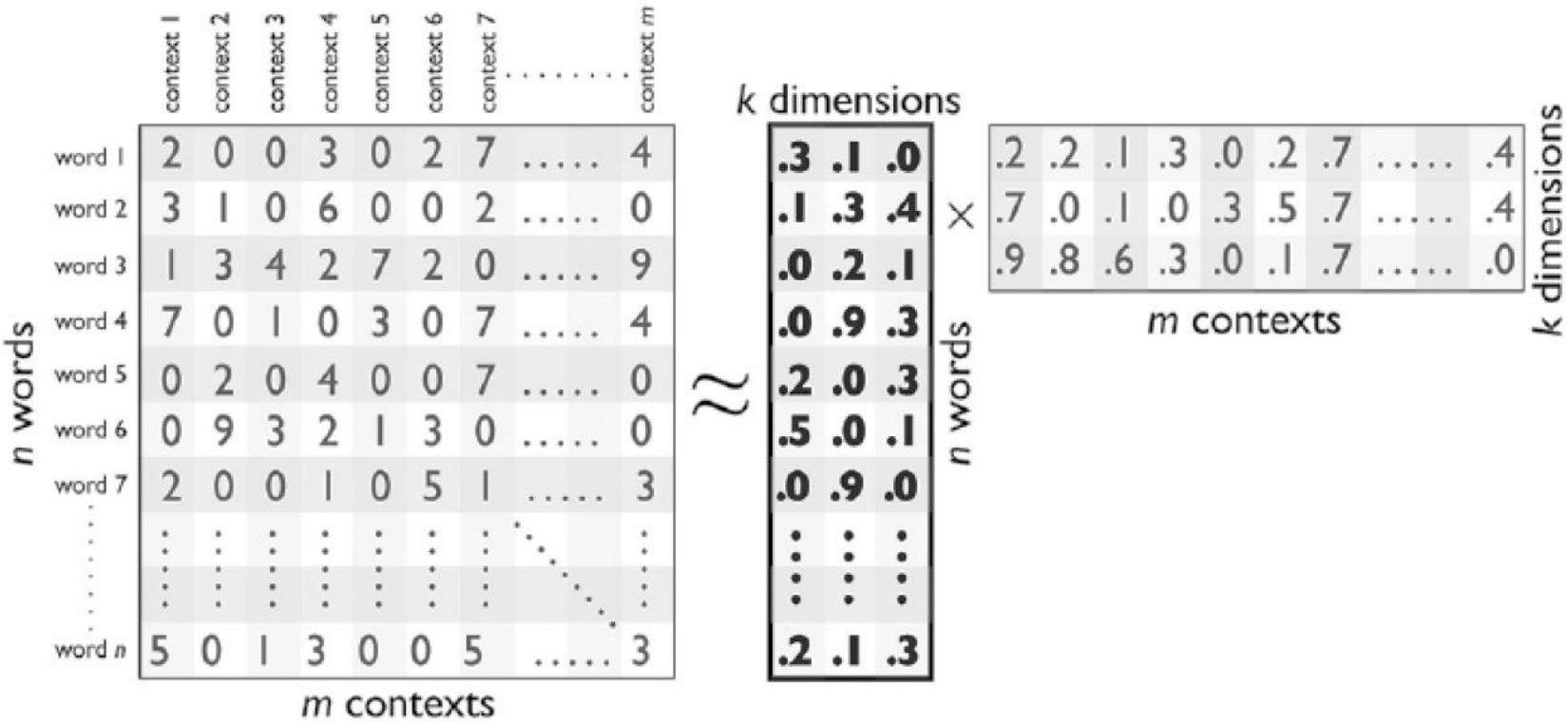
Birds ___ a feather flock together.

