Readme

1. Count the occurrences of each ngram
2. Build LM using ngram counts from the Q1:
    a. Unigram: $P(w_n|w_1^{n-1}) \sim P(w_n) = \frac{\#wn}{\#token}$
    b. Bigram: $P(w_n|w_1^{n-1}) \sim P(w_n|w_{n-1}) = \frac{CWn-1Wn}{CWn-1}$
    c. Trigram: $P(w_n|w_1^{n-1}) \sim P(w_n|w_{n-2}w_{n-1}) = \frac{CWn-2Wn-1Wn}{CWn-2Wn-1}$

3.Write a script ppl.sh that calculates the perplexity of a test data set given an LM using interpolation for smoothing

$P(w_n|w_{n-2}w_{n-1}) = \lambda_3 P_3(w_n|w_{n-2}w_{n-1}) + \lambda_2 P_2(w_n|w_{n-1}) + \lambda_1 P_1(w_n)$

4

lambda_1	lambda_2	lambda_3	perplexity
0.05	0.15	0.8	378.9072777164
0.1	0.1	0.8	358.3443256944
0.2	0.3	0.5	227.1302127488
0.2	0.5	0.3	205.9901976918
0.2	0.7	0.1	206.3451185266
0.2	0.8	0	231.4900511022
1.0	0	0	871.8097461884