Book Characterization using Project Gutenberg's Open Library with Goodreads Reviews

Information Processing and Retrieval

Group
68

Motivation

Massive bodies of text, such as books, have become readily available in large online e-book libraries.

Hard to:

- Look for context around well-known quotes
- Know a book's general rating
- Browse public domain texts

→ Information Retrieval System

Pipeline

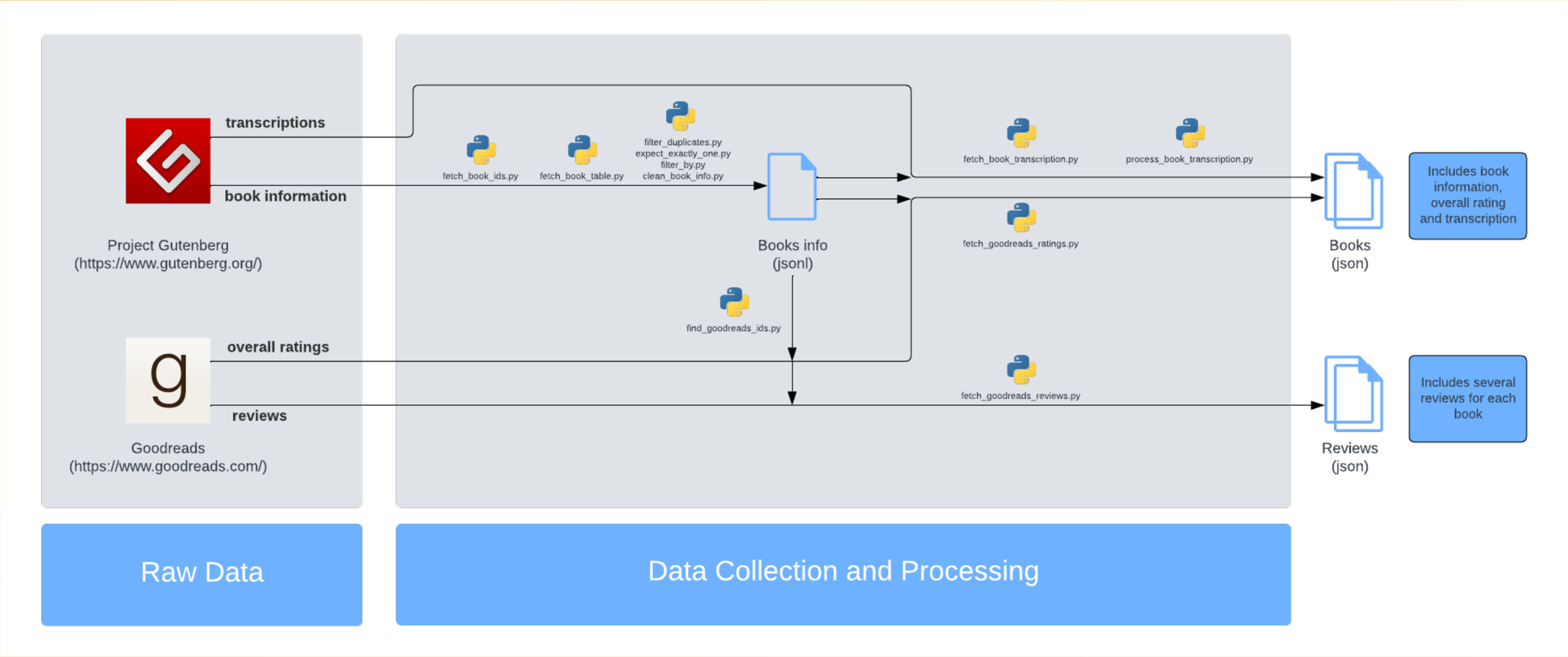


Fig. 1: Data Retrieval Pipeline

Conceptual Model

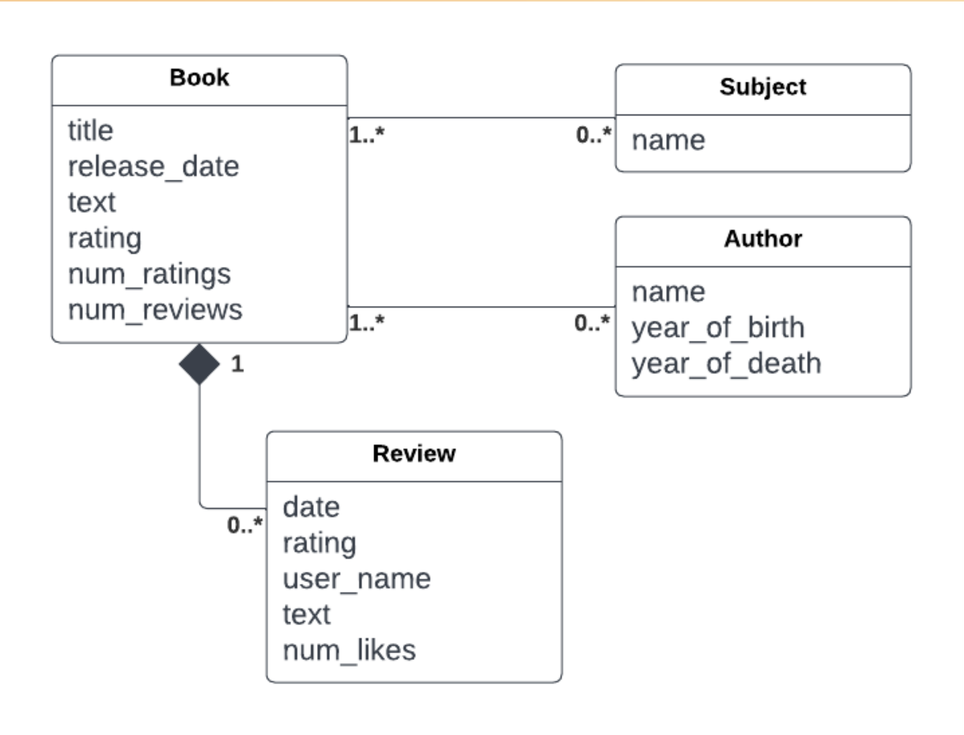


Fig. 2: Dataset Conceptual Model

Two main dataset groups:

Books, each with its own group of subjects and authors

Reviews, each associated with one and only one book

Data Characterization - Book Words

Table 1: Book word frequency

word	count	frequency
the	35999199	0.065195
of	19749394	0.035766
and	17721901	0.032095
to	14476953	0.026218
a	13785780	0.024966
in	10074152	0.018244
that	6193965	0.011217
was	5648257	0.010229
he	5636281	0.010207
I	5369173	0.009724

follows Zipf's Law

Table 2: Book word frequency (except *stop words*)