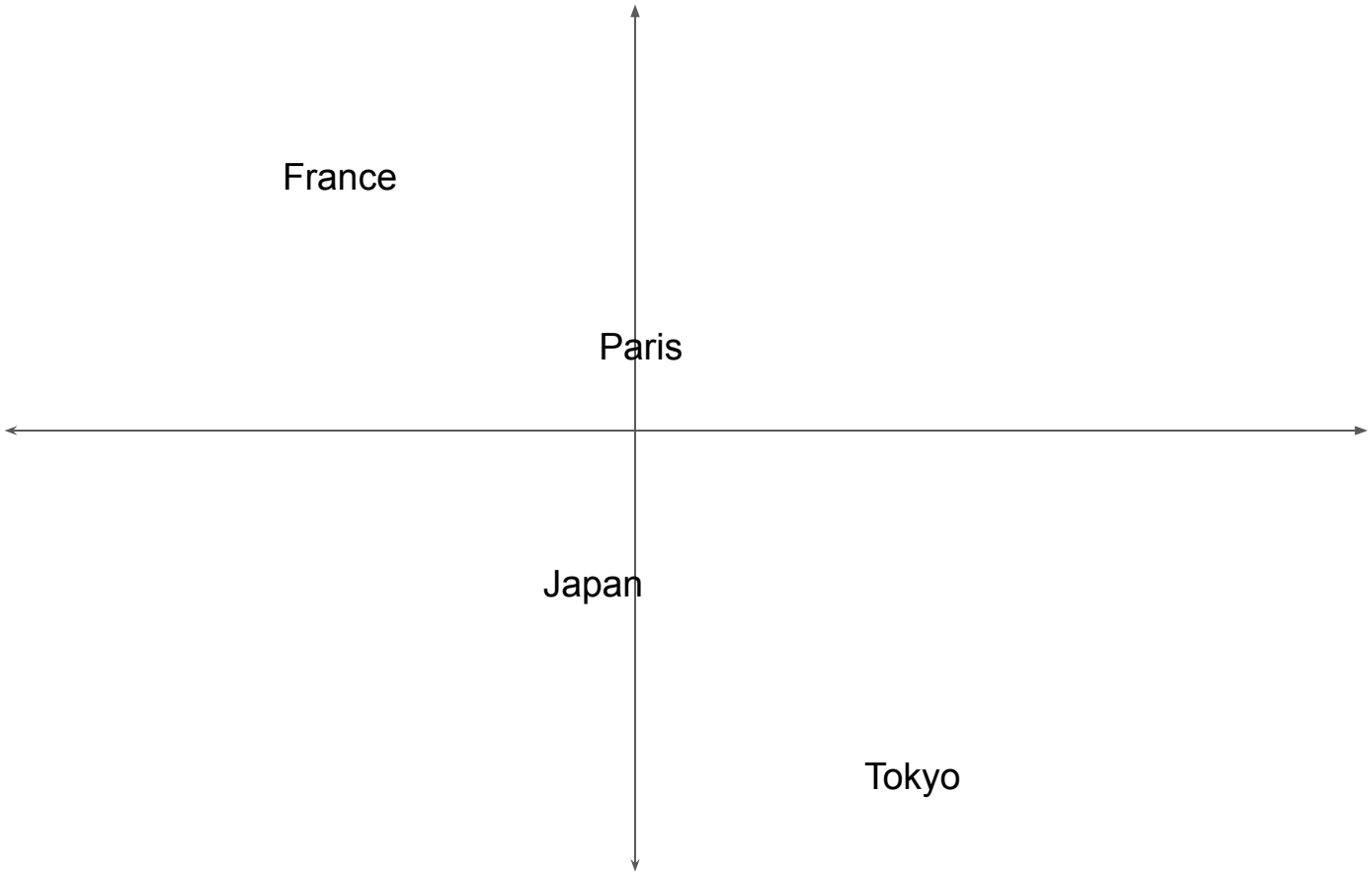
Birds of ___ feather flock together.

Birds of a ___ flock together.

Birds of a feather ___ together.

