Aditya Kaushika

Mini-Project 3 Writeup.

The data source that I used was Project Guttenberg. I choose three books to analyze for this project. The three books were: The Wolf Hunters, Death the Knight and the Lady, Seeing America First. I hoped to create new sentences from the three books and create a story of random words put together. After that, I hoped to find my favorite words within the new story that I had created.

I started by importing the words from the three stories and stripping them down to just lowercase letters and no punctuation. After that, I randomized the words and enumerated them. By enumerating them, I could use the key numbers to generate random words. From those random words, I created random sentences. Those sentences, I thought, would make sense, but they didn’t because they were just random words. After that, I wanted to create a keyword search engine. Basically a control f function for the new story that I created. It is designed to take in the body of text from a research paper as a text file and create a dataset, then traverse the dataset and create a dictionary detailing the number of occurrences of keywords as specified by the user. It is designed so that researchers can easily identify relevant or effective research papers from the occurrences of specified keywords. The main decision I needed to make was the words I needed to choose for the second portion of the project.

Since I did a text analysis, I can talk about the interesting findings I created from the randomization of the words into sentences. This portion of the program worked with no problem. The problem I occurred was with the second part where I needed to analyze the new story I created and find select words in the new story. Since you cannot iterate an integer, the second portion does not work as well as I would like. I wish that the results created more clear sentences so I could read it. Maybe that will be my next project.

From a process point of view, I find the code readable, and my doctests did their job very well. I picked doctests that would encompass as many situations as possible while remaining simple and not taking up a lot of space or becoming cumbersome and difficult to read or understand. My code runs smoothly and serves its intended purpose well. I could stand to reduce redundancy in my code or start to practice optimizing my code for shortest length. For its intended purpose, my project was very well scoped. For a future project, I would love to create a higher order program that could determine relevancy of the paper based on user specified parameters of keyword occurrence, without the user having to decide on their own after viewing the final keyword histogram. Additionally, I'd like to increase the scope of the program so that it can take in a multitude of research papers and determine/rate relevancy of the research papers with respect to each other and user specified parameters.