word	count	frequency
time	859451	0.003968
man	859450	0.003968
great	686541	0.003170
good	591981	0.002733
day	575192	0.002656
men	558979	0.002581
life	492075	0.002272
long	488925	0.002258
place	396396	0.001830
people	386386	0.001784

Data Characterization - Book Subjects

Table 3: Category frequency

category	count	frequency
short stories	485	0.024288
science fiction	407	0.020382
united states	348	0.017427
great britain	236	0.011818
england	233	0.011668
world war	232	0.011618
adventure stories	193	0.009665
fiction	185	0.009264
conduct of life	152	0.007612
detective and mystery stories	148	0.007411

Data Characterization - Authors

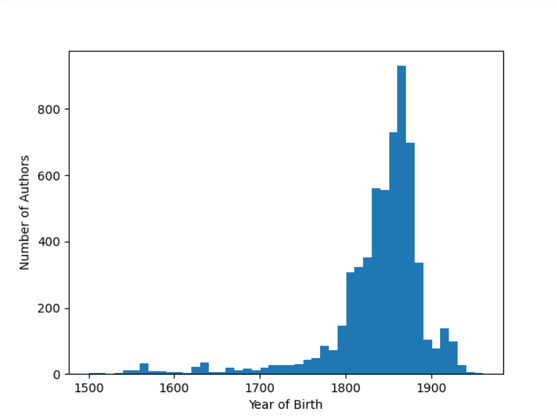


Fig. 3: Author Year of Birth Distribution

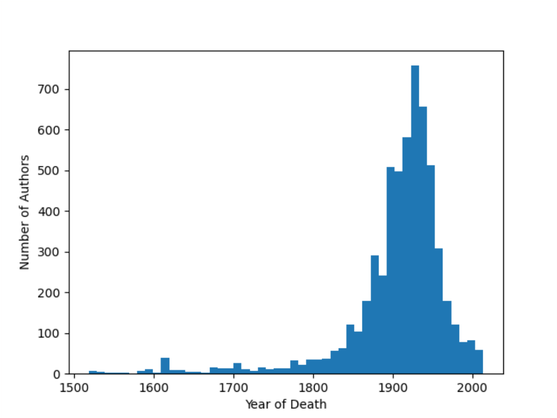


Fig. 4: Author Year of Death Distribution

Table 4: Author Book Count

author	count
Mark Twain	36
Charles Dickens	27
Edward Bulwer-Lytton	25
Horatio Alger	23
William Henry Giles Kingston	23
William Shakespeare	21
Georg Ebers	21
Honor de Balzac	19
George Manville Fenn	18
Arthur Conan Doyle	18

