Allison Basore

2/16/2017

Project 3: Jane Austen Analysis

Project Overview

My project was to analyze the words in the writings of Jane Austen throughout her career in the Georgian writing era. Text copies of all her works were used from the Project Gutenberg. Each book was striped of extra characters. Then a histogram of words used was made for each book and then adjusted for the length of each book with an augmented frequency formula. Lastly, her works were compared in order to see if her writing changed over her career.

Implementation

The general flow of this program is as follows: (1) requests a certain text file for a book, (2) removes the header, (3) clears everything except the words, (4) creates a histogram for the word usage adjusted for length, (5) repeats for several books, (6) graphs and stores data to compare frequencies of words. To accomplish these steps, several functions were created and the important concepts of these functions are outlined below.

To first obtain the text, I used a request download right from the Python code. However, since I did not want to risk reaching the max request download every time I re-ran the code, I decided to download the text files to my home directory. Also, since Austen is my favorite author, having copies of these books may come in handy when I am board in the airport. Furthermore, the header is removed from the text by using a boolean argument set to find the line that begins with '***' which is found at the beginning of text for every Gutenberg project book.

Next, I used line split methods to break the text into words and from there was able to remove the whitespace and punctuation. Using a for loop and get method, I was able to sort the words by their frequencies. Then I made another histogram from the first histogram, this time adjusting for the length of the book. This formula is simply taking each frequency value divided by the maximum frequency. With that information, I was able to index out the top five used words and compare those across all her works.

Results

After running my program for all of Jane Austen's works, I had the program organize them in order of dates and give the top five most used words and their use ratio. Then I put this data into MATLAB and made the following figure.

This figure shows that overtime, Austen was fairly consistent with her most frequent words like 'the', 'to','and', and 'of', but over her career, she used the word 'her' less. As a next step, it would be interesting to compare her use of words with those of other Georgian era authors.

Reflectio

Overall, ng data
from the sider the
insight I
worthwh dedicate
a little m and
compare