Chapter 5 Exercises

Erin M. Buchanan

2/14/2019

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.

Tokenization + Tagging

- Use the base `pos_tag` function to tokenize the following: "They wind back the clock, while we chase after the wind.".
  - You will likely need to `word_tokenize` the sentence first.
  - *Wind* is a heteronym - words with different pronunciation and meaning.
  - How did *wind* get tagged in this sentence?

Python Object Types

- How can you distinguish between a list, tuple, and dictionary in Python output?
- What is the difference between them when you are using them in Python (i.e. what do lists do that tuples can't? Why use a dictionary?)?

Dictionaries

- Create two dictionaries, d1 and d2, and add some entries to each.
- Now issue the command d1.update(d2).
- What did this do?
- What might it be useful for?

Explore Tagged Corpus

- Load the Brown Corpus.
  - Which word has the greatest number of distinct tags. What are they, and what do they represent?
  - How many words are ambiguous, in the sense that they appear with at least two tags?
  - Create a list of the 20 most frequent tags in the Brown Corpus. What do they represent?

Unigram Tagger

- Train your data on the Brown corpus using the science_fiction category. Because you are using different train and test sets, you do not need to split it into different sizes.
- Then evaluate your tagger on the Brown corpus data in the hobbies category.
- What were the results? How well did your training match the hobbies category?

- What might be a better corpus to use to train for the hobbies category (that's not specifically the hobbies data)?