Assignment 1

Text as Data

2022-09-08

Introduction

In this assignment, you are asked to produce analysis that completes a set of instructions. You can do this any way you like, as long as you show me your results and the code you used to get there. The easier this is for me to replicate, and the clearer the code is, the higher your mark will be. One option would be to make a copy of this file, add in code snippets, and submit the RMarkdown file along with the PDF of completed results. Another option would be to send me a link to an .ipynb notebook file on Github.

Getting and parsing texts

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*.

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

Manipulating data

- Create a histogram showing the number of lines per poem
- Create a document feature matrix treating each line as a document
- Create a separate document feature matrix treating each poem as a document
- Using one of these document feature matrices, create a plot that compares the frequency of words in each book.