

Dani Corum

12/9/25

Research Questions:

- What are the most common words in each chapter of *Dune*?
- What words are the most common across the entire novel?
- How does each chapter's sentiment change as the novel progresses?
- How many words are used only once?

My data comes from Frank Herbert's *Dune*, specifically a pdf version from [this link](#). From what I have done, I have used many variables, but the most important ones are `chTop5Words`, `chTop5Nums`, and `chSentiments`. These variables represent the top 5 words per chapter, the corresponding counts for those words, and the sentiments for each chapter. Unfortunately, the data these variables hold frequently needed to be reorganized to fit specific purposes. For example, I needed to merge `chTop5Words` into one big list, instead of a series of lists, so that I could enter it into a DataFrame.

I will clean the data by only working with the main 48 chapters, tokenizing each word, and removing stop words. I also chose to remove the word "said," since Frank Herbert rarely uses anything else in dialogue. After that, I will find the word counts per chapter and overall. I will then find the top 5 words in each chapter and in the entire novel. I plan to use bar graphs for modeling both the top 5 words per chapter and the sentiment analysis per chapter. To evaluate my results, I plan to observe the most notable chapters, especially in terms of the frequent words.

I have already done much of the initial work, such as setting up data, functions, and visualizations in my notebook. However, at the moment, it is somewhat disorganized. I believe I need to change the formatting of the variables the functions return, so that I don't need to reformat them later on. That has been the biggest challenge, alongside finding the best ways to put my data into a DataFrame.

I plan to reorganize my data and functions, both to make them work with less reformatting, and to make my code more understandable. I also need to figure out what I need to do to set up my GitHub repository. Lastly, another small change I would like to make is to change the colors for

the word frequency visualization, which is currently seemingly randomly assigned. This sounds simple on paper, but I suspect that it will be somewhat tedious getting colors that look different enough to represent 77 words.