Data Set Share #2 – New York Times Reviews of the Top 100 “Best Books” from Goodreads

Austin Smith

This data set builds off the Goodreads data (data share #1). Here, I attempt to pull the New York Times book review for the top 100 books on the “Best Books Ever” list of Goodreads, a compilation of the highest rated books from the Goodreads community. Only 31 books on the Goodreads list have been reviewed by the New York Times. Several of them (e.g., The Book Thief) have been reviewed multiple times by the New York Times.

The review texts are all saved in their own files, by book, and range in size from 3KB to 19KB. The text has been minimally cleaned.

In analyzing this data set, one could take a variety of approaches. For example, if you wanted to keep a narrow scope, you could look at books with several review results and compare and contrast them; how did different reviewers respond differently to the book?

Alternatively, one could analyze all of the reviews as an entire corpus and gain a feel for how the New York Times reviews books. What words come up most frequently, for example? What other books are referenced in reviews? How have the most frequently used terms changed for books that were reviewed decades ago compared to books reviewed more recently?

If I were to try to compare two different reviews of the same book, I would take the approach we did in class when comparing the text from the democratic and republican conventions. In doing so, I would compare the two texts on their number of tokens, average token length, number of unique tokens, most used words, and how the proportion of commonly used words differed in each review.