

L3: (Word) Embeddings

Data Mining

Sep 1, 2026

Jeff M. Phillips



Data X , raw data $x_i \in X$

① $x_i = \text{text}$ $\rightarrow$ vector $v_i \in \mathbb{R}^d$

word
(token)

document

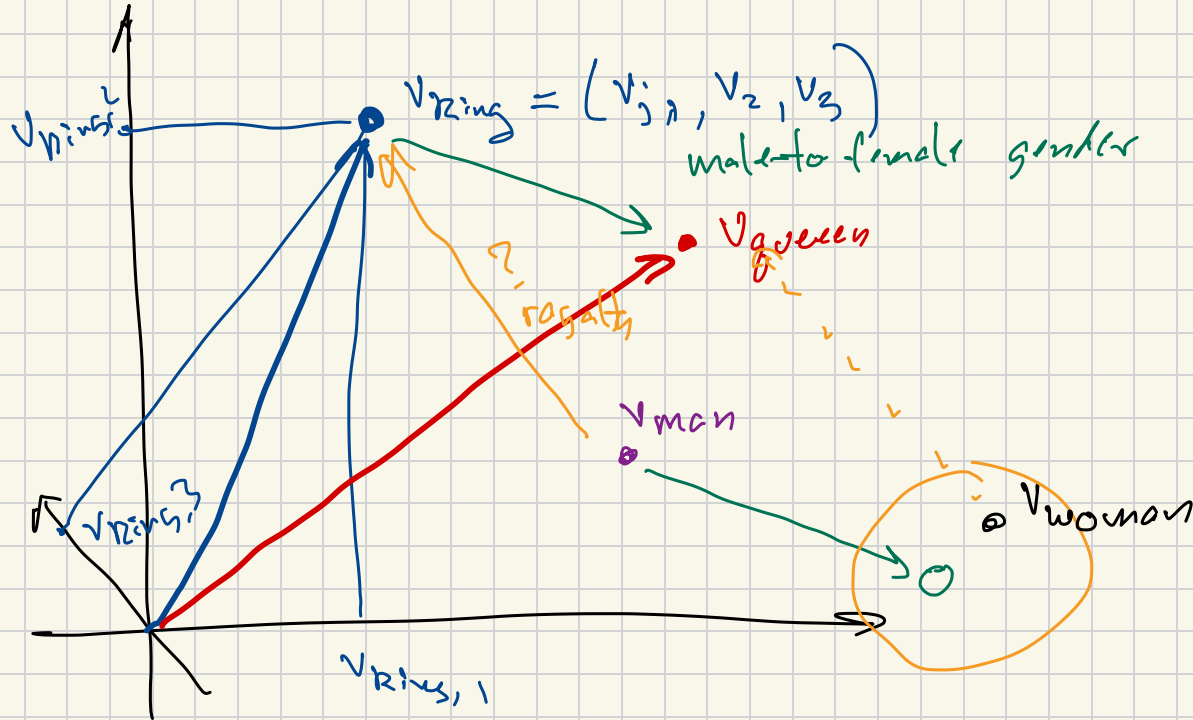
$$v_i = (v_{i1}, v_{i2}, v_{i3}, \dots, v_{id})$$
$$v_j = (v_{j1}, v_{j2}, v_{j3}, \dots, v_{jd})$$

historically: phonemes, strings, stems, thesaurus
wordnets, alphabetically, characters

② Distributional Hypothesis (Harris, Firth)
words are similar, if they used in similar context.

Word Vectors

$$v_j = v_{\text{king}} \in \mathbb{R}^{d=300}$$
$$j \in [m]$$



$m = 100,000$ words

works with large corpus, N words

word = "octopus"
 $j \rightarrow$ octopus

The scientist watched an octopus glide gracefully through the tank.
During our morning dive, a shy octopus hid beneath a rocky ledge.
My little sister drew a colorful octopus wearing eight yellow boots.
Near the reef, the clever octopus opened a tightly sealed glass jar.
After the storm, a stranded octopus slowly crawled back to the sea.

