Homework 2

1. Given the following bigram counts and probabilities from the Berkeley Restaurant Project corpus, compute the probability of the given sentences.

P(i | <s>) = 0.25
P(</s> | food) = 0.68

	i	want	to	eat	chinese	food	lunch	spend
i	0.002	0.33	0	0.0036	0	0	0	0.00079
want	0.0022	0	0.66	0.0011	0.0065	0.0065	0.0054	0.0011
to	0.00083	0	0.0017	0.28	0.00083	0	0.0025	0.087
eat	0	0	0.0027	0	0.021	0.0027	0.056	0
chinese	0.0063	0	0	0	0	0.52	0.0063	0
food	0.014	0	0.014	0	0.00092	0.0037	0	0
lunch	0.0059	0	0	0	0	0.0029	0	0
spend	0.0036	0	0.0036	0	0	0	0	0

   a. I want chinese food
   b. I want to eat chinese food
   c. I want to eat food

2. Now, using add-1 smoothing, recalculate the probabilities.

P(i | <s>) = 0.19
P(</s> | food) = 0.40

	i	want	to	eat	chinese	food	lunch	spend
i	0.0015	0.21	0.00025	0.0025	0.00025	0.00025	0.00025	0.00075
want	0.0013	0.00042	0.26	0.00084	0.0029	0.0029	0.0025	0.00084
to	0.00078	0.00026	0.0013	0.18	0.00078	0.00026	0.0018	0.055
eat	0.00046	0.00046	0.0014	0.00046	0.0078	0.0014	0.02	0.00046
chinese	0.0012	0.00062	0.00062	0.00062	0.00062	0.052	0.0012	0.00062
food	0.0063	0.00039	0.0063	0.00039	0.00079	0.002	0.00039	0.00039
lunch	0.0017	0.00056	0.00056	0.00056	0.00056	0.0011	0.00056	0.00056
spend	0.0012	0.00058	0.0012	0.00058	0.00058	0.00058	0.00058	0.00058

3. Which of the two probabilities you computed in the previous exercise is higher, unsmoothed or smoothed? Explain why.

4. Write a program to compute unsmoothed unigrams and bigrams.
   a. Use any language you want, but do not use any libraries other than math/probability ones (Java.Math, numpy, etc)
   b. Run your n-gram program on two different small corpora of your choice. Try and make them from different genres (i.e. a news article and a song lyric)
   c. Now compare the statistics of the two corpora. What are the differences in the most common unigrams between the two? How about interesting differences in bigrams?
   d. Add an option to your program to generate random sentences.