Assignment 3

Text as Data

2023-10-02

String manipulation

The following homework sets you a challenging task to give structure to raw text. You will need to look through the text to discover text patterns that will allow you to split, join, and extract in order to create the structure we are looking for. If you struggle to implement this, describe *in words* how you would plan to go about the task, thinking about how to split this into subtasks and how to describe these.

To start with, you are asked to retrieve *Songs of Innocence and of Experience* by William Blake from Project Gutenberg. It is located at https://www.gutenberg.org/cache/epub/1934/pg1934.txt. This is a collection of poems in two books: *Songs of Innocence* and *Songs of Experience*.

The goal of the task is to parse this into a dataframe where each row is a line of a poem (there should be no empty lines). The following columns should describe where each line was found:

- line_number
- stanza_number
- poem_title
- book_title

Substeps

Think about how to split this up into smaller tasks before bringing this all together. Remember that when working on loops it is often easier to work with a single item in the list first, before putting this into a loop. There will be other ways to approach this, but a step-by-step approach will do the following:

- Get the content of the book (removing publisher information, contents, and the copyright notice)
- Split the book into the two sub-books (Songs of Innocence and Songs of Experience)
- Split each book into poems
- Split each poem into stanzas (verses)
- Split each stanza into lines