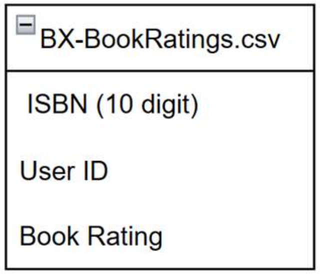
The Bookish Butterfly

Data Sources, Cleaning, and Preprocessing

Raw Data Sources

Book Ratings

Kaggle Book Crossing Dataset
BX-Book-Ratings.csv (1149779 values)
BX-Books.csv (271379 unique values)

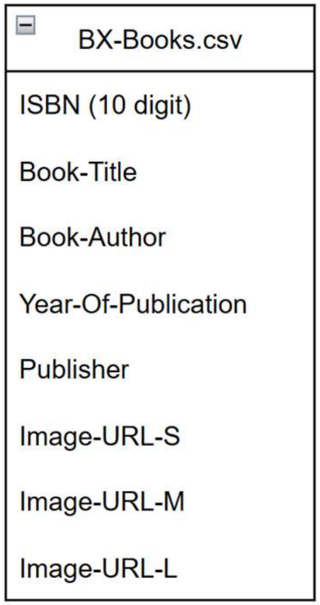


BX-BookRatings.csv

ISBN (10 digit)

User ID

Book Rating



BX-Books.csv

ISBN (10 digit)

Book-Title

Book-Author

Year-Of-Publication

Publisher

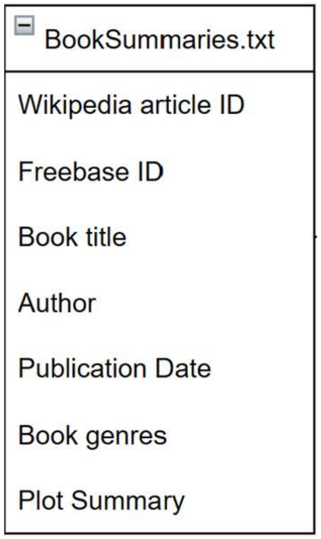
Image-URL-S

Image-URL-M

Image-URL-L

Plot Summaries

Kaggle CMU Book Summary
BookSumaries.txt
16,559 unique values



BookSummaries.txt

Wikipedia article ID

Freebase ID

Book title

Author

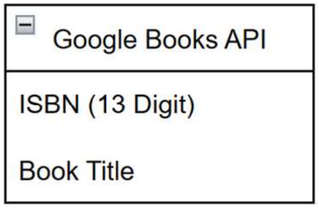
Publication Date

Book genres

Plot Summary

ISBN Matching

Google Books API



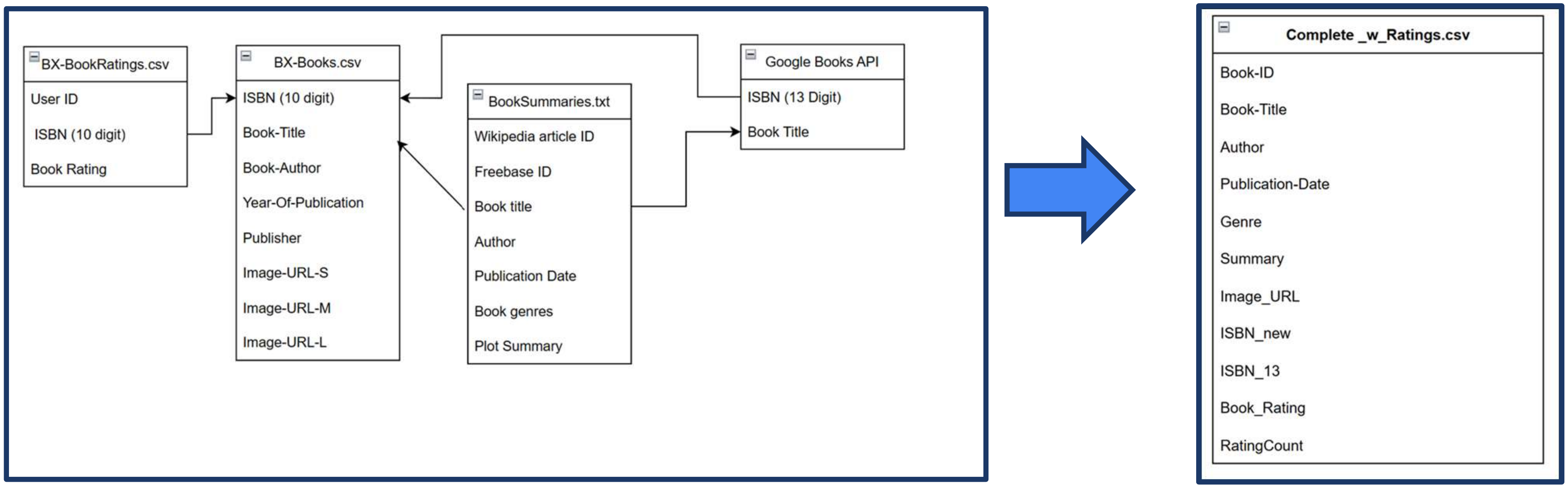
Google Books API

ISBN (13 Digit)

