Stony Brook University

Word Embedding in Text Analysis

Yongjun Zhang, Ph.D.
Dept of Sociology and IACS
https://yongjunzhang.com

Today's Agenda

1. Guest Speaker Dr. Yongren Shi (6:00-6:50PM)
2. Mini-Lecture on Word Embedding (7:00-7:50 PM)
3. Lab Tutorial on Word Embedding and Other Text Analysis (8:00-8:50PM)

Dr. Yongren Shi
Ass. Prof. of Soc and Criminology
@IOWA

Brief Summary on Text Analysis and Topic Modeling

Why should we care about text as data?

Texts are the conduits to access the minds of *actors* in specific social contexts.

Politicians, corporate elites, SMOs, etc.

In general, what kind of research can benefit from text analysis?

-Align the texts with a dimension of interest

- -Supervised learning; Dictionary methods; Topic modeling (bag-of-words approach)
- -Examples: Nationalist sentiment of politicians; Deviant behavior of community members; Classifications

-Explore the complex motives, behavior, or biases underlying the texts

- -Word2vec; Text networks; Topic modeling (relational approach)

Procedures of quantitative text analysis

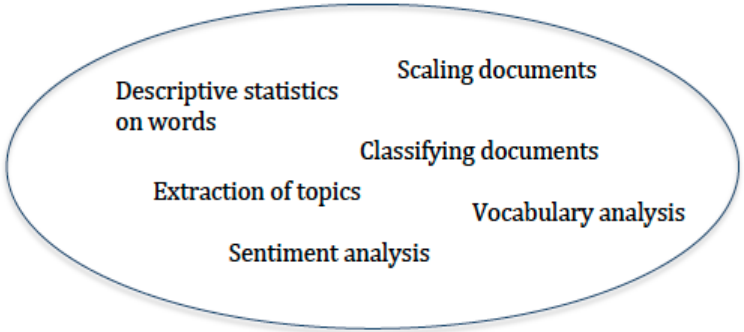
1. *Selecting texts*: Defining the corpus
2. *Conversion* of texts into a common electronic format
3. *Defining documents*: deciding what will be the documentary unit of analysis
4. *Defining features*. These can take a variety of forms, including tokens, equivalence classes of tokens (dictionaries), selected phrases, human-coded segments (of possibly variable length), linguistic features, and more.
5. *Conversion of textual features into a quantitative matrix*
6. A *quantitative or statistical procedure* to extract information from the quantitative matrix
7. *Summary* and interpretation of the quantitative results

When I presented the supplementary budget to this House last April, I said we could work our way through this period of severe economic distress. Today, I can report that notwithstanding the difficulties of the past eight months, we are now on the road to economic recovery.

In this next phase of the Government's plan we must stabilise the deficit in a fair way, safeguard those worst hit by the recession, and stimulate crucial sectors of our economy to sustain and create jobs. The worst is over.

This Government has the moral authority and the well-grounded optimism rather than the cynicism of the Opposition. It has the imagination to create the new jobs in energy, agriculture, transport and construction that this green budget will

	words										
docs	made	because	had	into	get	some	through	next	where	many	irish
t06_kenny_fg	12	11	5	4	8	4	3	4	5	7	10
t05_cowen_ff	9	4	8	5	5	5	14	13	4	9	8
t14_ocaolain_sf	3	3	3	4	7	3	7	2	3	5	6
t01_lenihan_ff	12	1	5	4	2	11	9	16	14	6	9
t11_gormley_green	0	0	0	3	0	2	0	3	1	1	2
t04_morgan_sf	11	8	7	15	8	19	6	5	3	6	6
t12_ryan_green	2	2	3	7	0	3	0	1	6	0	0
t10_quinn_lab	1	4	4	2	8	4	1	0	1	2	0
t07_odonnell_fg	5	4	2	1	5	0	1	1	0	3	0
t09_higgins_lab	2	2	5	4	0	1	0	0	2	0	0
t03_burton_lab	4	8	12	10	5	5	4	5	8	15	8
t13_cuffe_green	1	2	0	0	11	0	16	3	0	3	1
t08_gilmore_lab	4	8	7	4	3	6	4	5	1	2	11
t02_bruton_fg	1	10	6	4	4	3	0	6	16	5	3



Text Preprocessing: Remove Stopwords

a, able, about, across, after, all, almost, also, am, among, an, and, any, are, as, at, be, because, been, but, by, can, cannot, could, dear, did, do, does, either, else, ever, every, for, from, get, got, had, has, have, he, her, hers, him, his, how, however, I, if, in, into, is, it, its, just, least, let, like, likely, may, me, might, most, must, my, neither, no, nor, not, of, off, often, on, only, or, other, our, own, rather, said, say, says, she, should, since, so, some, than, that, the, their, them, then, there, these, they, this, tis, to, too, twas, us, wants, was, we, were, what, when, where, which, while, who, whom, why, will, with, would, yet, you, your

A List of Common English Stop Words

Text Preprocessing: stemming and lemmatization

Stemming the process for reducing inflected (or sometimes derived) words to their stem, base or root form. Different from lemmatization in that stemmers operate on single words without knowledge of the context.

Form	Suffix	Stem
studies	-es	studi
studying	-ing	study
niñas	-as	niñ
niñez	-ez	niñ

Lemmatization takes into consideration the morphological analysis of the words.

Form	Morphological information	Lemma
studies	Third person, singular number, present tense of the verb study	study
studying	Gerund of the verb study	study
niñas	Feminine gender, plural number of the noun niño	niño
niñez	Singular number of the noun niñez	niñez

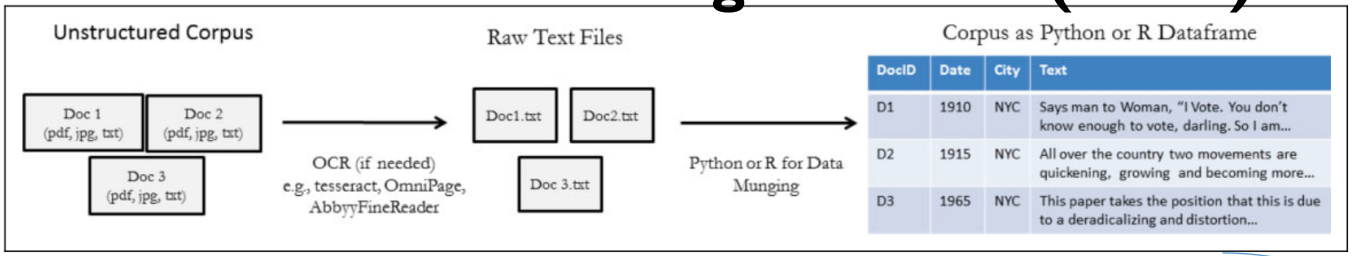
Issues with stemming approaches

The most common is probably the Porter stemmer

But this set of rules gets many stems wrong, e.g.

1. policy and police considered (wrongly) equivalent
2. general becomes gener, iteration becomes iter

Text Transformation--Bag of Words (BoW)



DocID	Date	City	Text
D1	1910	NYC	Says man to Woman, "I Vote. You don't know enough to vote, darling. So I am...
D2	1915	NYC	All over the country two movements are quickening, growing and becoming more...
D3	1965	NYC	This paper takes the position that this is due to a deradicalizing and distortion...

Figure 1. Corpus construction: From text to dataframe. This figure demonstrates a possible path from a collection of texts, saved in separate files, to a digital dataframe suitable for further computer-assisted text analysis techniques. Often historical texts are in the form of pdf or jpg images and thus require an intervening step using optical character recognition software. More contemporary texts are already digitized. Once digitized, the researcher can use Python or R to transform the separate files into one dataframe, with metadata attached to each text (in this example, date of publication and the city in which it was published).

Note: Nelson 2017

Construct a Document-Term Matrix (sparse)

DocID	Text	Term 1 (say)	Term 3 (all)
D1	Says man…	X tf=X/N1	0
D2	All over…	0	Y tf=Y/N2

Descriptive Statistical Methods for textual analysis

term frequency Some approaches trim very low-frequency words. Rationale: get rid of rare words that expand the feature matrix but matter little to substantive analysis

document frequency Could eliminate words appearing in few documents

inverse document frequency Conversely, could weight words more that appear in the most documents

tf-idf a combination of term frequency and inverse document frequency, common method for feature weighting

More on $tf-idf$

$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$ where $n_{i,j}$ is the number of occurrence of term t_i in document d_j, k is all the terms in document d_j.

$$idf_i = ln \frac{|D|}{|\{d_j : t_i \in d_j\}|}$$

Where:

$|D|$ is the total number of documents in the set.

$|\{d_j : t_i \in d_j\}|$ is the number of documents where the term t_i appears (i.e. $n_{i,j} \neq 0$)

$tf-idf_i = tf_{i,j} * idf_i$

Example:

We have 100 NYT news articles FROM dynamics of collective action, each with 1000 words. The first document contains 16 instances of the word "activism"; 40 of the news article contain the word "activism".

The term frequency is 16/1000 = 0.016

The inverse document frequency is 100/40 = 2.5, and ln(2.5) = 0.916

The tf-idf will then be 0.016 * 0.916 = 0.0147

If the word had only appeared in 15 of the 100 news articles, then the tf-idf would be 0.0304 (three times higher).

