Name: _____
CS159: Natural Language Processing
PReview #2
Due in class Feb 2, 2015

Review

Write a function that takes a text and generates all the bigrams in the text. It should then pick a random word in the text and iteratively pick a following word by randomly selecting a following word from the bigrams.
Generate 20-word "samples" for your choice of 2 texts from the book texts. On your worksheet, specify which texts you used, any additional cleanup you did, and the resulting automatically-generated texts.

■

Tag sets commonly have separate parts of speech for the comparative ("-er") and superlative ("-est") forms of adjectives. For example, the Penn Treebank Set identifies the following parts of speech:
JJ Adjective or ordinal number (third, regrettable, happy, exciting)
JJR Comparative adjective (fitter, happier)
JJS Superlative adjective (happiest, creepiest, proudest)
Do these categories match our definition of parts of speech as being categories of words that can show up in the same place? Give a context where a JJ would work, but a JJR or a JJS would not. Similarly, give a context where only a JJR would work, and one where only a JJS would work.
(You can use the nltk similar() and common_contexts() functions for ideas if you're stuck).

■

Preview

NLTK 2.4: Read in the texts of the State of the Union addresses, using the state_union corpus reader. Count occurrences of men, women, and people in each document. What has happened to the usage of these words over time?

■

NLTK 3.7: Write regular expressions to match the following classes of strings:
1. A single determiner (assume that a, an, and the are the only determiners).
2. An arithmetic expression using integers, addition, and multiplication, such as 2*3+8.

1.

2.

■