Assignment 1

L245

Due Wednesday, January 25

1. Do question #1 from chapter 1 of the textbook (p. 29). For part (a), discuss not only the difficulties encountered, but also adaptations you would make to the syllabary to at least partially accommodate English. (Note: it is okay to still only approximate English and not cover every single word perfectly.)

2. (a) Give the base ten numbers for the following base two numbers (show your work):
   i. 111011011
   ii. 11010 (=00011010)

   (b) Give the base two (binary) numbers for these base ten numbers (show your work):
   i. 241
   ii. 67

3. Do question #3 from chapter 1 of the textbook (p. 30).

4. Attempt to "break" one of the TTS systems mentioned in the notes.

   (a) Come up with example sentences to try; describe what you expect to go wrong; and analyze what the TTS system does well and what its limitations are.

   (b) Take some back-transliterated text from question #1 and see how well the TTS system does with pseudo-English words. Are there words which are misspelled but the system still gets right? Why?

5. [**moved to HW2:**] Here are some bigram probabilities (Figure 6.7, Jurafsky & Martin (2000), 1st edition). For example, $P(want|i) = 0.22$, whereas $P(i|want) = .0014$. Ignoring start & end probabilities, calculate the probabilities for the sentences (a) & (b) using a bigram model (show your work). Then, answer (c).

   (a) *i want to eat chinese lunch*

   (b) *i want to eat food*

		2nd word						
		i	want	to	eat	chinese	food	lunch
1st word	i	.0018	.22	.0020	.0028	.00020	.00020	.00020
	want	.0014	.00035	.28	.00035	.0025	.0032	.0025
	to	.00082	.00021	.0023	.18	.00082	.00021	.0027
	eat	.00039	.00039	.0012	.00039	.0078	.0012	.021
	chinese	.0016	.00055	.00055	.00055	.00055	.066	.0011
	food	.0064	.00032	.0058	.00032	.00032	.00032	.00032
	lunch	.0024	.00048	.00048	.00048	.00048	.00096	.00048

   (c) The sentence *i want to eat* is more likely than *i want to eat lunch*, yet both are good sentences. If we wanted a better grasp on the likelihood of a sentence actually appearing in the English language, what other properties might we need to account for in our model?

6. [**moved to HW2:**] Do question #7 from chapter 1 of the textbook (p. 30) ... with a few alterations/clarifications:

   - You need to ask 5 (or more) friends (or however many you need to sufficiently answer part b). You are still working with at least 10 bigrams.
   - Be sure to present your data (in a readable, organized format): you will lose points for not showing your bigrams and your friends' responses.
   - You also have a new part (c): Based on your data, describe how this modeling is similar to or different from n-gram language modeling.