

goodreads__webscraper

Benjamin Rosen

1/28/2019

GitHub repository: https://github.com/brosen255/ds_projects

This repository provides the scrapy spider to webscrape goodreads.com . It crawls through the most popular list pages to each list result page, each book page, and the author page for those books. It collects characteristics from each book, including title, year published, genre, number of reviews, number of ratings, average rating per book. It also collects information on the author and user interaction with their page.