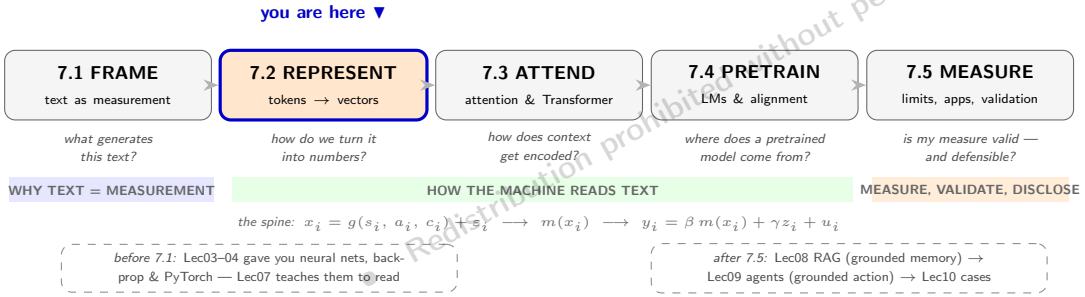
AI for Economic Research: Dynamic Models, Language, and Agents
Lecture 7.2b: Word Embeddings

Zhigang Feng

2026-07-09 05:57:06

Lecture 7 in One Map



Route: from counts to coordinates (part 2/2 of REPRESENT) — Word2Vec and GloVe, the geometry of meaning, what a direction in the space can measure, and where static vectors break.

From Counts to Coordinates

Learning a vector space in which meaning has geometry

Why Embed Tokens into Vectors?

- Neural networks operate on real-valued vectors, not discrete symbols.
- An **embedding** is a learned function

$$\mathcal{E}\colon \{1, \ldots, M\} \longrightarrow \mathbb{R}^d,$$

  mapping each token ID to a dense vector of dimension d (typically $d = 128$ to 4096).

- **What we want from embeddings:**
  - Semantically similar tokens should have nearby vectors
  - Algebraic operations (sums, differences) should reflect semantic relations
  - The mapping should be *learned* from data, not hand-designed

- **Bonus:** we will also need a *positional* signal so the model knows token order—more on this when we meet the Transformer (Act 3).

Three Approaches to Word Representation

Method	Type	Idea	Year
TF-IDF	Sparse, frequency	Re-weight word counts	1970s
Word2Vec	Dense, predictive	Predict neighbors in window	2013
GloVe	Dense, statistical	Factor co-occurrence matrix	2014
Transformer embeddings	Dense, contextual	Vary by surrounding tokens	2017+

Word2Vec: Dense Embeddings from Context

- **Distributional hypothesis** (Firth, 1957): *"You shall know a word by the company it keeps."*
- **Idea** (Mikolov et al., 2013): for each word in a corpus, look at a window of $\pm L$ neighboring words. Learn a vector for the center word that *predicts* its neighbors (*Skip-gram*) or learn one that *is predicted by* its neighbors (*CBOW*).
- **Training output:** a single static vector $\mathbf{u}_w \in \mathbb{R}^d$ for each word w in the vocabulary, with d typically 100–300.
- **Geometric magic:** after training, simple vector arithmetic captures analogies, e.g.

$$\mathbf{u}_{\text{king}} - \mathbf{u}_{\text{man}} + \mathbf{u}_{\text{woman}} \approx \mathbf{u}_{\text{queen}}.$$

  Words living on the same semantic axis differ by approximately the same vector.

Skip-Gram: The Idea in One Slide

Every word carries a vector; training nudges the vectors of words that appear together closer, and random pairs apart.

- The *score* of a (center word, nearby word) pair is just the dot product of their vectors, $\mathbf{u}_{\text{center}}^{\top}\mathbf{v}_{\text{context}}$ — large when the two point the same way. (Turning that score into a probability is a multinomial logit — familiar territory for economists.)
- Training sweeps the whole corpus and, over and over:
  - **pulls together** pairs that actually co-occur — push the score *up* (*inflation* near *persistent*);
  - **pushes apart** a handful of random pairs — push the score *down* (*inflation* vs. *basketball*).

  This push-apart shortcut is called *negative sampling*: it avoids re-scoring the entire vocabulary at every step.
- **Result:** words that keep similar company end up with similar vectors — meaning emerges from nothing but "who appears near whom."

