Data Set Share #1 – Goodreads Best Books Reviews

Austin Smith

Goodreads maintains a list of “Best Books Ever,” which is a ranked list of the highest rated books of all time, as rated by members of the Goodreads community. I pulled the first 100 books from that list and scraped all of the textual data from the websites that describe them, including a sampling of their reviews from Goodreads members.

I was prompted to seek out this data because 1) I love books, and 2) I plan to compare these reviews to the New York Times’ book reviews for the same books (or as many as have been reviewed by the NYT) because I think it will be interesting to see how a “professional” reviewer’s opinion compares to that of the masses.

There are 100 books in the data set. The textual information that is included for each includes a summary statement, various extraneous text from the website, and a sample of reviews from the Goodreads community. This textual data has been minimally cleaned. The files range from about 20 KB to over 170 KB.

There are a number of interesting questions one I could answer with this data. For example:

1. Is there a general “tone” to how people talk about these “best” books? For example, do words like “suspenseful” or “scary” or “sad” or “love” come up a lot? (Here, we’d be trying to get at the question of whether a love story vs. a horror story might be more likely to be loved by the Goodreads community).
2. One could dig into when the books on the list were published and figure out how many of them are modern versus classics.
3. One could pull some basic statistics about the reviews/summaries to ascertain how complex the language is that people use to talk about the book (e.g., by looking at their lexical diversity). Perhaps there’s a correlation between the complexity of a review and the complexity of the book. It might be interesting to observe that the way people write about a book mirrors the way the book is written; for example, is there a substantial different in the lexical diversity of reviews on “Crime and Punishment” compared to “Twilight?”
4. One could do a comparison with the list of best books ever and a movies list to see how many were made into movies and if they have some sort of defining features that make them especially “cinematic” (or not).
5. Does the word “high school” or “school” come up frequently in any of the reviews? That may tell us which of the books were likely to be assigned in school, or read as teenagers. That would also clue us into whether the book is more of a “fun” read or an academic pursuit.

If I were to try to answer #2 above, I would pull the published date from the textual data in this dataset (it’s included for every title), and then plot it on a timeline, perhaps using a density plot (from ggplot) to see how many books on the list were published per decade. I would likely have to use regular expressions to pull the published date from the text data. Alternatively, I could scrape the tables from the Wikipedia website of the most sold books (see github repo for the code on how to do this), and then match up the titles from Wikipedia with the titles in this dataset in order to more easily access their published dates.