Written First Draft

Research Questions:

The questions I want to answer are how sentiment changes over the course of the book, where I can see what chapters have positive, negative, or neutral emotions? Which characters have the biggest emotional impact on the story, where I can see how the sentiment shifts depending on what characters are a part of the scene? How does the emotional tone of the big events of the book compare with the other parts of the book, I can look at how sentiment changes around big plot points? How does the author's language change across the book, checking the language of the book?

Data Description:

My data is from an online pdf of Old Yeller by Fred Gipson, which I checked to make sure the content does match what the actual book says. The main variable is the sequence of words in the book, other variables include frequency, character sentiments, chapter sentiments, sentence length, and chapter number. The text file is about 40,000 words long. Chapters are denoted with ONE, TWO, and so on, which did trip me up at first. For cleaning the data, I converted all text to lowercase, removed punctuation and special characters, and tokenized the text into words and sentences. The text is chunked in chunks of 500. The pdf is read using pdf reader, concatenating text from all pages.

Methods:

Chapters are extracted using regex to match the chapter names. There is a line plot of sentiment per chapter. Sentiment scores are also printed for characters and plot points. Sentiment scores range from one to negative one, one being extremely positive and negative one being extremely negative. I have extracted the data, cleaned it, split it into chapters using regex, computed the average sentiment by chapter, character, and plot point, and plotted the sentiment by chapter on a line graph.

Progress and Next Steps:

I have completed most of what I set out to do, but it has some problems, which include it currently are it runs too slowly. It takes more than 5 minutes to run, but it does run, which is the first step and the most important part. I have split it into sections, but that didn’t solve anything, so I will be looking for more options on how to shorten the time it takes. I also need to clean it up, it has a lot of my notes and section breaks that are unnecessary, so it needs to be cleaned. Also, some of the plot points return “not found”, so I will need to either find the problem with it, or take out those points, probably find out the problem with it. My next steps will be to do as I previously stated, specifically speaking with Dr. Banuelos about how to shorten the time it takes my code to run. I will be going to office hours to also ask about the time shortening and some GitHub details.