Chapter 3 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.

Electronic Resources

Text Files

- Find a .txt file on Project Gutenberg to download using `request.urlopen()`.
- Read in the text file and use `word_tokenize` to create a list of the word tokens.
- Find the points in the book that contain the Project Gutenberg information and remove that text.

HTML Files

- Go to your favorite news organization and find an article you want to use as a corpus.
- Import that file using `bs4` package and HTML format clean up.

Regular Expressions

- Using your book or your html file, print a list of all *wh* word types that occur (wh-words in English are used in questions, relative clauses and exclamations: who, which, what, and so on).
  - Use the `set` function to get only a unique list of these words.
  - You do not have to use `.islower` to clean it up, let's look at the raw list to see how much variability there is in text.
  - Are any words duplicated in this list, because of the presence of case distinctions or punctuation?
- Pig Latin is a simple transformation of English text. Each word of the text is converted as follows: move any consonant (or consonant cluster) that appears at the start of the word to the end, then append ay, e.g. string → ingstray, idle → idleay. http://en.wikipedia.org/wiki/Pig_Latin
  - Write a function to convert a word to Pig Latin.
  - Use regular expressions to find the first vowel in a word, index that, and then rearrange the word to be vowel to end + beginning + ay.
  - Test your function on the following words: cheese, elephant, moose, thing

Stemmers

- Use the Porter Stemmer to normalize your tokenized book or html document, calling the stemmer on each word.
- Do the same thing with the Lancaster Stemmer and see if you observe any differences.

Tokenization

- Using your book or html file, tokenize the document into both words (`word_tokenize`) and sentences (`sent_tokenize`).
- Calculate the average length of the words avg_w and the average length of the sentences avg_s
- Calculate a readability index by using the formula: `4.71*avg_w + .5*avg_s - 21.43`.