

3.1 Report the 4-gram and 5-gram character counts

Total 4-gram character counts are 11651664

Total 5-gram character counts are 11651645

3.2 For calculating I have replaced low frequency words with 'UNK' and applied smoothing as instructed. Perplexity for file '03302_02.txt' is 1.6028679391499376

3.3 Score of files with highest perplexity is as follows

100101_01.txt	1.9651509541498489
00503_01.txt	1.8550437262621398
02803_02.txt	1.8076324619559627