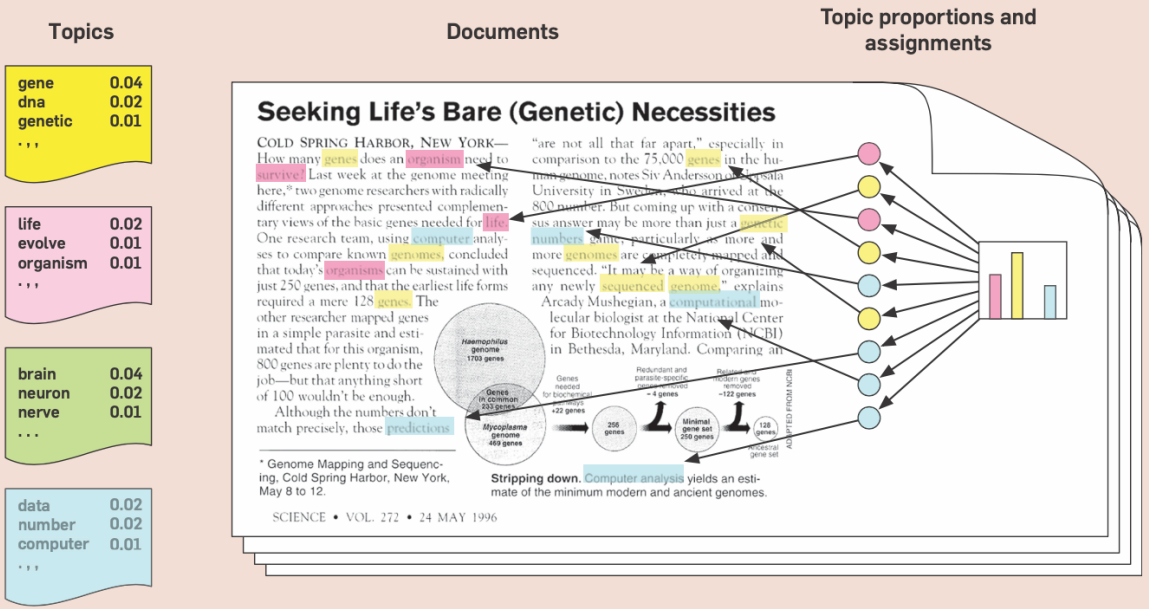
A high weight in $tf - idf$ is reached by a high term frequency (in the given document) and a low document frequency of the term in the whole collection of documents; hence the weights hence tend to filter out common terms

Topic Modeling – A non-tech Intro to LDA

1. Each document (text) within a corpus is viewed as a *bag-of-words* produced according to a mixture of themes that the author of the text intended to discuss.
2. Each theme (or topic) is a distribution over all observed words in the corpus, such that words that are strongly associated with the document's dominant topics have a higher chance of being selected and placed in the document bag.
3. The objective of topic modeling is to find the parameters of the LDA process that has likely generated the corpus.

Mohr and Bogdanov 2013

Figure 1. The intuitions behind latent Dirichlet allocation. We assume that some number of "topics," which are distributions over words, exist for the whole collection (far left). Each document is assumed to be generated as follows. First choose a distribution over the topics (the histogram at right); then, for each word, choose a topic assignment (the colored coins) and choose the word from the corresponding topic. The topics and topic assignments in this figure are illustrative—they are not fit from real data. See Figure 2 for topics fit from data.



We formally define a *topic* to be a distribution over a fixed vocabulary. For example, the *genetics* topic has words about genetics with high probability and the *evolutionary biology* topic has words about evolutionary biology with high probability. We assume that these topics are specified before any data has been generated.[a] Now for each document in the collection, we generate the words in a two-stage process.

- Randomly choose a distribution over topics.
- For each word in the document
  a. Randomly choose a topic from the distribution over topics in step #1.
  b. Randomly choose a word from the corresponding distribution over the vocabulary.

This statistical model reflects the intuition that documents exhibit multiple topics. Each document exhibits the topics in different proportion (step #1); each word in each document is drawn from one of the topics (step #2b), where the selected topic is chosen from the per-document distribution over topics (step #2a).[b]

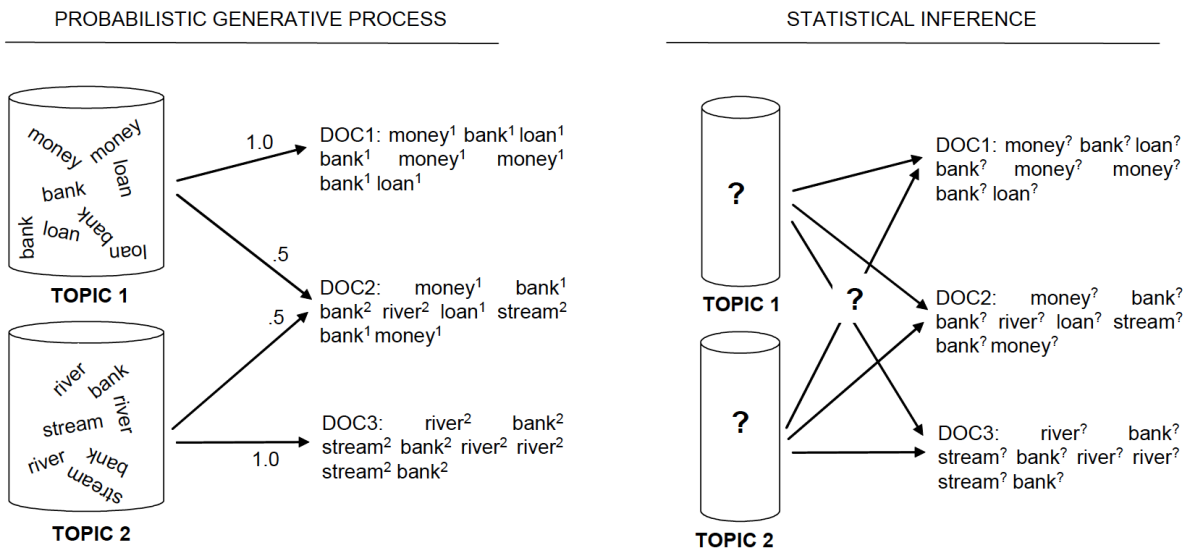


Figure 2. Illustration of the generative process and the problem of statistical inference underlying topic models

(from Steyvers and Griffiths 2007)

Key parameters:

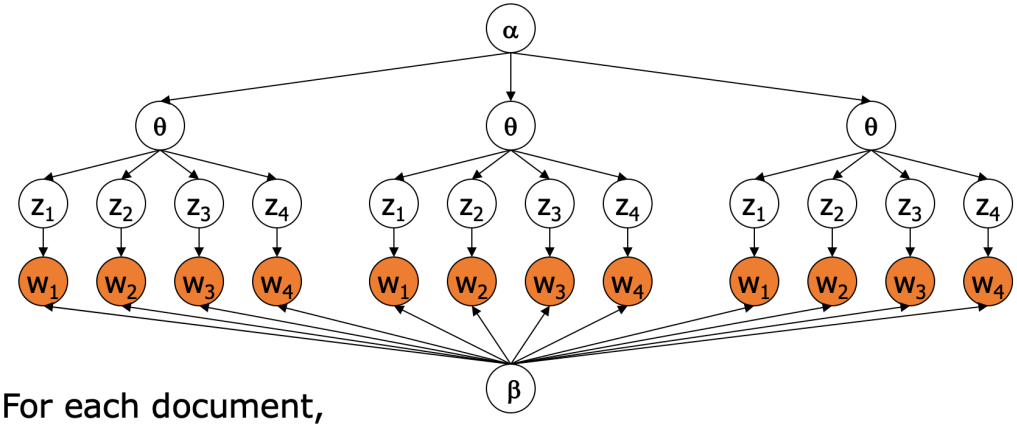
1. θ = matrix of dimensions N documents by K topics where θ_{ik} corresponds to the probability that document i belongs to topic k; i.e. assuming $K = 5$:

	T1	T2	T3	T4	T5
Document 1	0.15	0.15	0.05	0.10	0.55
Document 2	0.80	0.02	0.02	0.10	0.06
...					
Document N	0.01	0.01	0.96	0.01	0.01

2. β = matrix of dimensions K topics by M words where β_{km} corresponds to the probability that word m belongs to topic k; i.e. assuming $M = 6$:

	W1	W2	W3	W4	W5	W6
Topic 1	0.40	0.05	0.05	0.10	0.10	0.30
Topic 2	0.10	0.10	0.10	0.50	0.10	0.10
...						
Topic k	0.05	0.60	0.10	0.05	0.10	0.10

The LDA Model



- For each document,
- Choose θ~Dirichlet(α)
- For each of the N words wn:
  - Choose a topic z_n» Multinomial(θ)
  - Choose a word w_n from $p(w_n|z_n,\beta)$, a multinomial probability conditioned on the topic z_n.

$$f\left(x_1,\ldots,x_K;\alpha_1,\ldots,\alpha_K\right)=\frac{1}{\mathrm{B}(\boldsymbol{\alpha})}\prod_{i=1}^{K}x_i^{\alpha_i-1}$$

Structural Topic Model (STM)

The goal of the Structural Topic Model is to allow researchers to discover topics and estimate their relationship to document metadata.

