Group 53
Diana Freitas,
Diogo Samuel Fernandes,
Juliane Marubayashi

Monitorizado por João Damas

Goodreads books & reviews

PRI - Data Preparation

goodreads

Table of Contents

01. Data Gathering & Quality Assessment

02. Data Cleaning

03. Data Exploration

05. Conceptual Model

06. Pipeline

07. Conclusion

01

Data Gathering & Quality Assessment

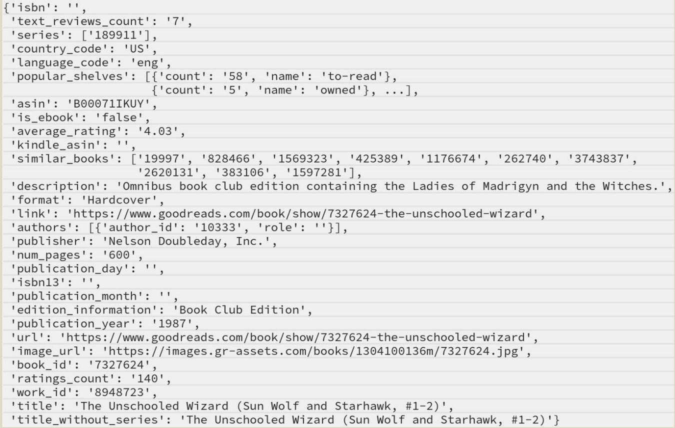
Data Gathering

Theme: Goodreads books & reviews

Datasets: 4 datasets in *json* that could be merged by matching book/author/review ids.

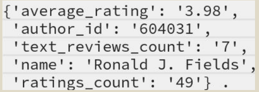
- **Books**: *2.36M* observations with book meta-data;
- **Authors**: 829K observations with details about each author;
- **Genres**: for each of the books, includes the list of genres;
- **Reviews**: reviews with spoilers.

books.json



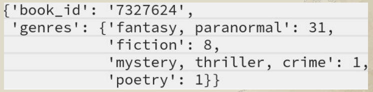
```
{'isbn': '',
 'text_reviews_count': '7',
 'series': ['189911'],
 'country_code': 'US',
 'language_code': 'eng',
 'popular_shelves': [{'count': '58', 'name': 'to-read'},
                     {'count': '5', 'name': 'owned'}, ...],
 'asin': 'B00071IKUY',
 'is_ebook': 'false',
 'average_rating': '4.03',
 'kindle_asin': '',
 'similar_books': ['19997', '828466', '1569323', '425389', '1176674', '262740', '3743837',
                   '2620131', '383106', '1597281'],
 'description': 'Omnibus book club edition containing the Ladies of Madrigyn and the Witches.',
 'format': 'Hardcover',
 'link': 'https://www.goodreads.com/book/show/7327624-the-unschooled-wizard',
 'authors': [{'author_id': '10333', 'role': ''}],
 'publisher': 'Nelson Doubleday, Inc.',
 'num_pages': '600',
 'publication_day': '',
 'isbn13': '',
 'publication_month': '',
 'edition_information': 'Book Club Edition',
 'publication_year': '1987',
 'url': 'https://www.goodreads.com/book/show/7327624-the-unschooled-wizard',
 'image_url': 'https://images.gr-assets.com/books/1304100136m/7327624.jpg',
 'book_id': '7327624',
 'ratings_count': '140',
 'work_id': '8948723',
 'title': 'The Unschooled Wizard (Sun Wolf and Starhawk, #1-2)',
 'title_without_series': 'The Unschooled Wizard (Sun Wolf and Starhawk, #1-2)'}
```

