Anika Payano

Software Design

Mini-Project 3

Text Mining: Lewis Carroll and Alice

Project Overview

I used Project Gutenberg to get two texts by Lewis Carroll, Alice in Wonderland and Alice's Adventures Under Ground, and I mined the text to see what common words showed up in both texts. I was hoping to just to create a basic code that would be able to compare the words from two different texts.

Implementation

In my code, there are only three really important functions that run the text analysis. It consists of: a processing function, a frequency function, and a comparing function. I was unable to use a doctest for the processing function and the frequency function, because both functions return some variation of dictionaries, and since the keys in a dictionary give the keys in random orders every time you run it, I am unable to create a doctest that will give me one correct answer.

Aside from that, the processing function's most important components are adding it to the dictionary and removing specific words and punctuation from the text. This would allow me to create a dictionary of just the words in the text and remove the very common words. I decided to do this because since I was looking at word frequency for this project, I wanted to find words that were not so common, like conjunctions. Additionally, the frequency function was just a basic function where it would return the most frequent words in the list, in highest to lowest order. There were many ways I could have found the frequencies of the words, maybe even easier, but I decided to do it this way because i feel like I have not really used the sorted() function, and I wanted to implement it so that I would understand the concept better. Lastly, the final function, compared two books and the 10 most frequent words that occurred in both. It was a pretty basic function in terms of compiling the list and then only returning the first 10 words.

Results

I did not really find anything interesting with this text analysis. It was clear before running the code that the most common word would be ‘Alice’. After seeing the results for the top 10 words in both books: 'Alice', 'little', 'about', 'would', 'could', 'thought', 'quite', 'their', 'began', and 'looked', it made sense that those were the most common words. It fits the genre of the book where it is most imaginary and about possibilities in Wonderland.

I would have wished to get some words like Jabberwocky or Bandersnatch, or some other nonsensical word. I would have somewhat expected some of those words to be pretty common. At the same time, Lewis Carroll has many that he uses occasionally, so again, it makes sense that it was not on the list.

Reflection

I think I could have explored many different options for counting the words in the text as well as saving the text without pickling. It also would be interesting to see where I could take this project and do other things with it, other than word frequencies. Unit testing was a little weird because of the dictionary usage. That can be another thing to improve. I will definitely use this project to keep building my knowledge and gain more confidence in coding. I wish I could have come up with a different idea. This one felt sort of boring but I could not think of anything else to do. Also, it would be interesting to try this project with all of Stephen King’s books. This is what I originally wanted to do, but the texts were not provided on Project Gutenberg.