Book Title

Used to augment CMU data with ISBN Numbers for matching with Book Ratings

Data Cleaning and Joining



BX-BookRatings.csv
User ID
ISBN (10 digit)
Book Rating

BX-Books.csv
ISBN (10 digit)
Book-Title
Book-Author
Year-Of-Publication
Publisher
Image-URL-S
Image-URL-M
Image-URL-L

BookSummaries.txt
Wikipedia article ID
Freebase ID
Book title
Author
Publication Date
Book genres
Plot Summary

Google Books API
ISBN (13 Digit)
Book Title

Complete _w_Ratings.csv
Book-ID
Book-Title
Author
Publication-Date
Genre
Summary
Image_URL
ISBN_new
ISBN_13
Book_Rating
RatingCount

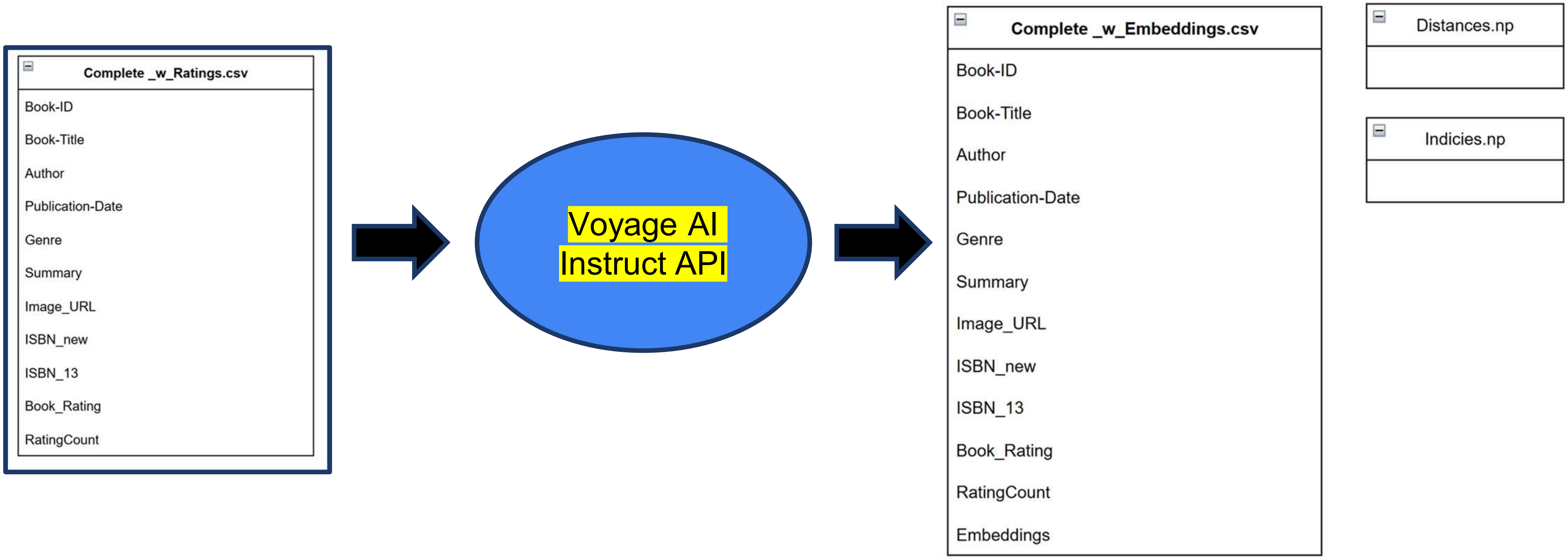
Data Limitations

- Small Number of Books can lead to surprising/frustrating results in app (e.g. no results returned for common book or author)
- BookSummaries.txt had no ISBN field for hard matching to ratings data
  - Used fuzzy/best matching on author field between BX-Books.csv and BookSummaries.txt → **6360** matches
  - Obtained ISBN (13 digit) for **7289** matches using Google API for total of **13649** matches
- Sparse Fields
  - Book genres (**11025** of 13649 books had genre marked)
  - Book Rating (**5115** of 13649 books had ratings)
  - Rating Count is close to 0 for most books

