For 1.3 Questions.

It's hard to debug the program with very big training data.
Make a small training file, 3-4 sentences, a few words each sentence.
Do it by hand
To check lgic error.

Part I - 3.4

$P(\text{Sam} \mid \text{am}) = C(\text{am}, \text{Sam}) / C(\text{am})$

$= 2 / 3$

↓ Add one smoothing

$= (2+1) / (3+10)$
$= 3/13$
$= 0.23077$

am Sam 2
am * 3
V = 10

</s>
I
am
sam
do
not
like
green
eggs
and

1.3 Questions.

1. 41738

2. 2468210

3. Word token: 1.6612 %
   Word type: 3.6058 %

4. Word token: 22.3679 %
   Word type: 99.8763 %

5. S = i look forvad to hearing your reply . </s>

unigram params: i look forvad to hearing your reply . </s>

$$P(S) = \log_2 P(i) + \log_2 P(look) + \log_2 P(forvad) + \log_2 P(to) + \log_2 P(hearing) + \log_2 P(your) + \log_2 P(reply) + \log_2 P(.) + \log_2 P(</s>)$$

$$= -8.4510 - 12.0326 - 12.4036 - 5.3973 - 13.5850 - 11.0432 - 17.5919 - 4.8689 - 4.6827$$
$$= -90.2561$$

bigram params: i look, look forvad, forvad to, to hearing, hearing your, your reply, reply ., . </s>

$\log_2 \frac{Count(i, look)}{Count(i)}$

$$P(S) = \log_2 P(i\ look) + \log_2 P(look\ forvad) + \log_2 P(forvad\ to) + \log_2 P(to\ hearing) + \log_2 P(hearing\ your) + \log_2 P(your\ reply) + \log_2 P(reply\ .) + \log_2 P(.\ </s>)$$

P(parameter) of these three are 0.

The result of the sentence under bigram is undefined. But if we treat $\log_2 P(\ )$ of these token as 0, we get ...

$$= -28.5463$$

bigram with add one smoothing

parameters: i look, look forvad, forvad to, to hearing, hearing your, your reply, reply ., . </s>

$P(S) = \log_2 P(i\ look) + \log_2 P(look\ forward) + \log_2 P(forward\ to) + \log_2 P(to\ hearing) + \log_2 P(hearing\ your) + \log_2 P(your\ reply) + \log_2 P(reply\ .) + \log_2 P(. \ </s>)$

$= \log_2 \frac{count(i, look)+1}{count(i)+|V|^2} + \dots\dots$

$= -26.6982 - 25.5687 - 24.0400 - 27.8409 - 30.6982 - 30.6982 - 30.6982 - 14.3594$

$= -210.6521$

6. S = i look forvad to hearing your reply . </s>

Perplexity $= 2^{-L}$ $\quad L = \frac{1}{M}\sum_{i=1}^{m} \log P(S_i)$

$|M| = 9$

$L = \frac{1}{9} \cdot (\log prob)$

perplexity $= 2^{-\frac{1}{9} \cdot \log prob\ of\ each\ model}$

unigram ppl $= 2^{-\frac{1}{9}\cdot(-89.7404)} = 1003.7300$

bigram ppl $= 2^{-\frac{1}{9}\cdot(-28.5463)} = 9.0118$

bigram with add one smoothing ppl $= 2^{-\frac{1}{9}\cdot(-210.6521)} = 11113338.12$

<s> included since Q7 not mentioned.

7. log prob

unigram = -28592.7144

bigram = -13763.2607

bigram w/ add one = -72145.7616

Perplexity/ unigram = 1283.6244

bigram = 31.3519

bigram w/ add one = 69707345.6550

Corpus:

<s> A unigram maximum likelihood model . </s>

<s> A bigram maximum likelihood model . </s>

<s> A bigram model with Add-One smoothing . </s>

V = { a 3 — Total = 7 + 7 + 8 = 22

unigram 1

maximum 2

likelihood 2

model 3

bigram 2

</s> 3

. 3

with 1

add-one 1

smoothing 1

}

unigram: P(a) = 3/22

P(unigram) = 1/22

P(maximum) = 2/22

P(likelihood) = 2/22

P(model) = 3/22

P(bigram) = 2/22

P(.) = 3/22

P(with) = 1/22

P(add-one) = 1/22

P(smoothing) = 1/22

bigram: P(a