Written Final Project

Research Questions:

The questions I have answered are how sentiment changes over the course of the book, where I can see what chapters have positive, negative, or neutral emotions? Sentiment is incredibly negative throughout the book, and spikes close to -0.5 at chapter 3, but every other chapter is very close to -1. 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? All character sentiments are negative on average throughout the book, but they are close to neutral, with Arliss being the closest to 0. 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? The events of the book are quite negative, though not as negative as the chapters themselves.

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:

I extracted the text from the pdf I found online using PDFreader. Since the PDF version has random line breaks and spacing issues, I concatenated every page into a single text string, to make sure the entire book was read. This made the run time incredibly long, so I had to separate them into 500 word chunks to make the run time more manageable. Chapters are extracted using regex to match the chapter names. Through the sentiment analysis pipeline, I used transformers to model sentiment analysis. I used a confidence score from -1 to 1, -1 symbolizing negative sentiment and 1 symbolizing positive sentiment. This let me use a plot to model the sentiment by 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. I used keywords to mtch the event or character with the sentiment of that section. My evaluation would be it works well for the most part, but if I had more time, I would want to investigate the chapter sentiment analysis a bit more, since they seem almost too negative in comparison with everything else.

Results:

The visualization I used is the line plot of sentiment by chapter, which shows the very negative sentiment of every chapter, with the highest being around -0.9 for chapter 3. This results in the novel's average sentiment score being -0.987, which seems pretty fitting for a sad and somber

novel. The characters sentiments are all negative as well, but they are much closer to neutral, with Arliss being the closest at -0.108, and Mama being the furthest at -0.376. Plot points are also all negative in sentiment, with a bit confusing most negative at Old Yeller Saves and most positive at Old Yeller Dies, I suspect it is because of the danger involved in Old Yeller saving people that it found the key words for negative sentiment, while Old Yeller dying had more good points on the family thanking and appreciating Old Yeller for what he did. If this book was analyzed for sadness and happiness instead of positive and negative, I suspect these events would be switched in sentiment, but that would be beyond the scope of my project. That and further analysis on the chapters sentiments would be what I would do if given more time.

Conclusion:
This project analyzed the sentiment of Old Yeller, showing an overall very negative sentiment score on average for the entire book, with chapter 3 being the highest sentiment score. The characters are close to neutral but still negative, and the plot points are also negative, closer to 0.5 on average. The line plot shows the difference in sentiment by chapter. The code still takes around 1-2 minutes to run, but it is still better than the previous over 5 minutes, which I would like to get down further, but that would be something I would do if I had more time.