The Crowdsourced "Classics" and the Revealing Limits of Goodreads Data

Melanie Walsh and Maria Antoniak
melanie.walsh@cornell.edu // maa343@cornell.edu
Cornell University, Information Science

Project Overview

- Goodreads is the largest social networking site for readers on the internet (90 million users) and a subsidiary of Amazon

- “**Classics**” is one of the most active Goodreads categories, with some of the most rated and reviewed books across the entire site

Most Rated Books on Goodreads

Title	Ratings (millions)	Publication Date
Harry Potter and the Sorcerer’s Stone	6.7m	1997
The Hunger Games	6.2m	2008
To Kill a Mockingbird***	4.3m	1960
The Great Gatsby***	3.6m	1925
The Fault in Our Stars	3.4m	2012
1984***	2.97m	1949
The Catcher in the Rye**	2.6m	1951
Animal Farm***	2.6m	1945

***Books in the top 100 most shelved “classics” are bolded

Motivating Questions (1st Set)

- Why are the classics so popular on Goodreads?
- Which books have readers “shelved” as classics most often?
- What do the classics mean to contemporary readers?
- What is the value of the classics to Goodreads and Amazon?

Project Overview

- When we turned to collect and analyze Goodreads reviews of the classics, we realized that Goodreads/Amazon manipulate this data in myriad ways

-The contemporary book world is increasingly governed by algorithms and data

"[H]ow are scholars to document, much less critique, algorithmic culture's self-reinforcing effects on cultural selection if denied access to the workings of the algorithm's engine-room?"

-Simone Murray, "Secret Agents: Algorithmic Culture, Goodreads and Datafication of the Contemporary Book World"

Motivating Questions (2nd Set)

- How does Goodreads’ default sorting algorithm influence reviews? How does Goodreads’ throttling to 300 visible reviews per book (in each sort setting) influence reviews?
- How might Goodreads reviewers’ perceptions of the classics be influenced by other Goodreads reviewers?
- How might Amazon/Goodreads be using all this Goodreads data?

Related DH Work

-Karen Bourrier and Mike Thelwall, “The Social Lives of Books: Reading Victorian Literature on Goodreads”

-J. D. Porter, “Popularity/Prestige”

-James F. English, Scott Enderle, and Rahul Dhakecha, “Mining Goodreads: Literary Reception Studies at Scale”

-Allison Hegel, “Social Reading in the Digital Age”

-Andrew Piper and Richard Jean So, “Study Shows Books Can Bring Republicans and Democrats Together”

Data

Our Goodreads “Classics” Dataset

- ~900 Goodreads reviews for each of 144 classic texts
  - 300 oldest, newest, and default reviews per book
  - Filtered to English-language reviews
- **127,855 total** Goodreads reviews (2007-2019)
- Code used to scrape these Goodreads reviews is available on Github: https://github.com/maria-antoniak/goodreads-scraper

Results

Themes in Reviews of Classics (Topic Modeling)

- School
- Editions & Translations
- Adaptations & Audiobooks
- Goodreads User Criticism
- Review Industry & Meta-Review Discourse
- Gender & Sexuality
- Race
- Family
- Life & Death
- War & Adventure
- Murder & Revenge
- The Future (Dystopias)
- Marriage
- Comedy
- Mystery & Suspense
- Children’s Literature

- Critical Status
- Plot & Characters
- Unlikeable Characters
- Beautiful Writing
- Length & Pace
- Enjoyable & Interesting
- Re-Readable
- Literary Language (Quotations)
- Conversational & Slangy Language
- Description & Dialogue (Quotations)
- Gushing & Loving Language
- Talking & Speaking
- Non-English Reviews

Results

The Classics According to Goodreads Reviewers

"I had been planning to read '1984' for a long time. It's one of those books that you are supposed to read in high school. My high school AP Lit teacher had us read Aldous Huxley's 'Brave New World' instead."
Andrew, "Andrew's Review of 1984," Goodreads, May 2, 2007

"This review is inspired by some of my GR [Goodreads] friends whose fearlessness about giving low stars to books they do not like has inspired me to change my rating of Lolita from three stars to two stars as that is what I really feel…I get that this a classic and book snobs who read this will sig[h] in indignation but I do not care. I just did not get it and still don't. I'd like to thank anti book snobs everywhere for giving me the courage to rate Lolita two stars. I will never forget you. Wow..is this what an Oscar speech feels like?"
Bren, "Bren's Review of Lolita," Goodreads, April 11, 2018

"Every so often I'll get into a classic. I guess because I feel like writing a really nasty review. Classics are great fodder for nasty reviews because 1. The people who made them are LONG dead…Saying bad stuff about a classic novel doesn't hurt the creator's feelings…2. Classics have such a pedestal in the literary world already that the opinion of one lone weirdo…is pretty irrelevant. It's not like bashing on this book is suddenly going to render it a Not A Classic or affect its sales. Frankly, I think that about everything I read, but with classics, it's a pretty rock solid premise."
Peter Derk's Review of The Phantom of the Opera," October 28, 2019

Algorithmic Echo Chamber

- The first 300 default Goodreads reviews for a given book develop into an echo chamber
  - Once a Goodreads review appears in this default sorting, it is more likely to be liked and commented on
- The Goodreads reviews that show up in the default sorting tend to be longer, more socially conscientious (e.g. include a spoiler alert), and written by a smaller set of Goodreads users
  - It also appears that Goodreads users may be more likely to go back and modify their reviews when they are prominently displayed by the default sorting algorithm

User Ethics

Many Goodreads users take pride in their reviews and craft them carefully, similar to a professional book reviewer. If we think of **Goodreads reviewers as creative artists or amateur critics**, as it seems the authors themselves do, then anonymizing their reviews (removing their names and/or paraphrasing the review text) would deprive them of proper creative credit. However, prior work has shown that even when internet users post on public platforms, **they have an expectation of privacy**. For these reasons, we have chosen not to publicly share our dataset, though we have shared the code that we used to collect data from the Goodreads website: https://github.com/maria-antoniak/goodreads-scraper.

For Goodreads reviews directly quoted in our forthcoming essay, **we have obtained explicit permission from each reviewer**. We messaged each of these selected reviewers on Goodreads, disclosed our affiliations and the project goals and structure, and asked for consent to publish parts of their review in this article. We offered users the option of being quoted in this essay and attributed by Goodreads user name or the option of being quoted in the essay but remaining anonymous.

DH2020 and Forthcoming Work

“The Goodreads ‘Classics’: A Computational Study of Readers, Amazon, and Crowdsourced Literary Criticism,” Melanie Walsh and Maria Antoniak, *Cultural Analytics / Post45*, forthcoming (late 2020)

We’re happy to take questions as part of the DH2020 panel, “Cultural Analytics and the Book Review: Models, Methods, and Corpora.”

Feel free to respond to our forum post on Humanities Commons or email us:

Melanie Walsh: melanie.walsh@cornell.edu

Maria Antoniak: maa343@cornell.edu