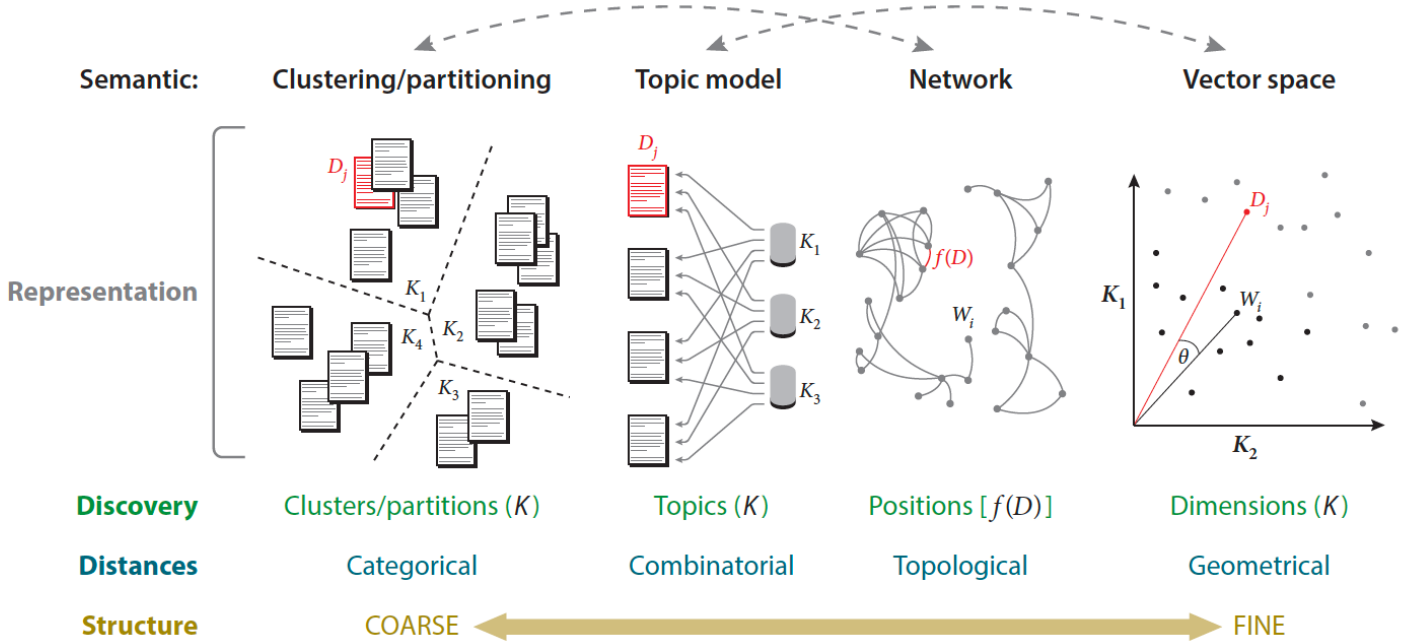
Topic prevalence: the proportion of a document devoted to a topic; *topical content:* the word rates used in discussing a topic.

Three critical differences between STM and LDA:

1. topics can be correlated;
2. each document has its own prior distribution over topics, defined by covariate *X* rather than sharing a global mean;
3. word use within a topic can vary by covariate *U* .

You can check structuraltopicmodel.com for more details.

Roberts et al. 2014 2016



Evans and Aceves 2016. Machine Translation: Mining Text for Social Theory. *Annual Review of Sociology*

Word Embeddings

How to Represent Texts?

Atomic Word Representation (One-hot coding)

apple	[0 0 0 0 0 **1** 0 0 0 0 0 0 0 0 0 0 0 ... 0 0 0 0 0 0]
orange	[0 0 0 0 0 0 0 0 0 0 0 0 0 **1** 0 0 0 0 ... 0 0 0 0 0 0]
car	[0 0 0 0 0 0 0 **1** 0 0 0 0 0 0 0 0 0 0 ... 0 0 0 0 0 0]

How to Represent Meaning in Text?

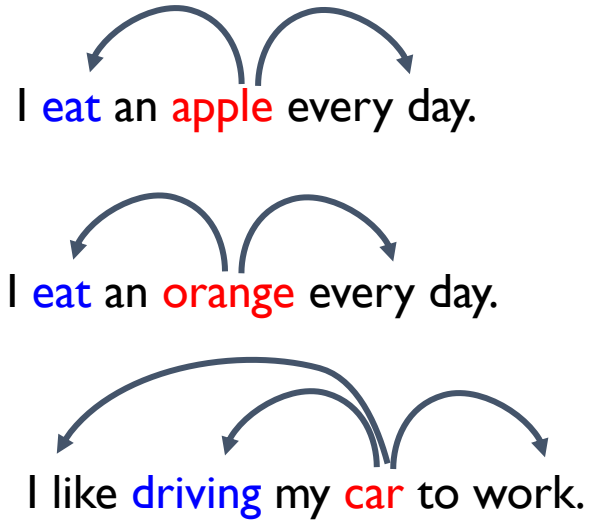
Distributional similarity-based representations

You can get a lot of value by representing a word by means of its neighbors

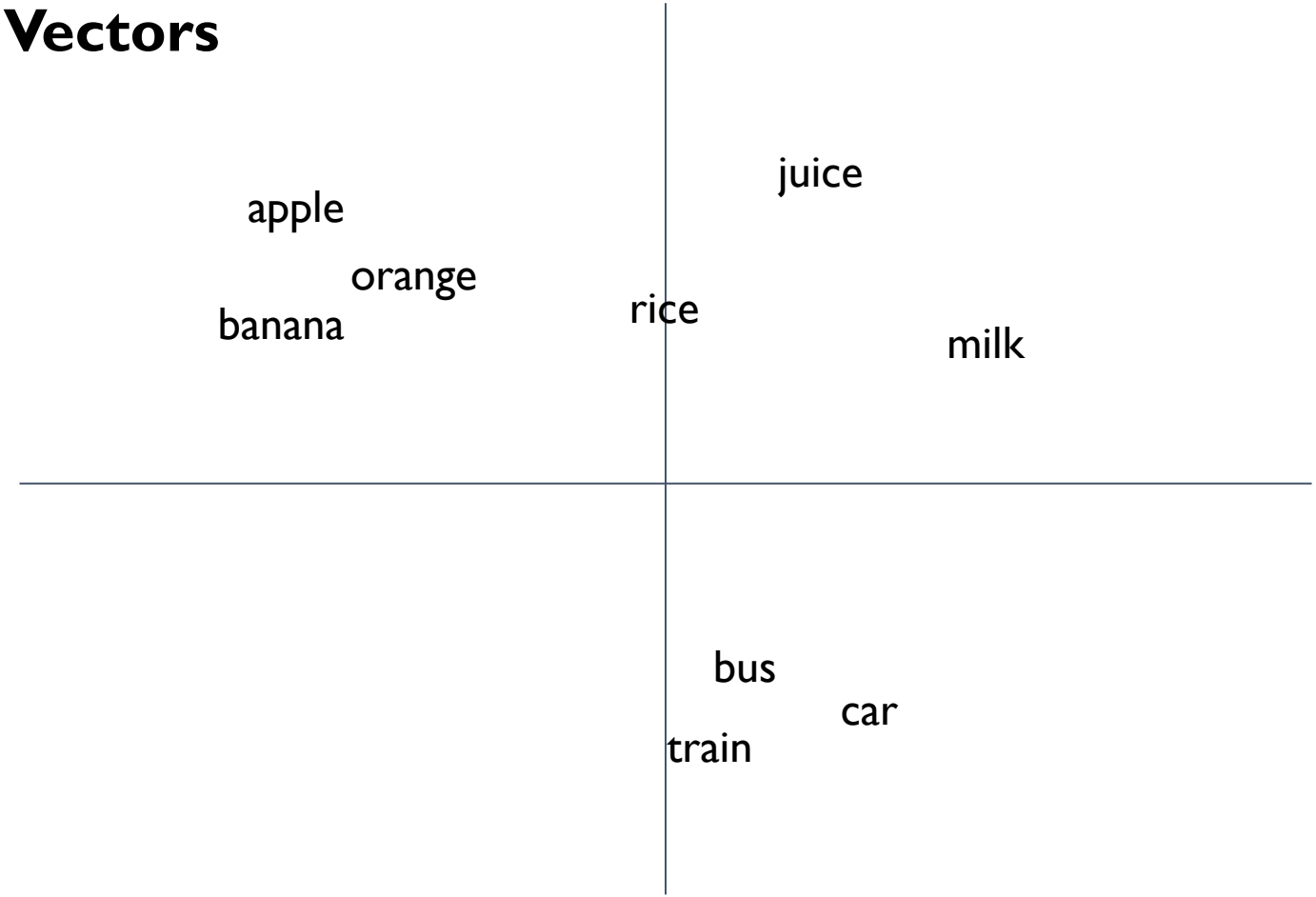
"You shall know a word by the company it keeps"
(J.R. Firth 1957)