Data Characterization - Review Ratings

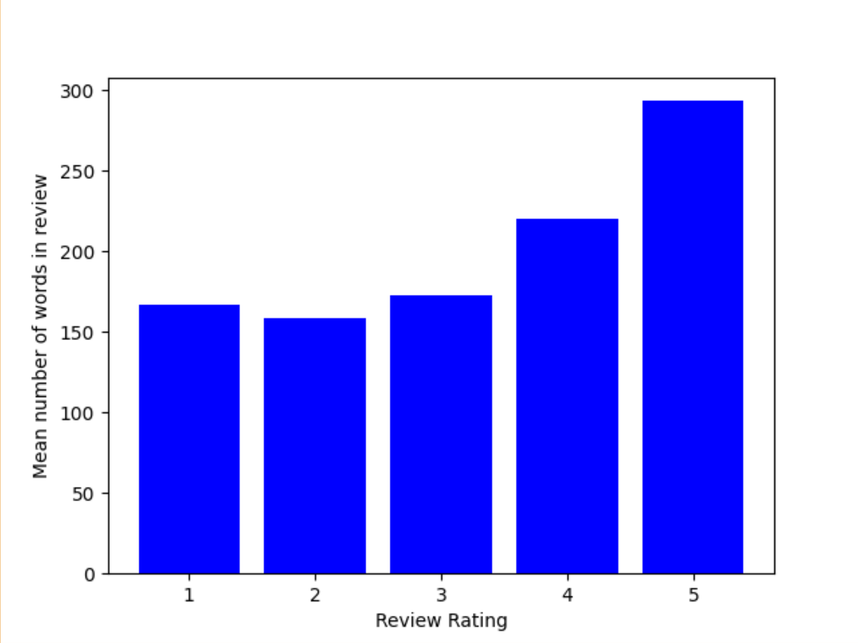


Fig. 5: Mean number of words in review per review rating

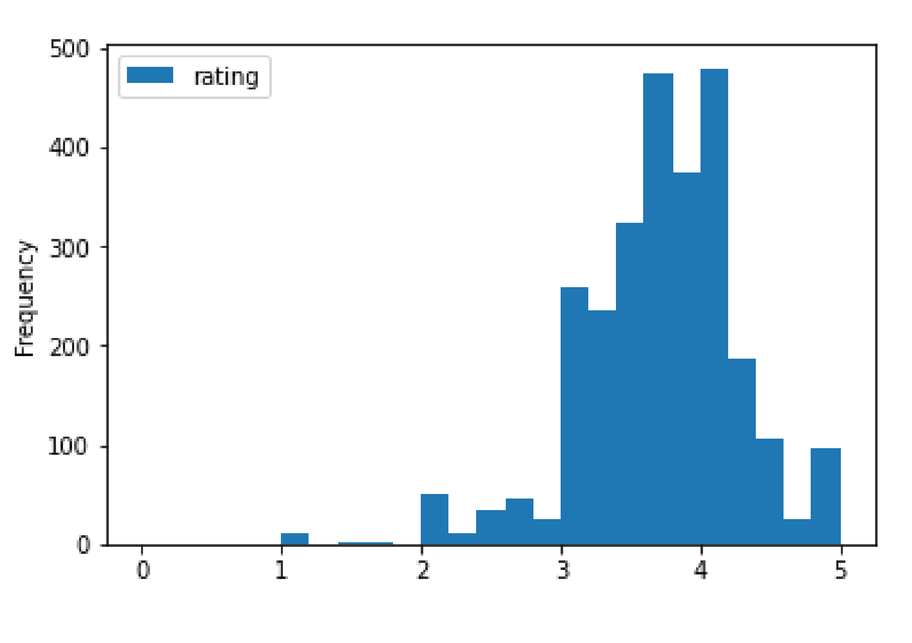
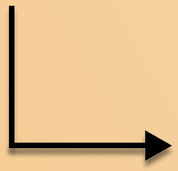


Fig. 6: Book rating distribution

Search Scenarios

There are 5 main search scenarios:

> I want to find excerpts about a specific concept or event.
> I want to find context on a book quote.
> I want to read a book from a certain time period.
> I want to understand what other people think of a book.
> I want to browse an extensive library of copyright-free literature.



All search scenarios will be solved with the Information Retrieval System

Conclusions and Future Work

Conclusions:

- Effective pipeline - sizeable and representative sample
- Added value of reviews
- Added value of indexing books

Future work:

- Information retrieval system based on dataset
- Respond to the prospective search scenarios
- Scrape more reviews