

Worksheet 1: Handling Novel Full-Texts

In this worksheet we are going to work with the full-text of a novel. To begin, we will focus on *Pride and Prejudice* by Jane Austen. For this novel, the file *fulltext.txt* contains an annotated version of the original text of the novel as provided on Project Gutenberg.

Task 1: Loading Text

Firstly, read the entire novel into a list of lines of text.

Display the total number of lines in the novel. How many lines of text in the novel full-text are not empty (i.e., containing non-whitespace symbols)?

Task 2: Splitting Chapters

Next, iterate over the lines of text from above to split the novel by chapter. Store the non-empty lines for each chapter in a separate list. Display the number of chapters that you found.

(Hint: The first line of each chapter has the format 'CHAPTER 1', 'CHAPTER 2' etc. A regular expression could be used to find lines matching this pattern).

Using the lists you created above, now create a list containing the count of lines per chapter. Use this to calculate the average number of lines per chapter in the novel.

Using the counts that you created above, create a line plot which shows the total number of lines in each chapter in the novel.

Bonus Task: Comparing Full-Texts

Next, we expand the process above to consider the full-texts for all the novels in our dataset:

1. *Pride and Prejudice* by Jane Austen
2. *Dracula* by Bram Stoker
3. *Frankenstein* by Mary Shelley

Load the full-text file for each novel and find the number of non-empty lines in each case. Visualise the line counts from above using a bar chart.