CSC 206: NLP
Spring 2015

Name (Print): _____

Problem Set 3: Working with NLTK

Due: Friday, March 17 2015, before class

Please create a folder for this problem set, and save all your Python files in this folder as well as a (one) pdf document containing the answers to all the questions requiring a written answer.

In this problem set you will work with corpora from NLTK and you will practice using frequency distributions and conditional frequency distributions. Chapters 1 and 2 of the NLTK book will be helpful.

Questions:

1. NLTK book, Chapter 2, Exercise 4: Read in the texts of the *State of the Union* addresses, using the `state_union` corpus reader.

    1. Write a function `changes` that counts the occurrences of *men*, *women*, and *people* in each *State of the Union* address and prints out the results.
    2. What has happened to the usage of these words over time?

    For part 1, submit a Python file called `changes.py`.

    Add your answer for part 2, to a document, which you will use for all answers to questions requiring a written answer.

2. NLTK book, Chapter 2, Exercise 17: Write a function called `non_stopwords` that finds the 50 most frequently occurring words of a text that are not stopwords. The function should take a text, such as `gutenberg.words('austen-emma.txt')` as its parameter value and it should return its results as a list.

    Note: See Chapter 2 Section 4 of the NLTK book for a description of the stopword corpus that comes with NLTK.

    Save your Python code in a file called verb+non-stopwords.py+.

3. NLTK book, Chapter 2, Exercise 19: Write a function called to create a table of word frequencies by genre, like the one given in the passage on the Brown Corpus in Section 2.1 for modals. Choose your own words and try to find words whose presence (or absence) is typical of a genre. Discuss your findings.

    Save your Python code in a file called `genres.py`.

    Add the discussion of your findings to the same document in which you answered part 2 of question 1.

4. NLTK book, Chapter 2, Exercise 10: Read the BBC News article: *UK's Vicky Pollards 'left behind'* (`http://news.bbc.co.uk/1/hi/education/6173441.stm`). The article gives the following statistic about teen language: "the top 20 words used, including *yeah*, *no*, *but* and *like*, account for around a third of all words."

    1. Use NLTK to find out how many word types account for a third of all word tokens, for a variety of text sources.
    2. What do you conclude about this statistic?

    Read more about this on LanguageLog, at `http://itre.cis.upenn.edu/~myl/languagelog/archives/003993.html`.

    For part 1, submit a Python file called `teen-language.py` that shows the code you used to answer this question.

    Add your answer to part 2 to the same document in which you answered part 2 of question 1.

Make sure that you use the function names and file names that are specified in the instructions.