ECE/CS/ME 539 – Fall 2024 — Activity 29

Practice Problems

Text Corpus

```
the cat sat on the mat the cat saw a rat the rat ran from the cat
```

Note: The corpus contains 20 words with several repeated words to ensure redundancy.

Problem 1: Unigram Counts and Probabilities

a. List all the unique words (unigrams) in the corpus and compute their counts.
b. Calculate the unigram probabilities for each word.

Problem 2: Bigram Counts and Probabilities

a. Generate a list of all bigrams in the corpus and compute their counts.
b. Calculate the bigram probabilities without any smoothing.

Problem 3: Trigram Counts and Probabilities

a. List all the trigrams in the corpus and compute their counts.
b. Explain why some trigrams might have zero counts and discuss how this affects probability estimation.

Problem 4: Applying Laplace (Add-One) Smoothing

a. Apply Laplace smoothing to the bigram counts obtained in Problem 2a.

b. Recalculate the bigram probabilities using the smoothed counts.

Problem 5: Parameter Counts and Memory Requirements

a. Calculate the total number of parameters required to store all possible unigram probabilities for this vocabulary.

b. Calculate the total number of parameters required to store all possible bigram probabilities for this vocabulary.

c. Discuss how the number of parameters increases with n-gram size (e.g., from unigram to bigram to trigram) and the challenges that arise.

Problem 6: Sample Calculations

a. Using the bigram probabilities (with and without smoothing), calculate the probability of the sentence: “the cat ran from the rat.”

b. Discuss the differences in the calculated probabilities with and without Laplace smoothing.