Unstructured Data for Economics

Lecture 3:
Word Embeddings

Stephen Hansen
University College London

Reading

Additional background material for the lecture: Jurafsky and Martin, *Speech and Language Processing*, Ch. 6 `https://bityl.co/QQRS`.

Introduction

Recall the one-hot encoding representation of vocabulary terms: V-dimensional, orthogonal vectors.

A word embedding is a low-dimensional vector representation of a word.

Ideally in this low-dimensional vector space words with similar meanings will lie close together.

The construction of word embeddings was an important precursor to the development of large language models.

Distributional Hypothesis

The distributional hypothesis states that words that share similar contexts share similar meanings.

Example from JM:

(6.1) Ongchoi is delicious sauteed with garlic.
(6.2) Ongchoi is superb over rice.
(6.3) ...ongchoi leaves with salty sauces...

And suppose that you had seen many of these context words in other contexts:

(6.4) ...spinach sauteed with garlic over rice...
(6.5) ...chard stems and leaves are delicious...
(6.6) ...collard greens and other salty leafy greens

Formalizing Local Context

Recall that $w_{d,n}$ is the nth word in document d.

The *context* of $w_{d,n}$ is a length-$2L$ window of words around $w_{d,n}$:

$$\mathrm{C}(w_{d,n}) = [w_{d,n-L}, w_{d,n-L+1}, \ldots, w_{d,n+L-1}, w_{d,n+L}]$$

Can truncate context appropriately if window stretches past beginning or end of text.

In line with distributional hypothesis, word embedding models seek to generate similar embeddings for words that share similar contexts.

GloVe

The GloVe model [Pennington et al., 2014] begins with a $V \times V$ matrix $\mathbf{W}$ of local word co-occurrences.

W_{ij} is the number of times term j appears within the context of i.

Assign to each term v an embedding vector $\boldsymbol{\rho}_v \in \mathbb{R}^K$.

$$\min \sum_{i,j} f(W_{i,j}) \left(\boldsymbol{\rho}_i^T \boldsymbol{\rho}_j - \log(W_{i,j})\right)^2$$

Terms that co-occur frequently will have more highly correlated embedding vectors.

Word2vec

Word2Vec

Word2vec [Mikolov et al., 2013a, Mikolov et al., 2013b] is a particularly well-known algorithm for the construction of word embeddings.

Important example of a neural-network-based language model that was scalable and effective.

Self-Supervised Learning

The 'meaning' of a word is an unobserved and subjective concept.

Difficult to directly formulate an objective function.

Important conceptual idea is to formulate word prediction tasks that are solved using word embeddings.[1]

Although word embeddings are formulated to solve prediction problems, they are nevertheless useful for the primary task of revealing meaning.

The approach of using auxiliary word prediction tasks to build high-quality embeddings is called self-supervised learning.

[1] See also [Bengio et al., 2003].

Prediction Tasks

There are two variants of word2vec which correspond to differing prediction tasks.

Skipgram model

1. Predict presence of each $w_{d,n-l} \in C(w_{d,n})$ given $w_{d,n}$.
2. Predict absence of randomly sampled words from the corpus given $w_{d,n}$.

Continuous Bag-of-Words model

1. Predict presence of $w_{d,n}$ given $C(w_{d,n})$.
2. Predict absence of randomly sampled words from the corpus given $C(w_{d,n})$.

Words and Context in Skipgram Model

"economic growth is weak but long-term productivity trends are strong"

Suppose $L = 2$.

Positive Examples		**Negative Examples**	
Word	Context	Word	Context
economic	growth	economic	down
economic	is	economic	towards
growth	economic	growth	inflation
growth	is	growth	mild
growth	weak	growth	very
is	economic	is	not
is	growth	is	can
is	weak	is	rate
is	but	is	how
.	.	.	.
strong	are	strong	many

The number of negative examples to sample per positive example is a modeling choice.

Parametrization of the Prediction Problems

Endow each word v in the vocabulary with an embedding vector $\boldsymbol{\rho}_v \in \mathbb{R}^K$ and a context vector $\boldsymbol{\alpha}_v \in \mathbb{R}^K$ where $K \ll V$.

The positive examples are modeled as

$$\Pr\left[w_{d,n-l} \in \mathrm{C}(w_{d,n}) \mid w_{d,n} \right] = \frac{\exp\left(\boldsymbol{\rho}_{w_{d,n}}^T \boldsymbol{\alpha}_{w_{d,n-l}}\right)}{1 + \exp\left(\boldsymbol{\rho}_{w_{d,n}}^T \boldsymbol{\alpha}_{w_{d,n-l}}\right)}$$

and the negative examples are modeled as

$$\Pr\left[w_{d,n-l} \notin \mathrm{C}(w_{d,n}) \mid w_{d,n} \right] = 1 - \frac{\exp\left(\boldsymbol{\rho}_{w_{d,n}}^T \boldsymbol{\alpha}_{w_{d,n-l}}\right)}{1 + \exp\left(\boldsymbol{\rho}_{w_{d,n}}^T \boldsymbol{\alpha}_{w_{d,n-l}}\right)}$$

Example

The first row of the table above would contribute the following elements to the loss function:

$$\Pr\left[\,\text{growth} \in \mathsf{C}(w_{d,n}) \mid w_{d,n} = \text{economic}\,\right] = \frac{\exp\left(\boldsymbol{\rho}_{\text{economic}}^{T}\boldsymbol{\alpha}_{\text{growth}}\right)}{1 + \exp\left(\boldsymbol{\rho}_{\text{economic}}^{T}\boldsymbol{\alpha}_{\text{growth}}\right)}$$

$$\Pr\left[\,\text{down} \notin \mathsf{C}(w_{d,n}) \mid w_{d,n} = \text{economic}\,\right] = \frac{1}{1 + \exp\left(\boldsymbol{\rho}_{\text{economic}}^{T}\boldsymbol{\alpha}_{\text{down}}\right)}$$

Loss function multiplies all such probabilities together and optimizes using gradient methods.