authors.json



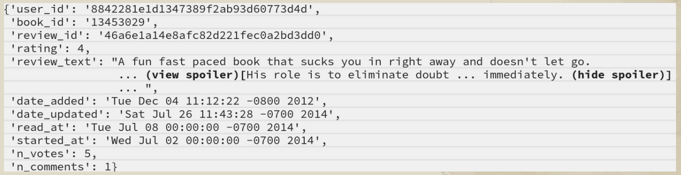
```
{'average_rating': '3.98',
 'author_id': '604031',
 'text_reviews_count': '7',
 'name': 'Ronald J. Fields',
 'ratings_count': '49'} .
```

genres.json



```
{'book_id': '7327624',
 'genres': {'fantasy, paranormal': 31,
            'fiction': 8,
            'mystery, thriller, crime': 1,
            'poetry': 1}}
```

reviews.json



```
{'user_id': '8842281e1d1347389f2ab93d60773d4d',
 'book_id': '13453029',
 'review_id': '46a6e1a14e8afc82d221fec0a2bd3dd0',
 'rating': 4,
 'review_text': "A fun fast paced book that sucks you in right away and doesn't let go.
                 ... (view spoiler)[His role is to eliminate doubt ... immediately. (hide spoiler)]
                 ... ",
 'date_added': 'Tue Dec 04 11:12:22 -0800 2012',
 'date_updated': 'Sat Jul 26 11:43:28 -0700 2014',
 'read_at': 'Tue Jul 08 00:00:00 -0700 2014',
 'started_at': 'Wed Jul 02 00:00:00 -0700 2014',
 'n_votes': 5,
 'n_comments': 1}
```

Data Quality Assessment

How was it assessed?

- Rich in textual data - reviews or book descriptions;
- Identify missing and inconsistent values;
- Validity of the data - type, unique and foreign-key constraints;
- Uniformity and consistency.

What issues were found?

- Date format inconsistency;
- Integer ids represented as strings;
- Missing genres for many books.

02

Data Cleaning

Data Cleaning

For each dataset

1 Books

- Subset of 11K observations;
- Discard books without description;
- Merged publication year, month, day ➜ date;
- Drop :
    - ...
    - references to other datasets;
    - similar_books.

2 Genres

- Subset of 11K observations;
- Drop the number of users that classified the book with each genre;
- Remove observations that represent books without any genre.

3 Authors

- Drop derived attributes (*text_reviews_count* and *ratings_count*);

4 Reviews

- Subset of 100K reviews;
- Drop *user_id* and details related with the management of each review (*date_updated*, *started_at*, *read_at*, …);
- Remove spoiler tags.

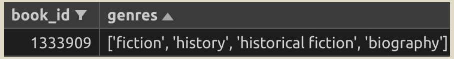
Data Cleaning

Combining the datasets

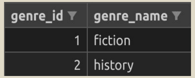
1 Books & Genres

Combine by book_id

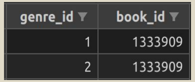
- Drop books that don't have genres;
- Drop genres that don't relate to any book;
- Map genre_name to genre_id;
- Many-to-many relation.



book_id	genres
1333909	['fiction', 'history', 'historical fiction', 'biography']



genre_id	genre_name
1	fiction
2	history

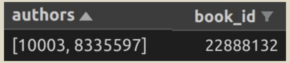


genre_id	book_id
1	1333909
2	1333909

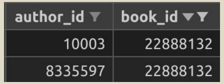
2 Books & Authors

Combine by author_id

- Drop books that don't have genres;
- Drop genres that don't relate to any book;
- Map genre_name to genre_id;
- Many-to-many relation.



authors	book_id
[10003, 8335597]	22888132



author_id	book_id
10003	22888132
8335597	22888132

3 Books & Reviews

Combine by book_id

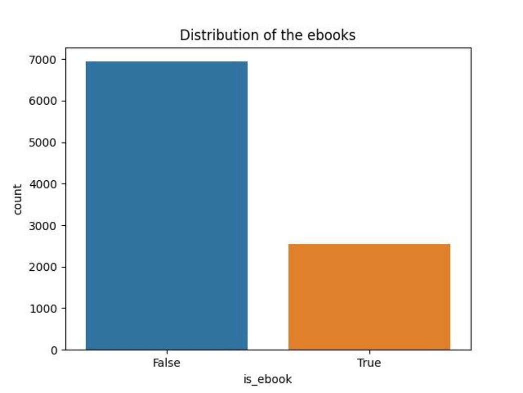
- Drop reviews that don't relate to any book;
- Many-to-one relation.

