CSCI366/CSCI780 Homework 1 Report

Part I:

Given corpus

```
corpus = [
    "<s> I am Sam </s>",
    "<s> Sam I am </s>",
    "<s> I am Sam </s>",
    "<s> I do not like green eggs and Sam </s>"
]
```

From the given corpus we get,

```
C(am, Sam) = 2,
C(am) = 3,
and V = 11

probability_sam_am = (c_am_sam + 1) / (c_am + vocabulary_size)
                   = (2+1)/(3+11)
                   = 3/14
                   = 0.21

So, the Probability of P(Sam|am): 0.21
```

1.3 Questions

1. **Number of Word Types in the Training Corpus:** 41,739 (including `</s>` and `<unk>`).
2. **Number of Word Tokens in the Training Corpus:** 2,468,210 (excluding `<s>`).
3. **Percentage of Unseen Word Types in the Test Corpus:** 3.74%.
    - Percentage of Unseen Word Tokens: 1.66%.
4. **Percentage of Unseen Bigram Types and Tokens in the Test Corpus:**
    - Types: 25.32%
    - Tokens: 20.96%.
5. **Log Probability for the Sentence "I look forward to hearing your reply.":**
    - Unigram Model: -94.94
    - Bigram Model: undefined
    - Add-One Bigram Model: -97.14.
6. **Perplexity for the Sentence "I look forward to hearing your reply.":**
    - Unigram Model: 721.01
    - Bigram Model: undefined
    - Add-One Bigram Model: 839.83.
7. **Perplexity of the Entire Test Corpus:**
    - Unigram Model: 11,715.10
    - Bigram Model: undefined
    - Add-One Bigram Model: 3,118.87.

Discussion of Differences

The preprocessing steps significantly reduced the number of unseen tokens and word types in the test corpus, leading to lower perplexity values for all models. The Bigram with Add-One Smoothing performed better than the Unigram MLE, likely due to its ability to address zero probabilities for unseen bigrams. However, the Bigram MLE struggled with unseen values, resulting in higher perplexity. This highlights the importance of preprocessing and model selection in improving language model performance.