Homework1

Prepared By:
Humaira Qadeer

PART I:

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
Count(am Sam) = 2
Count(am) = 3
V = 11

P(Sam | am) $\frac{= Count(am\ Sam) + 1}{Count(am) + V} = \frac{2+1}{3+11} = \frac{3}{14}$

PART II:

1. The total unique word types in training corpus including the end of sentence padding symbol </s> and the unknown token <unk>, and disregarding the start of sentence padding symbol <s> is equal to 39502.

2. Disregarding the start of sentence padding symbol <s>, the total number of tokens in the training corpus is equal to 221290.

3. Before mapping the unknown words to <unk> in training and test data,

    a the percentage of word tokens in the test corpus that did not occur in the training corpus is 1.873%

    b the percentage of word types in the test corpus that did not occur in the training corpus is 3.929%

4. Before mapping the unknown words to <unk> in training and test data,

    a the percentage of bigram tokens in the test corpus that did not occur in the training corpus is 25.4023%

    b the percentage of bigram tokens in the test corpus that did not occur in the training corpus is 28.194%

5.

Unigram Model

- Log_2(P(I)) = -8.400191146154114
- Log_2(P(look) = -11.982303986215943
- Log_2(P(forward) = -12.375607193140258
- Log_2(P(to) = -5.540690135458277
- Log_2(P(hearing) = -13.505537497185392
- Log_2(P(your) = -10.953683308276174
- Log_2(P(reply) = -17.623534706583843
- Log_2(P(.) = -4.811868102497583
- Log_2(P(</s>) = -4.62532894422938
- $Log_2P(S)$= -89.81874501974097

Bigram Model

- Log_2(P(<s>| i) = -5.631088893700847
- Log_2(P(i |look) = -8.875420257009424
- Log_2(P(look| forward) = -4.345774836841731
- Log_2(P(forward| to) = -2.2644156362321546
- Log_2(P(to| hearing) = -13.2203480948755
- Log_2(P(hearing| your) = undefined
- Log_2 (P(your| reply))= undefined
- Log_2 (P(. |</s>))= undefined
- $Log_2P(S)$= Undefined
  The log_2 probability for the sentence using the bigram model was undefined because the probability of P(hearing| your), (P(your| reply)), (P(. |</s>) equated to 0 which led to the log probability being undefined.

Bigram Model Smoothed

- Log_2 (P(<s>| i)) = -6.155272148946965
- Log_2 (P(i| look)) = -11.584868572205163
- Log_2 (P(look|forward)) = -10.482195722024208
- Log_2 (P(forward| to)) = -8.825392481587325
- Log_2 (P(to |hearing)) = -13.827425410027113
- Log_2 (P(hearing| your)) = -15.276596985010402
- Log_2 (P(your| reply)) = -15.309973651164077
- Log_2 (P(reply| .)) = -15.270039765748985
- Log_2 (P(. | </s>)) = -0.6693438414242925
- $Log_2P(S)$= -97.40110857813853

6. Compute the perplexity of the sentence above under each of the models.

- Perplexity under Unigram Model : 1009.8046830192555
- Perplexity under Bigram Model: undefined
- Perplexity under Bigram Smoothed Model: 1810.7170379443896

The perplexity of the bigram model was undefined due to the probability of the sentence being undefined however the perplexity under the unigram model was less than the perplexity of the Bigram model with smoothing.

The unigram model assigned a higher probability to the occurrence of words which therefore resulted in the perplexity under the model to be lower where in contrast the smoothed bigram model assigned an increased probability to unseen evens. In relation to the complexity of the models, the unigram model had high bias and low variance whereas the bigram smoothed model had a higher variance due to the probability mass of the zero probabilities allotting more probability space.

Compute the perplexity of the entire test corpus under each of the models. Discuss the differences in the results you obtained.

- Perplexity under Unigram Model : 1079.2881373075008
- Perplexity under Bigram Model: undefined
- Perplexity under Bigram Smoothed Model: 2640.1795213444884

The perplexity of the bigram model was undefined due to the probability of the sentence being undefined however the perplexity under the unigram model was less than the perplexity of the Bigram model with smoothing. The perplexity under the bigram smoothed model was much higher because of the weight it added to unseen events, redistributing the probabilities, and causing them to flatten. The unigram model is considered to be a better model based on its low perplexity in comparison to the other models.