04

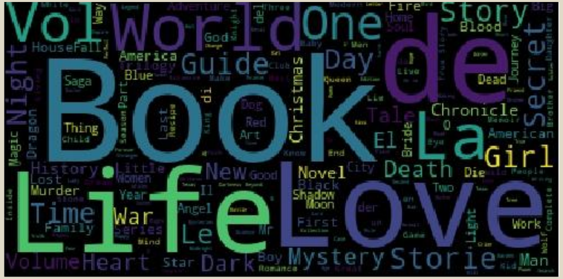
Data Exploration

Books

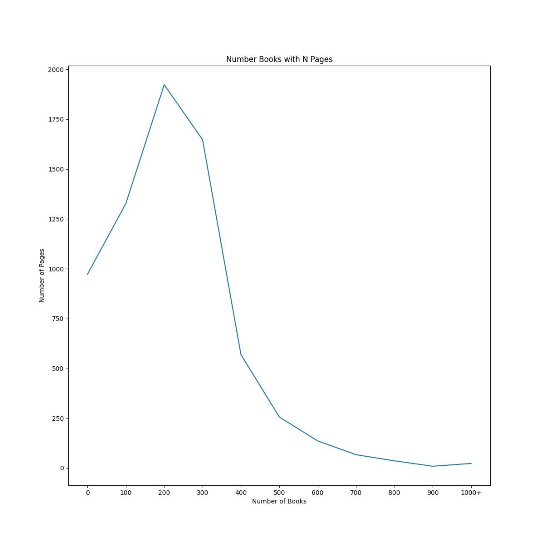
1



2



3



Topics:

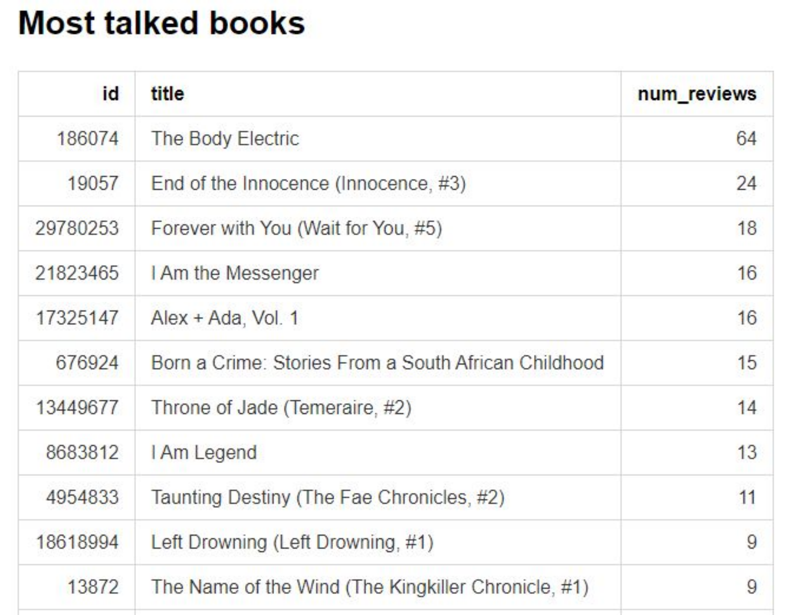
[1] Distribution of the books - ebooks and physical
[2] Books description word cloud
[3] Distribution of the number of pages

Reviews

Topics:

[1] 20 Most talked books
[2] Review text word cloud
[3] Reviews rating distribution

