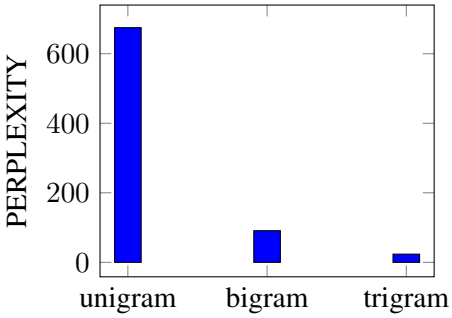
Report- Assignment 1 - NLU

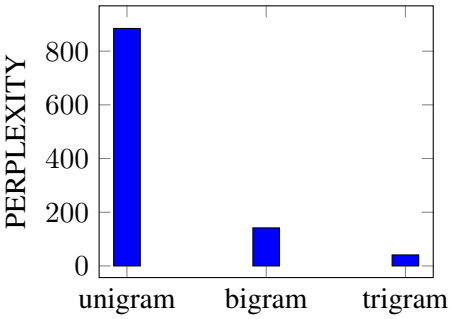
Anonymous ACL submission

QN1)
for the given set of training and test data the perplexity v/s ngrams plot is shown below.

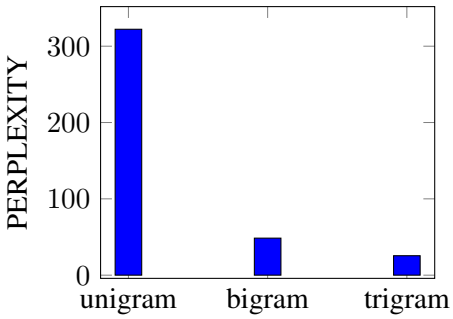
$S1$: Train: D1-Train, Test: D1-Test



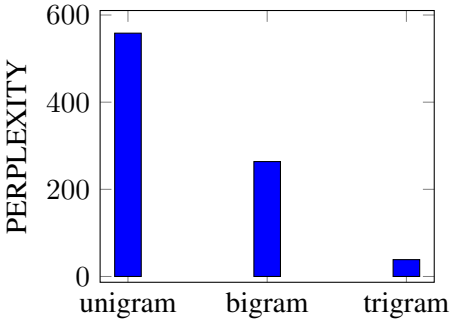
S2: Train: D2-Train, Test: D2-Test



S3: Train: D1-Train + D2-Train, Test: D1-Test



S4: Train: D1-Train + D2-Train, Test: D2-Test



we could see as the ngrams is inversely proportional to the perplexity

Q2) made use of smoothing techniques and using the below model is used for generating tokens.

$$P(w_1\ w_2) = \frac{c(w_1\ w_2)}{N}$$

$$P(w_1 \dots w_2) = \frac{c(w_1 \dots w_2)}{N}$$

- For any $n-$gram a:

$$P'(a) = \frac{C(a)+1}{N+V^n}$$

- For a test corpus $W = w1 \dots w_N$, the perplexity $PP(W)$ is

$$PP(W) = P(w_1 \dots w_n)^{\frac{-1}{N}} = \sqrt[N]{\frac{1}{P(w_1 \dots w_N)}}$$

- $$PP(W) = \sqrt[N]{\frac{1}{\Pi_{k=1}^{N} P(w_k|w_{k-1})}}$$

generated token by the model is shown below
"say shall unto lord come go thou one god make"