**Website:** UCSD Book Graph

* Link1: <https://sites.google.com/eng.ucsd.edu/ucsdbookgraph/home>
* Link2: <https://mengtingwan.github.io/data/goodreads.html>

**Associated GitHub Account:** <https://github.com/MengtingWan/goodreads>

**Overview:**

These datasets were collected in late 2017 from goodreads.com, where we only scraped users' public shelves, i.e. everyone can see it on web without login. User IDs and review IDs are anonymized. We collected these datasets for academic use only. Please do not redistribute them or use for commercial purposes.

We collected three groups of datasets: (1) meta-data of the books, (2) user-book interactions (users' public shelves) and (3) users' detailed book reviews. These datasets can be merged together by joining on book/user/review ids.

**Basic Statistics of the Complete Book Graph:**

* 2,360,655 books (1,521,962 works, 400,390 book series, 829,529 authors)
* 876,145 users; 228,648,342 user-book interactions in users' shelves (include 112,131,203 reads and 104,551,549 ratings)

**Meta-Data of Books**

* Detailed book graph (~2gb, about 2.3m books): goodreads\_books.json.gz
* Detailed information of authors: goodreads\_book\_authors.json.gz
* Detailed information of works (i.e., the abstract version of a book regardless any particular editions): goodreads\_book\_works.json.gz
* Detailed information of book series (Note: Unfortunately, the series id included here cannot be used for URL hack): goodreads\_book\_series.json.gz
* Extracted fuzzy book genres (genre tags are extracted from users' popular shelves by a simple keyword matching process): goodreads\_book\_genres\_initial.json.gz

**Book Shelves**

* Complete user-book interactions in 'csv' format (~4.1gb): goodreads\_interactions.csv.

**Citation:**

If you are using our datasets, please kindly cite the following papers:

* Mengting Wan, Julian McAuley, "Item Recommendation on Monotonic Behavior Chains", in RecSys'18. [bibtex]
* Mengting Wan, Rishabh Misra, Ndapa Nakashole, Julian McAuley, "Fine-Grained Spoiler Detection from Large-Scale Review Corpora", in ACL'19. [bibtex]

**Contact Email:** mengting.wan at microsoft.com