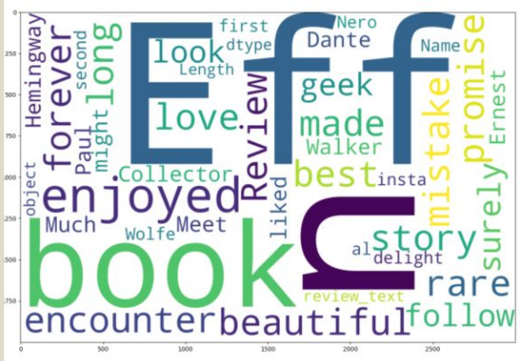
1



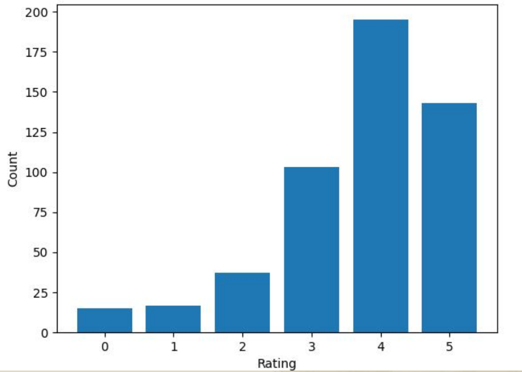
Most talked books

id	title	num_reviews
186074	The Body Electric	64
19057	End of the Innocence (Innocence, #3)	24
29780253	Forever with You (Wait for You, #5)	18
21823465	I Am the Messenger	16
17325147	Alex + Ada, Vol. 1	16
676924	Born a Crime: Stories From a South African Childhood	15
13449677	Throne of Jade (Temeraire, #2)	14
8683812	I Am Legend	13
4954833	Taunting Destiny (The Fae Chronicles, #2)	11
18618994	Left Drowning (Left Drowning, #1)	9
13872	The Name of the Wind (The Kingkiller Chronicle, #1)	9

2



3



Authors

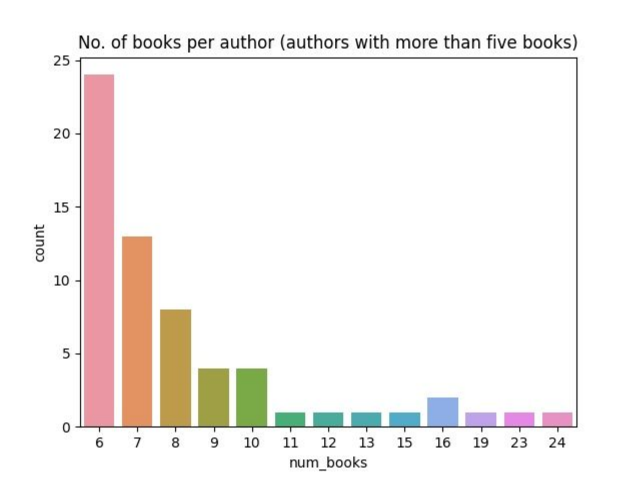
Topics:

[1] No. of books per author

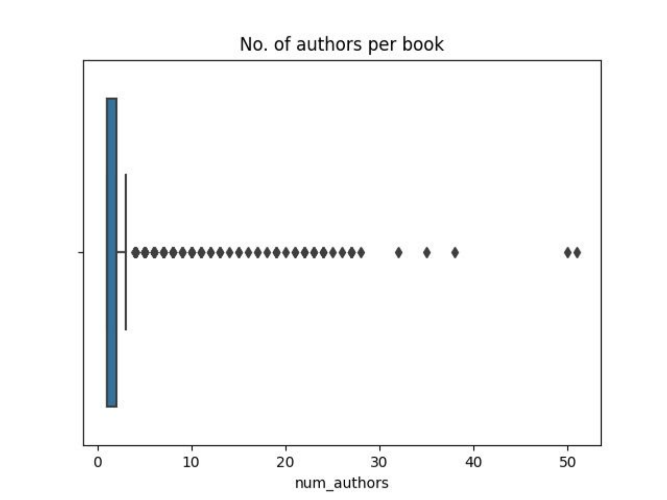
(authors with more than five books)

