HomeWork 3 - Part 2- I chose option 2 for part 2 of Homework 3.

Best model from part1:

The process included to extract lines from the training set and split the sentence to form words on white spaces.A dictionary is formed with keys as items and values are the frequency in the collection.We from a word2vec as a dictionary with keys as strings and values as integer index for each string based on the order of exploration.We ignore the words that happened once in training. We generate all the trigrams bigrams from the corpus.Thereby calculating the trigram sentence probability using trigram,bigram and unigram log probabilities.

I tried to improve the preprocessing process on part 1 model where the corpus is split on period and tokenized sentences using regular expression. In a way to get away with the corpus header and consider only the body of the corpus for training the model. Sentence segmentation was improvised by adding the <s> and </s> tags to the model. The model was further trained by splitting the corpus on white spaces into words.I also tried to improvise the smoothing using add one smoothing which is used to calculate the probability distribution of n-grams in a corpus.

Generation of random sentences using the best model from part 1 using Shannon game generation.

Shannon method aims to improvise on assigning probabilities to sentences to generate random sentences that are similar to the sentences from which the model was derived.

According to the Shannon algorithm, we pick the random bigram (<s>,w) based on the probability distribution over the bigrams returned by the bigramProb.I construct the Trigram from the bigram thus formed by using the bigram formed as the beginning two words of the sentence Now we sample a new random trigram (w,x,y) according to its probability where the prefix w and x matches the suffix of the first chosen bigram.I have defined a common procedure to generate random sentences.This logic is used to find all the trigrams having the first two prefix as w and x.Based on the trigram probability (defined from part1) of the trigrams thus formed random.choices helps to pick a trigram.The y of the trigram thus selected is appended to the randomly generated sentences.This process is continued until we randomly choose a </s> end of sentence marker which terminates the random sentences.

Result:

The 50 sentences thus formed from the generator, having a perplexity in the range of 300 to 1500 As well as the perplexity of those 50 sentences as a corpus is around 1300.