Contextual Representation

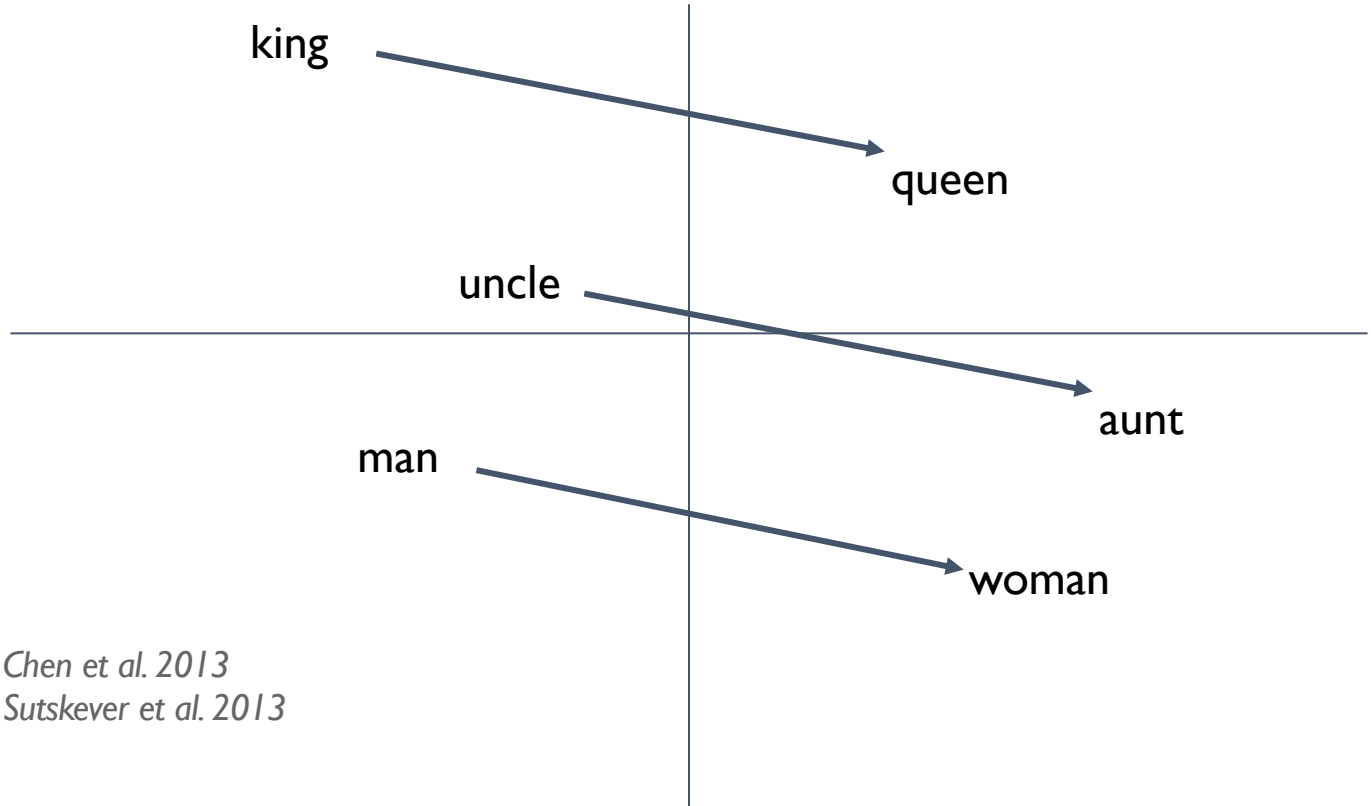
Word is represented by context in use



Word Vectors



Word Analogy



Mikolov & Chen et al. 2013
Mikolov & Sutskever et al. 2013

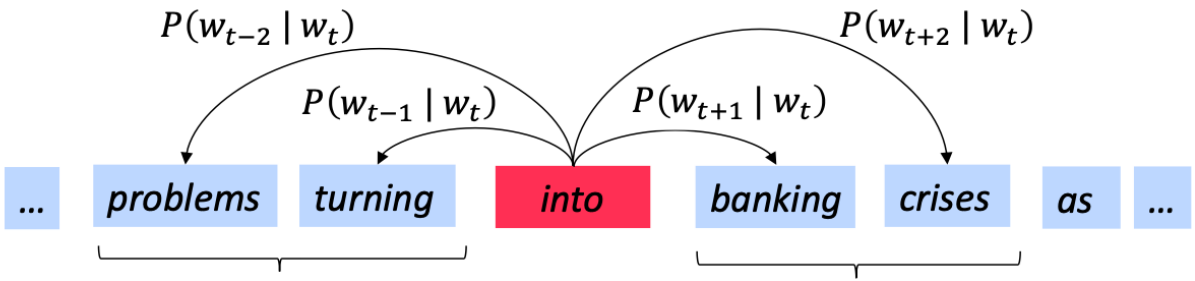
Google Books

word2vec

	D1	D2	D3	D4
W1	.02	.03	.5	.45
W2				
W3				
W4				

Basic Ideas of learning neural network word embeddings

We define a model that aims to predict between a center word and context words in terms of word vectors:



- $P(o|c) = \frac{\exp(u_o^T v_c)}{\sum_{w \in V} \exp(u_w^T v_c)}$
- Update vectors so you can predict better

Word Embeddings

Two Most Important Articles on word2vec by Mikolov et al. 2013

Efficient Estimation of Word Representations in Vector Space

Tomas Mikolov
Google Inc., Mountain View, CA
tmikolov@google.com

Kai Chen
Google Inc., Mountain View, CA
kaichen@google.com

Greg Corrado
Google Inc., Mountain View, CA
gcorrado@google.com

Jeffrey Dean
Google Inc., Mountain View, CA
jeff@google.com

Abstract

We propose two novel model architectures for computing continuous vector representations of words from very large data sets. The quality of these representations is measured in a word similarity task, and the results are compared to the previously best performing techniques based on different types of neural networks. We observe large improvements in accuracy at much lower computational cost, i.e. it takes less than a day to learn high quality word vectors from a 1.6 billion words data set. Furthermore, we show that these vectors provide state-of-the-art performance on our test set for measuring syntactic and semantic word similarities.

Distributed Representations of Words and Phrases and their Compositionality

Tomas Mikolov
Google Inc.
Mountain View
mikolov@google.com

Ilya Sutskever
Google Inc.
Mountain View
ilyasu@google.com

Kai Chen
Google Inc.
Mountain View
kai@google.com

Greg Corrado
Google Inc.
Mountain View
gcorrado@google.com

Jeffrey Dean
Google Inc.
Mountain View
jeff@google.com

Abstract

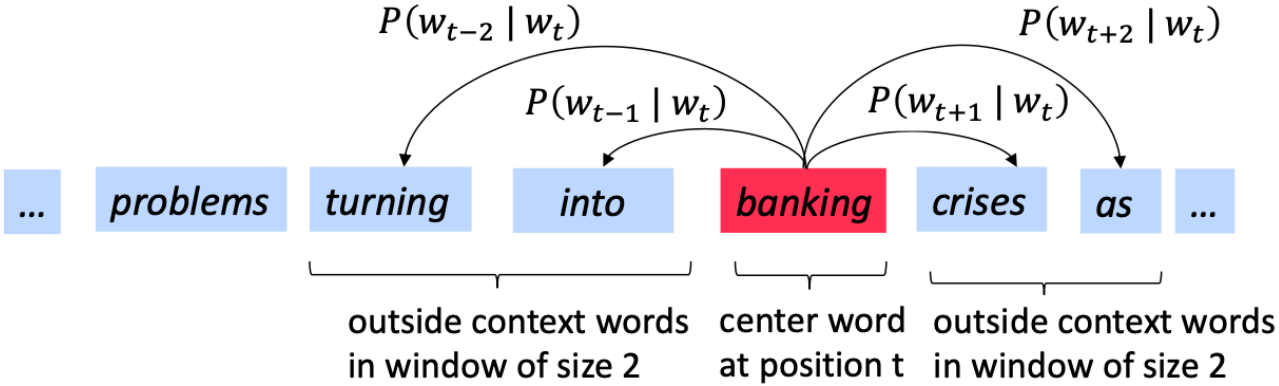
The recently introduced continuous Skip-gram model is an efficient method for learning high-quality distributed vector representations that capture a large number of precise syntactic and semantic word relationships. In this paper we present several extensions that improve both the quality of the vectors and the training speed. By subsampling of the frequent words we obtain significant speedup and also learn more regular word representations. We also describe a simple alternative to the hierarchical softmax called negative sampling.

Word2vec (Mikolov et al. 2013) is a framework for learning word vectors

Idea:

- We have a large corpus of text
- Every word in a fixed vocabulary is represented by a vector
- Go through each position *t* in the text, which has a center word *c* and context ("outside") words *o*
- Use the similarity of the word vectors for *c* and *o* to calculate the probability of *o* given *c* (or vice versa)
- Keep adjusting the word vectors to maximize this probability

Manning, CS224n, 2020



For each position $t = 1, \ldots, T$, predict context words within a window of fixed size m, given center word w_j.

Likelihood = $L(\theta) = \prod_{t=1}^{T} \prod_{\substack{-m \le j \le m \\ j \neq 0}} P(w_{t+j} \mid w_t; \theta)$

θ is all variables to be optimized

sometimes called *cost* or *loss* function

$$P(o|c) = \frac{\exp(u_o^T v_c)}{\sum_{w \in V} \exp(u_w^T v_c)}$$

The objective function $J(\theta)$ is the (average) negative log likelihood:

$$J(\theta) = -\frac{1}{T} \log L(\theta) = -\frac{1}{T} \sum_{t=1}^{T} \sum_{\substack{-m \le j \le m \\ j \neq 0}} \log P(w_{t+j} \mid w_t; \theta)$$

Manning, CS224n, 2020

② Exponentiation makes anything positive

① Dot product compares similarity of *o* and *c*.
$u^T v = u.v = \sum_{i=1}^{n} u_i v_i$
Larger dot product = larger probability

$$P(o|c) = \frac{\exp(u_o^T v_c)}{\sum_{w \in V} \exp(u_w^T v_c)}$$

③ Normalize over entire vocabulary to give probability distribution

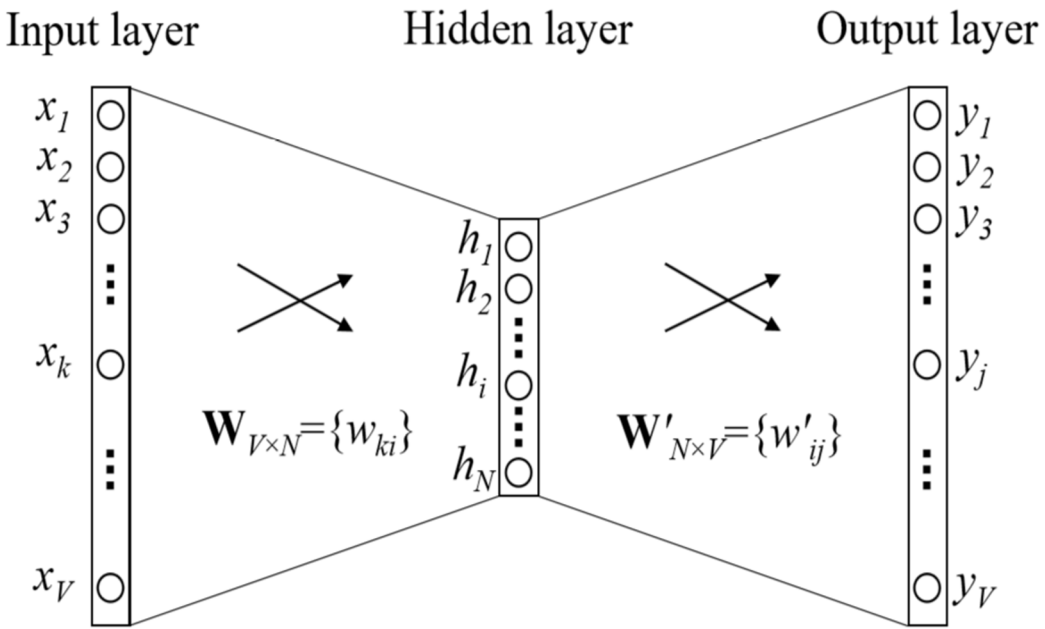
- This is an example of the **softmax function** $\mathbb{R}^n \rightarrow (0,1)^n$

