Report Of Homework 1

Part I:

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.

Answer:

|V|=11,

Am Sam = 2

Am = 3

Using add-one smoothing:

P (Sam | am) = (C (am, Sam) + 1)/ C(am)+|V|

= (2+1) / (3+11)

= 3/14

=0.21

The Probability of P (Sam | am) = 0.21

Part 2: 1.3 Questions answers:

1. Number of word types (unique words) are in the training corpus:
   Ans: Number of word types in the training corpus (including </s> and <unk>): 41740.
2. Number of word tokens in the training corpus:
   Ans: Number of word tokens in the training corpus (excluding <s>): 5036420
3. Percentage of unseen word types: 1.41%
   Percentage of unseen word tokens: 0.14%

4. Percentage of bigram types in the test corpus that did not occur in training: 1.41%.
   Percentage of bigram tokens in the test corpus that did not occur in training: 0.28%.

5. Log Probability for the sentence under Unigram model: -33.219280948873624
   Log Probability for the sentence under Bigram model: 0.0
   Log Probability for the sentence under Add-One Bigram model: -22.13488456773567
6. Perplexity for the sentence under Unigram model: 12.915496650148839
   Perplexity for the sentence under Bigram model: 1.0
   Perplexity for the sentence under Add-One Bigram model: 5.499999999999998
7. Perplexity for the test corpus under Unigram model: 68129206.91033816
   Perplexity for the test corpus under Bigram model: undefined
   Perplexity for the test corpus under Add-One Bigram model: 9.19657031621839