Birds of a feather flock ___.


context 1
context 2
context 3
context 4
context 5
context 6
context 7
context m
word 1
word 2
word 3
word 4
word 5
word 6
word 7
word n
n words
m contexts
k dimensions
n words
m contexts
k dimensions


France
Paris
Japan
Tokyo

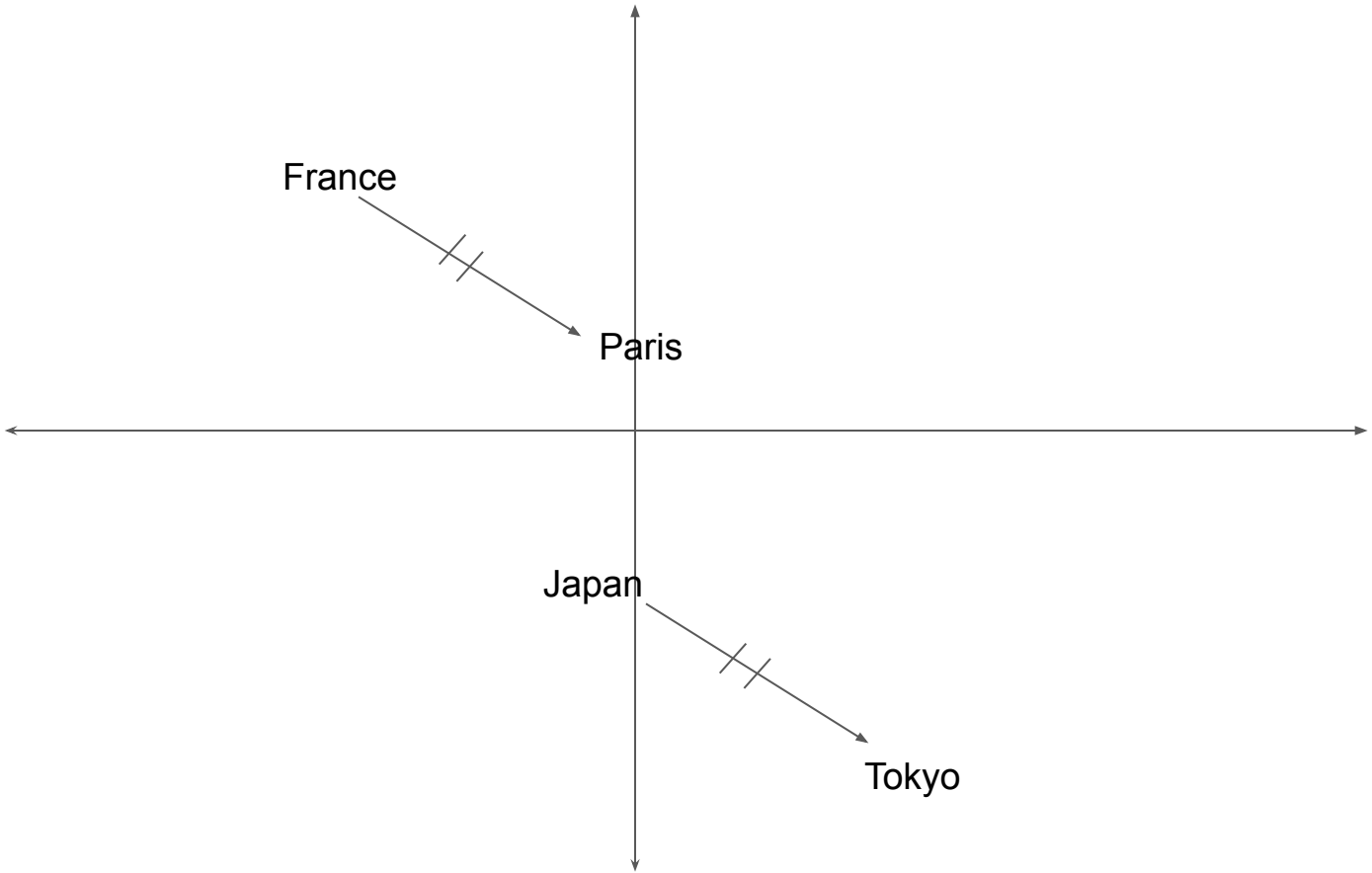

France
Paris
Japan
Tokyo

