Word Embeddings

1 - hot representation - represent every word as a $|V|$ vector (size of vocab)

with all 0's except for one 1 on the index of that word in the sorted English language

$$w^{aardvark} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

However, in this representation, inner products ≠ similarity

$$(w^{hotel})^T w^{motel} = (w^{hotel})^T (w^{cat}) = 0$$

So that's not a good vector representation of a word.

Try again.

Co-occurence matrix

$$X = \begin{bmatrix} & & & k & & \\ & & & \downarrow & & \\ & & & & & \\ i \rightarrow & & & & & \\ & & & & & \\ & & & & & \end{bmatrix}$$

# of times word i appears inside a window of interest ("context") of k

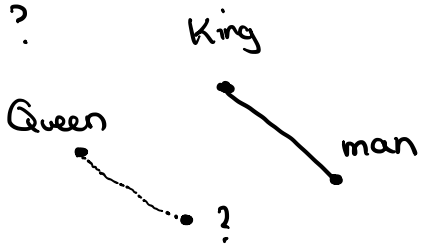
← X is very sparse since most words don't co-occur

Try taking SVD of X, take first k singular vectors & reconstruct X so it's not so sparse.

- SVD is computationally expensive
- requires hacks since X has imbalance in word frequency.

Word2vec - Mikolov et al 2013 - surprised everyone

King - Queen = man - ?



linear model!

Note: neural networks had been used since 2003 (Bengio et al)
and even the designers of Glove tried them (HSM)
but it didn't work back then.
works now though (ELMO, BERT, GPT-2, etc.)

GloVe - Global Vectors (Pennington Socher Manning 201?)

$\overline{\overline{X}}$ = matrix of word-word cooccurrence counts

X_{ij} = # times word j appears in the context of word i

$P_{ij} = P(j|i) = \frac{X_{ij}}{X_i}$ = prob word j appears in the context of word i

eg i = water $X_i = 20341$ $\quad P_{ij} = \frac{5311}{20341} \approx \frac{1}{4}$
j = fall $X_{ij} = 5311$

Consider i = ice
j = steam

there is a word related to ice but not steam: k = solid
we expect P_{ik} to be much larger than P_{jk}

that is $\frac{P_{ik}}{P_{jk}}$ should be large.

For a word related to steam but not ice, k = gas, the ratio $\frac{P_{ik}}{P_{jk}}$ ~ small, close to 0.

For words that are related to both ice & steam or water

to neither (aardvark), then $\frac{P_{ik}}{P_{jk}} = 1$.

This happens in real data.

Let's say we want to create a model for the ratio $\frac{P_{ik}}{P_{jk}}$. It must depend on words i, j, k:

$$F(\omega_i, \omega_j, \tilde{\omega}_k) = \frac{P_{ik}}{P_{jk}}$$

(ω_i, ω_j: word vectors; $\tilde{\omega}_k$: context word vector)

Because F is supposed to encode things in a vector space (which means we need linearity)

Choose

$$F(\omega_i - \omega_j, \tilde{\omega}_k) = \frac{P_{ik}}{P_{jk}}$$

Instead of choosing F to be a neural network, try this:

$$F((\omega_i - \omega_j)^T \tilde{\omega}_k) = \frac{P_{ik}}{P_{jk}} \quad (*1)$$

Then they wanted symmetry. Should be able to exchange word and context word since both words are in each other's context.

Some weird stuff follows:

want $+ \rightsquigarrow \times$, $- \rightsquigarrow \div$

$$\rightarrow F((\omega_i + \omega_j)^T \tilde{\omega}_k) = F(\omega_i^T \omega_k) F(\omega_j^T \tilde{\omega}_k)$$

$$\rightarrow F((\omega_i - \omega_j)^T \tilde{\omega}_k) = \frac{F(\omega_i^T \tilde{\omega}_k)}{F(\omega_j^T \tilde{\omega}_k)} \quad (*2)$$

$$(*1) = (*2) \Rightarrow F(\omega_i^T \tilde{\omega}_k) = P_{ik} = \frac{X_{ik}}{X_i}$$