Skip-Gram: Walkthrough on "central banks target inflation"

- **Sentence:** *"central banks target inflation"*, window radius $L = 2$, center word = *target*.
- Training pairs formed from this center word:

$$(\textit{target}, \textit{central}), \quad (\textit{target}, \textit{banks}), \quad (\textit{target}, \textit{inflation}).$$

- For each pair, the score is the dot product $\mathbf{u}_{\text{target}}^{\top}\mathbf{v}_{\text{context}}$. Gradient descent *increases* this score for observed pairs and *decreases* it for random (negative) pairs.
- **Across the whole corpus:** *target* appears near *rates*, *inflation*, *policy*, *Fed*. So do *raise* and *cut*. Consequently

$\mathbf{u}_{\text{target}}$ ends up close to $\mathbf{u}_{\text{raise}}$, $\mathbf{u}_{\text{cut}}$ and far from $\mathbf{u}_{\text{apple}}$.

- **Takeaway:** analogies emerge because consistent co-occurrences place related words at similar positions in $\mathbb{R}^d$.

GloVe: Count Once, Then Fit

Word2Vec slides a window and predicts neighbors; GloVe takes a shortcut — count how often each pair of words co-occurs across the whole corpus, then fit vectors to reproduce those counts.

- Let X_{ij} = the number of times words i and j appear near each other in the entire corpus. (*Fed* and *rates* might co-occur 820 times; *Fed* and *apple*, just 3.)
- Pennington et al. (2014) ask each pair of vectors to reproduce its co-occurrence count:

$$\underbrace{\mathbf{u}_i^\top \mathbf{v}_j}_{\text{how aligned the vectors are}} + \underbrace{b_i + b_j}_{\text{each word's baseline}} \approx \underbrace{\log X_{ij}}_{\text{how often they co-occur}}$$

- **The whole model in one sentence:** *two words that keep each other's company should get vectors that point the same way.* Everything else is just fitting that sentence by least squares.

GloVe: Read It Like a Gravity Equation

An economist has met this equation before — it is the gravity model of trade, term for term.

$$\underbrace{\log(\text{trade}_{ij})}_{\text{how much } i,\ j \text{ trade}} \approx \text{FE}_i + \text{FE}_j + (\text{pair closeness}) \qquad \Longleftrightarrow \qquad \log X_{ij} \approx b_i + b_j + \mathbf{u}_i^\top \mathbf{v}_j$$

- $\log X_{ij}$ — how much the two words "trade" with each other. We take logs because co-occurrence counts span orders of magnitude (3 vs. 820) — the same reason we use log GDP.
- b_i, b_j — **word fixed effects.** *the* co-occurs with everything simply because it is everywhere; the fixed effects soak up that baseline popularity so we don't mistake it for meaning.
- $\mathbf{u}_i^\top \mathbf{v}_j$ — what is *left over* once popularity is removed: the pair's **specific affinity**. This, and only this, becomes meaning.

The payoff: words that co-occur more than their popularity alone would predict get vectors pointing the same way; words that never meet stay at right angles. *Meaning becomes geometry, estimated like a regression.*

GloVe at Work: A Central-Bank Corpus

Watch "monetary-policy vocabulary" fall out of nothing but co-occurrence counts.

co-occurs with →	*rates*	*inflation*	*apple*
Fed	820	410	3
ECB	540	390	2
hawkish	210	160	0

What the fit does, row by row:

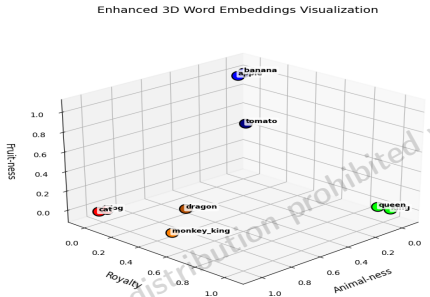
- *Fed* & *rates* co-occur 820 times ($\log X \approx 6.7$). The baselines $b_{\text{Fed}} + b_{\text{rates}}$ cannot explain a number that big — the only way to fit it is to make $\mathbf{u}_{\text{Fed}}$ and $\mathbf{v}_{\text{rates}}$ **point the same way**.
- *Fed* & *apple* co-occur just 3 times ($\log X \approx 1.1$) — about what the baselines predict for *any* two words. Nothing is left to explain, so their vectors end up **at right angles** (unrelated).
- *Fed*, *ECB*, and *hawkish* all keep the *same company* (rates and inflation, never apple), so all three get pulled into the **same corner** of the space — to the geometry they are near-synonyms.