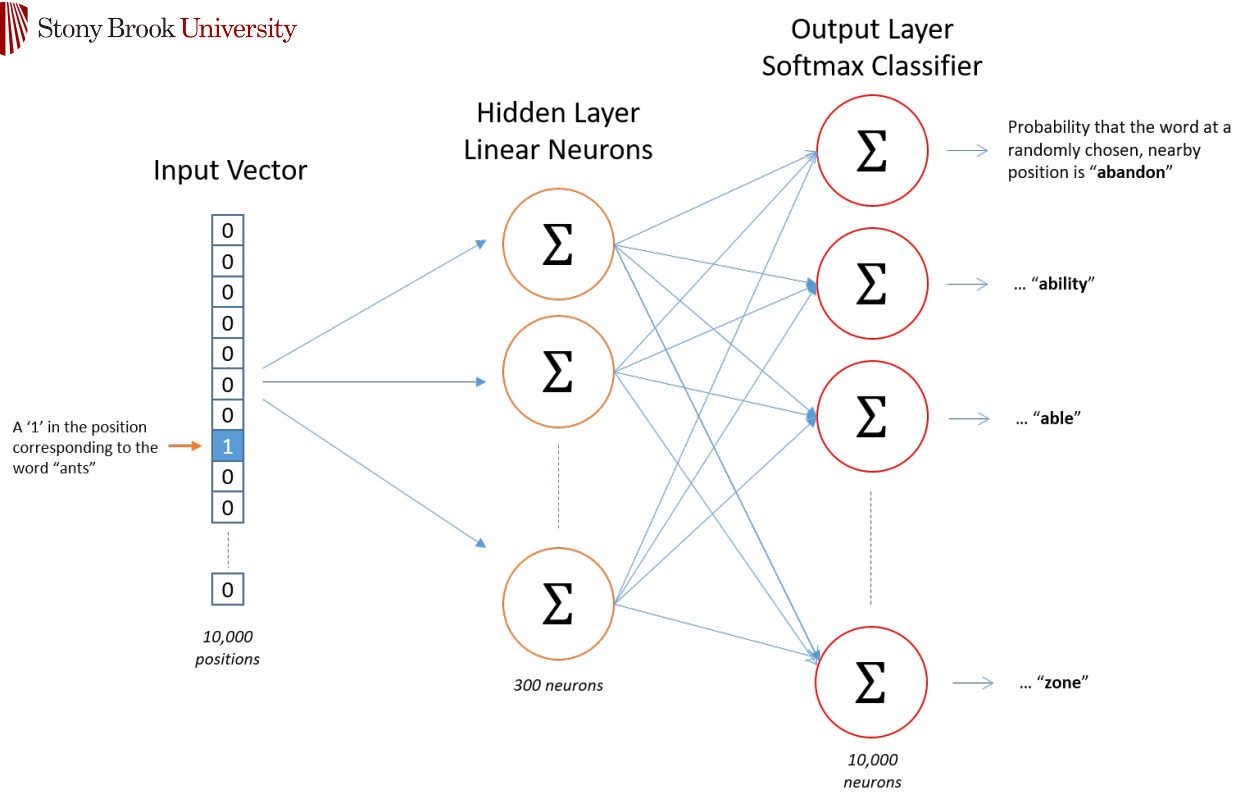
$$\text{softmax}(x_i) = \frac{\exp(x_i)}{\sum_{j=1}^{n} \exp(x_j)} = p_i$$

Open region

- The softmax function maps arbitrary values x_i to a probability distribution p_i
  - "max" because amplifies probability of largest x_i
  - "soft" because still assigns some probability to smaller x_i
  - Frequently used in Deep Learning

Manning, CS224n, 2020

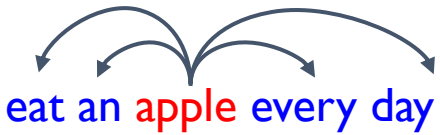

Input layer
Hidden layer
Output layer
x_1
x_2
x_3
x_k
x_V
h_1
h_2
h_i
h_N
y_1
y_2
y_3
y_j
y_V
$\mathbf{W}_{V\times N}=\{w_{ki}\}$
$\mathbf{W}'_{N\times V}=\{w'_{ij}\}$



When training this network on word pairs, the input is a one-hot vector representing the input word and the training output is also a one-hot vector representing the output word. But when you evaluate the trained network on an input word, the output vector will actually be a probability distribution (i.e., a bunch of floating point values, not a one-hot vector).
http://mccormickml.com/

Method 1: continuous bag-of-word (CBOW)

Method 2: skip-gram (SG)



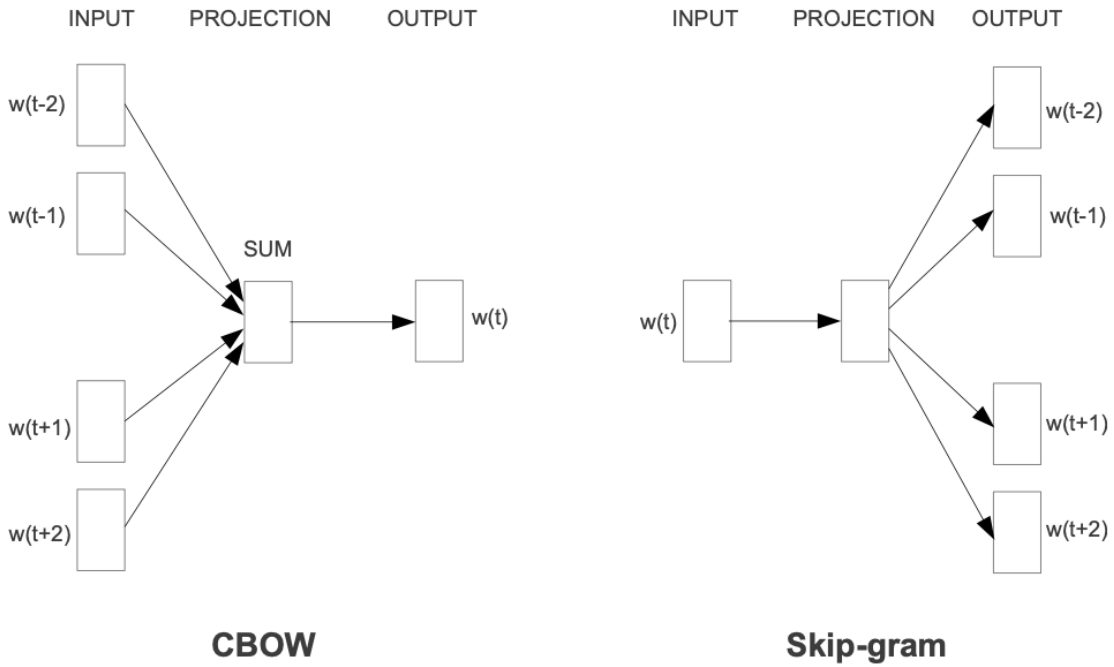
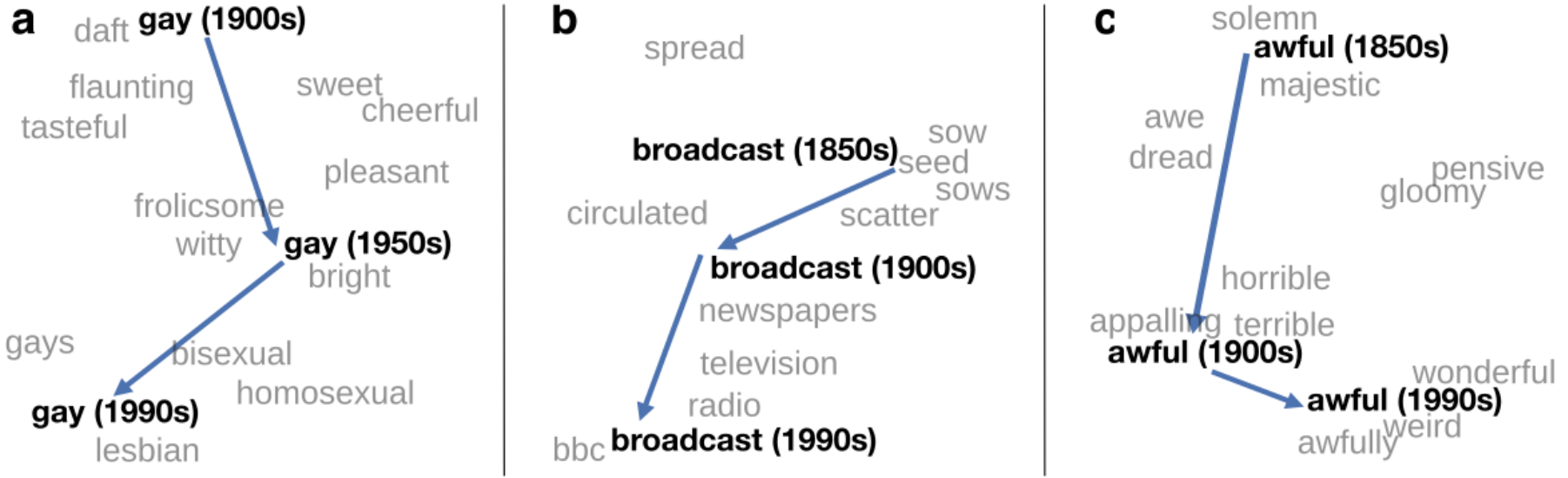


Figure 1: New model architectures. The CBOW architecture predicts the current word based on the context, and the Skip-gram predicts surrounding words given the current word.