Terms Close to Uncertainty in FOMC Transcripts

term	sim
uncertainties	0.741
anxiety	0.48
pessimism	0.479
skepticism	0.465
optimism	0.445
caution	0.442
gloom	0.437
uncertain	0.433
sensitivity	0.427
angst	0.426

term	sim
challenges	0.415
fragility	0.405
clarity	0.401
concerns	0.4
risks	0.397
disagreement	0.387
volatility	0.384
tension	0.383
certainty	0.382
skepticism	0.38

Terms Close to Risk

term	sim
risks	0.737
threat	0.609
danger	0.541
dangers	0.463
vulnerability	0.457
chances	0.451
breakout	0.433
probability	0.426
possibility	0.409
likelihood	0.406

term	sim
misdirected	0.385
odds	0.379
uncertainty	0.375
concern	0.371
prospect	0.37
instability	0.363
potentially	0.352
concerns	0.352
challenges	0.346
risking	0.342

Importance of Training Corpus

Relationships among words can vary depending on the training corpus.

Example of training word embeddings on Wiki/Newswire text and on Harvard Business Review.

team		**leader**	
HBR	Generic	HBR	Generic
teams	teams	leadership	leaders
project_team	squad	leaders	leadership
management_team	players	manager	party
executive_team	football	person	opposition
group	coach	strong_leader	led
staff	league	chief_executive	rebel

Embeddings and Cultural Attitudes [Garg et al., 2018]

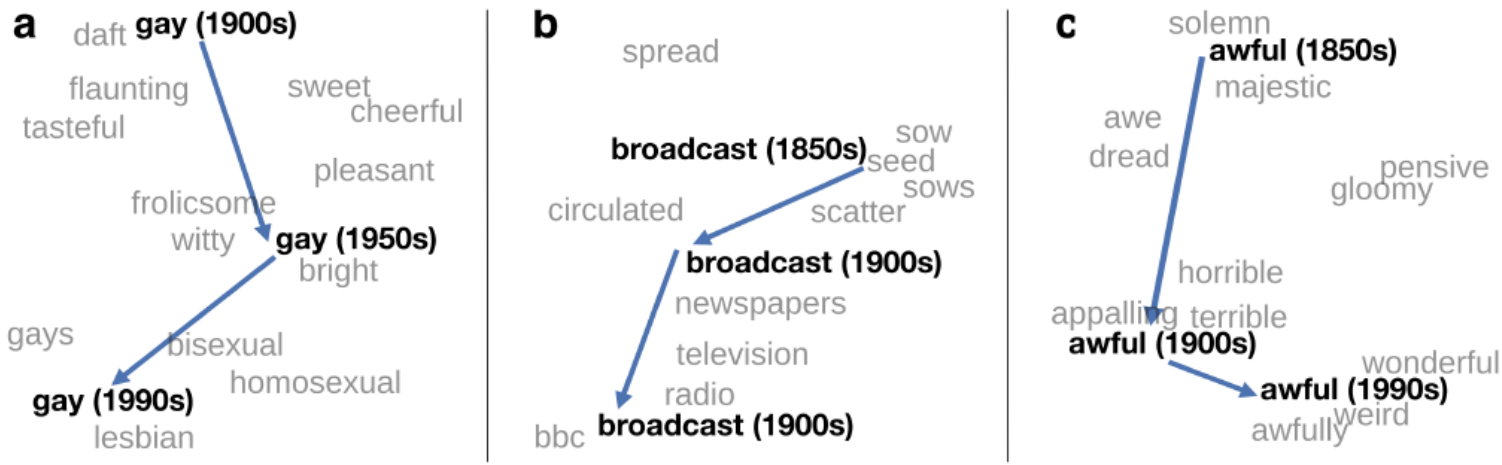
Table 2. Top adjectives associated with women in 1910, 1950, and 1990 by relative norm difference in the COHA embedding

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

Table 3. Top Asian (vs. White) adjectives in 1910, 1950, and 1990 by relative norm difference in the COHA embedding

1910	1950	1990
Irresponsible	Disorganized	Inhibited
Envious	Outrageous	Passive
Barbaric	Pompous	Dissolute
Aggressive	Unstable	Haughty
Transparent	Effeminate	Complacent
Monstrous	Unprincipled	Forceful
Hateful	Venomous	Fixed
Cruel	Disobedient	Active
Greedy	Predatory	Sensitive
Bizarre	Boisterous	Hearty

Evolution of Word Meanings [Hamilton et al., 2016]



Concept Detection

Expanding Dictionaries

One application of word embeddings is to augment human judgment in the construction of dictionaries.

Motivation is that economists are experts in which concept might be most important in a particular setting, but not in which words relate to that concept.

One can specify a set of 'seed' words and then find nearest neighbors of those words to populate a dictionary.

Strategy adopted by several recent papers:

1. [Hanley and Hoberg, 2019]
2. [Li et al., 2021]
3. [Bloom et al., 2021]
4. [Davis et al., 2020]

Embedding Dictionaries

Dictionaries provide a coarse representation of concepts in that some relevant terms might be missing altogether, and strength of association with concept isn't accounted for.

One strategy is to measure the association between documents and word lists in an embedding space rather than the bag-of-words space.

Recent example is [Gennaro and Ash, 2022] which studies emotional language in politics using the Congressional Record corpus.

Set A of words represents emotion, and set C of words represents cognition (both from LIWC).