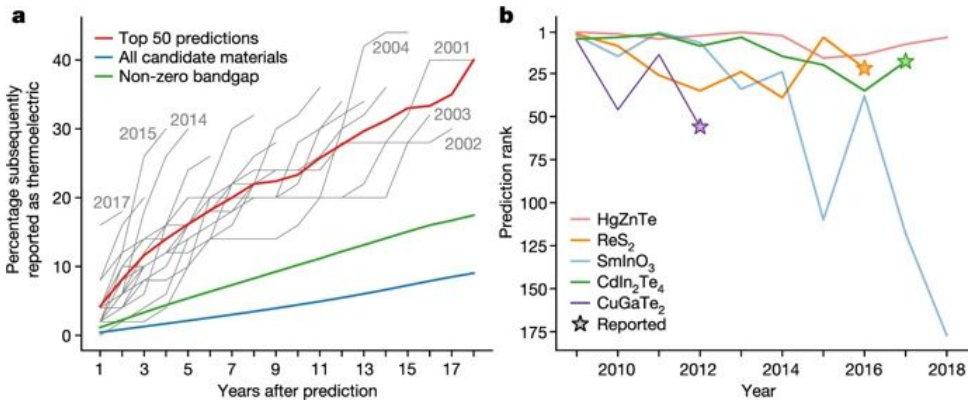
Table 8: *Examples of the word pair relationships, using the best word vectors from Table 4 (Skip-gram model trained on 783M words with 300 dimensionality).*

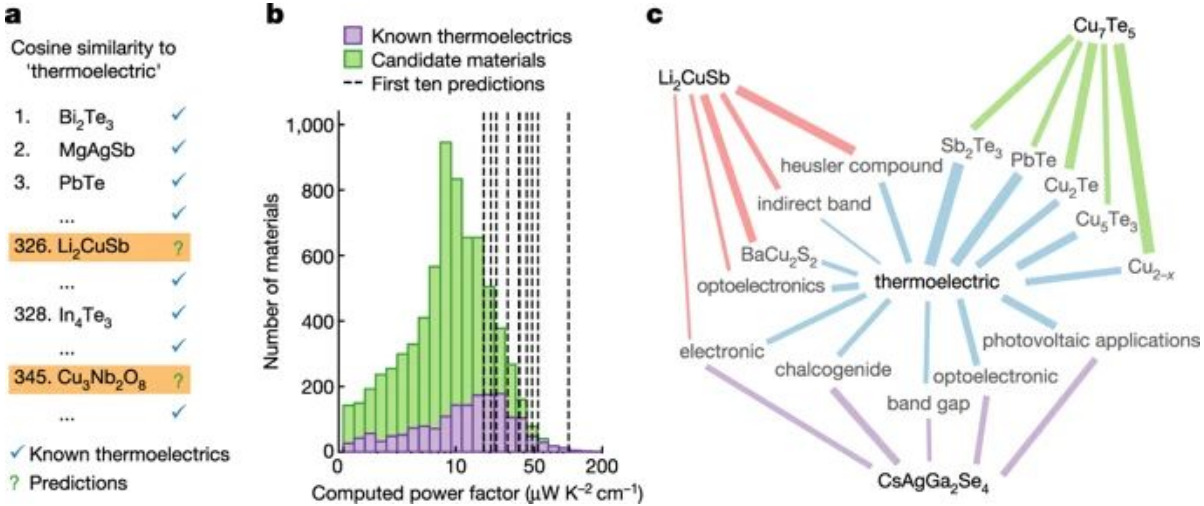
Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - 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