[2] Number of authors per book (outliers)

1

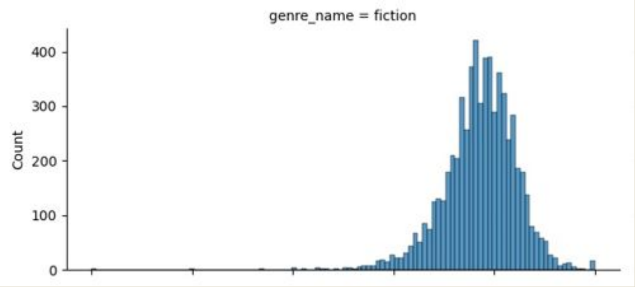


2

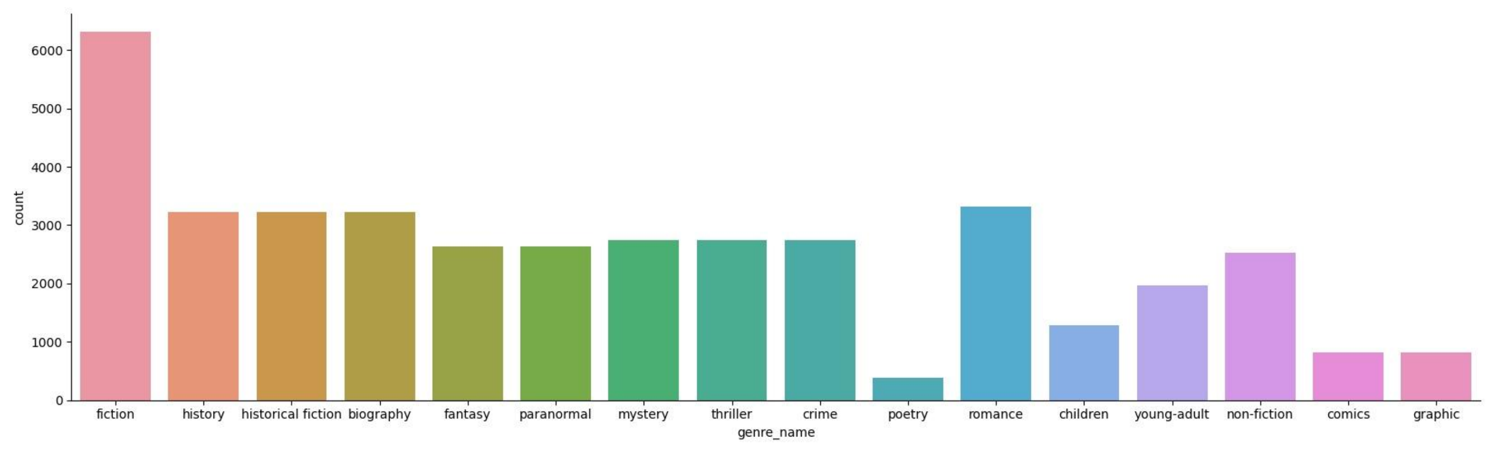


Genres

1



2



Topics:

