Chapter 2 Exercises

Erin M. Buchanan

1/15/2019

Get Started

- Create a Jupyter notebook with the following items. You can upload a compiled version of the notebook and the ipython or a script file.
  - Remember, use Markdown cells to answer text questions. Paste the questions into the cells so it's clear what you are answering.
- Import the nltk as shown in the lecture.

Basic Statistics

- Use the corpus module to explore austen-persuasion.txt.
  - How many word tokens does this book have?
  - How many word types?
- 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?

Conditional Frequency Distributions

- Define a conditional frequency distribution over the Names corpus that allows you to see which initial letters are more frequent for males vs. females (see the chapter for an example of the end letter differences).
- Write a program to generate a table of lexical diversity scores (i.e. token/type ratios), as we saw in 1.1.
  - Include the full set of Brown Corpus genres (nltk.corpus.brown.categories()).
  - Which genre has the lowest diversity (greatest number of tokens per type)?
  - Is this what you would have expected?
- Write a function that finds the 50 most frequently occurring words of a text that are not stopwords.

WordNet

Please note: these exercises are advanced - you will get credit for trying to find a solution, it is ok if it doesn't totally work.

- The polysemy of a word is the number of senses it has. Using WordNet, we can determine that the noun dog has 7 senses with: len(wn.synsets('dog', 'n')).
  - Compute the average polysemy of nouns.
  - You can get all noun synsets using list(wn.all_synsets('n')).
- Use one of the predefined similarity measures to score the similarity of each of the following pairs of words.
  - Rank the pairs in order of decreasing similarity.
  - Words: car-automobile, gem-jewel, journey-voyage, boy-lad.