Unsupervised word embeddings capture latent knowledge from materials science literature

Vahe Tshitoyan, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova, Kristin A. Persson, Gerbrand Ceder & Anubhav Jain

Nature **571**, 95–98(2019) | Cite this article

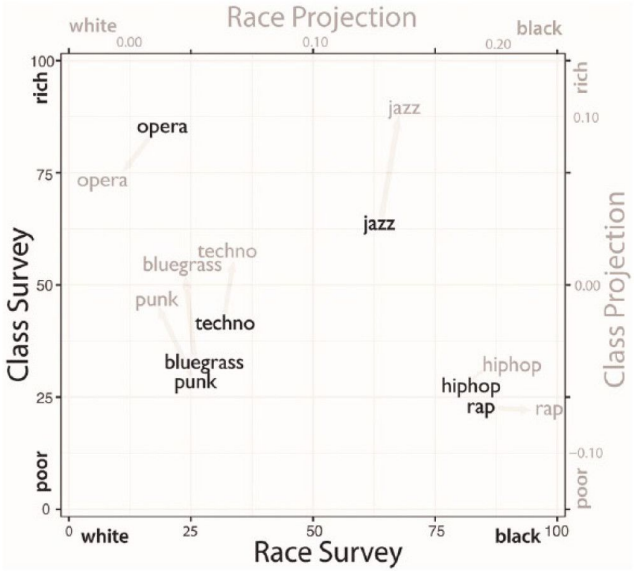


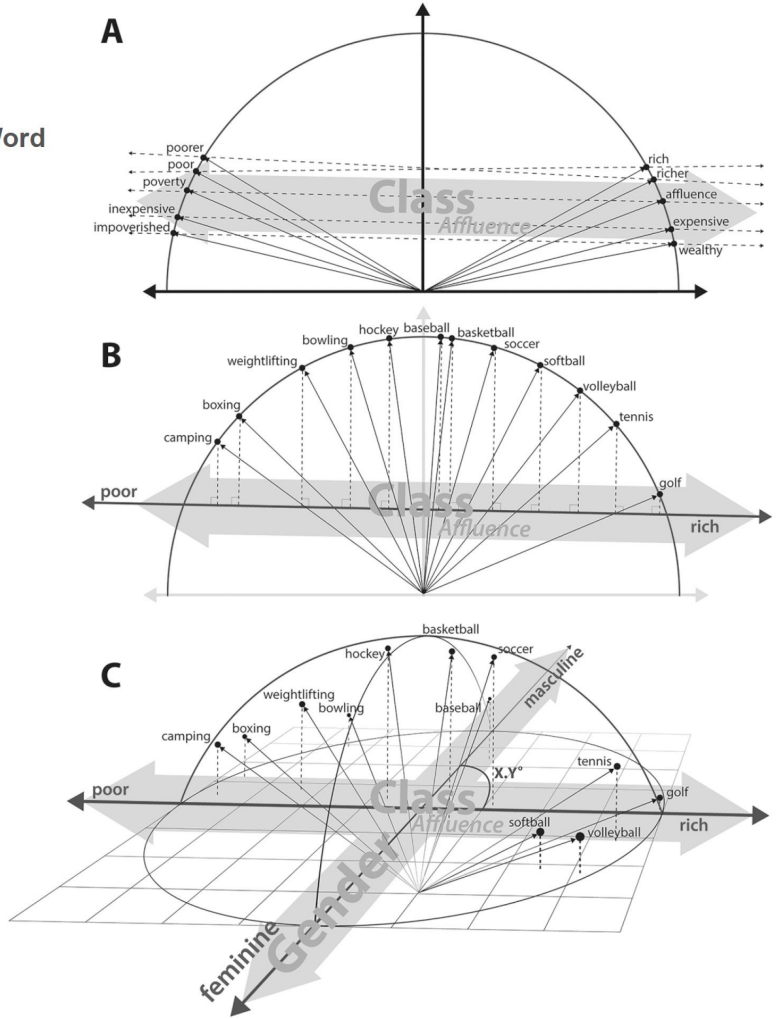


The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings

Austin C. Kozlowski, Matt Taddy, James A. Evans

First Published September 25, 2019 | Research Article | Check for updates
https://doi-org.stanford.idm.oclc.org/10.1177/0003122419877135



