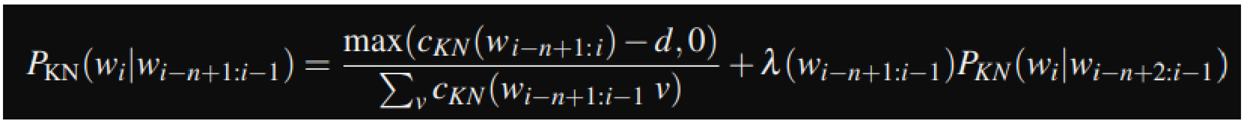
ASSIGNMENT – 1
SMOOTHING AND LSTM

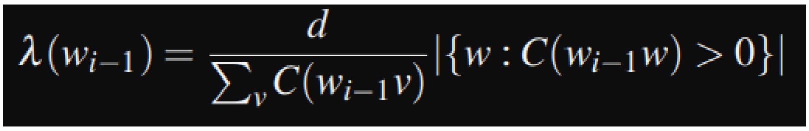
Name : Darur Lakshmipathi Balaji
Roll No : 2021114007

SMOOTHING

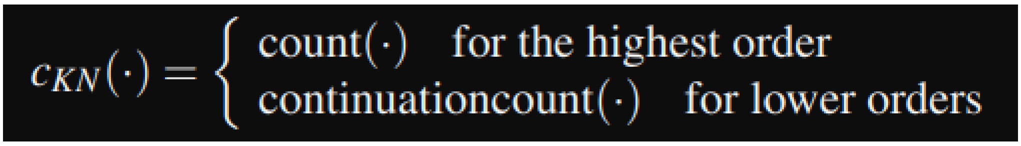
- KneserNey Smoothing
  The concept of P(continuation) is introduced here, the reference images used for the implementation of the algorithm are -



$$P_{\mathrm{KN}}(w_i|w_{i-n+1:i-1}) = \frac{\max(c_{KN}(w_{i-n+1:i})-d,0)}{\sum_v c_{KN}(w_{i-n+1:i-1}\ v)} + \lambda(w_{i-n+1:i-1})P_{KN}(w_i|w_{i-n+2:i-1})$$



$$\lambda(w_{i-1}) = \frac{d}{\sum_v C(w_{i-1}v)}|\{w : C(w_{i-1}w) > 0\}|$$



$$c_{KN}(\cdot) = \begin{cases} \text{count}(\cdot) & \text{for the highest order} \\ \text{continuationcount}(\cdot) & \text{for lower orders} \end{cases}$$

- WittenBell Smoothing
  the reference used for the implementation of the algorithm is here .

The average perplexities of the models on the different sets are as follows -

Model	Trainset	TestSet
LM1 - Pride and Prejudice - k	60.869935999999946	25.781
LM2 – Pride and Prejudice - w	65.061	14.164
LM3 - Ulysses - k	21595.400653925615	8945.631310559014
LM4 – Ulysses - w	294.009	409.956

It can be observed that relatively the perplexity scores for Ulysses is higher which shows the fact that the size of corpus.

LSTM

The implementation of the code is learnt from pytorch library LSTMs and DataLoaders. The resources used for the implementation and learning are – link