Homework 1

PART I:

Exercise 3.4 in Chapter 3:

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.

Ans: V = {<s>, I, am, Sam, </s>, do, not, like, green, eggs, and}

|V| = 11

P(Sam | am) =

PART II:

1. There are 41737 word types in the training corpus.
2. There are 2468210 word tokens in the training corpus.
3. 1.6612% of word tokens in the test corpus did not occur in training.

3.6058% of word types in the test corpus did not occur in training.

1. 28.7665% of bigrams in the test corpus did not occur in training.
2. Question 5 and 6

['I', 'look', 'forward', 'to', 'hearing', 'your', 'reply', '.']

Unigram:

p(I) = -8.340105,

p(look) = -12.040939,

p(forward) = -12.397288,

p(to) = -5.509641,

p(hearing) = -13.502940,

p(your) = -11.046083,

p(reply) = -17.474926,

p(.) = -4.751889,

The log probability of "I look forward to hearing your reply ." under Unigram is -85.063811

The perplexity is 1587.979235372888

Bigram:

p(I) = -8.340105,

p(forward|look) = -4.046963,

p(to|forward) = -2.148721,

p(hearing|to) = -13.080762,

p(your|hearing) doesn't exist in training corpus, no log probability.

p(reply|your) doesn't exist in training corpus, no log probability.

p(.|reply) doesn't exist in training corpus, no log probability.

Because some bigrams don't exist in training corpus, no log probability, total probability is zero.

Bigram with add-1 smoothing:

p(I) = -8.369362,

p(forward|look) = -10.463203,

p(to|forward) = -8.945034,

p(hearing|to) = -13.814196,

p(your|hearing) = -16.579286,

p(reply|your) = -16.592720,

p(.|reply) = -16.576470,

The log probability of "I look forward to hearing your reply ."

under Bigram with add-a smoothing is -91.340271

The perplexity is 2735.3958083856537

Bigram with discounting and Katz backoff:

p(I) = -8.340105,

p(forward|look) = -4.068337,

p(to|forward) = -2.156025,

p(hearing|to) = -13.206293,

p(your|hearing) = -13.489689,

p(reply|your) = -19.255481,

p(.|reply) = -6.867366,

The log probability of "I look forward to hearing your reply ."

under Bigram with with discounting and Katz backoff is -67.383296

The perplexity is 343.2014797068245

Question 7:

--------------------------------------------------

Unigram:

Because there are unknown token, no log probability for entire test corpus under Unigram.

Bigram:

Because some bigrams don't exist in training corpus, no log probability, total probability is zero.

Bigram with add-1 smoothing:

The log probability of entire test corpus

under Bigram with add-a smoothing is -30756.930083

The perplexity is 34164165.2082008

Bigram with discounting and Katz backoff:

The log probability of entire test corpus

under Bigram with with discounting and Katz backoff is -23092.198550

The perplexity is 453084.6581278604