Quantifying Shakespeare

Importing a File and Analyzing Text

Setting the scene:

Your English Professor asks you to determine the most commonly used letters in all of Shakespeare's work

… but that includes:

- 38 plays
- 154 sonnets
- many, many poems

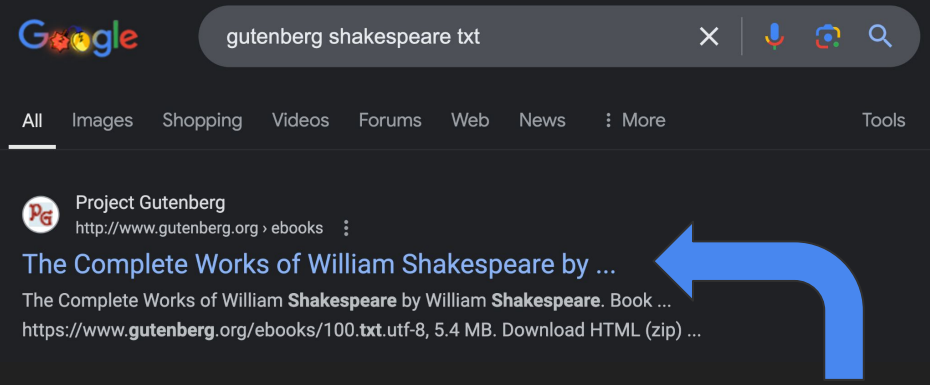
Let's write code to accomplish this!

The steps

1. Acquire all of Shakespeare's work
    a. Save the text in a file we can "read" with Python (✨new functionality!✨)
2. Keep track of the number of occurrences of each letter in the text
    a. What COMP110 concepts might we need to do this?
    b. What data structure could we use to store these data (of each letter and its associated occurrences)?
3. Print our findings!

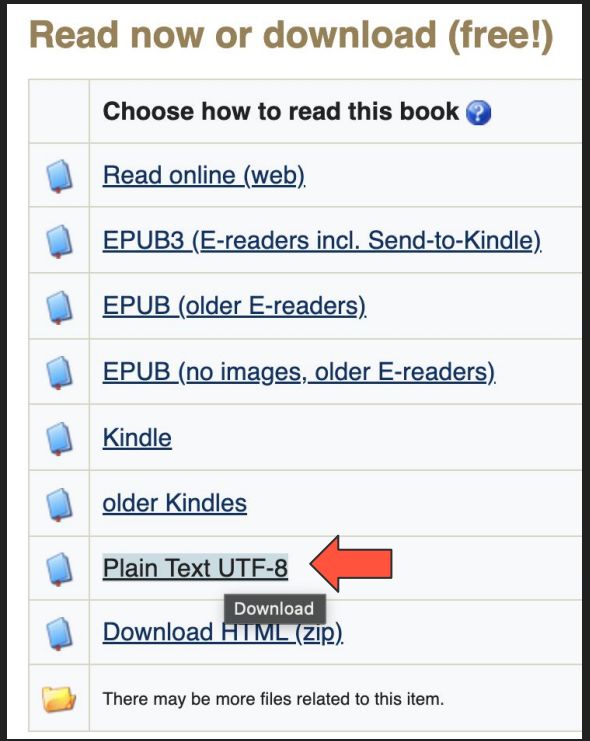
First, we need the data:

1. Google “gutenberg shakespeare txt”



and click on the first result

2. Click on the “Plain Text UTF-8”



First, we need the data:

You should see a looooooong page of text, starting with this:

```
The Project Gutenberg eBook of The Complete Works of William Shakespeare

This ebook is for the use of anyone anywhere in the United States and
most other parts of the world at no cost and with almost no restrictions
whatsoever. You may copy it, give it away or re-use it under the terms
of the Project Gutenberg License included with this ebook or online
at www.gutenberg.org. If you are not located in the United States,
you will have to check the laws of the country where you are located
before using this eBook.

Title: The Complete Works of William Shakespeare

Author: William Shakespeare

Release date: January 1, 1994 [eBook #100]
               Most recently updated: October 29, 2024

Language: English



*** START OF THE PROJECT GUTENBERG EBOOK THE COMPLETE WORKS OF WILLIAM SHAKESPEARE ***
The Complete Works of William Shakespeare

by William Shakespeare
```

3. Select all of the text
   a. Ctrl+A on Windows or command+A on Mac
4. Copy it
   a. Right click → copy

Then, in VS Code:

5. Create a new folder called “shakespeare”
6. In that folder, create a new file called “shakespeare.txt” and paste the copied text into it!

Scroll through the `.txt` file.

Do you notice anything that might hinder our ability to count the occurrence of each letter in Shakespeare's works?

We need to remove the extra text!

From top to bottom, delete lines 1-80 ("THE SONNETS") and 195961 ("*** END OF THE [...]") onward

.ipynb files: Jupyter Notebooks

With Jupyter Notebooks, we can write text (Markdown) and Python code in “chunks” to analyze and manipulate data in individual steps.

Let’s get to work! →