Emotionality of speech i is

$$Y_i = \frac{\text{sim}(\boldsymbol{d_i}, \boldsymbol{A}) + b}{\text{sim}(\boldsymbol{d_i}, \boldsymbol{C}) + b}$$

Results

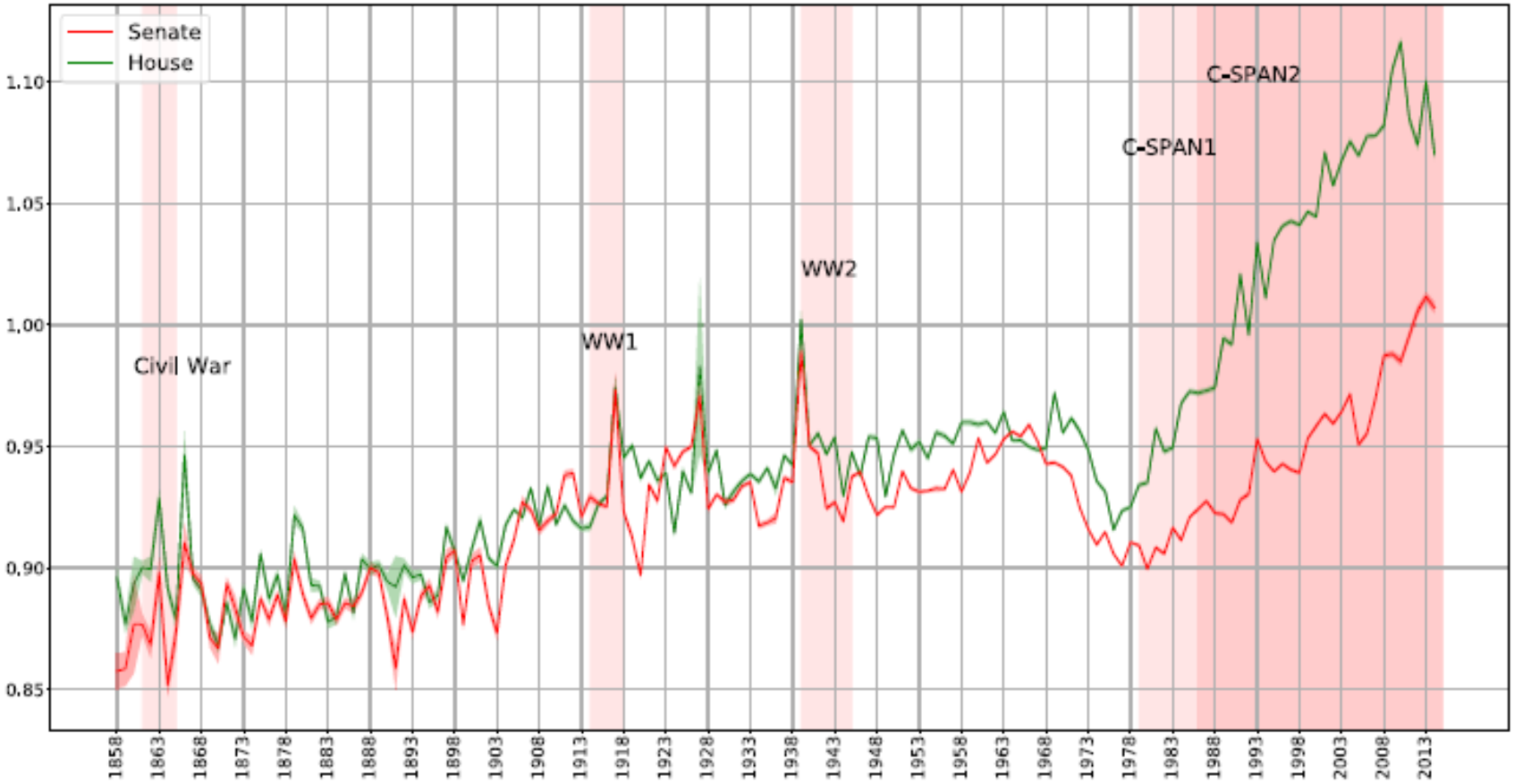
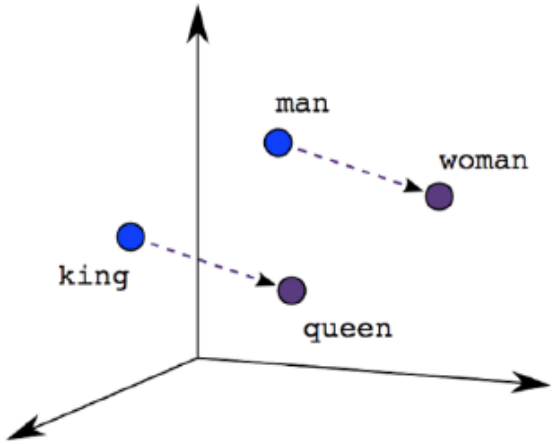


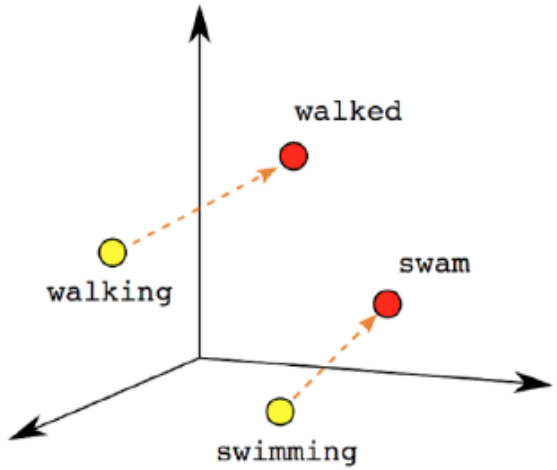
Fig. 2. *Emotionality in U.S. Congress by Chamber, 1858–2014.*
Notes: Time series of emotionality in the Senate (red) and the House of Representatives (green).