co-occurrence vector $b_j \in \mathbb{R}^m$

$b_{ij} = \# \text{ times word } j \text{ has word } i \text{ in context}$

probability $p(j) = \frac{\# \text{ words } j \text{ occurs}}{N}$

$$p(j) = b_{ij} / N$$

co-occurrence vectors $b_j \in \mathbb{R}^m$

$b_{ij} = \# \text{ times } i \text{ occurs in context } b_j$

point-wise mutual information b_{ij}

$\log \left(\frac{p(i,j)}{p(i) \cdot p(j)} \right)$

> 1 if similar

< 1 if not together

$$v_{j,i} = \max \left(0, \log \left(\frac{p(i,j)}{p(i) \cdot p(j)} \right) \right)$$

unlikely
round to 0

for documents: similar ideas tf-idf, BM25

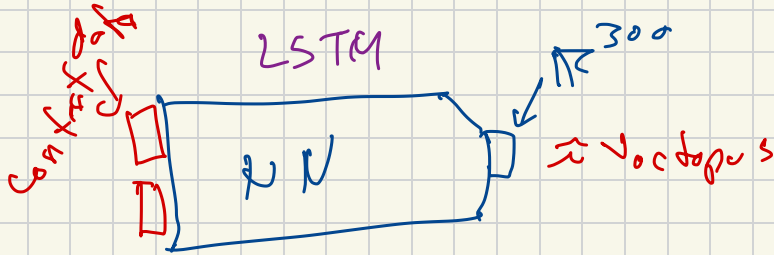
Self-supervised Learning

$$v_j \in \mathbb{R}^{300}$$

Suddenly, Mia spotted an octopus. waving eight arms from the shipwreck.

As I peered into the cave, eight curling arms emerged, followed by a curious octopus.

next word prediction



→ Word2Vec (Google)
2013

→ GloVe (Stanford)
2014

each word "bank" 1 representation

$$v_{\text{bank}} \in \mathbb{R}^{300}$$

bank = store money

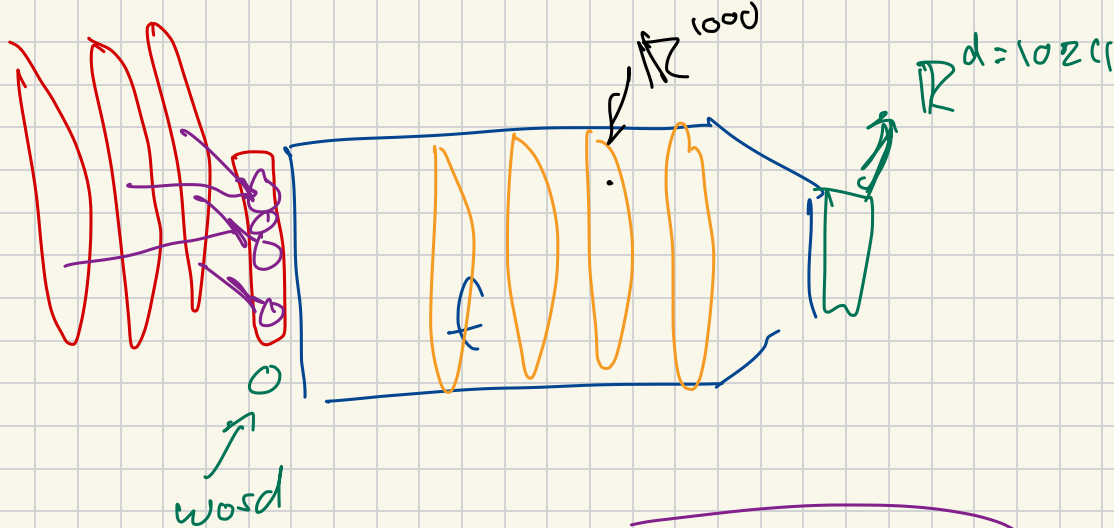
bank = side of a river

→ 2017 ELMo (Washington, AI²)

NN $\left[\boxed{\text{word}} + \boxed{\text{context}} \right] \rightarrow v_i \in \mathbb{R}^d$

word $\rightarrow \boxed{} \rightarrow v \in \mathbb{R}^{300}$

→ 2017/2018 BERT (Google, Transformers)
context, only hebrew



Bert + Attention Transformers

$f[\text{context} + \text{word}]$

```
import gensim.downloader as api
from transformers import AutoTokenizer, AutoModel
```

```
# One fixed vector for "bank"—no context
static = api.load("glove-twitter-25")["bank"]
```

→ GloVe

```
# A contextual vector for "bank" in a sentence
text = "She deposited money at the bank."
tokenizer = AutoTokenizer.from_pretrained("bert-base-uncased")
model = AutoModel.from_pretrained("bert-base-uncased")
```

```
inputs = tokenizer(text, return_tensors="pt")
output = model(**inputs).last_hidden_state
i = tokenizer.convert_ids_to_tokens(inputs.input_ids[0]).index("bank")
contextual = output[0, i].detach().numpy()
```

→ Bert