INLP Assignment - 1 Report

Name: Ishan Kavathekar
Roll no: 2022121003

Generation:

We change the value of n for a n-gram model and observe how it affects the generation capability of the model. These results are for the unsmoothed models. Hence, there will be a significant number of instances where the context will not be present in the n-gram model. In this case, we output the <EOS> tag </s> and stop. The model generates a maximum of 20 tokens. This number can be changed by changing the max_tokens in the generate_sequence function of the class N_Gram_model. It should be noted that the punctuation marks were discarded after tokenization for better generation. Hence, the generated sequence will also not contain any punctuation marks.
The following are the results for n=1 to n=5:

Prompt: ‘I am not where I want to be……’

Results:
Unigram model: 'I am not where I want to be </s>'

Bigram model: 'I am not where I want to be in the whole of the whole of the whole of the whole of the whole of the whole of the'

Trigram model: 'I am not where I want to be in London and when at last that he had been brought up for the sake of discovering them To be'

Four gram model: 'I am not where I want to be told why my views were directed to Longbourn instead of to yours A house in town I conclude They are'

Five gram model: "I am not where I want to be told whether I ought or ought not to make our acquaintance in general understand Wickham's character They are gone off"

It should be noted that as we increase the value of N from 1 to 5, the quality of generated text also increases. The generated text tends to capture more intricate patterns and dependencies within the given training data. This increase in N allows the model to consider a longer history of words, resulting in more contextually coherent and semantically meaningful output.

OOD (out of data) scenario is when the model encounters a context which is drastically different from the training data. The probabilities assigned to such contexts are zero in the unsmoothed n-gram models. Hence, these models are not expected to generate anything. Such cases have been handled by outputting the <EOS> tag.
Following are the results for a OOD scenario from various n-gram models:

Prompt: 'Earth is the third planet in the solar system….'

Results:
Unigram model: 'Earth is the third planet in the solar system </s>'

Bigram model: 'Earth is the third planet in the solar system </s>'

Trigram model: 'Earth is the third planet in the solar system </s>'

Four gram model: 'Earth is the third planet in the solar system </s>'

Five gram model: 'Earth is the third planet in the solar system </s>'

We can observe that increasing the value of N for N-gram models does not improve the quality of generated text in OOD cases.

After the smoothing techniques used in the assignment, we use the stored probabilities to generate sequences. If the context is not present then we output the <UNK> token.
Following are results of sequence generation for smoothed probabilities:

Prompt: 'What are you doing here'

Results: 'What are you doing here Stephen It flows purling widely flowing floating foampool flower unfurling They talk excitedly Little piece of original verse written by'

Prompt: 'King Macbeth was'
Result: 'King Macbeth was <UNK> <UNK> <UNK> <UNK> <UNK> <UNK> <UNK> <UNK> <UNK> <UNK> <UNK> <UNK> <UNK> <UNK> <UNK> <UNK> <UNK> <UNK> <UNK> <UNK>'

We can see that the OOD scenario only generated <UNK> tokens.

For the next word prediction task, the top k choices for next word are printed with their corresponding probabilities. If the choices are lower than k, then all the available choices are printed.
If the context is not present in the training corpus, then the <UNK> token is printed with the corresponding probability.

Perplexity scores:

LM type	Perplexity score
LM 1 Train set	6.419386145345034
LM1 Test set	6.401843201204255
LM 2 Train set	794798764.6620748
LM2 Test set	928416383.7422663
LM 3 Train set	inf
LM 3 Test set	inf
LM 4 Train set	inf
LM 4 Test set	inf

We observe that the train and test sets of LM3 and LM4 have perplexity scores inf. This can be attributed to the Ulysses corpus. This corpus has no punctuation marks for chapter 18. Thus the model treats the complete chapter as one sentence which leads to inf perplexity.

For the LM1 train and test set, we observe that there isn’t much difference between the perplexity scores. This can be attributed to the fact that the probability of unseen tokens is quite high because the number of trigrams occuring once is significantly higher than other trigrams.

For the LM2 train and test set we see that the perplexity scores for the test set are higher than the train set perplexity scores.