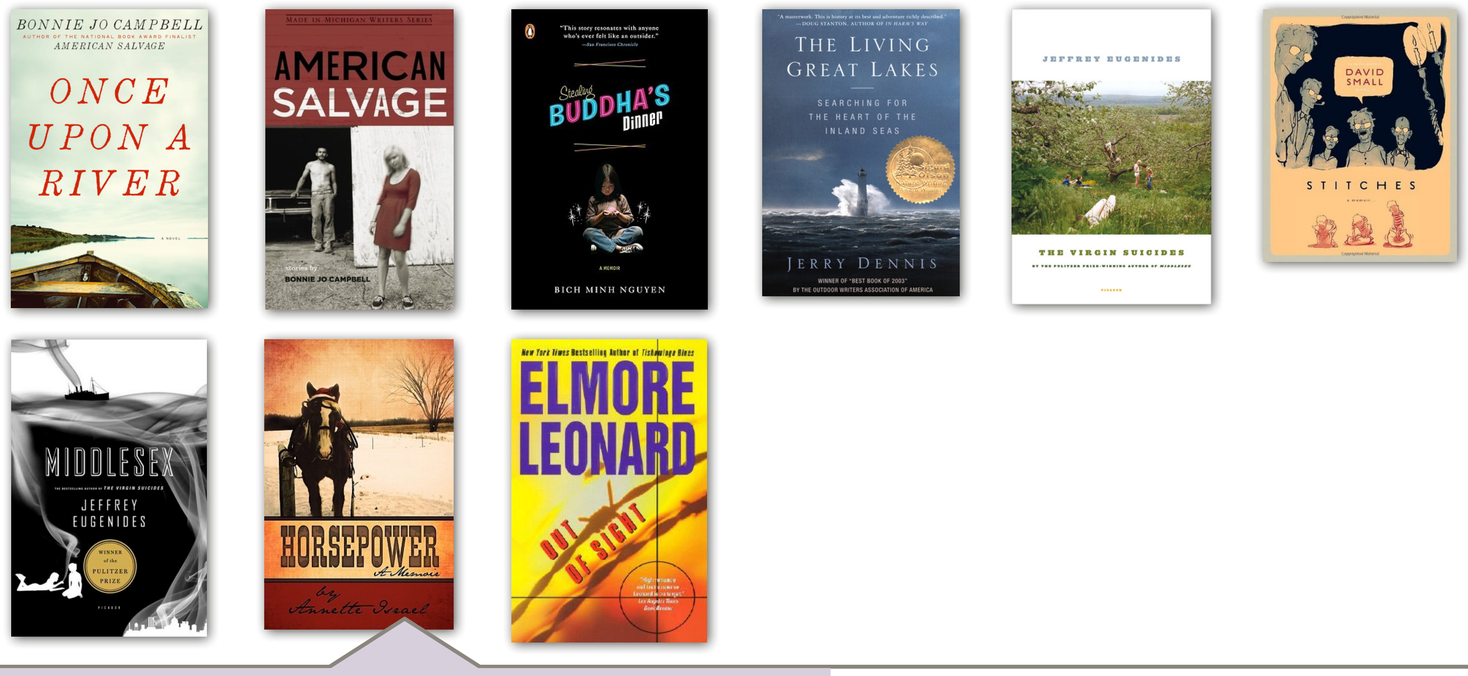
(mini) Book Genome Project

Miriam Boon • Christina Burghard • Noah Liebman • nliebman@u.northwestern.edu

goodreads

Emily Anne's Books: michigan (7) ×



Machines are replacing booksellers

But most machines use collaborative filtering, which tends to suggest books from what is already popular, and based on everything you like rather than what you want now.

Readers enjoy curating book lists

Goodreads is a book-oriented social networking site with over 12 million members who have added over 420 million books to their virtual bookshelves.

A concept-based alternative

We explore directly comparing books to reader-curated virtual shelves based on their librarian-assigned subjects.

How did it go?

While data exploration suggests that an attribute-driven approach is feasible, system performance proved unsatisfactory.

What is a shelf?

