Assignment 02

By Shivam Bansal, Roll No. 170101063

Link to Colab notebook

Preparing dataset

- NLTK library used for sentence segmentation and word tokenization
- Sentence segmentation is done using Punkt sentence segmentation
- Word tokenization is done as in the Penn Treebank

Model performance

Using only 31000 sentences for sake of RAM usage and time of computation.

Held out dev/validation sets (5 different sets)

Perplexity is in range of 240 to 400

Log likelihood is in range -59000 to -67000

Independent test set

Perplexity is 245

Log likelihood is -67000

Only Laplace smoothing (on test set)

Perplexity is 2000

Log likelihood is -93000

Overall using only Laplace smoothing results in a relatively worse model.