Homework_1

Part_1

Let P* = smooth probability, find P*(Sam | am).

P*(Sam | am) = $\frac{C(am, Sam)+1}{C(am) + V}$;

C(am, sam) = 2;
C(am) = 3;
V(total unique words) = 11;

So, P*(Sam | am) = $\frac{2+1}{3+11} = \frac{3}{14}$

Part_2

Question 1:
The number of word types in the training corpus is: 83044

Question 2:
Total number word tokens in the training corpus is: 2468210

Question 3:
Percentage of word types in the test corpus did not occur in training is: 3.61%
Percentage of word tokens in the test corpus did not occur in training is: 1.6%

Question 4:
Percentage of bigrams types not occur in training is:22.52%
Percentage of bigrams tokens not occur in training is:16.11%

Question 5:
Sentence to compute: "I look forward to hearing your reply . "

```
--------------------------------------------------------------------------------
Unigram Model:
All the parameter's probability:
<s>: 0.03893762581720342
i: 0.0028576323587245593
look: 0.00023868764625945 7
forward: 0.00018456434637354423
```

```
to: 0.02065563174351007
hearing: 8.137963795795515e-05
your: 0.00047387090619536564
reply: 5.061891356236445e-06
.: 0.03422383683577278
</s>: 0.03893762581720342

The log base 2 of all the parameter's probability:<s>: -4.682691269922203
i: -8.450963962476674
look: -12.032588480668233
forward: -12.403588495460756
to: -5.597321004705777
hearing: -13.584972612278131
your: -11.043218291645285
reply: -17.591892026217923
.: -4.868854680279238
</s>: -4.682691269922203

Number of tokens is: 10
Sum of all log probability is : -94.93878209357642
The average log probability is: -9.493878209357643
------------------------------------------------------------------------------

------------------------------------------------------------------------------
Bigram Model:
All the parameter's probability:
<s> i: 0.02006
i look: 0.0020438751873552256
look forward: 0.05546492659053834
forward to: 0.2109704641350211
to hearing: 0.00011310511235107827
hearing your: 0
your reply: 0
reply .: 0
. </s>: 0.9430336541743464

The log probability can not compute due to below parameter have 0 probability:
hearing your
your reply
reply .
------------------------------------------------------------------------------

------------------------------------------------------------------------------
Bigram Model with Add One Smoothing:All the parameter's probability:
<s> i: 0.014152773760221251
i look: 0.0003056359264843718
```

```
look forward: 0.0008027956176803929
forward to: 0.002368938478667709
to hearing: 6.329981959551416e-05
hearing your: 2.3839038809955182e-05
your reply: 2.3279634975323588e-05
reply .: 2.395094845755892e-05
. </s>: 0.6393973756682326

The log base 2 of all the parameter's probability:<s> i: -6.1427713594772495
i look: -11.675898242214917
look forward: -10.282679638245058
forward to: -8.721543552387656
to hearing: -13.947439086458202
hearing your: -15.35631440692812
your reply: -15.390572037471506
reply .: -15.34955768662052
. </s>: -0.6452152721298866

Number of tokens is: 9
Sum of all log probability is : 0
The average log probability is: -10.834665697992568
--------------------------------------------------------------------------------


Question 6:
The perplexity for the sentence in each model are:
Unigram Model: 721.0113746656128
Bigram Model: undefined
Bigram Model with Add One Smoothing: 1826.2463800257967


Question 7:
The perplexity of the test corpus in each model are:
Unigram Model: 469.88004030959013
Bigram Model: undefined
Bigram Model with Add One Smoothing: 893.7963798229065
```