The punchline: nobody told the model what a central bank is. Rows with similar co-occurrence patterns are forced to similar coordinates, and a "monetary-policy" cluster emerges purely from counting.

The Geometry of Meaning

Distances, directions, analogies — and the biases baked in

A 3D Toy of Word Embeddings



The three axes loosely correspond to "Animal-ness", "Royalty", and "Fruit-ness". Words like *cat*, *king*, and *apple* cluster along the dimension that fits them best. Real LLM embeddings live in hundreds of dimensions, but the same intuition holds.

We compare words by the *direction* of their vectors — **cosine similarity** $\cos(\mathbf{u}, \mathbf{v}) = \mathbf{u}^\top \mathbf{v} / (\|\mathbf{u}\| \|\mathbf{v}\|)$ — not raw distance: frequent words grow longer vectors in training, and direction strips that out.

Analogies and the Bias Problem

- **Geometric magic** (Mikolov et al., 2013): simple vector arithmetic captures analogies.

$$\mathbf{u}_{\text{king}} - \mathbf{u}_{\text{man}} + \mathbf{u}_{\text{woman}} \approx \mathbf{u}_{\text{queen}} \qquad \mathbf{u}_{\text{Paris}} - \mathbf{u}_{\text{France}} + \mathbf{u}_{\text{Italy}} \approx \mathbf{u}_{\text{Rome}}$$

- Words on the same semantic axis differ by roughly the same vector — gender, capital city, monetary stance. (Caveat: the clean pattern weakens for specialized, low-frequency vocabulary — validate on your own terms.)
- **Bias is encoded too** (Bolukbasi et al., 2016, *NeurIPS*; Caliskan, Bryson & Narayanan, 2017, *Science*):
  - $\mathbf{u}_{\text{man}} - \mathbf{u}_{\text{woman}}$ has a large component on *programmer* vs. *homemaker*.
  - Word embeddings reproduce IAT-style biases at human magnitudes.
  - For economic text: occupation/gender, race/credit-risk, region/sentiment correlations all transfer from the training corpus into the geometry.
- **Lesson:** embeddings are a mirror of the corpus. Checking that a sentiment vector tracks *policy* stance and not *speaker identity* is a real research step (returns in T5b) — and the encoded bias can itself become the research object.

Four Generations of Representation, One Table

Every step toward "understanding" trades away interpretability.

	Counts / TF-IDF	Word2Vec	Contextual (BERT)
Captures meaning?	no (orthogonal)	yes (distributional)	strong (context-aware)
Same word, many senses?	no	one static vector	varies with context
Interpretable / auditable?	high (word = axis)	low (latent space)	very low
Typical econ use	EPU, dictionaries	culture / network scores	sentiment, stance

Rule: when validity rests on an auditable, line-by-line evidence chain, the fancier representation is often the *worse* choice. Pick the rung your research design needs — not the newest one.

From Analogies to Constructs: The Hawk–Dove Direction (Thread B)

Define an economic concept as a direction in the space.

- Candidate monetary-stance axis: $\hat{\mathbf{v}} = \mathbf{u}_{\text{hike}} - \mathbf{u}_{\text{cut}}$. Score any FOMC passage by the cosine of its (averaged) vector with $\hat{\mathbf{v}}$.
- The same trick, generalized: culture scores from seed-word directions (Li, Mai, Shen & Yan 2021); time-varying product-market networks from 10-K business-description similarity (Hoberg & Phillips 2016).
- **Validation idea:** a genuine stance direction should co-move with policy — check the $\hat{\mathbf{v}}$-scores against subsequent fed-funds-rate changes before using them in a regression.

The separability assumption. *king* − *man* + *woman* works because the difference is (mostly) one clean dimension — everything else cancels. But *hike* co-occurs with overheating and *cut* with recessions, so $\hat{\mathbf{v}}$ may encode **stance and the business cycle at once** — like a bond that bundles duration and credit risk in one price. Purge or control for the cycle dimension before calling it "policy stance"; ask which institutional features make the coupling worse.