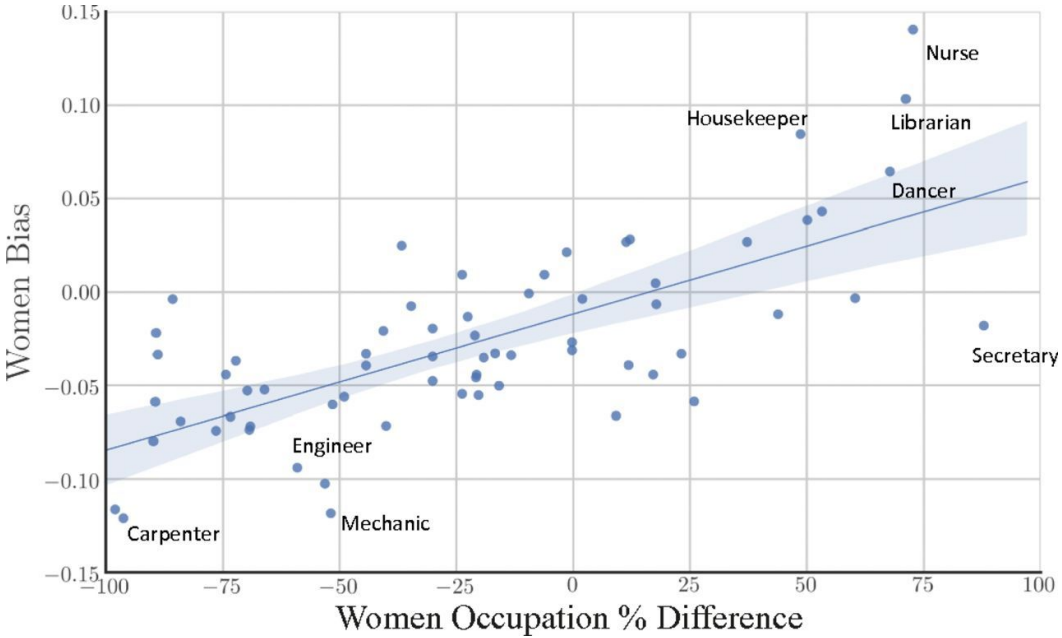


RESEARCH ARTICLE

Word embeddings quantify 100 years of gender and ethnic stereotypes

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou

PNAS April 17, 2018 115 (16) E3635-E3644; first published April 3, 2018
https://doi.org/10.1073/pnas.1720347115



	1910	1920	1930	1940	1950	1960	1970	1980	1990
1910	1.00	0.71	0.68	0.71	0.65	0.67	0.54	0.45	0.45
1920	0.71	1.00	0.71	0.68	0.65	0.66	0.55	0.51	0.46
1930	0.68	0.71	1.00	0.73	0.71	0.70	0.60	0.52	0.53
1940	0.71	0.68	0.73	1.00	0.74	0.69	0.56	0.53	0.51
1950	0.65	0.65	0.71	0.74	1.00	0.71	0.58	0.51	0.49
1960	0.67	0.66	0.70	0.69	0.71	1.00	0.62	0.54	0.50
1970	0.54	0.55	0.60	0.56	0.58	0.62	1.00	0.63	0.56
1980	0.45	0.51	0.52	0.53	0.51	0.54	0.63	1.00	0.62
1990	0.45	0.46	0.53	0.51	0.49	0.50	0.56	0.62	1.00