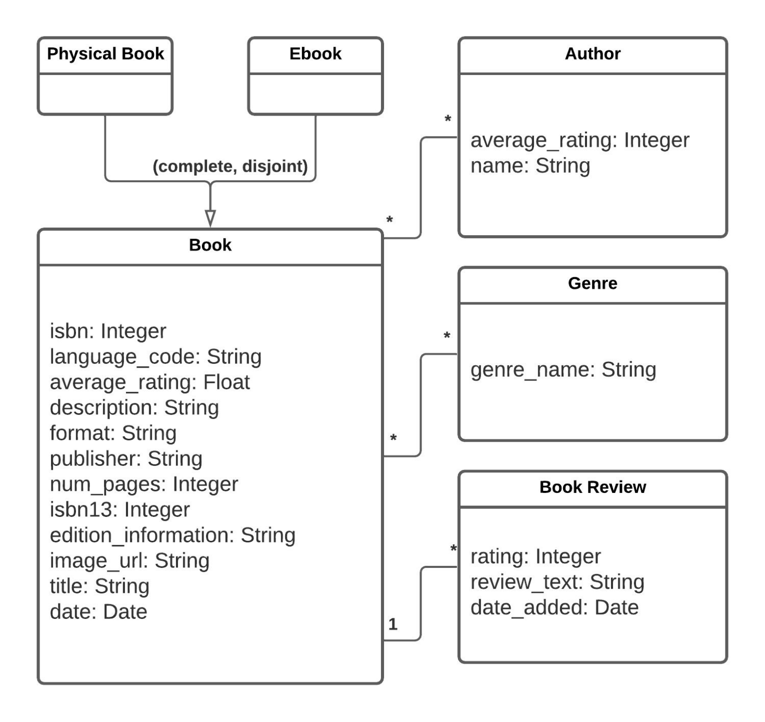
[1] Distribution of ratings in fiction books

[2] Number of books per genre

05

Conceptual Model

Conceptual Model



06

Pipeline

Pipeline

Data Gathering:

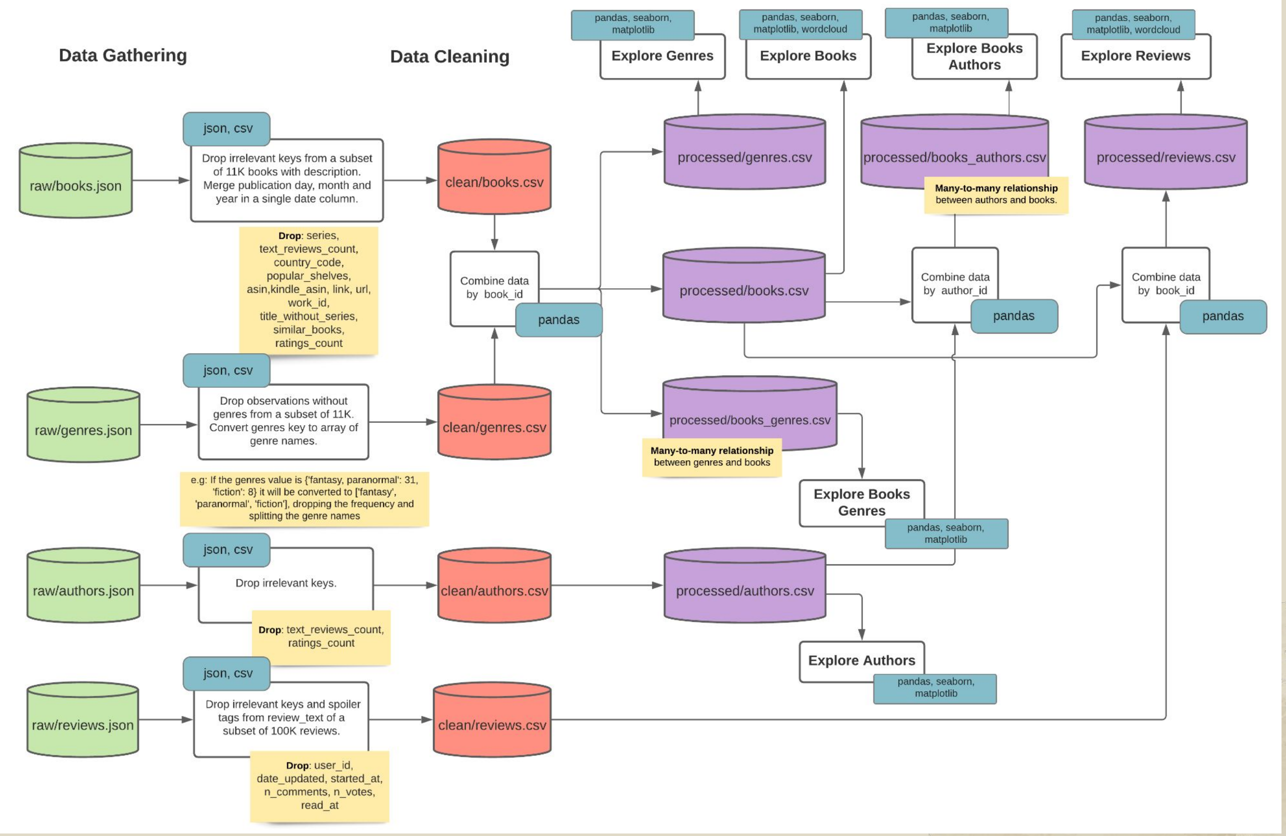
- Original *json* datasets obtained.

