# SNLP Assignment 3 Report

## Basics of Language Modeling:

### 1. N-gram probabilities:

The distribution w.r.t "the" word as prior has higher expected value (15.3528) than the distribution with "in" (13.241). This is solely due to the reason that the frequency of "the" word is very high throughout the corpus and the conditional probability of Top 20 words with "the" as the preceding word is more compared to the one in the "in" distribution.

### 2. Perplexity:

The perplexity for bigram model will be less, as we are considering last 2 words occurrence, and hence we get better probability of the next words.
We observed perplexity of bigram model to be more than the unigram model. And as we know perplexity is the measure to see how good the language model is. Hence bigram model is better.

Smoothing is important as for test data there can be many unseen words, and we don't have any conditional probability defined for them so we need to use smoothing otherwise it will be difficult to access for new words. We assign it a smoothed probability. Here we use Lidstone smoothing.