Report

Name - Srija Mukhopadhyay
Roll Number - 2021114002

Models Implemented

Kneser Ney

This model uses the interpolated Kneser Ney smoothing technique to calculate perplexities and probabilities of words.

The formula used for calculating probability was

$$p_{KN}(w_i|w_{i-n+1}^{i-1}) = \frac{\max\{c(w_{i-n+1}^{i}) - D, 0\}}{\sum_{w_i} c(w_{i-n+1}^{i})} + \frac{D}{\sum_{w_i} c(w_{i-n+1}^{i})} N_{1+}(w_{i-n+1}^{i-1}\bullet)\, p_{KN}(w_i|w_{i-n+2}^{i-1})$$

where each term has it's own expected meaning.

It was observed that the model gave a perplexity of 430 and 530 respectively for the Pride and Prejudice and Ulysses test corpus which was decent but can probably be improved.

Witten Bell

This model uses the Witten Bell smoothing technique to calculate the perplexities and probabilities of words.

The formula used for calculating the probability was

$$p_{WB}(w_i|w_{i-n+1}^{i-1}) = \lambda_{w_{i-n+1}^{i-1}}\, p_{ML}(w_i|w_{i-n+1}^{i-1}) + (1 - \lambda_{w_{i-n+1}^{i-1}})\, p_{WB}(w_i|w_{i-n+2}^{i-1})$$

where the 1 - lambda term was further expanded as follows

$$1 - \lambda_{w_{i-n+1}^{i-1}} = \frac{N_{1+}(w_{i-n+1}^{i-1}\bullet)}{N_{1+}(w_{i-n+1}^{i-1}\bullet) + \sum_{w_i} c(w_{i-n+1}^{i})}$$

with the rest of the terms having it's own standard meaning.

It was observed that the model gave a perplexity of 82 and 105 respectively for the Pride and Prejudice and Ulysses test corpus which was decent.