CS432/532: Final Project Report

Project Title: Extraction and Analysis of **Goodreads Books Collection** to gain insights about the data and correlation between different fields.

Team Member(s): Diksha Aswal, Spoorthi Sathya Narayana Murthy

I. N+1 NOSQL QUERIES

Preprocessing: We have pre-processed genre and publication date fields to extract the necessary information for the query. The field 'genre' was a string of words which described different genres that a book belonged to. We converted it into a genre array so that it is convenient for us to use it in the queries. Similarly we have extracted decade and publication year from the publication date field.

Query 1: No. of books published and their rating trend in every decade for different genres. From this data, we have first calculated the number of books published per decade in the top seven genres using a non tabular column 'genre' and then find the trend, how the quality of the books in a particular genre has changed over time(decade by decade). This will help us to know whether new authors are able to match the standard of writing quality which has been set by the experienced authors, using the columns.
Column : publication_date, rating_score, num_ratings, genre,title/isbn
Unstructured column : genre, title

Description: We have deduced the quality of the genre by taking the product of average rating and average number of ratings that the book has got in each decade.

Query 2: How interested are people(through current readers and want to read fields) in particular genres in every decade. This will help us to analyze how the popularity of the different book genres changed with time(decade by decade), among the readers, using the columns: current_readers, want_to_read.
Column : publication_date, current_readers, want_to_read
Unstructured column : genre

Description: We have deduced the popularity of the genre by taking the sum of the average number of current readers and average number of people who want to read the book.

Query 3: In this query we have tried to find the relation between the length and quality of the title and description with people's interest(popularity of the book) towards the book, in different decades of time.
For example: whether the positive and negative sentiment of the description affects the reader's decision to choose a book.
Column : current_readers, want_to_read, publication_date
Unstructured Column: title, description

Description: Popularity of the book here means the current_readers, want_to_read attributes count. We calculated the **Pearson correlation coefficient** between title length and popularity metrics. The statistics used are average and median which we wrote from scratch. We also tagged the description positive, negative or neutral **sentiment** on the basis of frequency of positive and negative words present. Additionally, We tried to find if there is any relationship between popularity metrics and the sentiment of the description.

II. NoSQL Database and Dataset

NoSQL Database: MongoDB

Dataset: Top Goodreads Books Collection (1980-2023) https://www.kaggle.com/datasets/cristaliss/ultimate-book-collection-top-100-books-up-to-2023

There are a total of 4400 book records and 20 attributes for each book.

III. Project Outcome

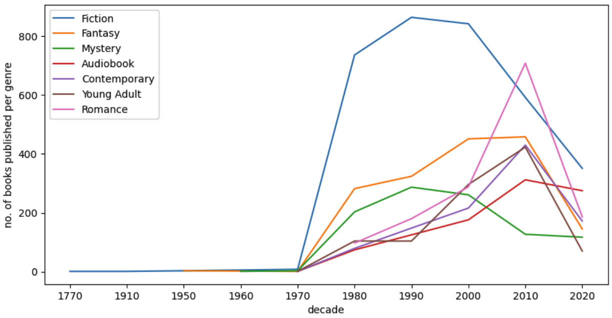


Figure 1: This graph is part of the first query which shows the number of books published in each of the top seven genres vs decade. As we can see there are more books which belong to the genre ‘Fiction’ till 1990 and there is a sudden decrease after 2000. Similarly we can also observe a steady increase in the number of audio books released.

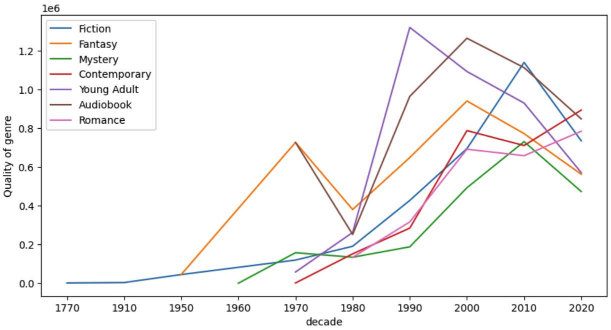


Figure 2: In this image we can see the quality of each of the top seven genres over a period of time decade-wise. The attributes rating and number of ratings for the book indirectly represent the quality of the genre. We have considered the product of the average of these fields as quality. It is observed that quality of books in genres like contemporary and Romance has steadily increased over the years, on the other hand quality of books in genres like audio books and young adult books has taken a hit after 2000.

