Programming assignment 2 questions

Please turn in a separate file with your answers to these questions in the Dropbox. Do not zip the file with your program file.

You will be comparing the output of your 3 models in these questions. You should be able to answer the questions by simply looking at a printout of your results. Your test sentences will have very different probabilities depending on how many words are in the sentence. For this reason, only compare model outputs for the same sentence; for example, never compare the output of the unigram model for Sentence 1 to the bigram output for Sentence 2.

Please review how to interpret log10 numbers before beginning. Or reconvert to probabilities before answering the questions (but keep them as log10 in the output of the program you turn in).

Please be concise. You should only need a sentence or two to answer each question.

1. Considering only the unigram output and bigram output for the test sentences that have results that are not ‘Undefined’ (i.e., sentences in which all the bigrams were seen in the training file), which is the better model? Why?
2. Considering the unigram and bigram output for ALL the sentences, which model is better? Why?
3. Considering the bigram output and the smoothed bigram output for test sentences that have results that are not ‘Undefined’, which model gets the higher percentage? Why?
4. Considering the bigram and smoothed bigram output for ALL of the sentences, which model is better? Why?