



# Quantifying Shakespeare

## Importing a File and Analyzing Text

# Announcements

## Final exam logistics:

- Official exam date/time: 4–7pm on Saturday, Dec 6
- Makeup exam (for exam excuses): Sunday, Dec 7 at 12pm
  - Please [complete our internal form](#) to let us know you'll be there!
- We will share location logistics and seat assignments via email closer to the exam

Want to be a COMP110 UTA in Spring 2026?

[Apply by December 10!](#)



## 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”

A screenshot of a Google search results page. The search bar at the top contains the query "gutenberg shakespeare txt". Below the search bar, there are links for "All", "Images", "Shopping", "Videos", "Forums", "Web", "News", and "More". A "Tools" button is also present. The main search results section shows a link to "Project Gutenberg" with the URL "http://www.gutenberg.org › ebooks". Below this, a snippet of text reads "The Complete Works of William Shakespeare by ...". A large blue curved arrow points from the text "and click on the first result" below to the "Project Gutenberg" link.

and click on the first result

2. Click on the “Plain Text UTF-8”

A screenshot of a Project Gutenberg download page. At the top, it says "Read now or download (free!)". Below that, a section titled "Choose how to read this book" lists several options: "Read online (web)", "EPUB3 (E-readers incl. Send-to-Kindle)", "EPUB (older E-readers)", "EPUB (no images, older E-readers)", "Kindle", "older Kindles", "Plain Text UTF-8", and "Download HTML (zip)". A red arrow points to the "Plain Text UTF-8" link, which is highlighted with a black box. Below the list, it says "There may be more files related to this item."

# 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](http://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 (or Ctrl+C on Windows or command+C on Mac)

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!
  - a. Right click → Paste (or Ctrl+V on Windows or command+V on Mac)

# 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 (up to “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” or “cells” to analyze and manipulate data in individual steps.

Let's get to work! →