Relationship Among Concepts

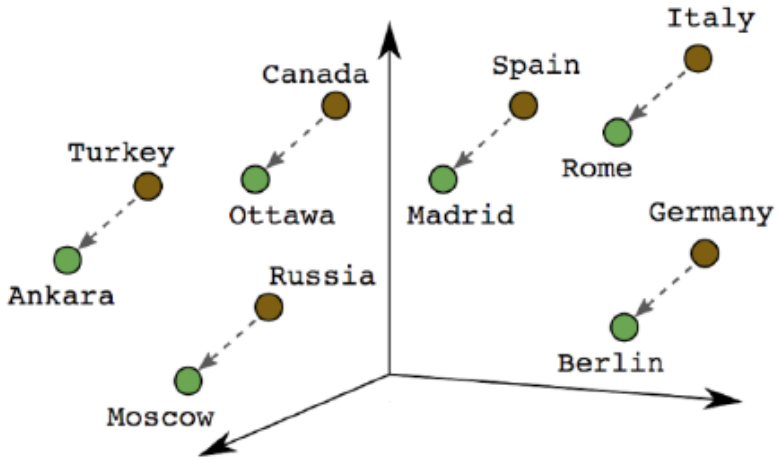
Directions Encode Meaning



Male-Female



Verb Tense



Country-Capital

Word Embeddings and Cultural Attitudes

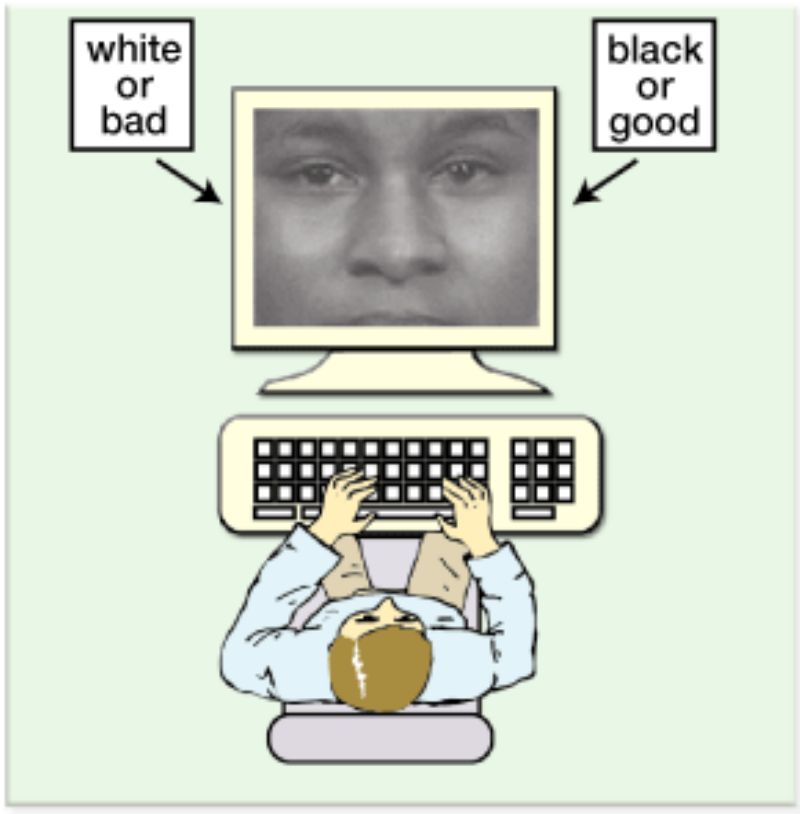
Because word embeddings appear to capture semantically meaningful relationships among words, there is interest in using them to measure cultural attitudes.

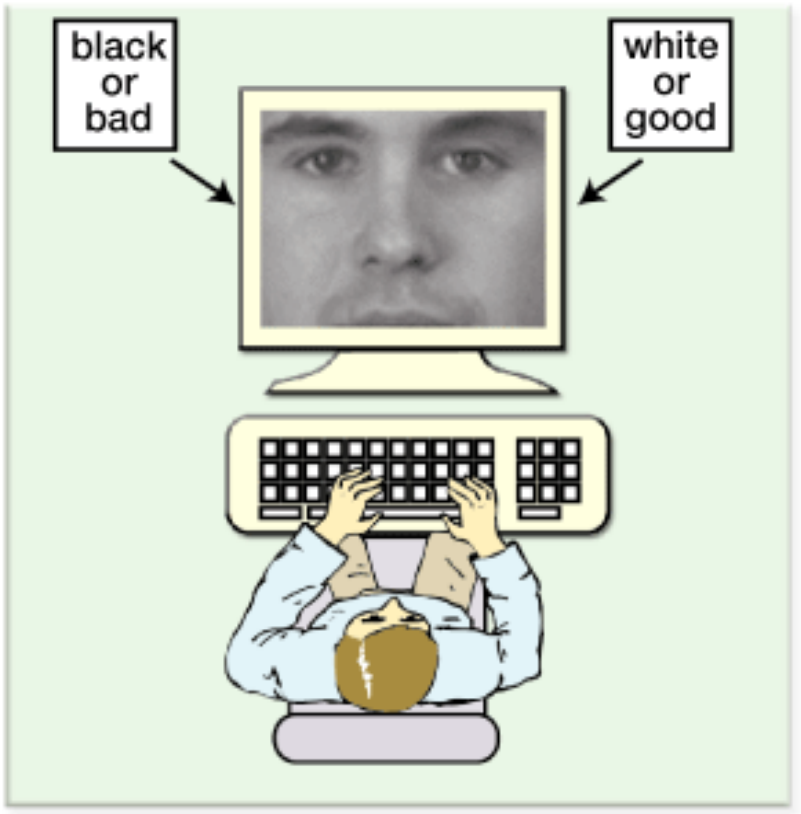
In psychology there is a long-standing Implicit Association Test that measures participants' time to correctly classify images depending on word combinations.

The hypothesis is that reaction times are shorter when word combinations more naturally belong together, which allows a measure of bias.

[Caliskan et al., 2017] have use word embeddings to ask whether similar biases exist in natural language.

Implicit Association Test





Word-Embedding Association Test

The Word-Embedding Association Test (WEAT) measures whether two sets of target words X, Y (e.g. male, female words) differ in their relative similarity to two sets of attribute words A, B (e.g. career, family words).

Let $\cos(\mathbf{x}, \mathbf{y})$ be cosine similarity between vectors $\mathbf{x}$ and $\mathbf{y}$.

