The Language Modeling problem

- Create a probability distribution over all sequences of words
  - finite vocabulary: Σ
  - infinite set of sequences: Σ^*
    - Any sentence/sequence of words $\boldsymbol{e} = \boldsymbol{w_1 w_2 \dots w_n}$ is an element of Σ^*

$$\sum_{e \in \Sigma^*} P_{LM}(e) = 1$$

$$P_{LM}(e) \geq 0 \quad \forall e \in \Sigma^*$$

Our First Attempt

- Assume we have N training sentences
- Let $w_1 w_2 \dots w_n$ be a sentence, and $\mathrm{count}(w_1, w_2, \dots, w_n)$ be the number of times it appeared in the training data.
- Define a language model:

$$P(w_1, \dots, w_n) = \frac{\mathrm{count}(w_1, \dots, w_n)}{N}$$

Unigram Language Model

"I have a dog whose name is Lucy. I have two cats, they like playing with Lucy."

- $\sum_{w \in \Sigma} \text{count}(w) =$
- $P(Lucy) =$
- $P(cats) =$

$$P(w_k | w_{1:k-1}) \approx P(w_k)$$

$$\hat{P}(w) = \frac{\text{count}(w)}{\sum_{v \in \Sigma} \text{count}(v)}$$

Bigram Language Model

"I have a dog whose name is Lucy. I have two cats, they like playing with Lucy."

- $P(\text{have} \mid \text{I}) =$
- $P(\text{two} \mid \text{have}) =$
- $P(\text{eating} \mid \text{have}) =$

$$P(w_k | w_{1:k-1}) \approx P(w_k | w_{k-1})$$

$$\hat{P}(w_2 \mid w_1) = \frac{\text{count}(w_1 w_2)}{\text{count}(w_1)}$$