SoftDes Project 3
Anil Patel

Overview

For this project I used archive.org as my data source and obtained txt files for two books, Eragon and Inheritance. I used a series of different functions to analyze these texts. I chose these two texts because they are part of the same series. Eragon, the first, was written when the author was 15, and Inheritance, the last, was written when the author was 23. I hoped to analyze aspects of the texts to find writer development over time between the texts.

Implementation

Due to the nature of the txt files and the NLTK library, I had to perform some pruning operations for accurate analysis later. I decided to do this after attempting to analyze before pruning and encountering many problems. The function cleaned up the text by separating any extraneous punctuation and deleting non-sentence-ending symbols from each word in the list. This would allow me to accurately analyze the length of words and the length of sentences by being able to search for sentence-ending punctuation in the lists.

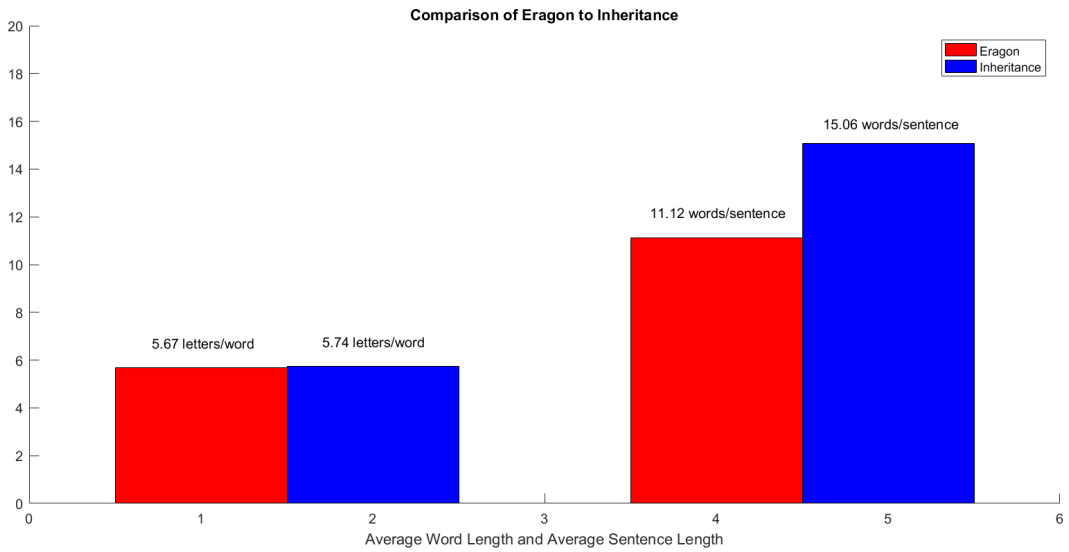
One design decision I made was how I separated letters and symbols. I created a string with all the letters of the alphabet in both lower and uppercase and then use an if ___ in “alphabetstring” to determine if it was a letter. This is much faster than writing many if and elifs to check every individual letter possibility.

To analyze the cleaned-up strings I looked at average length of words. For more meaningful results I removed all words less than 4 letters from the lists containing all the words in the texts. This also removed extraneous punctuation. With the updated list I added up all the lengths of words and the divided by total number of words.

My next analysis was on sentence lengths. To find this I ran through all the words in the text and added to a new list every time a sentence-ending punctuation appeared (.?!). I found the length of this list to find total number of sentences and then divided the total number of words in each text by this number.

Results

I found that my predictions were both somewhat correct. The following graphic shows the numbers I found for differences in average word length and sentence length.



In terms of word length, there is a small, yet evident increase in the average length of each word. It is expected that this change would be small because the vocabulary of an experienced adult writer does not contain many extra long words that would be frequently used in comparison to the number of words in the text. A more prominent difference was found in sentence length. I expected that the complexity of sentences written by the author would increase with the author's age which would increase the average length of sentence.

Reflection

In hindsight, I am proud of the analyses I was able to complete yet recognize that there is a lot more I could have done to look at similarities and differences between the two that would've added to the complexity of my project. I really enjoyed working with new libraries such as NLTK, and am excited to continue utilizing new ones in future softdes projects. Due to the length of time spent planning, writing, and debugging the code for the tests I did, I was unable to expand upon my analyses as I would have liked to, but am very satisfied that the functions I did write were successful.