Slp3-chapter3-Exercises-20200428

3.1 Write out the equation for trigram probability estimation (modifying Eq. 3.11). Now write out all the non-zero trigram probabilities for the I am Sam corpus on page 33.

【1】

LaTeX equation :

\[
P(w_n|w_n-1)=\frac {C(w_{n-1}w) }{C(w_{n-1})}
\]

$$P(w_n|w_n-1) = \frac{C(w_{n-1}w)}{C(w_{n-1})}$$

【2】

Because <s> I am Sam </s>, so :

P(<s> I am Sam </s>)

=P(am | <s> I) P(Sam | I am) P(</s> | am Sam)

=0.5x0.5x1

=0.25

3.2 Calculate the probability of the sentence i want chinese food. Give two probabilities, one using Fig. 3.2 and the ‘useful probabilities’ just below it on page 35, and another using the add-1 smoothed table in Fig. 3.6. Assume the additional add-1 smoothed probabilities P(i|<s>) = 0.19 and P(</s>|food) = 0.40.

【1】2-grams

P(<s> i want chinese food </s>)

=P(i | <s>)P(want | i)P(chinses | want)P(food | chinese)P(</s> | food)

=0.25x0.33x0.0065x0.52x0.68

=0.000189618

【2】2-grams（add-1 smoothing）

P*(<s> i want chinese food </s>)

=P*(i | <s>) P*(want | i) P*(chinses | want) P*(food | chinese) P*(</s> | food)

=0.19x0.21x0.0029x0.052x0.40

=0.00000240676

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

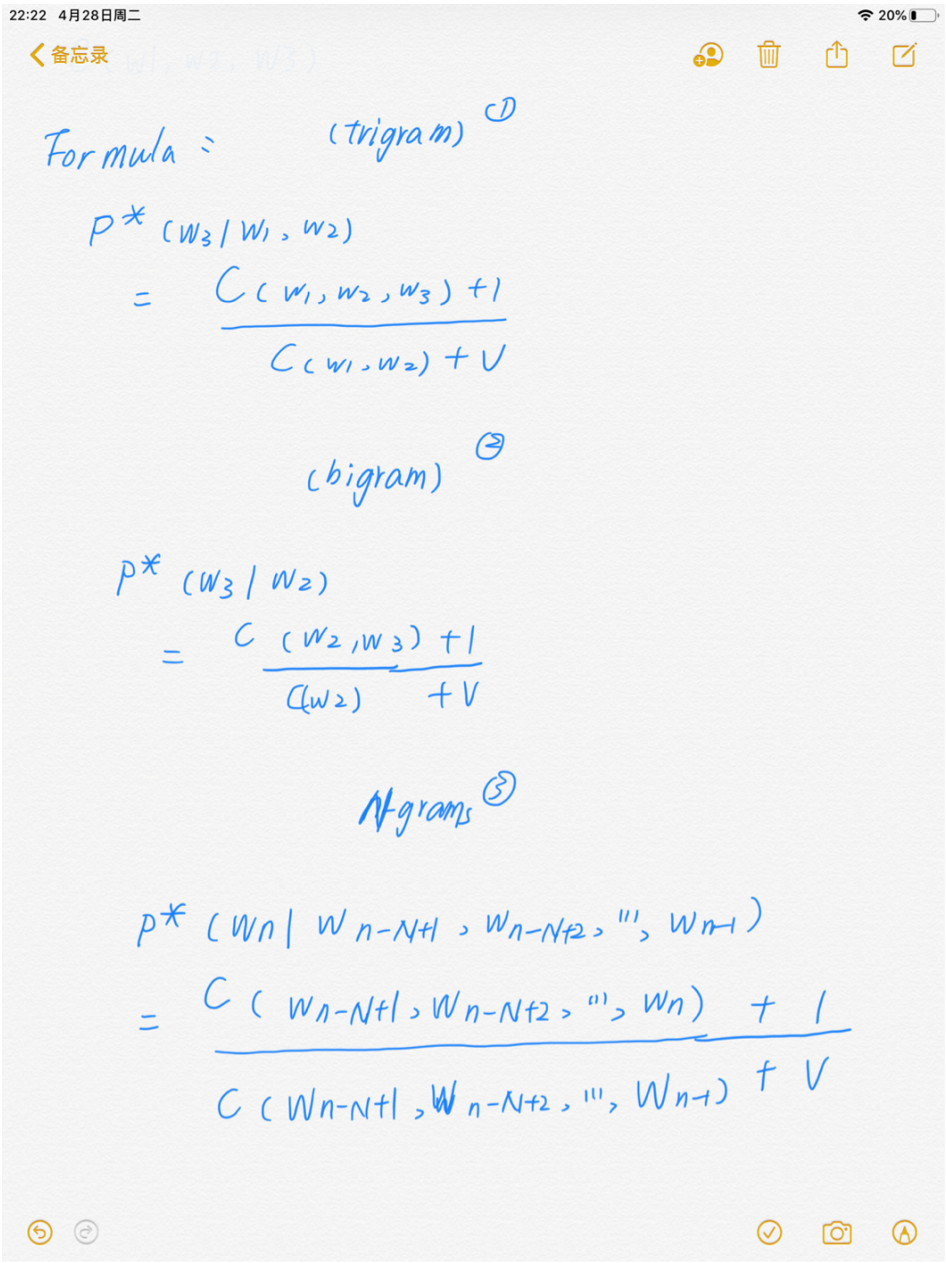
Answer: Unsmoothed. The sharp change in counts and probabilities occurs because too much probability mass is moved to all the zeros.

3.4 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.

$$
\begin{aligned}
P^*_{Laplace}(Sam \mid am) &= \frac{C(am\ Sam)+1}{C(am)+V} \\
&= \frac{2+1}{3+25} \\
&= \frac{3}{28} \\
&= 0.1071428
\end{aligned}
$$

3.5 Suppose we didn't use the end-symbol </s>. Train an unsmoothed bigram grammar on the following training corpus without using the end-symbol </s>:
<s> a b
<s> b b
<s> b a
<s> a a
Demonstrate that your bigram model does not assign a single probability distribution across all sentence lengths by showing that the sum of the probability
of the four possible 2 word sentences over the alphabet {a,b} is 1.0, and the
sum of the probability of all possible 3 word sentences over the alphabet {a,b}
is also 1.0.

3.6 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.



3.7 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>
If we use linear interpolation smoothing between a maximum-likelihood bigram model and a maximum-likelihood unigram model with $\lambda 1 = 0.5$ and $\lambda 2 = 0.5$, what is P(Sam|am)? Include <s> and </s> in your counts just like any other token.

$$\begin{aligned}\hat{P}(sam|am) &= \lambda_1 P(sam|am) + \lambda_2 P(sam) \\ &= 0.5 \times \frac{2}{3} + 0.5 \times 1 \\ &= \frac{5}{6}\end{aligned}$$

3.8 Write a program to compute unsmoothed unigrams and bigrams.

3.9 Run your n-gram program on two different small corpora of your choice (you might use email text or newsgroups). Now compare the statistics of the two corpora. What are the differences in the most common unigrams between the two? How about interesting differences in bigrams?

3.10 Add an option to your program to generate random sentences.

3.11 Add an option to your program to compute the perplexity of a test set.

3.12 Given a training set of 100 numbers consists of 91 zeros and 1 each of the other digits 1-9. Now we see the following test set: 0 0 0 0 0 3 0 0 0 0. What is the unigram perplexity?