2252654 NLP Lab3 Math Exercise

Phuong Huynh

February 2025

1 Problem 1

The trigram probability estimation is given by:

$$P(w_3|w_1, w_2) = \frac{\text{Count}(w_1, w_2, w_3)}{\text{Count}(w_1, w_2)}$$

Trigram Probabilities

The non-zero trigram probabilities are:

$$P(\text{am}| < s >, I) = \frac{1}{2}$$

$$P(\text{Sam}|I, am) = \frac{1}{2}$$

$$P(< /s > |am, Sam) = \frac{1}{1} = 1$$

$$P(I| < s >, Sam) = \frac{1}{1} = 1$$

$$P(am|Sam, I) = \frac{1}{1} = 1$$

$$P(< /s > |I, am) = \frac{1}{2}$$

$$P(do| < s >, I) = \frac{1}{2}$$

$$P(not|I, do) = \frac{1}{1} = 1$$

$$P(like|do, not) = \frac{1}{1} = 1$$

$$P(green|not, like) = \frac{1}{1} = 1$$

$$P(eggs|like, green) = \frac{1}{1} = 1$$

$$P(and|green, eggs) = \frac{1}{1} = 1$$

$$P(Sam|eggs, and) = \frac{1}{1} = 1$$

$$P(</s> |and, Sam) = \frac{1}{1} = 1$$

2 Problem 2

The probability of the sentence is calculated as:

$P(\text{i want chinese food}) = P(i \mid \langle s\rangle) \times P(\text{want} \mid i) \times P(\text{chinese} \mid \text{want}) \times P(\text{food} \mid \text{chinese}) \times P(\langle /s\rangle \mid \text{food})$

$P_{unsmoothed}(\text{i want chinese food}) = 0.19 \times 0.33 \times 0.0065 \times 0.52 \times 0.40 = 0.0000848$

$P_{smoothed}(\text{i want chinese food}) = 0.19 \times 0.21 \times 0.0029 \times 0.52 \times 0.40 = 0.0000241$

3 Problem 3

The unsmoothed probability is higher than the smoothed probability. This is because Laplace smoothing redistributes some probability mass to unseen or rare events, which reduces the probability of the observed events. The bigrams in question 2 are all more frequent bigrams therefore it is discounted resulting in a smaller probability.

4 Problem 4

The unique words in the corpus are: $\langle s\rangle$, I, am, Sam, $\langle /s\rangle$, do, not, like, green, eggs, and. So, the vocabulary size $V = 11$.

$$\text{Count}(\text{am}, \text{Sam}) = 2$$

$$\text{Count}(\text{am}) = 3$$

$$\text{Smoothed Count}(\text{am}, \text{Sam}) = \text{Count}(\text{am}, \text{Sam}) + 1 = 2 + 1 = 3$$

$$\text{Smoothed Count}(\text{am}) = \text{Count}(\text{am}) + V = 3 + 11 = 14$$

$$P(\text{Sam} \mid \text{am}) = \frac{\text{Smoothed Count}(\text{am}, \text{Sam})}{\text{Smoothed Count}(\text{am})} = \frac{3}{14} \approx 0.214$$

5 Problem 5

$$\text{Count}(\text{am}, \text{Sam}) = 2$$

$$\text{Count}(\text{am}) = 3$$

$$\text{Count}(\text{Sam}) = 4$$

$$P_{\text{bigram}}(\text{Sam} \mid \text{am}) = \frac{\text{Count}(\text{am}, \text{Sam})}{\text{Count}(\text{am})} = \frac{2}{3} \approx 0.6667$$

$$P_{\text{unigram}}(\text{Sam}) = \frac{\text{Count}(\text{Sam})}{\text{Total words}} = \frac{4}{25} = 0.16$$

$$P(\text{Sam} \mid \text{am}) = \frac{1}{2} \cdot 0.6667 + \frac{1}{2} \cdot 0.16 = 0.33335 + 0.08 = 0.41335$$

6 Problem 6

- Probability of 0:

$$P(0) = \frac{91}{100} = 0.91$$

- Probability of each digit 1 through 9:

$$P(1) = P(2) = \cdots = P(9) = \frac{1}{100} = 0.01$$

$$Perplexity(0000030000) = \sqrt[10]{\frac{1}{0.91^9 \times 0.01}} = 1.725$$