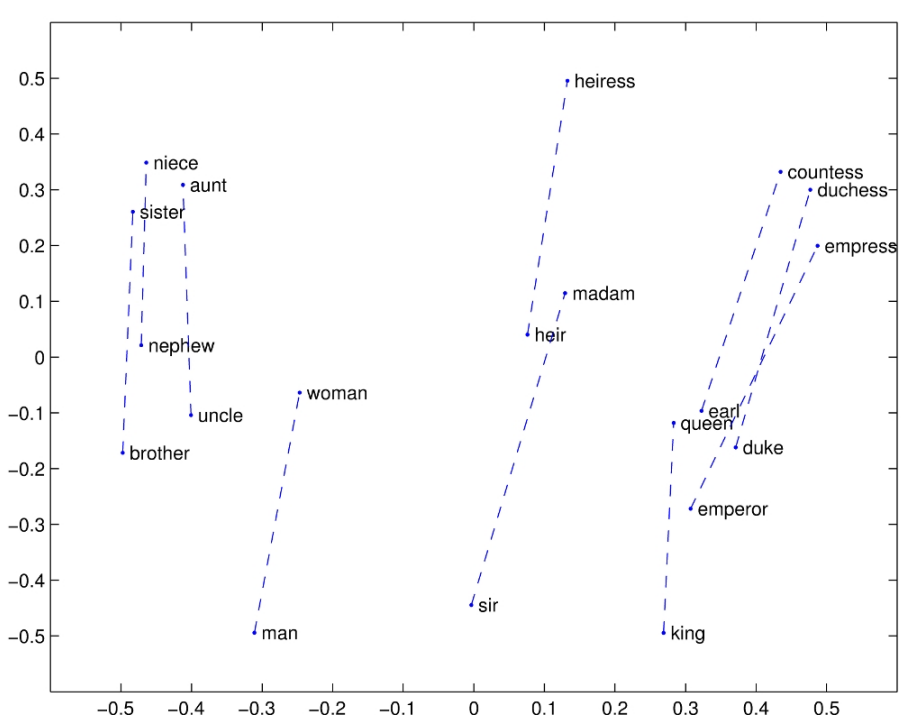


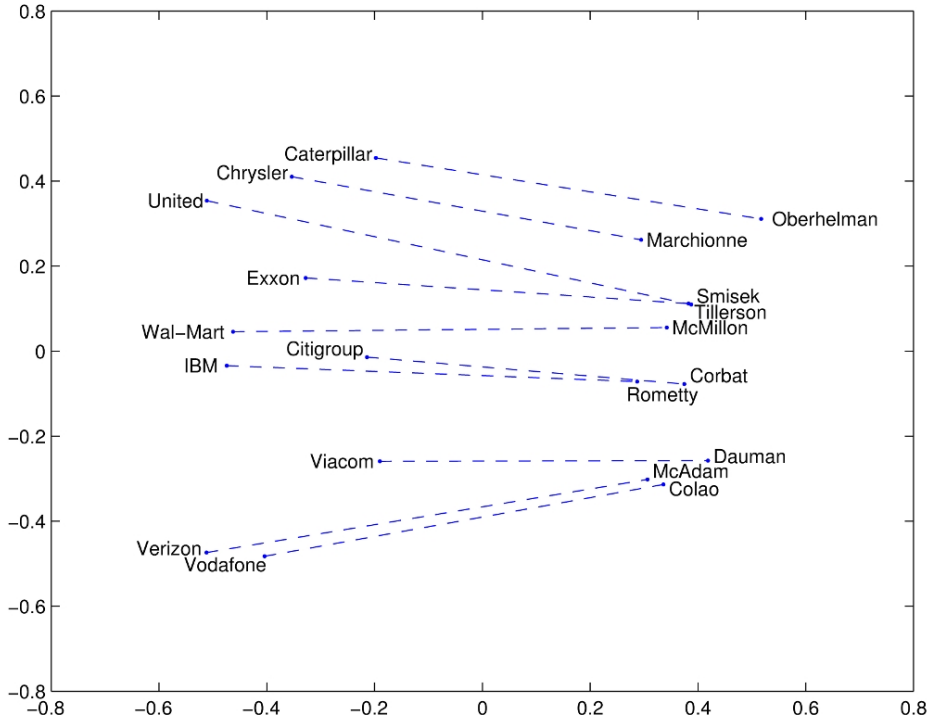
Changes in word meaning can be visualized by projecting historical word vectors into a 2-D space. Gay shifted in meaning over the last century, from meaning "showy" or "cheerful" to denoting "homosexuality".

Broadcast used to refer to the act of throwing seeds, but then this motion became associated with the throwing of newspapers, and eventually broadcast developed its current meaning of "disseminating information."

Awful underwent a process known as pejoration; it used to literally mean "full of awe", but over time it became more negative and now signifies that something is "upsetting."

https://nlp.stanford.edu/projects/histwords/


niece
aunt
sister
nephew
uncle
brother
woman
man
heiress
madam
heir
sir
countess
duchess
empress
earl
queen
duke
emperor
king


Caterpillar
Chrysler
United
Oberhelman
Marchionne
Exxon
Smisek
Tillerson
Wal−Mart
McMillon
Citigroup
IBM
Corbat
Rometty
Viacom
Dauman
McAdam
Colao
Verizon
Vodafone

wevi: word embedding visual inspector：

https://ronxin.github.io/wevi/

Some Pre-trained Models

Google Word2Vec

https://code.google.com/archive/p/word2vec/

Word2Vec Tutorial - The Skip-Gram Model

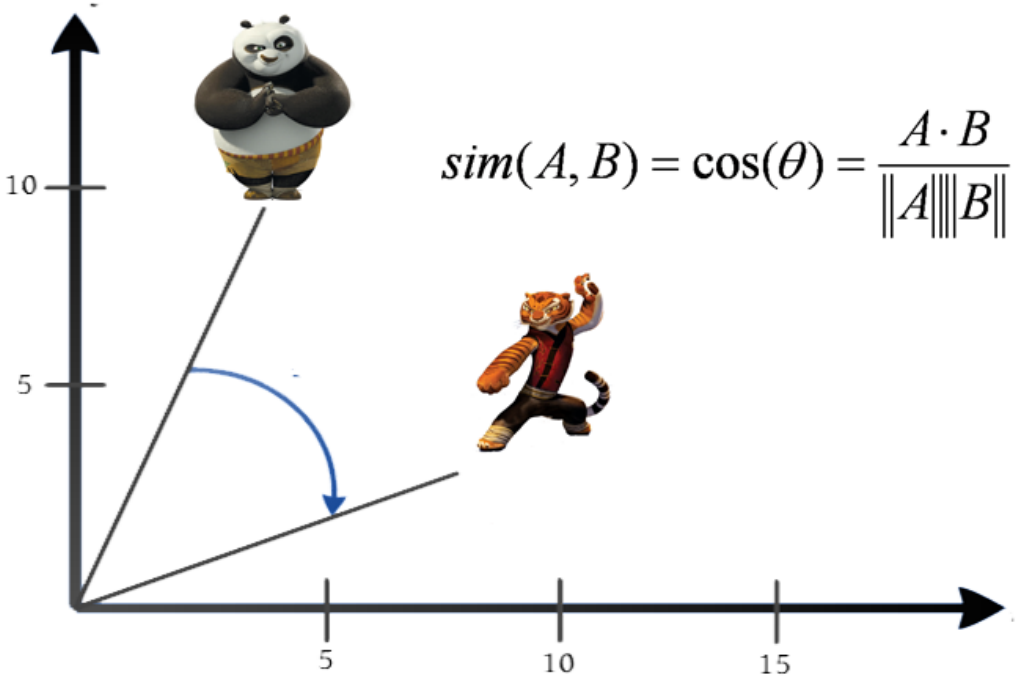
http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

Improvements and pre-trained models for word2vec:

https://nlp.stanford.edu/projects/glove/

https://fasttext.cc/ (by Facebook)

Cosine Similarity



$$sim(A,B) = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$

Some Examples of Using Word Embeddings

Garg et al. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *PNAS*.

Gender and ethnic stereotypes.

