NLP Assignment #3

Contributions

Words are lowered to decrease the perplexity of the models
Punctuations are removed to decrease the perplexity of the models
Vocabulary has been restricted to top 5000 words and rest are considered as OOV words.

This smoothing technique applied is basically deleting unigrams with frequency 1 which basically considers words occurring with frequency 1 as unknown words.

Techniques implemented

- Perplexity reduction
- Good Turing Smoothing
- Backoff as in this EMNLP paper [Jeffrey Dean etal (Google Research), EMNLP 2007] http://www.aclweb.org/anthology/D07-1090.pdf

Optimizations implemented

- Extremely fast method for language modelling combining Good Turing and Stupid Backoff
- Time required for a sentence of n length is O(n) i.e. every query is processed in O(1)!
- Preprocessing required for the same is also O(n) per sentence, i.e. O(1) per query!
- Extremely memory efficient too. 1 word stored for predicting next word n-gram.
- Still theoretically implements the full version of the original model with backoff and smoothing.

How to run:-

>> python 201402004_assignment3.py