Height of women's movements in 1960s-70s

Table 2. View popup

Top adjectives associated with women in 1910, 1950, and 1990 by relative norm difference in the COHA embedding

1910	1950	1990
Charming	Delicate	Maternal
Placid	Sweet	Morbid
Delicate	Charming	Artificial
Passionate	Transparent	Physical
Sweet	Placid	Caring
Dreamy	Childish	Emotional
Indulgent	Soft	Protective
Playful	Colorless	Attractive
Mellow	Tasteless	Soft
Sentimental	Agreeable	Tidy

Some complications

There are additionally “Doc2Vec” algorithms and structural embedding models

Mittens: a “warm start” version of GloVe for small corpora

Contextualized word embeddings (e.g. ELMo and BERT) embed sentences, paragraphs, words, etc. simultaneously

SotA (to my knowledge) is BERT (with the caveat that ALBERT is a faster-to-train version)

There are also embeddings of nodes in networks (e.g. node2vec) as well as entire networks (e.g. GraphWave)

Some decisions to be made...

What algorithm to use (Word2Vec, GloVe, BERT, etc.)?

What corpus do you train on (local vs. large corpus)?

In many models, you have to select the size of the context (“window size”) and how many dimensions you want words represented in

(Some) publicly available materials

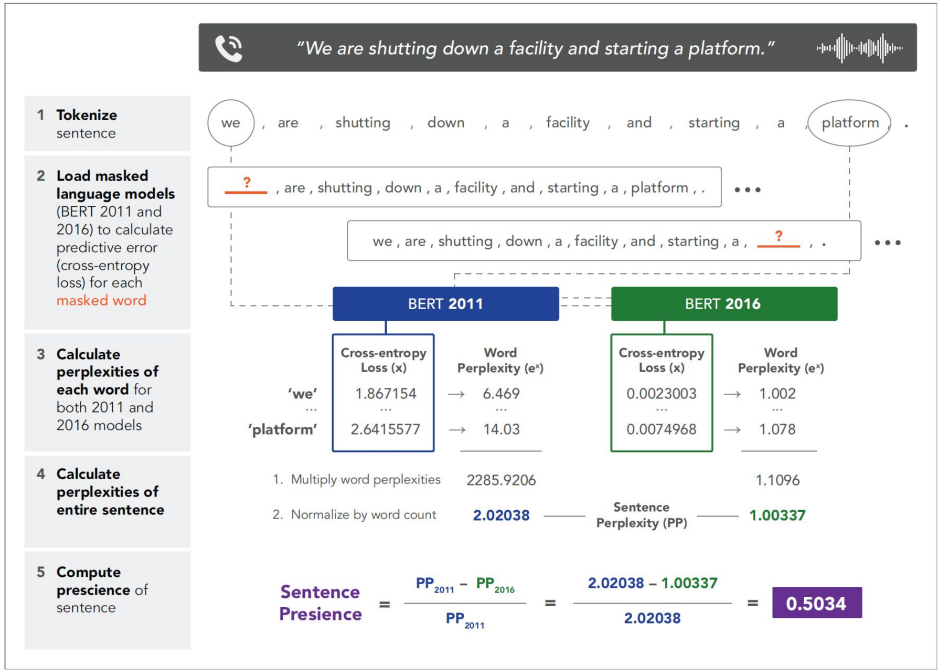
Pre-trained Word2Vec vectors are publicly available (https://code.google.com/archive/p/word2vec/) and can be easily accessed with the gensim library in Python

Pre-trained GloVe vectors trained on multiple corpora are also publicly available (https://nlp.stanford.edu/projects/glove/) as CSVs with words as the rows and dimensions as columns

WARNING: embeddings take up a non-trivial amount of storage on your machine and can take a couple of minutes to load into memory

Some Future Directions (Here at Stanford)

Using contextualized word embeddings to measure firm prescience



Some Future Directions (Here at Stanford)

