Software Design 3 Writeup

Gabriel Butterick

February 2015

1 Overview

This project makes use of the Gutenburg Project's free online books and uses characterization by word usage frequency. Using this information and tool together, I wanted to discover if there was a difference between word frequency and writer quality. Word frequency information is useful in that it can show how effective a writer is. If they use the same word many times in one work, they are usually considered to be poor writers by virtue of their sub par vocabulary. This word frequency program helps differentiate the high vocabulary writers from those who struggle to word their thoughts.

2 Implementation

The code consists of the main function, which is responsible for calling all the others and doing some of the analysis, and a variety of helper functions. The main code iterates through the unique words and adds up the number of times each word is used, as well as using the separation of duties to keep the main code cleaner and more well organized than if I had put everything in one function. The helper functions ranged in responsibility from analysis to cleaning up the strings so they could be better analyzed. Placing the strip function for removing the Gutenberg Project text helped substantially as it allowed me to hide 300 lines of useless code. The helper functions all rely on each other for outputs to refine and focus the algorithm. They drop all their results in the main function which then pushes them where they need to go. The output of the whole system is a single number instead of the entire list of words in the book, so the post-analysis is made much easier.

3 Results

The results from this experiment into the capability of writers based on their word usage proved my early hypothesis false. If one were to follow the results as I said, a writer of little import would beat out Arthur Conan Doyle instead of the other way around. This is obviously false, and leads me to believe that the more words are repeated, the better the author. Moby Dick claims the top number of words used more than ten times and, by popular viewpoint, is the most sophisticated of the three works. That being said, I believe overall I have learned you cant judge an author by their word usage. Even authors with upper level vocabularies seem to repeat the same high level words over and over instead of diversifying.

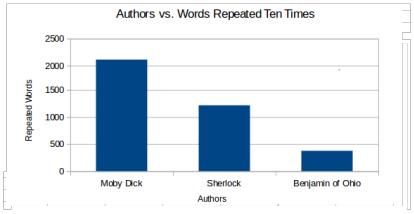

Figure 1: Graph Depicting Results

4 Reflection

Overall, I would say this project was successful. That being said, I know I could have done much more with it than I did. I was caught in an unfortunate amount of work and had to do what I consider barely sufficient for this project just because I have no time. From a technical aspect, I struggled for a while with compiling times. I eventually came to realize the cause was the fact that I was printing everything all the time. I learned that printing is useful for proof of concept success, but especially with large amounts of data, printing the whole result should be avoided. This topic is really interesting to me, and I think I will refine it more when I have time, just to experience all the other data collection tools and sources I had to overlook.