Fall Semester 2017
Iowa State University

LING 120 - Language and Computers - In class exercise

Date: 1st September 2017
Author: Dr. Sowmya Vajjala

Instructions: Work in groups of 2–3 people and address the questions given below. This counts as your attendance for today. Spend about 20-25 minutes working on this exercise, and we will use the rest of the class to discuss what you learnt.
The aim of this exercise is two fold:

- Breaking the ice among you (yes, I need to wait for 2 weeks!)
- To make you think a little bit on what you learnt so far and what you will encounter in the coming weeks

Note: You won't be graded for the correctness of your answers. Please feel free to note down your thoughts as they stand now, with your current knowledge of this subject. Don't make use of the textbook or any other resources except the thoughts of your team mates while doing this. Write the names of your team members on the sheet (along with your notes) and return it back to me before you leave. One problem sheet per team will do.

Task: We have briefly discussed about a few computer technologies where language plays a role. We did not discuss in detail about the operations, but got introduced to some terms. In the last class, I mentioned a word n-gram. This idea is useful in solving several problems in the interface of language and computers. Discuss among yourselves and make a note on the following:

- How are word n-grams useful in doing spelling and grammar error detection and correction?
- Are word n-grams useful in automatic speech recognition? If so, how? if not, why not?
- Are word n-grams useful in text to speech conversion? How? if not, why not?
- Can we use character n-grams anywhere? Do you think it makes sense to use their frequencies in isolated word spelling correction?

(If you did the readings, you will have some intuition about these!)

Background information:

- n-grams are sequences of n-words.
- In the sentence: ”I have a book”, ”I” is a word unigram/1-gram. ”I have” is a word bi-gram/2-gram ”I have a” is a word trigram/3-gram. ”I have a book” is a word 4-gram. and so on.
- ”I hav” is a character 5 gram with 5 characters - I, space, h, a, v
- If I have a large collection of text files (called **corpus**), I can compile large dictionaries of such word and character n-grams of any arbitrary n, and how frequently are they seen in this corpus.
- An example of such a large text corpus is: Wikipedia
- a piece of extra information: if a word W appears 1000 times in entire wikipedia, then its frequency is 1000. If there are a total of 10000 words in wikipedia (counting multiple appearances of the same word as different words), the ”probability” of seeing W in wikipedia is 1000/10000 i.e, 0.1 - if you don’t know probability, that is fine. You can still answer this exercise.