Problems with count vectorizers:

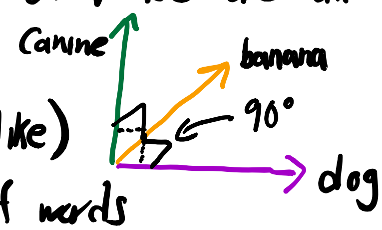
① each word is equidistant in its vector representation:

intuitively, we know dog should be more similar to canine than banana.

dog: [0 0 1]
banana: [1 0 0]
canine: [0 1 0]

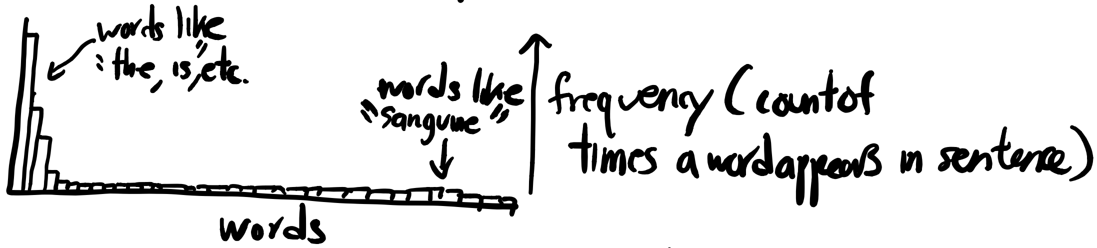
all dot products are the same (0).

The angles made by a pair of vectors are all orthogonal:



② counts skew towards high frequency stopwords (the, is, on, I, like)

Remember the long tail distribution of words we saw in Week 1 (Zipf's Law):



Problems with rule-based domain trees:

- color
  - purple
    - violet
    - mauve
    - orchid
  - red
  - green

(i.e. hyponyms and hypernyms)

- extremely labor intensive.
- slow to adapt to new words
- difficult to maintain
- "subjective"; two SMEs (subject matter experts) can have two totally valid trees.

So... what do we do?

We make a new assumption: the distributional hypothesis: "You shall know a word by the company it keeps" - JR Firth (British linguist).

Continuous Bag of Words

I used class notes to study for the test.

$$\underbrace{t\text{-}3 \quad t\text{-}2 \quad t\text{-}1 \quad t \quad t{+}1 \quad t{+}2 \quad t{+}3}_{\text{context window } (m=3)}$$

$$P(x_t = \text{study} \mid x_{t-3} = \text{class}) \times$$

$$P(x_t = \text{study} \mid x_{t-2} = \text{notes}) \times$$

$$\dots$$

$$\boxed{P(x_t = \text{study} \mid x_{t+3} = \text{test})}$$

probability that study is the target word given that the context word is test.

How would we find this probability?

these words represent the targets

these words represent the context

$$
\begin{array}{c|cccccc}
 & \text{ant} & \text{art} & \text{bad} & \cdots & \text{study} \cdots & \text{zebra} \\
\hline
\text{ant} & & & & & & \\
\text{art} & & & & \cdots & & \\
\text{bad} & & & & & & \\
\vdots & & & & & & \\
\text{giraffe} & 4 & 6 & 2 & & 5 & 7 \\
\vdots & & & & & & \\
\text{test} & 3 & 2 & 5 & & 9 & 6 \\
\vdots & & & & & & \\
\text{zebra} & 1 & 2 & 3 & & 4 & 0
\end{array}
$$

$$\frac{\text{count(test, study)}}{\text{count test}} =$$

$$\frac{9}{3+2+5+9+6} = \frac{9}{25} = \underline{.36}$$

However, in practice, we typically do not use CBOW, but rather <u>skipgram</u>, which is the opposite of CBOW:

	Inputs	Target
CBOW	context words	target word
skipgram	target word	context words

$P(X_{t-3}=\text{class} \mid X_t=\text{study})$

$P(x_{t-1}=t_0 \mid x_t=\text{study}) \times$

$\vdots \qquad \vdots$

$P(X_{t+3}=\text{test} \mid X_t=\text{study})$

$$\Rightarrow \boxed{\prod_{\substack{-M \le m \le M \\ m \ne 0}}} P(X_{t+m} \mid X_t)$$

"for each context word in the window"

$M=3$

context window size is 3.

We do this for each target word:

for each target word

for each context word

$$\prod_{t=1}^{T} \prod_{\substack{-M \le m \le M \\ m \ne 0}} P(X_{t+m} \mid X_t, \theta)$$

context word

target word

word vectors (our parameters, will learn more)

This becomes our objective function that we optimize (the higher this value the more "correct" our model is):

$$\boxed{J(\theta) = \prod_{t=1}^{T} \prod_{\substack{-M \le m \le M \\ m \neq 0}} P(x_{t+m} \mid x_t, \theta)}$$

So what exactly is θ? Its actually two matrices:

context word vector

study's context vector

$$\begin{array}{l}
\text{ant} \\ \text{art} \\ \\ \text{study} \\ \\ \text{zebra}
\end{array}
\begin{bmatrix}
4.3 & 0.1 & 0.4 \\
\vdots & \vdots & \vdots \\
\boxed{-0.4 \quad 1.2 \quad 3.4 \quad \ldots} & & \\
\vdots & \vdots & \vdots \\
-0.2 & 1.2 & 2.1
\end{bmatrix}$$

$\xleftrightarrow{\qquad}$

$D = 300$

target word vector

study's target vector

$$\begin{bmatrix}
1.2 & -1.1 & -3.2 \\
\vdots & \vdots & \vdots \\
\boxed{1.2 \quad 0.4 \quad 1.5} & & \\
\vdots & \vdots & \vdots \\
-1.1 & -2.1 & 3.2
\end{bmatrix}$$

$\xleftrightarrow{\qquad}$

$D = 300$

Each word has <u>both</u> a <u>context</u> vector and a <u>target</u> vector.