E1 246 Assignment-1

Rahul Chittimalla
Sr no: 06-02-01-10-51-17-1-14585
rahulc@iisc.ac.in

1 Task 1

For this task, each dataset is divided as train, dev and test sets in the ratios 90%, 5% & 5% respectively. NGram model is used on the dataset for language modelling with $n = 4$. For the words that appear in a test set in an unseen context, *backoff* smoothing technique is used to keep the language model from assigning zero probability to these events.

For the four settings given, namely $S1, S2, S3$ & $S4$, the metric used for comparing the language model is **perplexity**. First the model is trained on first 90% of the text from each category of the corpus and later tested on last 5% of each category. As n increases, the perplexity decreases. This can be shown empirically by running the model on a corpus. In this case, the model is run on *editorial* category to support the claim. The code is run on a machine with an *intel i7* processor and 8GB RAM. The corpus(brown and gutenberg) is taken from NLTK library, so all the scripts require NLTK support. The results are shown in the Table 1 and Figure 1.

n	*perplexity*
3	289.09281327
4	245.523032621
5	174.911659907
6	133.364452404

Table 1: Perplexity for various n

When the language model is evaluated for the given settings, the setting $S4$ tends to give better results in terms of perplexity as shown in Table 2 and Figure 2.

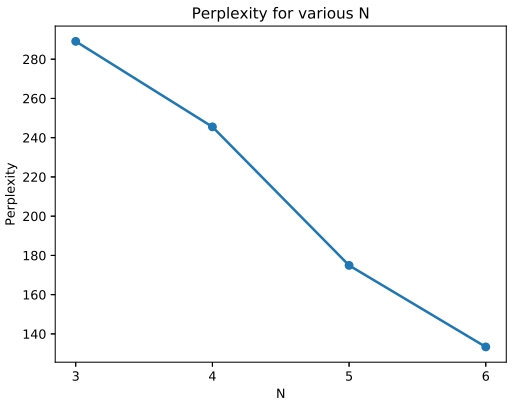

Figure 1: Perplexity

setting	*perplexity*
$S1$	220.578614463
$S2$	156.542050361
$S3$	284.39328995
$S4$	150.035393653

Table 2: Perplexity for various settings

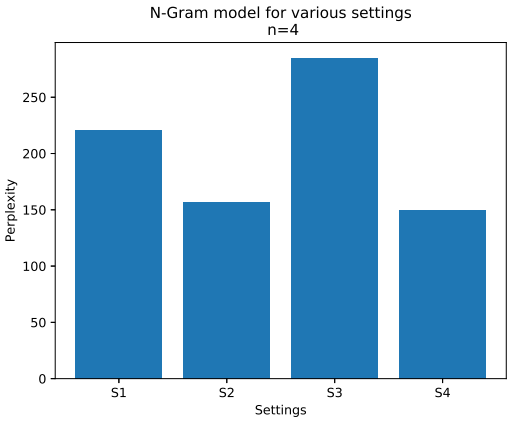

Figure 2: Perplexity for settings

We can observe that the $S4$ has the lowest perplexity among all the settings.

2 Task 2

With the best model in Task 1, the sentences are generated using that language model. Some of the sentences generated are:

- as the junior mates were hurrying to execute the warrants
- upon it in truth in judgment and in lovingkindness and
- of course who keep it alive and preserve it so
- the box would break open a hamper and produce filets
- miles a day under sub freezing temperature conditions attendants inactivation
- always wished to be a christian means to say yes

The generate_sentence.sh script is included on the github repository. generate_sentence.sh script requires two more files generate.py and my_prob.pkl.