Homework 5: Strings & NLTK

L435/L555

Due Tuesday, October 25

1. Write a program that checks if a word that the user enters is a palindrome. Make sure that your program works for words with an odd number of characters as well as for words with an even number.

2. It turns out that in the task of authorship attribution, it really helps to use character n-grams to classify texts. Take the file `vm.pos` and print out all the unique *character* 4-grams (i.e, 4-gram types), along with their frequencies.

   - Ignore lines starting with *%%*. Ignore POS information.
   - When you hit a sentence boundary (indicated by an empty line), stop the 4-gram calculation and start over.
   - You should use the NLTK `FreqDist` utility to store the 4-grams. If I were you, I would first make sure I could print out all the character 4-grams in the order they appear in the text. Only after that would I worry about storing them.

3. NLTK Book, ch. 1, #22 (`http://www.nltk.org/book/ch01.html`):

   Find all the four-letter words in the Chat Corpus (`text5`). With the help of a frequency distribution (`FreqDist`), show these words in decreasing order of frequency.