

## Worksheet 2: Word Frequencies

A common type of analysis of text data is to look at the frequency with which different words appear in the text. In this worksheet we will look at applying **word frequency analysis** to novel full-texts. 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: Text Preparation

Before applying word frequency analysis, we need to perform a number of preparation steps to the full-text. Firstly, read the entire novel into a single string. Print the length of this string.

Next, convert the text that you have loaded into all lowercase.

---

### Task 2: Finding Words

Using the text prepared from Task 1, we now can start to identify the words in the novel's text.

Firstly, split the text into a list of all words appearing in the text. We can define a word as a substring that is separated by whitespace characters (e.g. spaces, tabs etc) and/or punctuation symbols. (Hint: We can do this a number of different ways, including by using regular expressions)

Next, filter out any words from the list which contain less than 2 characters (symbols). Report the number of filtered words and the number of remaining words.

Report the number of *unique words* that appear in the list of remaining words.

---

### Task 3: Counting Words in Full-Texts

Using the remaining words from Task 2, count the number of times that each word appears in the list (i.e., the word frequencies).

Display the top-20 most common words. (Hint: a Python Counter might be useful here)

Many of the words we see above are common stop-words which frequently appear in the English language and might not convey much information about the novel itself. An example set of stop-words is given below.

```
["am", "an", "and", "are", "as", "at", "be", "been", "but", "by", "can",  
"could", "do", "did", "for", "from", "had", "has", "have", "how", "i",  
"if", "in", "is", "it", "its", "me", "must", "my", "no", "not", "of",  
"on", "one", "or", "our", "say", "said", "shall", "so", "some", "such",  
"that", "than", "the", "them", "there", "this", "these", "to", "was",  
"were", "what", "when", "where", "which", "who", "why", "will", "with",  
"would", "you", "your"]
```

Remove all of the stop-words from the current list of word frequencies and display the top-20 most common remaining words. Visualise the frequencies for the top-20 words from above using a horizontal bar chart.

Now try removing additional stopwords and regenerating the chart above to see how it affects the top-20 word visualisation.



---

## Bonus Task: Comparing Word Frequencies in Full-Texts

Next, we expand the word frequency analysis to consider the full-texts for two different novels in our dataset:

1. *Dracula* by Bram Stoker
2. *Frankenstein* by Mary Shelley

Load and prepare the two full texts using the steps that we saw previously.

For each text, identify a list of the top-30 most common words. You should filter short words (less than length 2) and common stop-words as part of this process.

From the top-30 word lists, identify:

1. The top words common to both novels
2. The top words unique to *Dracula*
3. The top words unique to *Frankenstein*