DS5001 Exploratory Text Analytics
Final Project - The Harvard Classics
Source File Manifest

Derek Liu (hdl5hz@virginia.edu)
May 11, 2021

Provenance: The source files were downloaded from Project Gutenberg's website. The entire collection of The Harvard Classics was found on archive.org, however they were unreadable due to file type conversion and scanning errors (including many alien symbols and spelling mistakes). The collection was from The Harvard Classics bookshelf on Project Gutenberg, sorted by popularity. The top 39 books from the collection were downloaded (the criteria was more than 1000 downloads), and 22 were successfully converted into the F1 and F2 formats.
The link to the bookshelf is: https://www.gutenberg.org/ebooks/bookshelf/40

Source Files Link (.zip): https://virginia.box.com/s/5puwdfiyw1qg5ing6qkesastdsffgnnq

Description: The Harvard Classics is a collection of books, also known as Dr. Charles William Eliot's Five-Foot Shelf of Books. Dr. Eliot was a former president of Harvard University, and he published this collection in the early 1900s for the general public to read. The collection was marketed and known for "a liberal arts experience at Harvard University" at the time.

Format: All the source files were of PlainText UTF-8 format when downloaded from the Project Gutenberg website.