Goodreads users organize their books onto "shelves", many of which represent literary concepts such as the one above. Such shelves are unlikely to be complete. **Our goal is to find and suggest missing books.**

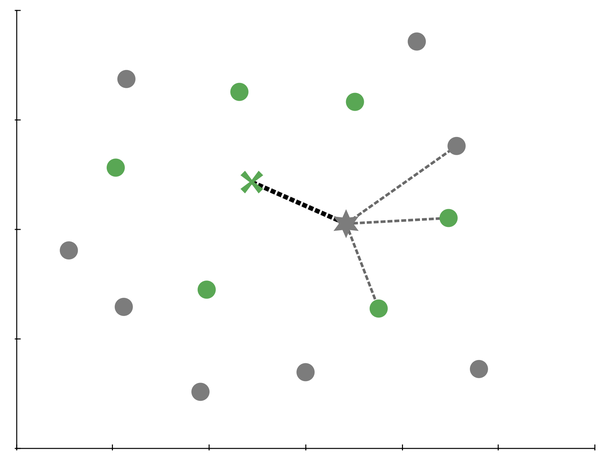
What is a book?

The Library of Congress assigns subject headings to every book. We use these subjects, plus book length, to describe books.

May we suggest…

Recommending books using k-nearest neighbor

1. Randomly pick a book ✕, from the shelf.
2. Find the closest unclassified book, ★.
3. Find the k closest books to the candidate book. Here, $k = 4$.
4. A majority vote of the closest books determines the class of the candidate.
5. If the majority is positive, suggest it!



Measuring distance

237 pages
Fiction
Vampires
Romance
High schools

$$d^2(A,B) = \left(\frac{n-m}{nm}\right)^2 \sum_{\text{common attrs}} w_i^2 + \frac{1}{n^2} \sum_{\text{attrs only in } A} w_i^2 + \frac{1}{m^2} \sum_{\text{attrs only in } B} w_i^2$$

Each attribute has weight w, which is inversely proportional to how frequently that attribute appears in the world.

349 pages
Fiction
Wizards
Magic
High schools

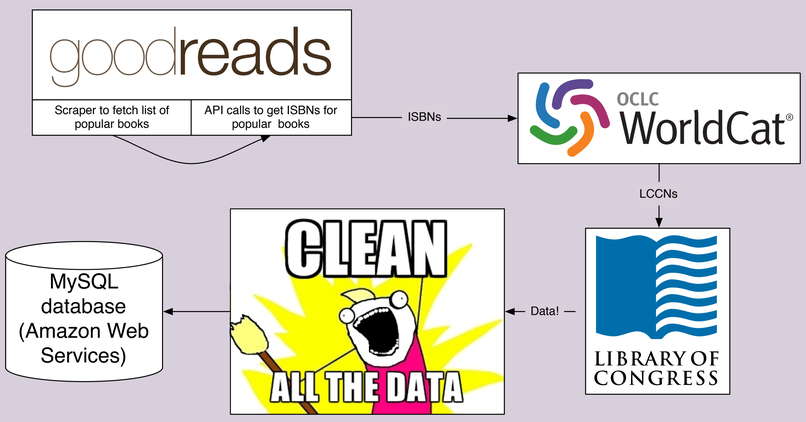
Book A has n attributes

Book B has m attributes

Some data about our data

2,165 books
915 subject heading attributes
9,768 book–attribute relationships
1,220 lines of code written
2,343,612 pre-calculated distances

Where the data came from

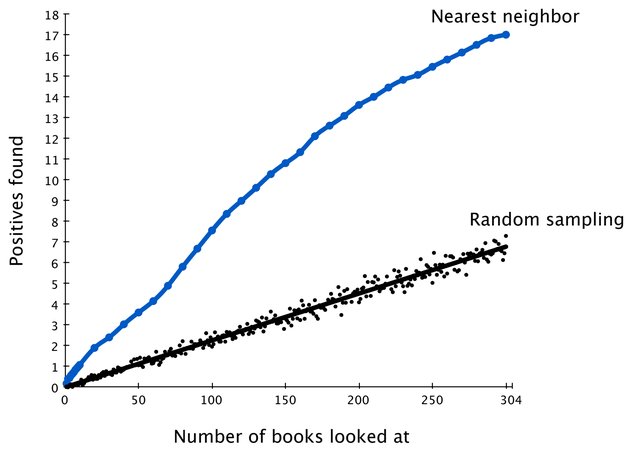


Evaluation

Shelves as book clusters

Books that represent the same concept are near each other in our concept space.

This shows the average number of positive (i.e., on the shelf) books found in a sample of k books. The bottom line was sampled randomly; the top line is from neighbors nearest to shelf members.



Making book suggestions

For $k = 1$, 2, and 3, we make better suggestions than a baseline of suggesting a book at random, but still achieve low recall and low precision.

Out of 41 possible books to correctly suggest, we never suggested more than 11. Conversely, we made many incorrect suggestions.