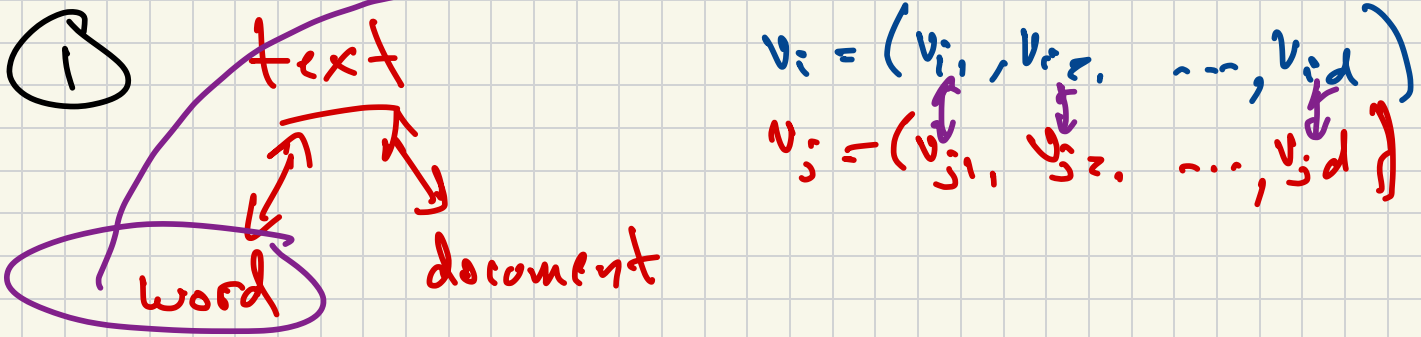
L3: Word Embeddings

Aug 25, 2025
Data Mining

Jeff Phillips

Data X, raw data $x_i \in X$

raw data $\longrightarrow$ vector $v_i \in \mathbb{R}^d$



2 Distributional Hypothesis

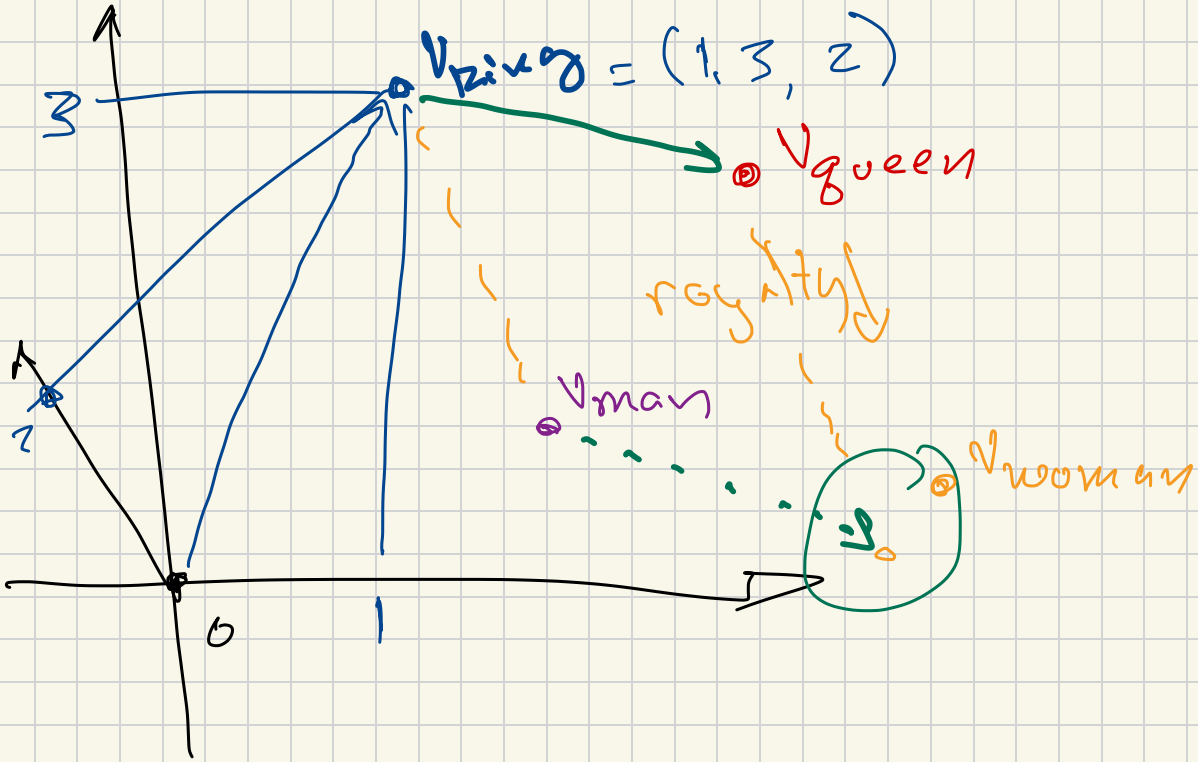
alts: alphabetical, decay-decimal, word-net

↳ words are similar if they tend to have similar context.

Word Vectors

v_j , $v_{king} \in \mathbb{R}^{d=300}$

$j \in [m]$



$m = 100{,}000$ words

word = "octopus"

Corpus of text w/ N words.

"... I saw octopus [redacted] teacher this weekend"

"the larger octopus ate the fish at the zoo."

Co-occurence vector $b_j \in \mathbb{R}^m$

b_{ji} = # times word j has word i in its context.

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

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

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

b_{ji} = # times word j has word i in its context.

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

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

pointwise mutual information of i, j

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

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

vector $v_j = (v_{j,1}, v_{j,2}, \ldots v_{jm}) \in \mathbb{R}^m$

similar

tf-idf

→ BM25

Self-Supervised Learning

(X, y) y_{i-1} y_{i+1}

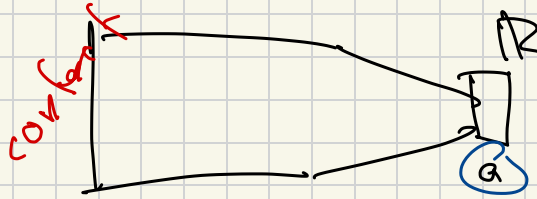
labels come from expert

next word prediction

"The fish swam away ... from the octopus, right to the shark."

Mask out a word j

use context to predict its value.



$\langle a, v_j \rangle$

2013, 2014

Word2vec GLoVe

= each word has 1 representation

"bank"

↳ 2019 Elmo (LSTM)

f_θ[context + word] $\Rightarrow v \in \mathbb{R}^d$

↳ Best (Transformer, attention)

context text