

Worksheet 3: Working with Novel Metadata

In this worksheet we will start with a single novel, *Pride and Prejudice*, by Jane Austen, and we will look at the **metadata** associated with this novel, which represents the "data about our data".

The metadata takes two different forms. For each novel, this is stored in two different files in the JSON data exchange file format:

1. Metadata relating to the novel and the author of the novel. This is stored in the file *metadata.json*
2. Metadata relating to the characters in the novel. We refer to this as the "character dictionary". This is stored in the file *characters.json*

Task 1: Reading Basic Novel Metadata

Start by loading the metadata relating to the novel and the author of the novel from the JSON file *metadata.json*. The JSON data from the file should be parsed into a Python data structure, so that we can access it more easily.

Once you have loaded and parsed the novel data, extract and display:

1. The novel title and publication year information.
2. The full name of the novel's author.

Task 2: Reading a Character Dictionary

Next, load the data about all of the characters in the novel. Again, the JSON data from the file should be parsed into a Python data structure, so that we can access it more easily. In this case the structure will be a Python dictionary. We will refer to this as a character dictionary.

How many characters have been identified for this novel?

Display the **definitive names** for the characters in the character dictionary.

Display the **aliases** provided for a single character, the protagonist of the novel, Elizabeth Bennet.

Next, display the **attributes** provided for Elizabeth Bennet.

Task 3: Counting Character Attributes

We have character attributes associated with most of the characters in our character dictionary. First, count the number of characters with either the attribute female or male.

Next, count the number of times that each character attribute appears in the dictionary and display the top 20 most common attributes in the dictionary. (Hint: a Python Counter might be useful here)

Create a bar chart which visualises the attribute counts for the following character attributes:

```
["mother", "father", "wife", "husband", "son", "daughter", "brother", "sister"]
```

Bonus Task: Comparing Novel Metadata

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

3. *Pride and Prejudice* by Jane Austen
4. *Dracula* by Bram Stoker
5. *Frankenstein* by Mary Shelley

Load in the character dictionary for each of the three novels and store them in separate Python dictionaries.

Count the total number of characters in each novel. Display this information visually using a bar chart.

Finally, use the character attributes from each novel to calculate the ratio of female-to-male characters in each novel. Display the results as a new bar chart.