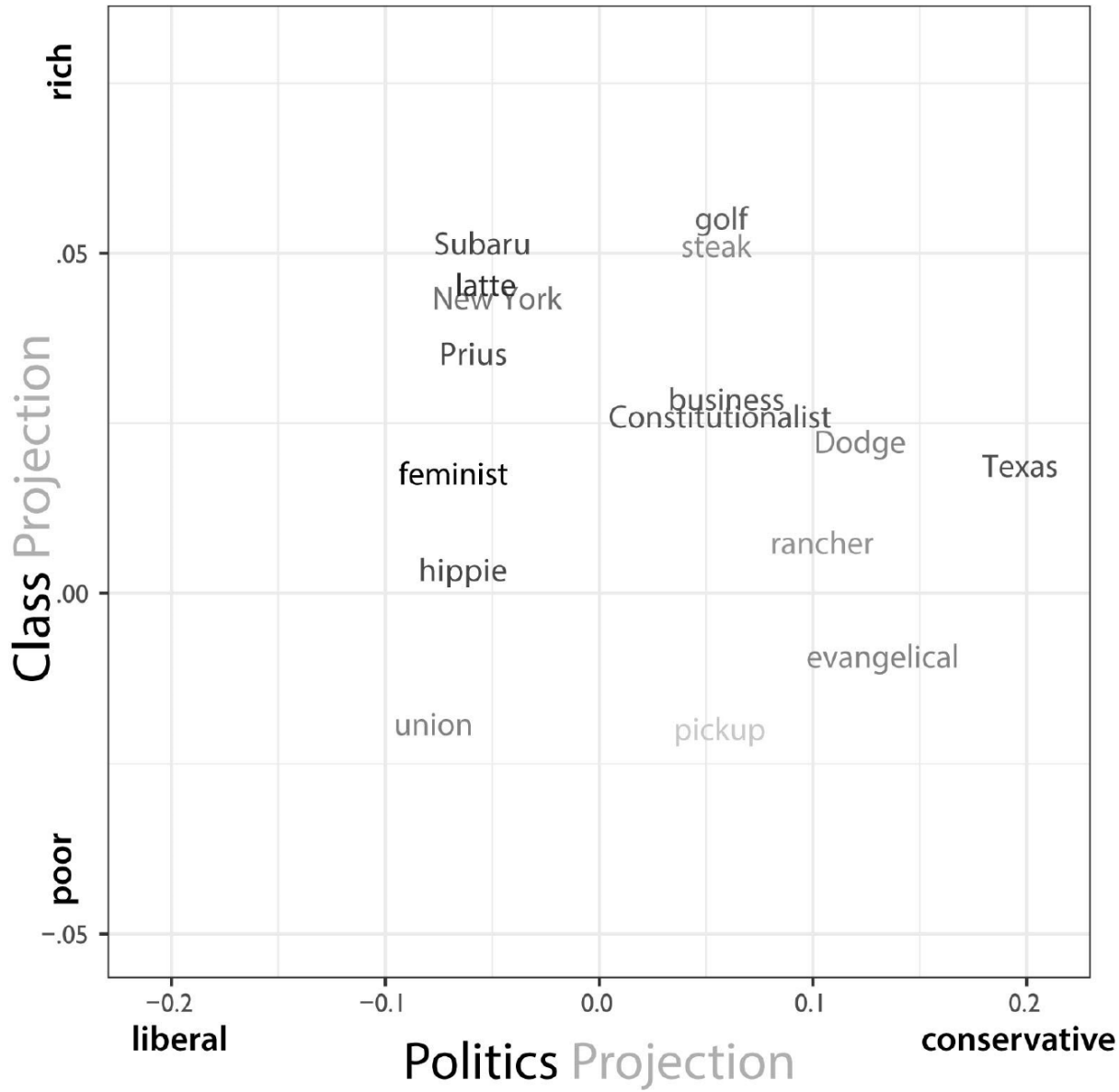
Let $s(\mathbf{w}, A, B) = \text{mean}_{\mathbf{a} \in A} \cos(\mathbf{w}, \mathbf{a}) - \text{mean}_{\mathbf{b} \in B} \cos(\mathbf{w}, \mathbf{b})$.

$$\text{WEAT} = \frac{\sum_{\mathbf{x} \in X} s(\mathbf{x}, A, B) - \sum_{\mathbf{y} \in Y} s(\mathbf{y}, A, B)}{\text{std}_{\mathbf{x} \in X \cup Y}\, s(\mathbf{x}, A, B)}$$

IAT vs WEAT

Target words	Attribute words	Original finding				Our finding			
		Ref.	*N*	*d*	*P*	N_T	N_A	*d*	*P*
Flowers vs. insects	Pleasant vs. unpleasant	(5)	32	1.35	10^{-8}	25 × 2	25 × 2	1.50	10^{-7}
Instruments vs. weapons	Pleasant vs. unpleasant	(5)	32	1.66	10^{-10}	25 × 2	25 × 2	1.53	10^{-7}
European-American vs. African-American names	Pleasant vs. unpleasant	(5)	26	1.17	10^{-5}	32 × 2	25 × 2	1.41	10^{-8}
European-American vs. African-American names	Pleasant vs. unpleasant from (5)	(7)	Not applicable			16 × 2	25 × 2	1.50	10^{-4}
European-American vs. African-American names	Pleasant vs. unpleasant from (9)	(7)	Not applicable			16 × 2	8 × 2	1.28	10^{-3}
Male vs. female names	Career vs. family	(9)	39k	0.72	$<10^{-2}$	8 × 2	8 × 2	1.81	10^{-3}
Math vs. arts	Male vs. female terms	(9)	28k	0.82	$<10^{-2}$	8 × 2	8 × 2	1.06	.018
Science vs. arts	Male vs. female terms	(10)	91	1.47	10^{-24}	8 × 2	8 × 2	1.24	10^{-2}
Mental vs. physical disease	Temporary vs. permanent	(23)	135	1.01	10^{-3}	6 × 2	7 × 2	1.38	10^{-2}
Young vs. old people's names	Pleasant vs. unpleasant	(9)	43k	1.42	$<10^{-2}$	8 × 2	8 × 2	1.21	10^{-2}

Language and Culture [Kozlowski et al., 2019]