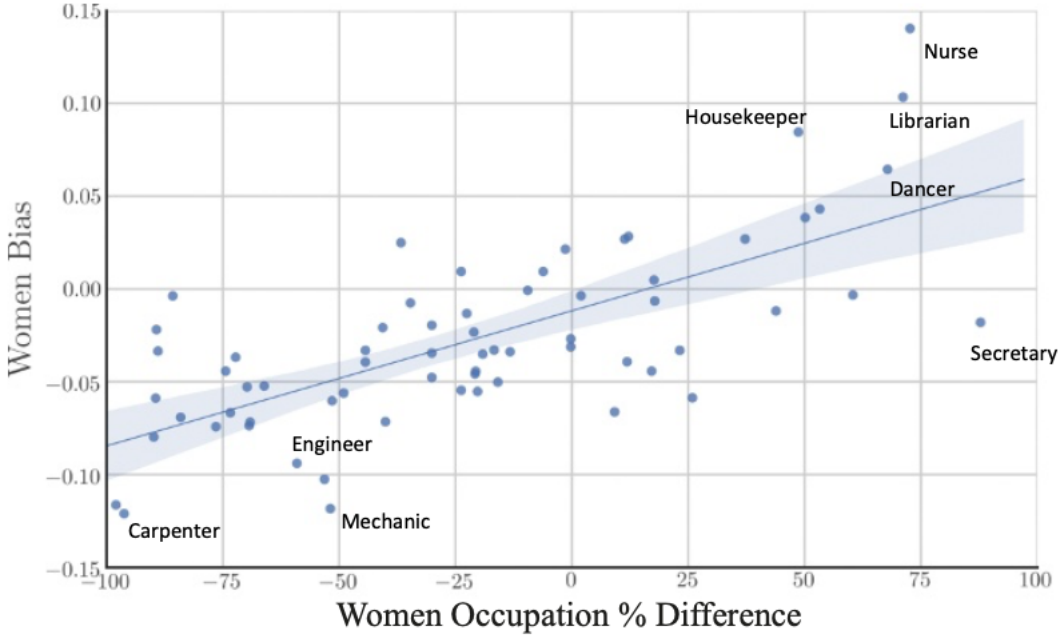
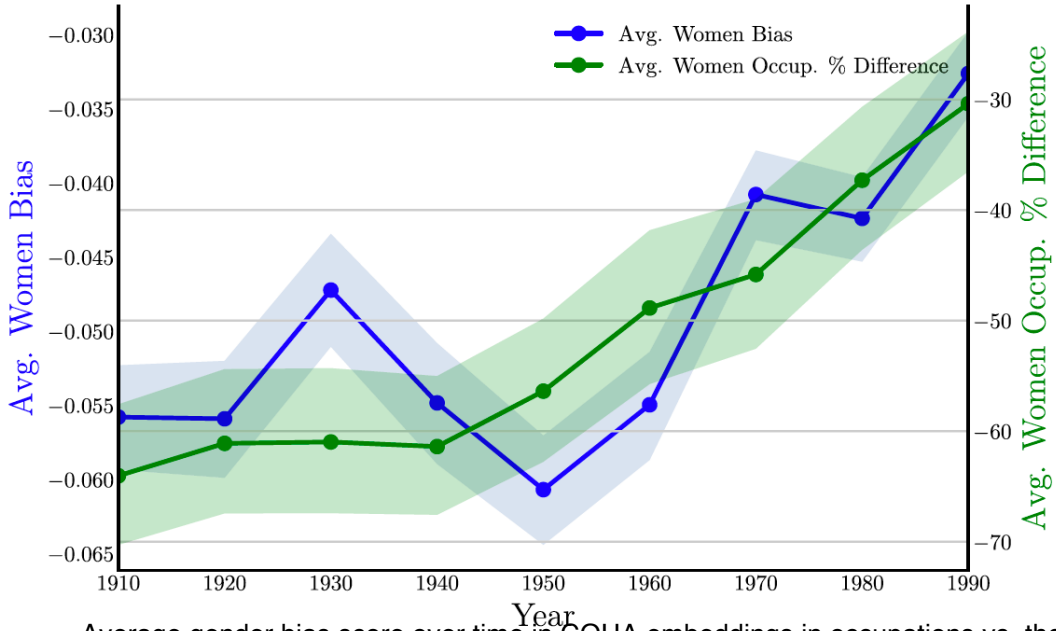
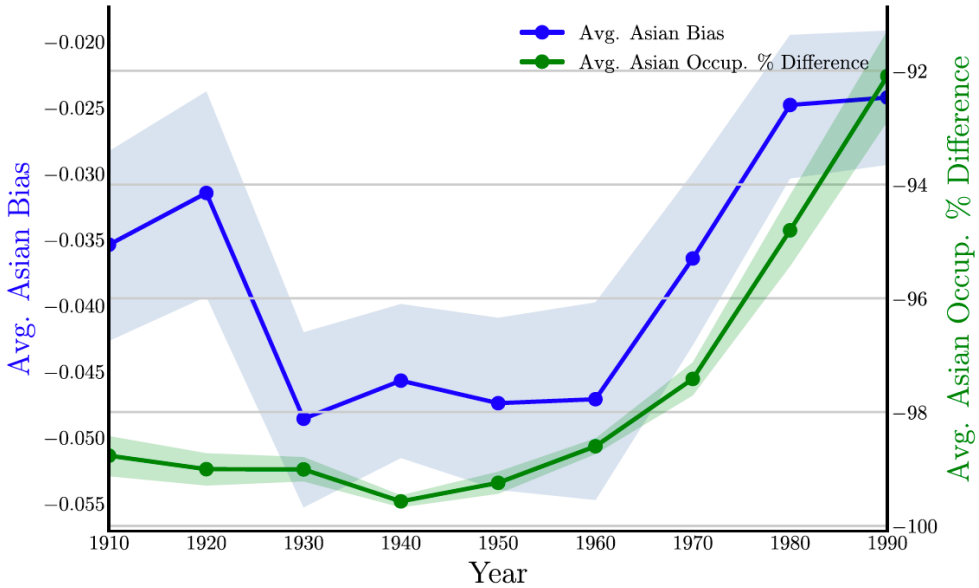


Fig. 1. Women's occupation relative percentage vs. embedding bias in Google News vectors. More positive indicates more associated with women on both axes. $P < 10^{-10}$, $r^2 = 0.499$. The shaded region is the 95% bootstrapped confidence interval of the regression line. In this single embedding, then, the association in the embedding effectively captures the percentage of women in an occupation.



Average gender bias score over time in COHA embeddings in occupations vs. the average percentage of difference. More positive means a stronger association with women. In blue is relative bias toward women in the embeddings, and in green is the average percentage of difference of women in the same occupations. Each shaded region is the bootstrap SE interval.



Average ethnic (Asian vs. White) bias score over time for occupations in COHA (blue) vs. the average percentage of difference (green). Each shaded region is the bootstrap SE interval.

Kozlowski, Taddy and Evans. (ASR 2019)

> “We show that dimensions induced by word differences (e.g., man – woman , rich – poor , black – white , liberal – conservative) in these vector spaces closely correspond to dimensions of cultural meaning, and the projection of words onto these dimensions reflects widely shared cultural connotations”

Table 1. List of word pairs averaged to construct cultural dimensions of gender, class, and race

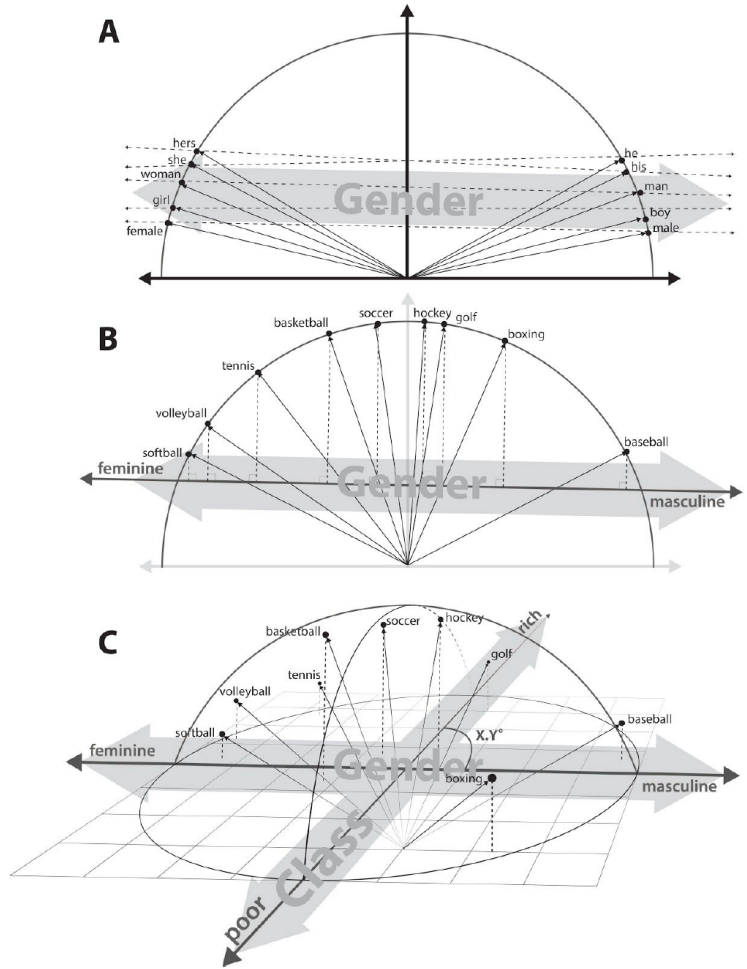
Gender	Class	Race[†]
man – woman	rich – poor	black – white
men – women	richer – poorer	blacks – whites
he – she	richest – poorest	Blacks – Whites
him – her	affluence – poverty	Black – White
his – her	affluent – impoverished	African – European
his – hers	expensive – inexpensive	African – Caucasian
boy – girl	luxury – cheap	
boys – girls	opulent – needy	
male – female		
masculine – feminine		