Using word embeddings to interrogate the social meaning of age

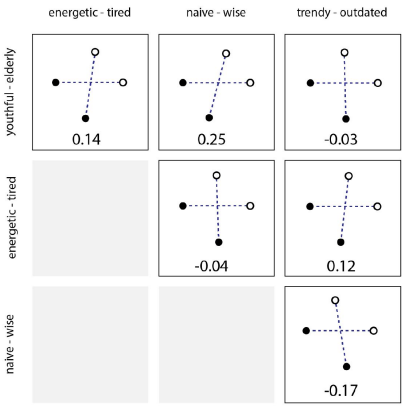


FIGURE 3: Angles between the difference vectors of four age-related word pairs in a 300-dimensional GloVe model trained on COCA. Closed circles represent the older-typed word, and open circles represent the younger-typed word. Numbers in each box represent the cosine similarity, describing the angle between the vectors that connect the two words, and can be interpreted as a marker of the strength of the association in meaning.

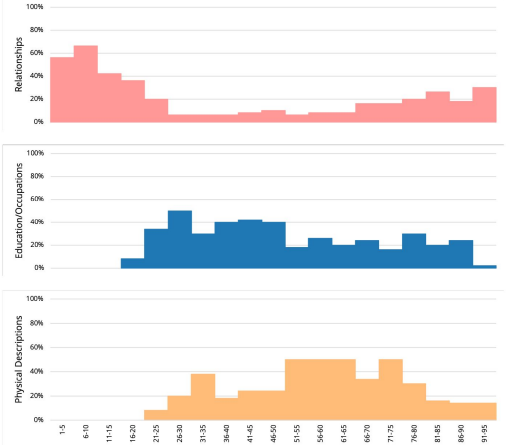
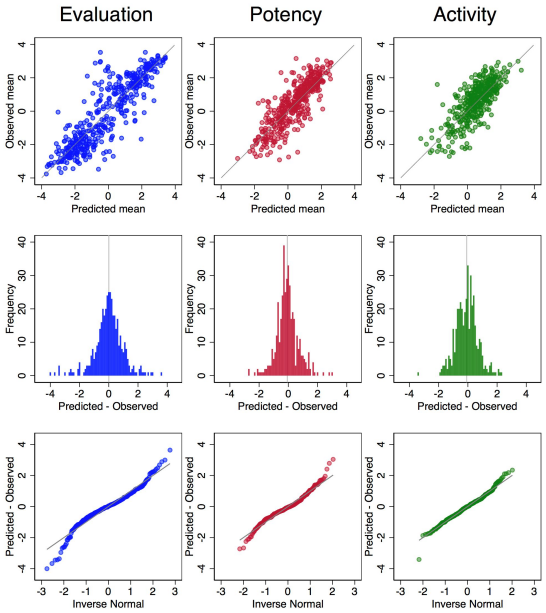


FIGURE 6: Percent of the 10 nearest neighbors to each years-old age word ("x-year-old") that pertain to three axes of meaning: words that describe relationships (e.g. "mother" or "divorced"); education/occupation (e.g. "homemaker," "retired," "sophomore", or "Harvard-educated"); and physical descriptions (e.g. "gray-haired" or "grizzled"). I display rates of nearest neighbors in buckets of 5-year years-old age intervals. See Supp. Fig 7 and Supp. Table 1 for full word list and coding beyond these three categories.

Some Future Directions (Here at Stanford)

Using word embeddings to replace survey responses for the affective meaning of concepts (Code: https://github.com/AustinVL/TowardsAMoreComputationalACT/blob/master/AnalysisScripts/first%20analysis/AnalysisScript.ipynb)



Thank you!

Survey Links

KSS Survey:

https://bit.ly/ctasskss

Lightning Talk sign-up:

https://bit.ly/ctasslightning

RSVP for Jen Pan’s 1/24 talk:

https://bit.ly/ctassjenpan