Goodreads Querying - Brief ReadMe

Tahmeed Tureen

Fall 2017

This document briefly highlights what I did for this small scale project regarding Goodreads data. For access to the database, contact me at tureen@umich.edu

Steps:
1. Open a connection to a SQLite database using Python
2. Read in data from CSV data files obtained from www.goodreads.com
3. Use SQL to create multiple tables in a database
4. Populate the tables using the data acquired from the CSV using petl
5. Extract info from the authors table to create a new table and link it to the books table appropriately
6. Query database to answer the following questions

(Q1) : Who are the top ten highly rated authored based solely on the Goodreads data?

(A) :

1. Lane T. Dennis
2. Bill Watterson
3. Ronald A. Beers
4. Kelly Jones
5. Steve Oliff
6. Lee Loughridge
7. Daniel Vozzo
8. James E. Talmage
9. Hafez

10. Angie Thomas

Surprising, right? We got this list because we used the five star ratings used in Goodreads and averaged those for each author. These authors do not have that many reviews, so their average ratings are high (for example: if they have like 1-3 positive reviews and 0 negative reviews).

(Q2) : Who are the top ten most popular authors based on people's "to-read lists"?

(A) :

1. Stephen King
2. Neil Gaiman
3. J.K. Rowling
4. Cassandra Clare
5. George R.R. Martin
6. Rick Riordan
7. Jane Austen
8. John Green
9. Brandon Sanderson
10. James Patterson

This looks reasonable! J.K. Rowling is #1 in my books though (no pun intended)