1. Estimating N-gram Probabilities

   Table 1 shows several statistics of two different corpora, corpus 1 and corpus 2, including some of their unigram and bigram counts. In both corpora, the beginning and end of each sentence are marked with the start and end tokens <s> and </s>.

   For both corpora, compute the Maximum Likelihood Estimate (MLE) for the unigram and bigram probabilities and enter them in table 2.

Corpus 1		**Corpus 2**	
# sentences	3600	# sentences	5100
# tokens	60000	# tokens	90000
size of vocabulary	10900	size of vocabulary	11300
Unigrams	Count	Unigrams	Count
Lisa	4	Lisa	1
likes	40	likes	15
to	2700	to	3040
run	2	run	10
Bigrams	Count	Bigrams	Count
<s> Lisa	3	<s> Lisa	1
Lisa likes	1	Lisa likes	1
likes to	20	likes to	4
to run	1	to run	5
run </s>	1	run </s>	1

Table 1: Statistics of corpora 1 and 2

	Corpus 1	**Corpus 2**
Unigrams	$P(w)$	$P(w)$
Lisa		
likes		
to		
run		
Bigrams	$P(w_2 \mid w_1)$	$P(w_2 \mid w_1)$
<s> Lisa		
Lisa likes		
likes to		
to run		
run </s>		

Table 2: Unigram and bigram probabilities for corpora 1 and 2

2. Creating a Language Model

   Given is the following text corpus:

   <s> ain't no sunshine </s>
   <s> when she's gone </s>
   <s> it's not warm </s>
   <s> when she's away </s>
   <s> ain't no sunshine </s>
   <s> when she's gone </s>

   Hint: "ain't", "she's" and "it's" are considered as one word.

   (a) Create both a unigram and a bigram model from the given corpus.

   (b) What is the most frequent unigram and bigram, respectively that includes no start or end token? What is the most likely next word after *"she's"* in each model?

   (c) Compute the probability of each of the following sentences using first your unigram and then your bigram model:

      i. <s> ain't no warm </s>

      ii. <s> she's not gone </s>

   (d) Recalculate the probabilities of sentences i. and ii. in the bigram model using add-one smoothing. Additionally, compute their perplexity.

   (e) What problem does add-one smoothing address? What other methods do you know that tackle the same issue? Explain them briefly.