Kirsnaut

H.W. #1

PART II.

1. How many word types are there in the training corpus (including padding symbols and unknown token)

Ans: There are **41739** unique word in the training corpus.

1. How many word tokens are there in the training corpus?

Ans: There are **2568210** word-tokens in the training corpus.

1. What percentage of word tokens and word types in the test corpus that did not occur in training corpus (before you mapped the unknown words to <unk> in training and test data)?

Ans: The percentage of word tokens and word types in the test corpora that did not occur in the training corpus is: **2.960%**

1. What percentage of bigrams (bigram types and bigram tokens) in the test corpus that did not occur in training corpus (treat <unk> as a token that has been observed)?

Ans: The percentage of bigrams types and tokens in the test corpus that did not occur in the training corpus is: **42.935%**

1. Compute the log probabilities of the given sentence under the three models (ignore capitalization and pad each sentence as described above). Please list all of the parameters required to compute the probabilities and show the complete calculation. Which of the parameters have zero values under each model? (Use log base. Map words not observed in the training corpus to the <unk> token.

Ans: Log Probability:" <s> i look forward to hearing your reply . </s>"

Unigram: **-94.939**

Bigram: **(Undefined)**

Bigram Add-One: **-97.140**

**“Undefined” happens when the log probability result is ‘undefined’. This can only mean that the bigram probability is 0, which only happens when a bigram token does not exist. In the program, log prob of ('hearing', 'your') is zero.**

1. Compute the perplexities of the sentence above under each of the models.

Ans: Perplexity of: " <s> i look forward to hearing your reply . </s>"

Unigram: **721.011**

Bigram: **(Undefined)**

Bigram Add-One: **839.831**

**By definition, “Lower perplexity corresponds to a better fit of the language model”. Unigram has the best fit of the three models. The bigram model is undefined, so perplexity is undefined since we cannot compute (cannot divide by zero)**.

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

Ans: Perplexity of Entire Test Corpus

Unigram: **1000.200**

Bigram: **(Undefined)**

Bigram Add-One: **1365.080**

**The M=2969 therefore I expect the perplexity to be high. The Bigram model remains undefined because it has a tuple with log probability of zero. Although the Add-one smoothing solved the zero probability tuples, it is still worse than the Unigram Model. Overall, the Unigram Model has the lowest perplexity therefore it is the best model for our test corpus.**