Xi Wang (xw2ez@virginia.edu)
DS 5001 Exploratory Text Analytics
May 6, 2024

Manifest

Provenance

The source files I used for this final project are from Project Gutenberg (https://www.gutenberg.org/). Project Gutenberg is a volunteer-driven initiative dedicated to digitizing and archiving cultural works, particularly literary texts. This website is a library that contains over 70,000 free eBooks. Overall, Project Gutenberg plays an essential role in promoting literacy, preserving cultural heritage, and democratizing access to knowledge by providing free and open access to a vast collection of public domain texts.

Location

https://virginia.box.com/s/2ik5xfc0m6q3c27ahseh92xvlebbsfum

Description

The general subject matter of Project Gutenberg encompasses a broad range of literary texts and written works that fall into the public domain, including classic literature, historical texts, philosophical and religious works, reference works, scientific and technical literature, foreign language texts, etc. The corpus is diverse and contains virtually all areas of human knowledge and creativity that have entered the public domain.

For this project, I specifically chose nine books from three different authors. The nine books are *Emma*, *Mansfield Park*, *Pride and Prejudice*, *Adam Bede*, *Israel Potter*, *The Apple - Tree Table and Other Sketches*, *The Piazza Tales*, *Middlemarch*, and *The Mill on the Floss*, and the three authors are Jane Austen, Herman Melville, and George Eliot. These books collectively still cover a diverse range of subjects, including romance, social commentary, historical events, moral dilemmas, and philosophical reflections, etc.

Format

The source files are all in plain text format (txt). The basic structure of the plain text format is that data are arranged in rows, with several values stored on each row.

The basic conceptual structure of a plain text format is that the data are arranged in rows, with several values stored on each row. There is also a file header, which indicates general information about the dataset or metadata.