N-gram Language Models Exercises

Following exercises will refer to Chapter 3. N-gram Language Models (SLP3). See: https://web.stanford.edu/~jurafsky/slp3/3.pdf

Problem 1: Write out the equation for trigram probability estimation (modifying Eq. 3.11). Now write out all the non-zero trigram probabilities for the I am Sam corpus on page 4.

Problem 2: Calculate the probability of the sentence `i want chinese food`. Give two probabilities, one using Fig. 3.2 and the "useful probabilities" just below it on page 6, and another using the add-1 smoothed table in Fig. 3.6. Assume the additional add-1 smoothed probabilities P(i|<s>) =0.19 and P(</s>|food) =0.40.

Problem 3: We are given the following corpus, modified from the one in the chapter:

```
<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>
```

Using a bigram language model with add-one smoothing, what is P(Sam|am)? Include `<s>` and `</s>` in your counts just like any other token.

Problem 4: We are given the following corpus, modified from the one in the chapter:

```
<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>
```

If we use linear interpolation smoothing between a maximum-likelihood bi-gram model and a maximum-likelihood unigram model with $\lambda_1 = \frac{1}{2}$ and $\lambda_2 = \frac{1}{2}$, what is P(Sam|am)? Include <s> and </s> in your counts just like any other token.

Problem 5: You are given a training set of 100 numbers that consists of 91 zeros and 1 each of the other digits 1-9. Now we see the following test set: 0 0 0 0 0 3 0 00 0. What is the unigram perplexity?