Soln to (*2) is $F = \exp$ since $\exp(a+b) = e^a e^b$

$$\exp(a-b) = \frac{e^a}{e^b}.$$

So, $F(w_i^T \tilde{w}_k) = \exp(w_i^T \tilde{w}_k) = P_{ik}$

$$w_i^T \tilde{w}_k = \log P_{ik} = \log(X_{ik}) - \log(X_i)$$

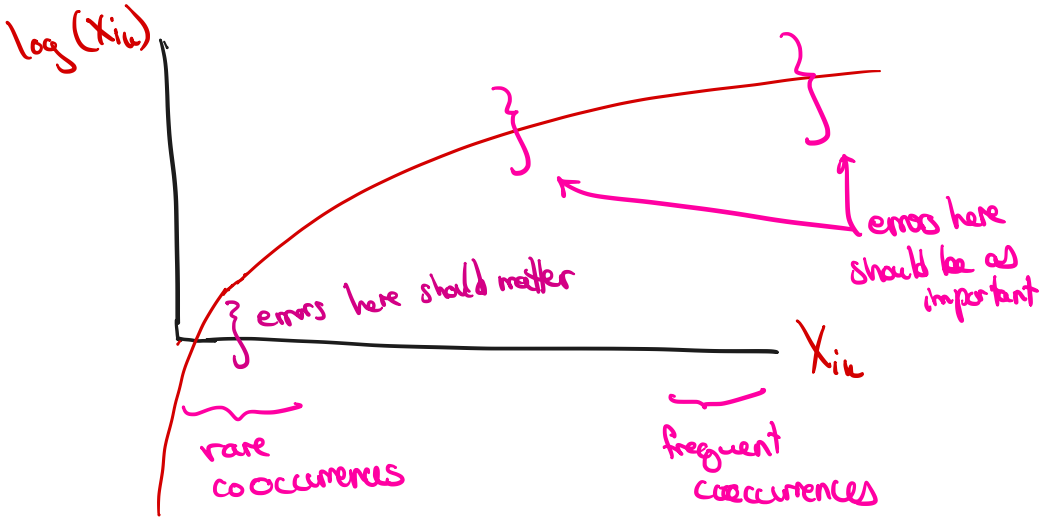
but this isn't symmetric. So they did this:

$$w_i^T \tilde{w}_k + b_i + \tilde{b}_k = \log(X_{ik})$$

Symmetric in form at least.

But then there are issues

- $\log(0)$ ← ugh, happens often too...
- forcing $w_i \tilde{w}_k + b_i + \tilde{b}_k$ to equal $\log(X_{ik})$ will have issues:

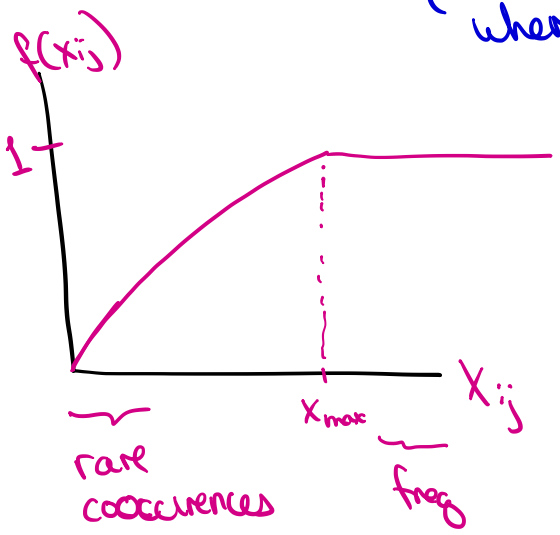


Form objective as weighted least squares:

$$J = \sum_{i,j=1}^{|V|} f(X_{ij}) \underbrace{\left(w_i^T \tilde{w}_j + b_i + \tilde{b}_j \right)}_{\text{model}} - \underbrace{\log X_{ij}}_{\text{data}}\Big)^2$$

($|V|$ ← vocab size)

where $f(0) = 0$ (no more $\log 0$ .... make sure $\lim_{x \to 0} f(x) \log^2 x$ is finite!)



$f(x)$ nondecreasing so that rare occurences aren't over weighted

$f(x)$ small ish for large x so frquent cooccurences aren't overweighted

They chose $f(x) = \begin{cases} \left(x / x_{max}\right)^{\alpha} & \text{if } x < x_{max} \\ 1 & \text{otherwise} \end{cases}$

$x_{max} = 100$

$\alpha = 3/4$

So to get an embedding of a word, compute:

$$\min_{\{w_i, \tilde{w}_i, b_i, \tilde{b}_i\}_i} J$$

w_i and $\tilde{w}_i$ should be the same but they aren't!

Use $w_i + \tilde{w}_i$ as the embedding for i

Trained on wikipedia

Tasks:

Named Entity Recognition

[Jen]$_{\text{person}}$ bought 300 shares of [Acme Corp]$_{\text{organization}}$ in [2019]$_{\text{time}}$

Word Analogies

Athens is to Greece as Berlin is to ____?

Worse Sense Disambiguation

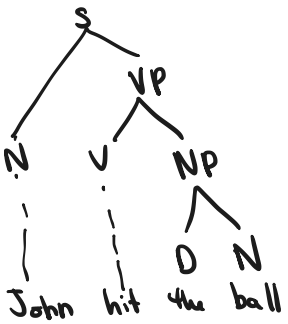
bank → I withdrew cash at the bank
bank → I saw the riverbank

bridge, bass, tie

POS tagging
part of speech

Did you book the artist? Did you leave your book behind?

Parsing



Coreference resolution

The project leader, is not helping. The fool, thinks only of himself

Other tasks:

- Text generation
- Text summarization
- Question Answering
- Machine Translation