Data Preprocessing - Book Similarity Measures

Voyage-AI-Instruct API

- Ranked #1 in MTEB leaderboards on hugging face in Semantic Textual Similarity(STS)



Complete _w_Ratings.csv

Book-ID
Book-Title
Author
Publication-Date
Genre
Summary
Image_URL
ISBN_new
ISBN_13
Book_Rating
RatingCount

Voyage AI
Instruct API

Complete _w_Embeddings.csv

Book-ID
Book-Title
Author
Publication-Date
Genre
Summary
Image_URL
ISBN_new
ISBN_13
Book_Rating
RatingCount
Embeddings

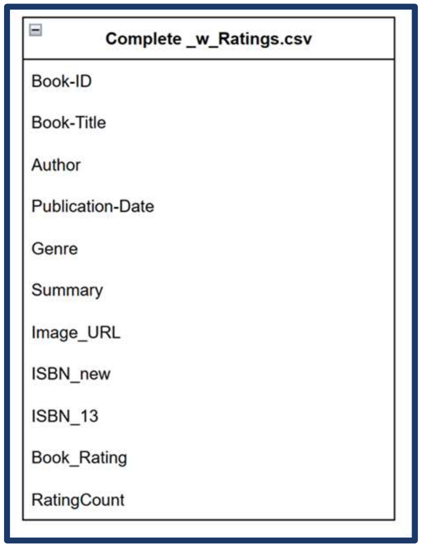
Distances.np

Indicies.np

Data Used In Production

Simple Field Matching

(e.g. Author2 Search)



Complete _w_Ratings.csv

Book-ID

Book-Title

Author

Publication-Date

Genre

Summary

Image_URL

ISBN_new

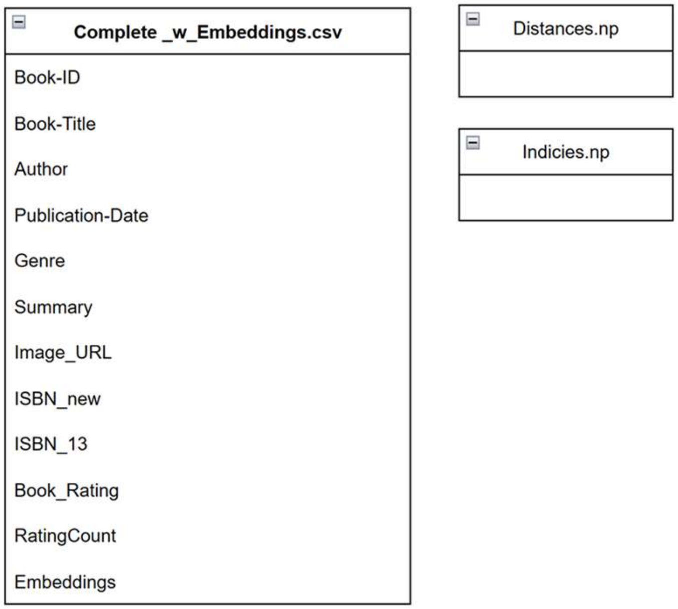
ISBN_13

Book_Rating

RatingCount

Semantic or Semantic + Keyword Search

(resource and memory intensive, only use when necessary)



Complete _w_Embeddings.csv

Book-ID

Book-Title

Author

Publication-Date

Genre

Summary

Image_URL

ISBN_new

ISBN_13

Book_Rating

RatingCount

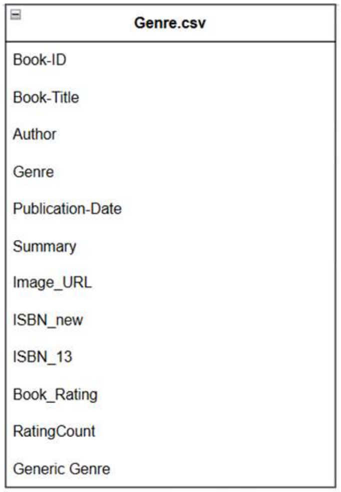
Embeddings

Distances.np

Indicies.np

Search on Standardized Genre

(one book can span multiple Generic Genres and thus appear in multiple rows)



Genre.csv

Book-ID

Book-Title

Author

Genre

Publication-Date

Summary

Image_URL

ISBN_new

ISBN_13

Book_Rating

RatingCount

Generic Genre