Group Name: Diverse Romance Novels

Group Members: Cristyn Filla

Topic of Curation Protocol: Diverse romance novels in English

Target Audience: US romance readers

Statement of Goals:

The goal of this project is to create a curated database on GitHub of diverse romance novels in English. There is limited to no datasets on common open data repositories or GitHub for this type of data. I would like to curate and develop a protocol for this data so that it is more accessible to romance readers trying to navigate the overly white cis romance novel market to find diverse titles. I plan to scrape the GoodReads list linked in the relevant data to obtain the data for this project. There are a number of guides and scripts written for how to perform this process in Python.

I would like to learn more about data curation with book data as well as how to scrape websites for data to forward data collection/management work.

Link to Relevant Data: https://www.goodreads.com/shelf/show/diverse-romance

Potential User Community

This repository is meant to fill a niche gap in the book world and publishing industry and has four user groups that the repository creator envisions interacting with the repository both as contributors and as data users. These three groups are:

- Readers
    - Primarily romance readers.
    - Readers in the United States or interested in US genre fiction or any English language reader.
- Publishers
    - Primarily publishers in the genre fiction and/or romance novel sub-industry.
        - i.e. Avon, Berkeley, Zebra
- Researchers
    - Primarily researchers interested in diversity in literature and/or publishing, in genre fiction issues, in romance novel issues, or in DEI more broadly.
- Librarians
    - Primarily public librarians, though school, private, and academic librarians are also considered potential users for this repository.

User Stories

The segments of the user community all have different user stories that describe their need for the data in the repository, however they share one common story in why they contribute to the repository.

This contribution story is centered on accessibility. The user contributing to the collection has a desire to make the data on diverse romance novels more accessible to other, no matter their role in the user community. The repository has been built to close the accessibility (particularly the easy accessibility) of this type of data and those looking to contribute data to the collection are looking to provide greater accessibility to the user community for these diverse novels that is not being provided elsewhere.

The user stories for the segments of the user community looking to access and use the data in the repository are more disparate. These user stories are described below:

User Segment	Goal	User Story
Reader	Find more diverse romance novels for personal reading	“As an avid romance reader I am finding an over-representation of white authors and characters and I would like to expand my reading to more diverse books but finding a guide that is longer than a 20 book recommendation list is hard. I would like to see a long list of what is available where I can find books that have tropes I like as well as being diverse.”
Publisher	Find data about competitors and the industry	“As an editor at Avon I want to see what other publishers are publishing in the diverse romance space. Knowing the quantity per year of diverse romance novels others are publishing, the authors, and the subgenres or tags will help me get a sense for the field and if I am acquiring on a similar cadence and style for this subset of romance novels.”
Researcher	Find high level data on diverse romance publishing	“As a researcher studying diversity in publishing today I want to gather more data on what publishers are publishing diverse romance and look for trends across publisher and time. I will compare this data against other datasets that focus on diverse publishing in other fiction genres.”
Librarian	Find data to build professional repertoire	“As a reader advisory and circulation librarian I am looking to have more resources to provide better reader advisory to patrons looking for romance novels. I am conscious of DEI issues

		within publishing and would like to have a reliable list of diverse romance that I can use as a tool in my advisory role.”

Through these stories the importance of trope/theme/tag was brought to the forefront. For publishers, readers, and librarians, knowing not only if a title is diverse but also what the subgenre of that title is is very important. Given the discussions circulating in the romance reader spaces this is to be expected. Fortunately, the genre has a relatively standardized set of tropes/subgenres that can be used to complete metadata fields in this repository in order to meet this shared need.

Deposit Criteria

Contents

- All submissions must be datasets pertaining to romance novels published by BIPOC or LGBTQ+ authors.
- All submissions must be datasets in English pertaining to romance novels published in English.
- All submissions must contain at a minimum metadata on author name, title, publisher, published year, and subgenre/tropes. See the required metadata table below and the Metadata and Standards file for more information.
- All submissions must include files that follow the naming conventions laid out in the Data Governance section of this protocol.

Type

- Submissions may contain quantitative or qualitative data, so long as they meet the content criteria.
- Submissions must contain data recorded in alphanumeric character strings.

Formats

- The repository accepts both tabular and text file data submissions.
- Non-proprietary, openly-documented formats are highly recommended and encouraged.
    - i.e. JSON, XML, etc for text files, .csv for tabular files
- Submissions must be 1GB or less in size.

Licensing

- The depositor may only submit data that they know they are at liberty to share on an open access platform.
- Once data is shared with the repository, the repository has the license to publish the data in the repository’s standard dissemination information package (DIP) to anyone who requests the data. There will be no barriers to data access (data will be open access).

Metadata Label	Data Definition
author_name	Name of the primary author of the book
book_title	Title of the book
pub_date	Year the book was originally published
publisher	Name of the original publisher
genres_tags	Subgenre name and relevant theme/trope tags
diversity_label	BIPOC Author, BIPOC MC, LGBTQ+ Author, LGBTQ+ MC are the four options for this data. (MC = Main Character)

Metadata and Data Standards for Submission

This repository accepts submissions for book data only. Therefore, the book schema for linked data (https://schema.org/Book) can be used to meet most of the metadata and data requirements for this repository. This repository does not require all elements of this schema to be included. Instead the following property must be included in each submission, other included properties are not required but are accepted.

Repository Label	Book Schema Property
author_name	author
book_title	name
pub_date	datePublished
publisher	publisher and/or publisherImprint
genres_tags	genre and/or keywords

The repository requires one additional piece of metadata to be completed by the depositor. Each title within the data set should indicate a diversity label that makes the title appropriate for inclusion in the repository. These labels are not mutually exclusive. Each title can include any combination of these labels and as many as are applicable to the book, but each title must include at least one. These labels are as follows:

Diversity Label	Label Definition
BIPOC Author	Author of the book identifies as BIPOC
BIPOC MC	Main character of the book identifies as BIPOC
LGBTQ+ Author	Author of the book identifies as LGBTQ+
LGBTQ+ MC	Main character of the book identifies as LGBTQ+

Rules Governing Data Quality, Preservation, and Transfer

Datasets received into this repository should follow the format guidelines for Data Deposit. Once submitted into the collection all data will be stored in .csv files. Any transformation required to convert the data into .csv from the datasets received in the submission information package (SIP) will be undertaken by the data curators for the repository. The repository will also undertake processes to normalize the data within the SIP. The depositor is only required to ensure data existence and accuracy for the metadata points required by the repository (see Metadata and Standards). Normalization will include:

1) Listing author as [last name],[first name]
2) Listing pub_date as [YEAR] and removing any day or month data
3) Standardizing publisher name based on publishing industry standards
    a) I.e. Simon & Schuster will be normalized to “Simon & Schuster” from any other representation such as “SS”, “S&S” or “Simon and Schuster”

Any data provided by the depositor outside of the required metadata fields will be included as “Additional Information” for each observation that additional data is provided for.

Given that book data is inherently not sensitive there is no cleaning of sensitive data out of the datasets submitted to the repository.

Datasets submitted to the repository will be stored and preserved by the repository in the format that it is submitted in the SIP. When data is accessed through the DIP, however, the data across all datasets stored in the repository will be collated into a single .csv file with no depositor information tied to the observations.