

1. 1-gram and 2-gram Frequency Distributions.

Unigram Frequencies:

I: 3 love: 4 studying: 2 natural: 1
language: 1 processing: 1 in: 1 university: 1
deep: 1 learning: 1 Anhui: 2 We: 1

Bigram Frequencies: This part ignores the beginning and end markers.

I love: 3 love studying: 2 studying natural: 1
natural language: 1 language processing: 1 processing in: 1
in university: 1 studying deep: 1 deep learning: 1
love Anhui: 2 We love: 1

2. Probability Distributions.

Unigram Probabilities $P(\text{word}) = \text{count} / 19$.

I: $\frac{3}{19}$ love: $\frac{4}{19}$ studying: $\frac{2}{19}$ natural: $\frac{1}{19}$
language: $\frac{1}{19}$ processing: $\frac{1}{19}$ in: $\frac{1}{19}$ university: $\frac{1}{19}$
deep: $\frac{1}{19}$ learning: $\frac{1}{19}$ Anhui: $\frac{2}{19}$ We: $\frac{1}{19}$

Bigram Probabilities $P(w_2 | w_1) = \text{count}(w_1 w_2) / \text{count}(w_1)$

$P(\text{love} | \text{I}) = 1$ $P(\text{studying} | \text{love}) = \frac{1}{2}$ $P(\text{natural} | \text{studying}) = \frac{1}{2}$

$P(\text{language} | \text{natural}) = 1$ $P(\text{processing} | \text{language}) = 1$

$P(\text{in} | \text{processing}) = 1$ $P(\text{university} | \text{in}) = 1$ $P(\text{deep} | \text{studying}) = \frac{1}{2}$

$P(\text{learning} | \text{deep}) = 1$ $P(\text{Anhui} | \text{love}) = \frac{1}{2}$ $P(\text{love} | \text{We}) = 1$

3. Sentence Probabilities

Sentence 1: "I love deep learning."

$$\text{Unigram } P = \frac{3}{19} \times \frac{4}{19} \times \frac{1}{19} \times \frac{1}{19} = \frac{12}{130321}$$

$$\text{Bigram: } P = \frac{3}{19} \times 1 \times 0 \times 1 = 0$$

Sentence 2: "I love Anhui University."

$$\text{Unigram: } P = \frac{3}{19} \times \frac{4}{19} \times \frac{2}{19} \times \frac{1}{19} = \frac{24}{130321}$$

$$\text{Bigram: } P = \frac{3}{19} \times 1 \times \frac{1}{2} \times 0 = 0$$

Handling New words:

Add-one Smoothing:

$$P(\text{love} | \text{I}) = \frac{1+3}{13+3} = \frac{4}{16} \quad P(\text{deep} | \text{love}) = \frac{1+0}{12+4} = \frac{1}{16}$$

$$\text{Bigram: } P(\text{learning} | \text{deep}) = \frac{1+1}{12+1} = \frac{2}{13} \quad P(\text{Anhui} | \text{love}) = \frac{1+2}{12+4} = \frac{3}{16}$$

$$\text{Sentence 1: } P = \frac{3}{19} \times \frac{4}{16} \times \frac{1}{16} \times \frac{2}{13} = \frac{1}{2470}$$

$$\text{Sentence 2: } P = \frac{3}{19} \times \frac{4}{16} \times \frac{3}{16} \times \frac{1}{14} = \frac{3}{5320}$$

4. Next Word Prediction with 2-gram Model

After "I love":

Next words: "studying" (probability = 0.5)

"Anhui" (probability = 0.5)

Prediction: Tie between "studying" and "Anhui".

After "deep":

Next words: "learning" (probability = 1.0)

Prediction: "learning".