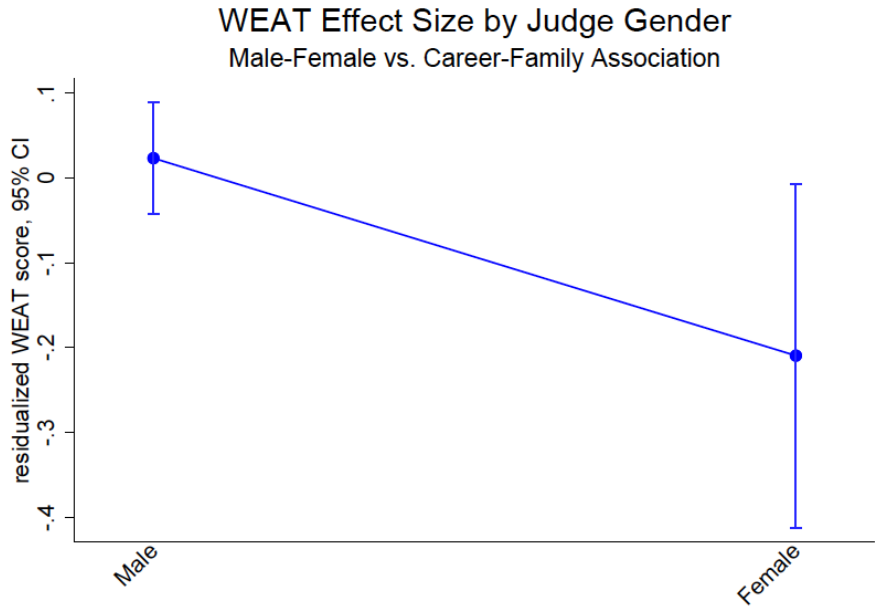
Does Language affect Decisions?

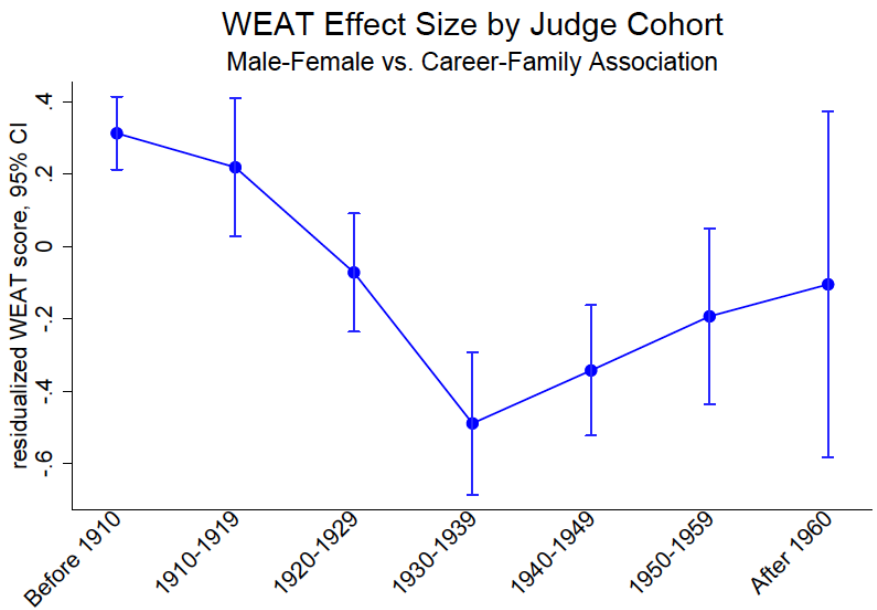
[Ash et al., 2024] use a measure similar to WEAT to measure linguistic gender bias among judges using written opinions.

They then match judge-specific bias scores with individual judge decisions to see whether the two are related.

Data is the universe of US appellate court decisions from 1890-2013.

WEAT and Judge Characteristics





Effects of WEAT

Judges with higher lexical bias are:

- Less likely to cast vote in favor of women's interests
- More likely to vote more conservatively across all issues
- Less likely to cite women in their opinions
- More likely to reverse female district judges

Document Similarity

Embedding-Based Similarity

Several papers use the distance between documents as captured by average embedding vectors.

[Kogan et al., 2019] measures distance between patents and occupation descriptions to proxy exposure of jobs to technical change.

[Hansen et al., 2021] measures distance between O*NET occupation descriptions and job postings to proxy skill demand.

Word2Vec Summary

Word2Vec introduces several ideas that remain influential:

1. Words as low-dimensional embedding vectors.
2. Self-supervised learning using auxiliary word-prediction tasks to build informative representations of language.
3. Neural network estimation in place of statistical models.
4. Surprising behavior of estimated latent meaning space.

One important limitation: vectors are built using local context but they do not vary with local context.

Embeddings for Non-Textual Data

Contextual Data

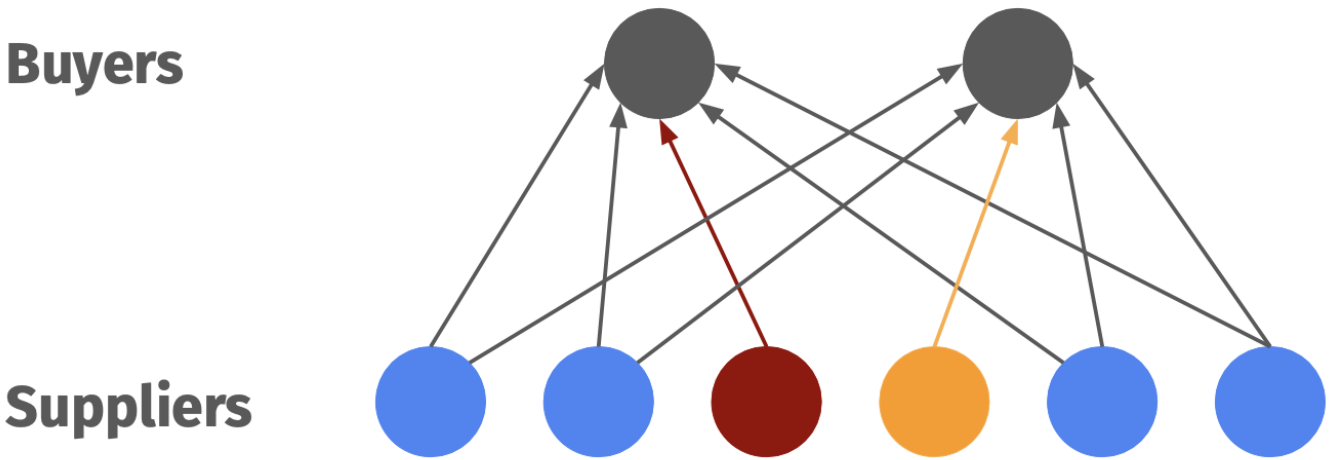
The ideas above are relevant for more general contextual data.

