WriteUp Questions Answers : NLP Assignment - 1 Dxp170020

Writeup Question 1.1:
Data type for vocabulary : Set. Vocabulary has been initialized with start_word token (start word token = ‘~’ and stop word token is ‘~~’).Vocabulary has the set of all characters used by this model.
Data type for ngram_count : dictionary with key value pairs where key is the ngram tuple generated and count is the number of times that ngram has generated
Data type for context_count : dictionary with key value pairs where key is the contexts tuple and count is the number of times the context has been observed
There is another variable called word_list() used for key value pairs where key is the word count and value is the number of times the word has been observed in the training data

Writeup Question 1.2:
We use 1/|V| in order to avoid the scenario for novel context appearances. In the case of new unseen context, our count of context will be 0 during and eventually the final probability will be giving out error for math.domain .1/|V| will depict the probability of that new word since |V| is the number of unique words in the data. Since the context is unknown, we can generalise the probability of new words to be 1/|V| .

Writeup Question 1.3:
Dauphin is an unseen word, hence we don't have any special condition to handle the unseen words during our testing. This will create final probability to be 0 which will result in the following python error : “ValueError: math domain error” since log(0) is not defined.
The probability of the sentence 1 is : -13.86

Writeup Question 2.1:
Now we have tackled the case of unseen words and unseen context as well. So, everytime we use probability of unknown word for any word not in training data. This helps to get rid of the 0 coming in numerator or denominator of the ngram-word probability formula.
Log probability of the sentence 2 is -30.970

Writeup Question 2.2:
For n = 3,
The delta values for estimating the log probability of sentence 1 is 1, The log probability achieved from this is as follows : -78.16
The delta value for estimating the log probability of sentence 1 is .6. The log probability is : -76.05
The delta value for estimating the log probability of sentence 2 is 1. The log probability achieved from this is as follows : -34.41
The delta value for estimating the log probability of sentence 2 is .6. The log probability achieved from this is as follows : -33.01

Adding delta, adds smoothing and moves some of the probability mass from the seen to the unseen events. So for the unseen data, we help to avoid zero probability for that word.
So, delta basically discounted some non-zero counts in order to get the probability mass that will be assigned to the zero counts
Add-1 smoothing has made a change to the probability space. It is not good for n-gram because it can cause a sharp change in counts and probabilities due to too much probability mass moving to zero-occuring words.

Writeup Question 2.3 : Bonus

WriteupQuestion 2.4 :
The log probability of sentence 1 using NgramInterpolator is -69.98
The log probability of sentence 2 using NgramInterpolator is -38.21
The ngraminterpolatar takes into account all the n-1 grams and hence gives a more smoothes probability. Here, since we are given equal weight to all the unigrams, bigrams, and trigrams, the probability is almost the same. We can adjust this probability using different lambadas. For.example, we can make the λs for those trigrams higher and thus give that trigram more weight in the interpolation. Like smoothing, interpolation helps to avoid the problem of zeroes if we havent observed them in our training data.

Writeup Question 3.1 :
The perplexity with smoothing is .991 (delta = .5) tested on sonnets.txt file
The perplexity without smoothing will also include the stop tokens. We get the following issues with perplexity :

1. Numeric underflow is problem
2. Perplexity is undefined if model assigns zero probability to test set.
3. For unsmoothed models, we will get zero probabilities which we lead to undefined log probabilities and hence give an error related to math.domain.

Writeup Question 3.2:
perplexity for shakespear train data 0.9999999455852715
perplexity for warpeace train data 0.9999999492420862
The perplexity will be high when we have test data unsimilar to training data.
HOwever, the perplexity of warpeace is less than perplexity of shakespear because the warpeace has more matching vocabulary, i.e. more similar to the sonnets.txt

Writeup Question 3.3:
We can use Bayesian decision theory. If we want to classify a text D in an author category, we can pick category which has the largest posterior probability given text. We can learn a separate language model for each author. We can do this by training on data set from the different authors. In order to categorize a new text, we can feed new text to each language model. Once

we have fed in, we can evaluate the likelihood of new text under the model and pick the winning author.
Writeup Question 4.1:
' beards do d are but bound commit deeds all a',
' inclinable blood d a botcher country any express em a',
' affections came ever because and better a another boasting but',
' came any and breathe as a do capitol cominius and',
' measles doing altitude back country corn ever better disdain considering'
These sentences are random and have less meaning in English. They do not make sense.

Writeup Question 4.2 :
Yes. the sentences are more meaningful since it is returning the same word for the same context. Hence it is likeliest word for a given context. The newly generated sentences are more meaningful as compared to the randomly generated sentences.
‘The is his they say it and so are here’

Writeup Question 5:
Writeup Question 5.1 : How long did this homework take you to complete (not counting extra credit)?
It took me 5 total days to understand and complete the homework. The language was a bit confusing and it took me 3-4 reads every time to debug a statement.

Writeup Question 5.2: Did you discuss this homework with anyone?
Yes , Vinit