CSC 206: NLP
Spring 2015

Name (Print): _____

Problem Set 7

Due: Friday, May 29 2015, before class

Please create a folder for this problem set, and save all files you need to submit in this folder. Zip up the folder and submit the .zip file on Nexus.

Questions:

1. Hidden Markov Models

    1. Download the completed implementation of the ice-cream example from Nexus. The file is called `viterbi_ice_cream_alternative.py`.
    2. Run the example that is provided. This is the example we looked at in class. The sequence of observations is `3 1 3` and the viterbi algorithm tells us that the most likely sequence of states (temperatures) that would lead to this sequence of observations is `Hot Hot Hot`. Find your way around the code. Make sure you understand how to read the output and the specification of the HMM model.
    3. Now change the sample input sequence such that Jason eats only one ice cream on the first day. I.e., the input sequence now is `1 1 3`. What is the most likely sequence of states for this sequence of observations? Is that what you would have expected for this input sequence?
    4. What is the probability of the first day being hot (even though Jason only eats one ice cream on that day)? How does it compare to the probability of the first day being cold? Explain where these (maybe somewhat unexpected) probabilities come from?
    5. How about the second day (with input sequence `1 1 3`)? Is it more likely to be hot or cold?
    6. Now, look again at the most likely sequence of states leading to the observations `1 1 3`. How can you explain this result in the light of your answer to the previous question?

2. Noun phrase chunking

    Use the `nltk.app.chunkparser()` tool you used in class to develop a regular expression grammar (i.e. a set of regular expressions) that describes English noun phrases. Try to achieve at least 70% or greater in precision as well as recall, when run on the 100 sentence development set.

    Save your grammar in a file called `np-chunk-grammar.chunk`.

What to Submit:

- a pdf document containing the answers to Question 1
- your NP chunking grammar from Question 2