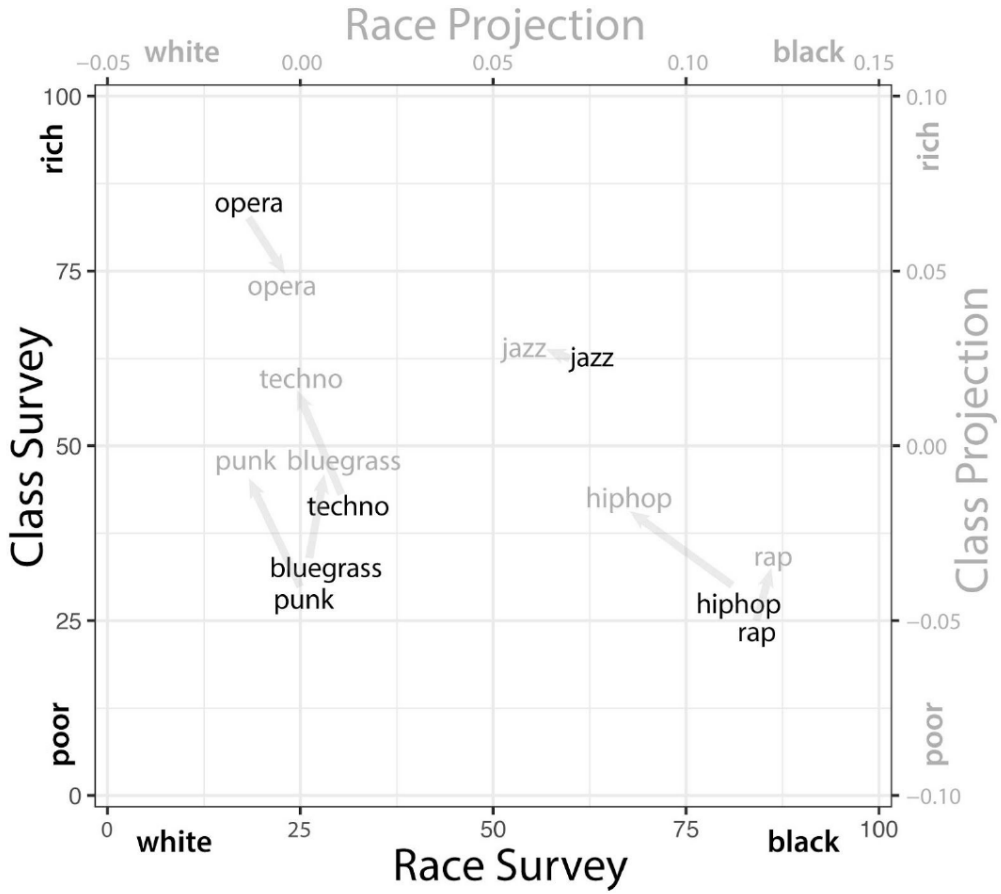
[†]Google Ngrams text was reduced to lowercase in preprocessing for all analyses to increase the word count for rare words. Because of this, the list of word pairs for used for ecological validation of Google Ngrams is *black-white, blacks-whites, african-european,* and *african-caucasian*.

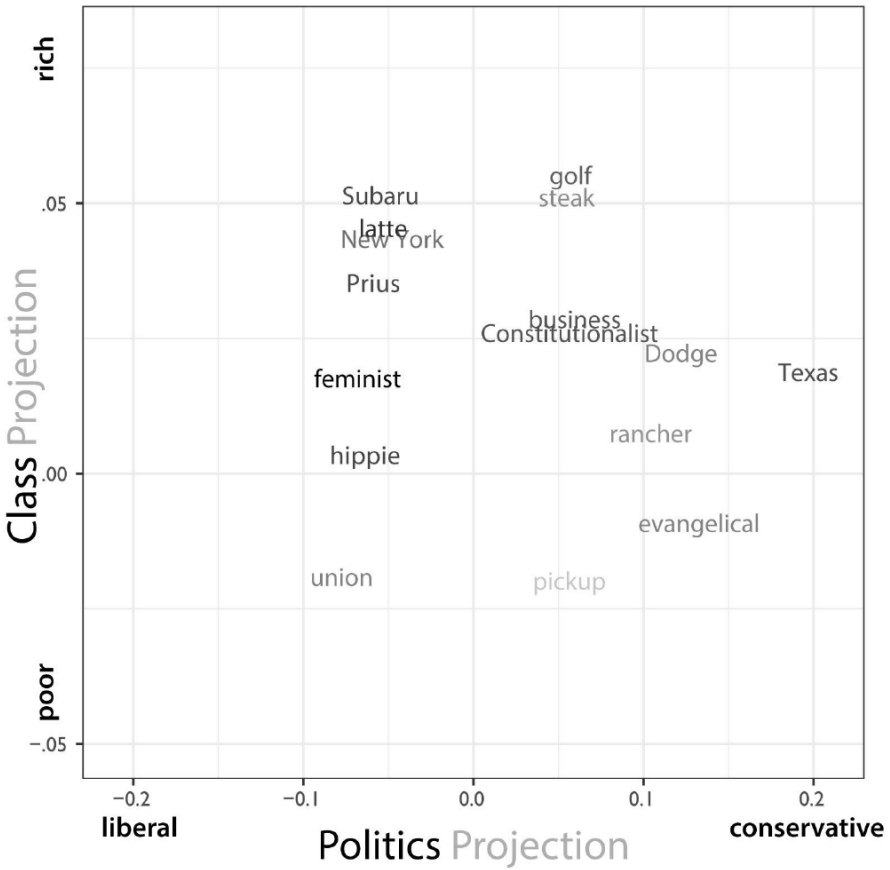
Table 2. Pearson correlations between survey estimates and word embedding estimates for gender, class, and race associations

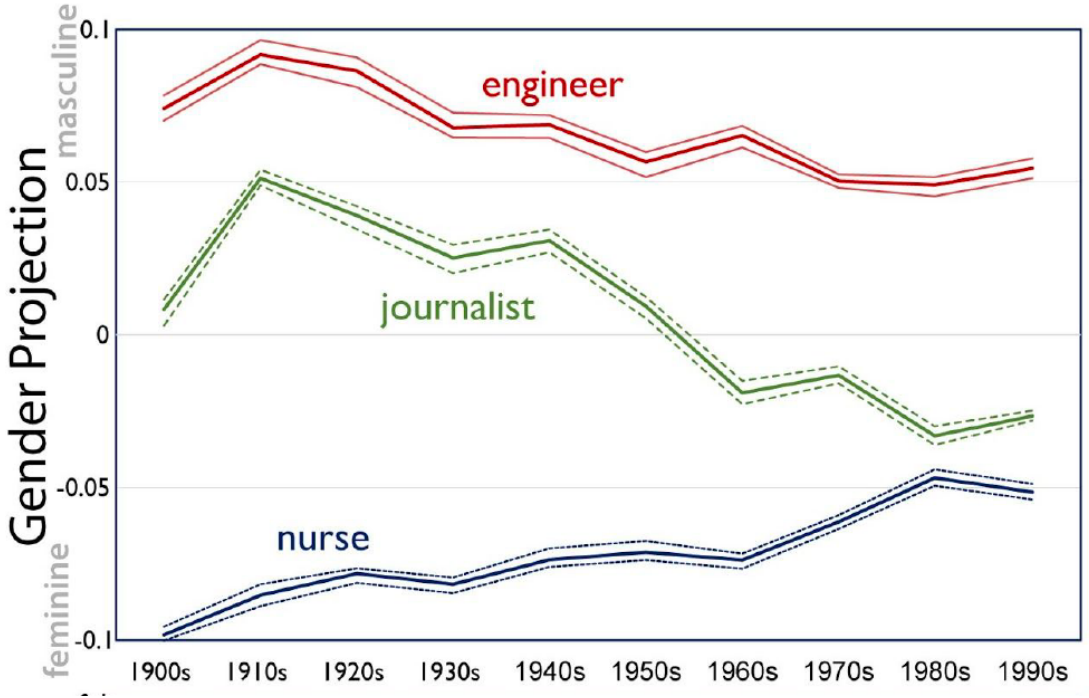
	Gender	Class	Race
Google News *word2vec* Embedding	0.88	0.52	0.70
Common Crawl *GloVe* Embedding	0.90	0.40	0.42
Wikipedia *fastText* Embedding	0.87	0.51	0.49
Google Ngrams *word2vec* Embedding[†]	0.78	0.60	0.17

N=59, except [†]N=56 where three words measured in the survey did not occur frequently enough in the text to appear in the word embedding.


A
hers
she
woman
girl
female
Gender
he
his
man
boy
male
B
basketball
soccer
hockey
golf
boxing
tennis
volleyball
softball
baseball
feminine
Gender
masculine
C
basketball
soccer
hockey
rich
golf
tennis
volleyball
softball
baseball
X.Y°
feminine
Gender
boxing
masculine
poor
Class


Race Projection
white
black
−0.05
0.00
0.05
0.10
0.15
rich
poor
Class Survey
100
75
50
25
0
opera
opera
techno
jazz
jazz
punk
bluegrass
techno
bluegrass
punk
hiphop
rap
hiphop
rap
Class Projection
0.10
0.05
0.00
−0.05
−0.10
0
25
50
75
100
white
black
Race Survey


rich
.05
.00
−.05
poor
Class Projection
golf
Subaru
steak
latte
New York
Prius
business
Constitutionalist
Dodge
Texas
feminist
rancher
hippie
evangelical
union
pickup
−0.2
−0.1
0.0
0.1
0.2
liberal
conservative
Politics Projection


Gender Projection
masculine
feminine
0.1
0.05
0
-0.05
-0.1
engineer
journalist
nurse
1900s
1910s
1920s
1930s
1940s
1950s
1960s
1970s
1980s
1990s

Thank you!

Yongjun Zhang, Ph.D
Assistant Professor
Dept of Sociology and
Institute for Advanced Computational Science
Stony Brook University
Yongjun.Zhang@stonybrook.edu
https:yongjunzhang.com