[Magnolfi et al., 2023] embeds products based on surveys comparing product similarity.

See also [Ruiz et al., 2020].

Firm embeddings: general idea

- Exploit information on suppliers selling to customers (firm-to-firm).
- Model each observation (firm) conditional on its context (other suppliers).
- Predict the probability that any firm i is supplying any potential customer j.
- Firm A is similar to firm B if they tend to co-supply the same customers.



Model setup

Setup

- Set of producers: $i = 1, ..., N$.
- Sales relationship from supplier i to customer j $a_{ij} \in \{0, 1\}$.
- Adjacency matrix of the production network: $a_{ij} \in G = (N \times N)$.

Notes

- This is all the information the algorithm gets!
- No firm characteristics, sector of activity, value of trade, ...

The probability of supplying a customer

The conditional probability that i is supplying j, given all other suppliers to j is

$$i \in S_j | S_j^{-i} \sim Bernoulli(p_i) \text{ where } p_i \equiv \Pr\left(i \in S_j | S_j^{-i}\right)$$

Then the set of firms S_j that supply a given customer j is

$$P(S_j) = \prod_{i \in S_j} \Pr\left(i \in S_j | S_j^{-i}\right)$$

Aggregating over all possible customers in the economy $j \in C$:

$$P(G) = \prod_{j \in C} \prod_{i \in S_j} \Pr\left(i \in S_j | S_j^{-i}\right)$$

Mapping high-dimensional to low-dimensional

Enter embeddings

- Parameterize Pr $\left(i \in S_j | S_j^{-i}\right)$ using low-dimensional continuous vectors.
- For each supplier i: embedding $\rho_i = [\rho_i, ..., \rho_{iK}]$ and context $\alpha_i = [\alpha_i, ..., \alpha_{iK}]$.
- Jointly determine the conditional probability relating each firm to its context.

Simplest parameterization of exponential family embeddings

- Linear combination of embeddings vectors: $\left(\rho_i^T \sum_{s \in S_j^{-i}} \alpha_s\right)$.
- Sigmoid link function to map outcomes to probabilities $\sigma(\cdot)$.

Embeddings

The **conditional probability of** i **supplying** j, given the set of suppliers to j is

$$\Pr\left(i \in S_j | S_j^{-i}\right) = \underbrace{\sigma}_{\text{link function}} \underbrace{\left(\frac{1}{\#S_j^{-i}} \rho_i^T \sum_{s \in S_j^{-i}} \alpha_s \right)}_{\text{linear embedding}}$$

The **probability of a supplier** g **not supplying** to customer j is

$$\Pr\left(g \neq S_j | S_j\right) = 1 - \Pr\left(g \in S_j | S_j\right) = 1 - \sigma\left(\frac{1}{\#S_j} \rho_g^T \sum_{s \in S_j} \alpha_s \right)$$

Note: computing for all g is expensive $\rightarrow$ negative sampling (Mikolov et al, 2013).

Data sources

NBB B2B Transactions (Dhyne, Magerman and Rubinova, 2015)

- Universe of firm-to-firm transactions (2014).
- 14 million links across all economic activities and regions.
- Sales value m_{ijt} from seller i to customer j in year t.

Annual accounts

- Firm characteristics: sales, inputs, employment, labor cost, capital, ...

Crossroads Bank of Enterprises

- Firm official seat's postal code, main NACE code.

Prodcom

- Production of manufacturing goods: PC8, year, value, quantity, unit.

Data construction

Observables

- Retain only $m_{ij} = 0/1$ for the Bernoulli embeddings.
- Keep all firm info for correlations embeddings and firm observables.

Sample selection

- Drop firms with only one supplier (no context for the supplying firm).
- Drop all transactions less than 1% of customer purchases.

Final sample

- 7.3 million transactions across 840k customers and 480k suppliers.
- Number of suppliers per customer: mean (8.7); median (7).

Embeddings and geographical space

For all firms i and j in the same NACE4 sector, estimate:

$$\cos \text{sim}_{ij} = \beta_0 + \beta_1 \ln \text{distance}_{ij} + \varepsilon_{ij}$$

	Log distance$_{i,j}$
% of sectors with $\beta_1 < 0$	82.6%
% of sectors with $\beta_1 > 0$	0.9%
% of sectors with $\beta_1 = 0$ at 95%CI	16.5%
N	430

- Similarity between suppliers i and j decreases in distance (for 82.6% of sectors).
- I.e. they tend to co-supply very different sets of customers.
- Hence, for these sectors, suppliers need to be close to their customers.

