Language Models and Smoothing

This question requires you to train some language models on a training corpus and to test them on two smaller corpora. Stater code is also provided with this assignment. Each sentence must be surrounded by a start of sentence and end of sentence marker (<s> ... </s>). Preprocess to add these sentence markers. These markers will allow your models to generate sentences that have realistic beginnings and endings.

Implement the following models:
1. **UnigramModel**: an unsmoothed unigram model
2. **SmoothedUnigramModel**: a unigram model smoothed using Laplace (add-one) smoothing,
3. **BigramModel**: an unsmoothed bigram model,
4. **SmoothedBigramModelLI**: a bigram model smoothed using Linear Interpolation

For each of the four language models, you need to implement the following methods:

generateSentence(self): returns a sentence sent that is generated by the language model. sent is a list of the form [<s> w1, ….., wn </s>]. You can assume that <s> starts each sentence (with probability 1). The following words (<s> w1, ….., wn </s>) are generated according to your language model's distribution. The number of words (n) is not fixed. Instead, you stop generating a sentence as soon as you generate the end of sentence symbol </s>.

getSentenceProbability(self, sen): returns the probability of the sentence sen (which is again a list of the form [<s> w1, …. , wn </s>]) according to the model.

Please use the provided generateSentencesToFile method and your unigram and bigram language models to generate 20 sentences (saved as unigram output.txt, smooth unigram output.txt, bigram output.txt, and smooth bigram kn output.txt).

Perplexity
You need to compute the perplexity (normalized inverse log probability) of the two test corpora according to all of your models. Evaluate the models on the test corpora. Do you see a difference between the two test domains?

Questions
1. When generating sentences with the unigram model, what controls the length of the generated sentences? How does this differ from the sentences produced by the bigram models?
2. Consider the probability of the generated sentences according to your models. Do your models assign drastically different probabilities to the different sets of sentences? Why do you think that is?
3. Generate additional sentences using your bigram and smoothed bigram models. In your opinion, which model produces better / more realistic sentences?
4. For each of the four models, which test corpus has a higher perplexity? Why? Make sure to include the perplexity values in the answer.