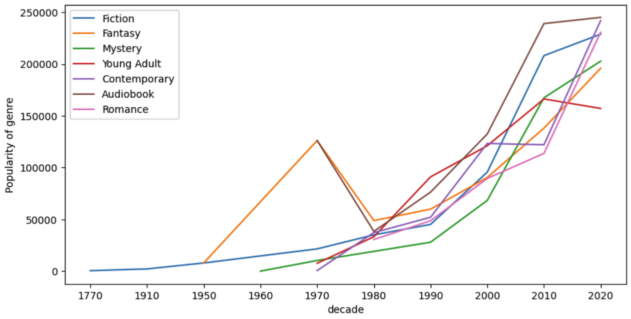


Figure 3: The graph helps us to understand the popularity of top seven genres over a period of time decade-wise. Overall the popularity of all the genres has been steadily increased, this shows that people are reading and willing to read more books over time. One special observation is that the popularity of audiobooks has increased and it has not seen a big decline even though the we don't have much data after 2023.

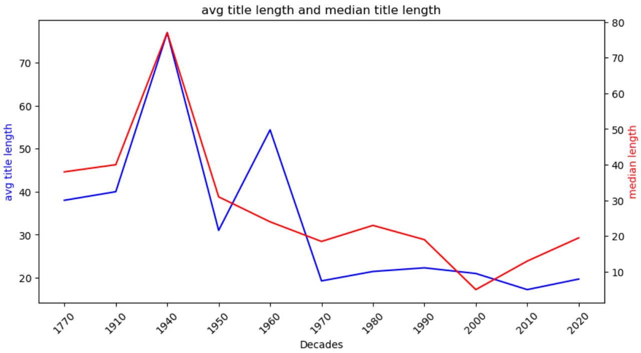


Figure 4: The graph reveals a recent trend among authors towards shorter, crisp titles. Discrepancies between the average and median values, particularly notable in the 1960s, suggest that a few books had notably lengthy titles during that period.

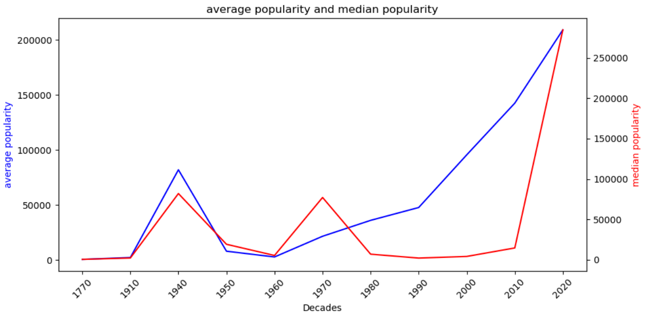
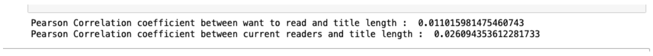


Figure 5: Over time, newer books tend to be more popular among readers. The high average popularity post-1980 suggests certain books have earned significant followings, surpassing those published earlier.

Figures 4 and 5 suggest a potential correlation between book title length and popularity, indicating a preference for shorter titles among readers. To investigate this hypothesis, we computed the correlation between title length and popularity metrics.



```
Pearson Correlation coefficient between want to read and title length :  0.011015981475460743
Pearson Correlation coefficient between current readers and title length :  0.026094353612281733
```

The correlation coefficient revealed no significant direct correlation between title length and popularity metrics. The observed relationship in Figures 4 and 5 may be influenced by attributes such as genre and decade, suggesting an indirect relationship between title length and popularity metrics through these factors.

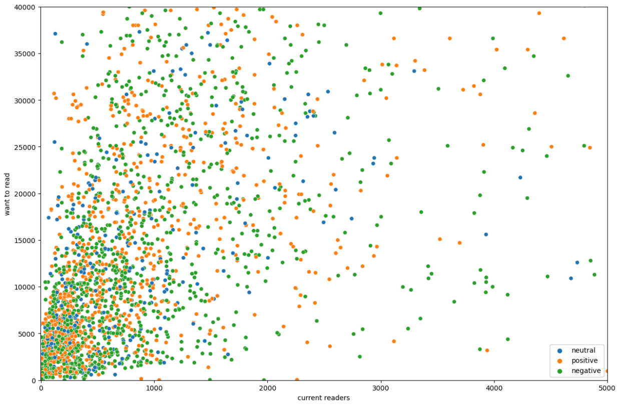


Figure 6: This graph is about the popularity metrics for different kinds of descriptions. It shows that descriptions with negative sentiments are more than the positive ones. This is because the top 2 genres are Fantasy and Fiction, the words in these two genres are strong enough to make the whole sentiment of description negative. Negative descriptions correlate with lower interest from readers, particularly in want_to_read and current_readers. Positive descriptions are more evenly distributed, indicating readers' preference for books evoking positive sentiments.

References

[1] https://www.kaggle.com/datasets/cristaliss/ultimate-book-collection-top-100-books-up-to-2023

[2] https://www.mongodb.com/docs/

[3] https://matplotlib.org/

[4] https://seaborn.pydata.org/

[5] https://pymongo.readthedocs.io/en/sta