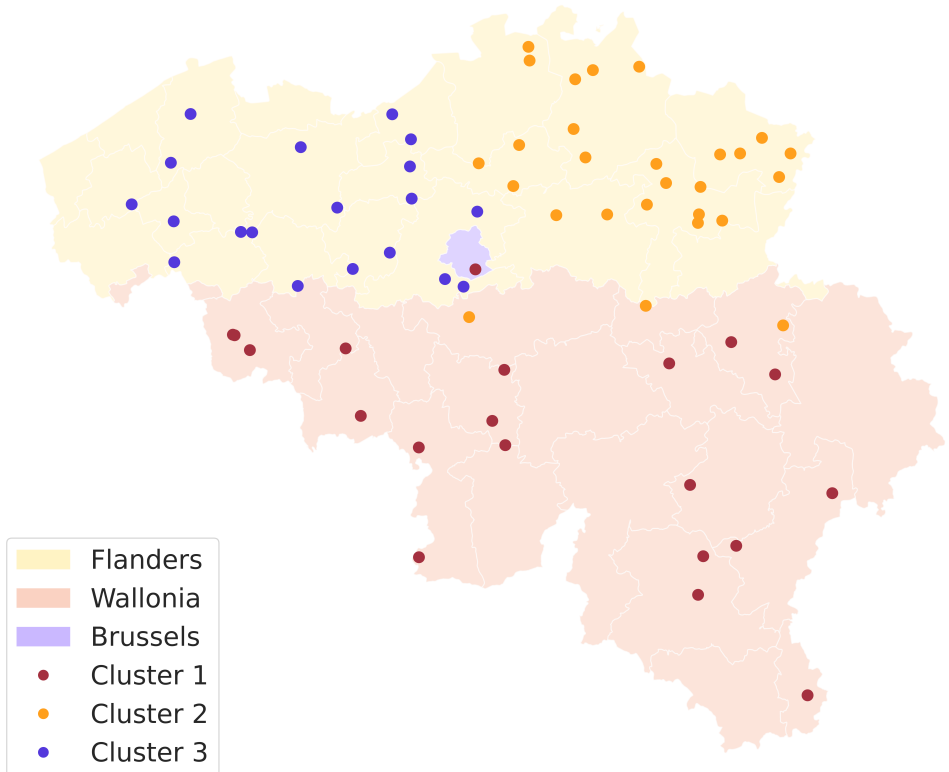
Case study: ready-mix concrete

Syverson (2008):

- Perhaps no other manufacturing industry faces greater transport barriers.
- It is not nationwide, but a set of quasi-independent local geographic markets.
- The average shipment distance is 32 miles (vs 550 miles for all commodities).

Case study: ready-mix concrete

- All single establishment/single product ready-mix concrete suppliers (NACE 2363).
- Group into clusters (k-means clustering on their ρ embeddings).



Embeddings and product space

Prodcom: values and quantities for each 8-digit product of a firm (8,132 firm-products in 2014).

8-digit code	**Description**
24.20.12.10	Casing, tubing and drill pipe, of a kind used in the drilling for oil or gas, seamless, of stainless steel
24.20.12.50	Casing, tubing and drill pipe, of a kind used in the drilling for oil or gas, seamless, of steel other than stainless steel
13.20.20.14	Woven fabrics of cotton, not of yarns of different colours, weighing $\leq$ 200 g/m^2, for clothing
13.20.20.42	Woven fabrics of cotton, not of yarns of different colours, weighing $>$ 200 g/m^2, for clothing

Embeddings and product space

Product-share embeddings

- Sparse vector representation for each firm: a firm as a "bag of products".
- Compare similarity of this "bag of products" to embeddings cosine similarity.
- I.e. two firms are similar if they sell the same products (sales weighted).

Text embeddings

- Text description of Prodcom products: a firm as a "bag of words".
- Compare both cosine similarities.

Measure	Product-share embeddings	Text-based embeddings
Pearson correlation	0.085***	0.115***
Spearman rank correlation	0.101***	0.083***
Kendall rank correlation	0.082***	0.057***

References I

Ash, E., Chen, D. L., and Ornaghi, A. (2024).
Gender Attitudes in the Judiciary: Evidence from US Circuit Courts.
American Economic Journal: Applied Economics, 16(1):314–350.

Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003).
A neural probabilistic language model.
The Journal of Machine Learning Research, 3(null):1137–1155.

Bloom, N., Hassan, T. A., Kalyani, A., Lerner, J., and Tahoun, A. (2021).
The Diffusion of Disruptive Technologies.
Working Paper 28999, National Bureau of Economic Research.

Caliskan, A., Bryson, J. J., and Narayanan, A. (2017).
Semantics derived automatically from language corpora contain human-like biases.
Science, 356(6334):183–186.

Davis, S. J., Hansen, S., and Seminario-Amez, C. (2020).
Firm-Level Risk Exposures and Stock Returns in the Wake of COVID-19.
Working Paper 27867, National Bureau of Economic Research.

References II

Garg, N., Schiebinger, L., Jurafsky, D., and Zou, J. (2018).
Word embeddings quantify 100 years of gender and ethnic stereotypes.
Proceedings of the National Academy of Sciences, 115(16):E3635–E3644.

Gennaro, G. and Ash, E. (2022).
Emotion and Reason in Political Language.
The Economic Journal, 132(643):1037–1059.

Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016).
Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change.
In Erk, K. and Smith, N. A., editors, Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Hanley, K. W. and Hoberg, G. (2019).
Dynamic Interpretation of Emerging Risks in the Financial Sector.
The Review of Financial Studies, 32(12):4543–4603.

Hansen, S., Ramdas, T., Sadun, R., and Fuller, J. (2021).
The Demand for Executive Skills.
Technical Report 28959, National Bureau of Economic Research, Inc.

References III

Kogan, L., Papanikolaou, D., Schmidt, L., and Seegmiller, B. (2019).

Technology, Vintage-Specific Human Capital, and Labor Displacement: Evidence from Linking Patents with Occupations.

Kozlowski, A. C., Taddy, M., and Evans, J. A. (2019).

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

American Sociological Review, 84(5):905–949.

Li, K., Mai, F., Shen, R., and Yan, X. (2021).

Measuring Corporate Culture Using Machine Learning.

The Review of Financial Studies, 34(7):3265–3315.

Magnolfi, L., McClure, J., and Sorensen, A. T. (2023).

Triplet Embeddings for Demand Estimation.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a).

Efficient Estimation of Word Representations in Vector Space.

arXiv:1301.3781 [cs].

References IV

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013b).
Distributed Representations of Words and Phrases and their Compositionality.
arXiv:1310.4546 [cs, stat].

Pennington, J., Socher, R., and Manning, C. (2014).
GloVe: Global Vectors for Word Representation.
In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Ruiz, F. J. R., Athey, S., and Blei, D. M. (2020).
SHOPPER: A probabilistic model of consumer choice with substitutes and complements.
The Annals of Applied Statistics, 14(1):1–27.