Data Cleaning:

1. Modifies the data of each dataset individually.
2. Combines the datasets.

Data Exploration:

- Use of data visualization and statistical techniques to better characterise.



Data Gathering

Data Cleaning

raw/books.json

json, csv

Drop irrelevant keys from a subset of 11K books with description. Merge publication day, month and year in a single date column.

Drop: series, text_reviews_count, country_code, popular_shelves, asin,kindle_asin, link, url, work_id, title_without_series, similar_books, ratings_count

clean/books.csv

raw/genres.json

json, csv

Drop observations without genres from a subset of 11K. Convert genres key to array of genre names.

e.g: If the genres value is {'fantasy, paranormal': 31, 'fiction': 8} it will be converted to ['fantasy', 'paranormal', 'fiction'], dropping the frequency and splitting the genre names

clean/genres.csv

Combine data by book_id

pandas

raw/authors.json

json, csv

Drop irrelevant keys.

Drop: text_reviews_count, ratings_count

clean/authors.csv

raw/reviews.json

json, csv

Drop irrelevant keys and spoiler tags from review_text of a subset of 100K reviews.

Drop: user_id, date_updated, started_at, n_comments, n_votes, read_at

clean/reviews.csv

processed/genres.csv

processed/books.csv

processed/books_genres.csv

Many-to-many relationship between genres and books

processed/authors.csv

processed/books_authors.csv

Many-to-many relationship between authors and books.

Combine data by author_id

pandas

processed/reviews.csv

Combine data by book_id

pandas

Explore Genres

pandas, seaborn, matplotlib

Explore Books

pandas, seaborn, matplotlib, wordcloud

Explore Books Authors

pandas, seaborn, matplotlib

Explore Reviews

pandas, seaborn, matplotlib, wordcloud

Explore Books Genres

pandas, seaborn, matplotlib

Explore Authors

pandas, seaborn, matplotlib

Conclusions

Project Goals

The goals proposed, namely the **Collection**, **Preparation** and **Characterisation** of the datasets, were achieved with success.

Results

The final datasets are free of irrelevant attributes and, when combined, represent the main entities of the Goodreads domain - **Books**, **Authors**, **Reviews** and **Genres**; as well as the relations between them.

Technologies

The technologies that were selected, particularly some Python libraries like **json**, **csv**, **pandas**, **matplotlib** and **seaborn**, had a great impact on the success of this phase, given that they are intuitive and provide methods to deal with most of the Data Processing tasks.

Future Work

As future work, an information retrieval tool will be used on the project's datasets and the data will be explored with free-text queries.

Bibliographic References

1. UCSD Book Graph. https://sites.google.com/eng.ucsd.edu/ucsdbookgraph/books
2. Mengting Wan, Julian McAuley, "Item Recommendation on Monotonic Behavior Chains", in RecSys'18.
3. Mengting Wan, Rishabh Misra, Ndapa Nakashole, Julian McAuley, "Fine-Grained Spoiler Detection from Large-Scale Review Corpora", in ACL'19.