Natural Language Processing
Lab 2.1

This lab sheet is to practice the concepts taught this week far with a focus on n-gram language models.

1. Calculate the probability of the sentence *i want chinese food*. Give two probabilities, one using Fig. 3.2 and another using the add-1 smoothed table in Fig. 3.7. Assume the additional add-1 smoothed probabilities $P(i| < s >) = 0.19$ and $P(< /s > |food) = 0.40$.

	i	want	to	eat	chinese	food	lunch	spend
i	0.002	0.33	0	0.0036	0	0	0	0.00079
want	0.0022	0	0.66	0.0011	0.0065	0.0065	0.0054	0.0011
to	0.00083	0	0.0017	0.28	0.00083	0	0.0025	0.087
eat	0	0	0.0027	0	0.021	0.0027	0.056	0
chinese	0.0063	0	0	0	0	0.52	0.0063	0
food	0.014	0	0.014	0	0.00092	0.0037	0	0
lunch	0.0059	0	0	0	0	0.0029	0	0
spend	0.0036	0	0.0036	0	0	0	0	0

Figure 3.2 Bigram probabilities for eight words in the Berkeley Restaurant Project corpus of 9332 sentences. Zero probabilities are in gray.

	i	want	to	eat	chinese	food	lunch	spend
i	0.0015	0.21	0.00025	0.0025	0.00025	0.00025	0.00025	0.00075
want	0.0013	0.00042	0.26	0.00084	0.0029	0.0029	0.0025	0.00084
to	0.00078	0.00026	0.0013	0.18	0.00078	0.00026	0.0018	0.055
eat	0.00046	0.00046	0.0014	0.00046	0.0078	0.0014	0.02	0.00046
chinese	0.0012	0.00062	0.00062	0.00062	0.00062	0.052	0.0012	0.00062
food	0.0063	0.00039	0.0063	0.00039	0.00079	0.002	0.00039	0.00039
lunch	0.0017	0.00056	0.00056	0.00056	0.00056	0.0011	0.00056	0.00056
spend	0.0012	0.00058	0.0012	0.00058	0.00058	0.00058	0.00058	0.00058

Figure 3.7 Add-one smoothed bigram probabilities for eight of the words (out of $V = 1446$) in the BeRP corpus of 9332 sentences. Previously-zero probabilities are in gray.

Ans:

```
Unsmoothed:
P(<s> i want chinese food </s>)
= P(i<s>)P(want|i)P(chinese|want)P(food|chinese)P(<\s>|food)
=0.19 x 0.33 x 0.0065 x 0.52 x 0.40
= 0.0000847704

Smoothed:
P(<s> i want chinese food)
= P(i<s>)P(want|i)P(chinese|want)P(food|chinese)P(<\\s>|food)
=0.19 x 0.21 x 0.0029 x 0.052 x 0.40
=0.00000240676
```

2. Which of the two probabilities you computed in the previous exercise is higher, unsmoothed or smoothed? Explain why.

Ans: Add-one smoothing makes a big change to the counts given the redistribution of the probabilities. For example the probability of *i want* decreasing after smoothing, which in turn has an impact on the product of the respective bigram probabilities.

3. We are given the following corpus, modified from the example Dr Seuss corpus:

```
<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.

Ans:

```
P(Sam|am) = 0.214
```

4. Suppose we train a trigram language model with add-one smoothing on a given corpus. The corpus contains V word types. Express a formula for estimating $P(w3|w1, w2)$, where w3 is a word which follows the bigram (w1,w2), in terms of various n-gram counts and V. Use the notation c(w1,w2,w3) to denote the number of times that trigram (w1,w2,w3) occurs in the corpus, and so on for bigrams and unigrams.

Ans:

$P(w_3|w_1, w_2) = \frac{c(w_1,w_2,w_3)}{c(w_1,w_2)}$

$P^*(w_3|w_1, w_2) = \frac{c(w_1,w_2,w_3)+1}{c(w_1,w_2)+V}$