BioE 131/231 Lab Assignment 3

Stephen Martis

September 6, 2018

Shakespeare by the numbers: statistics of word counts in text

In this lab you will use your Python skills to take a text (we'll use Romeo and Juliet) and plot a histogram (bar graph) of the words in the play. The skills you use to complete this assignment (string comprehension, dictionaries, plotting, etc) will be applicable to biological analyses we'll do in this class. There is a version of last year's lab with some background on Python if you don't feel confident with some of the data structures and methods you will need. As always with coding questions, Google is your best friend (and learning to formulate the right search terms is a practiced skill). Also please ask either of the GSIs any questions you might have and communicate if you are feeling unsure about any part of the assignment.

The following steps should guide your analysis:

1. Read in the datafile 'romeoandjuliet.txt'. The file is in the bCourses folder for Lab 3.

2. Create a dictionary in which the keys are unique words and the values are the number of times that word appears in the play.

3. Plot a bar graph that shows the frequency with which each word appears. What are the ten most frequent words? What about the least frequent? Comment on what this might mean.

4. **Bonus:** sort the dictionary according to decreasing word count. Plot the word count vs. the rank of the word (i.e. the most frequent word will be rank=1, the second most frequent word will be rank=2, etc). Try plotting this on a log-log scale (i.e. log[word count] vs. log[rank]). What does this look like? Can you try to interpret this?

Please upload your Jupyter notebook to github and submit a link to it on bCourses.