Deliverable #2 - EDA

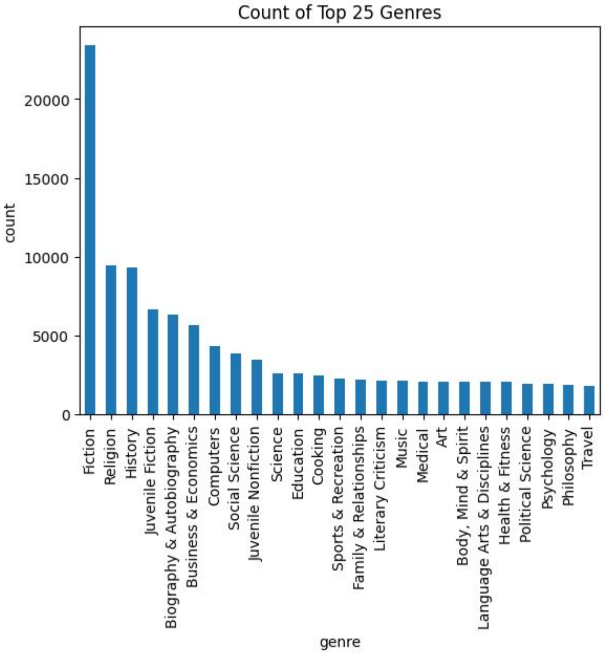
Spring 24 COMS W4995 AML Group 11

1. Introduction

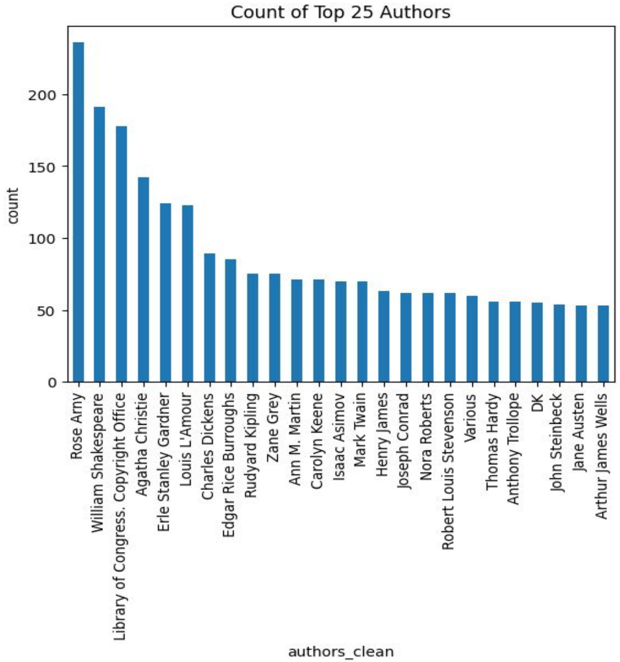
1. **Data Exploration**: EDA for some important features from the reviews/books details dataset and their insights/explanations
2. **Cleaning and sampling**: Data Cleaning/Merging for the ML models to be implemented
3. **Ready for the Modeling**: Machine Learning techniques to be implemented

2. Insights from Data Exploration

Bar plot of different features



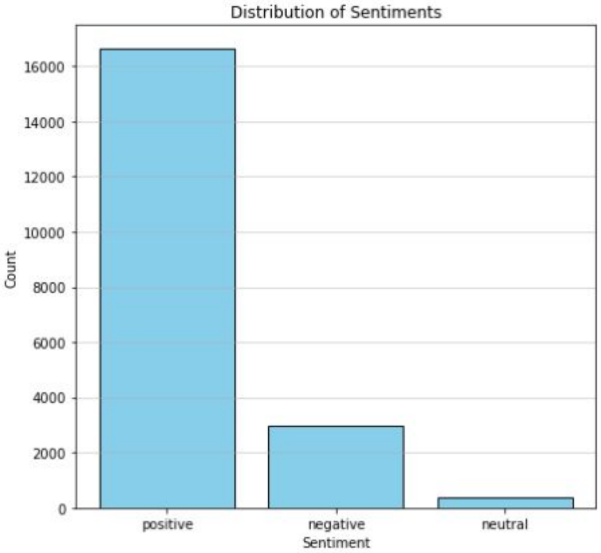
Fiction is the genre with the greatest count of masterpieces.

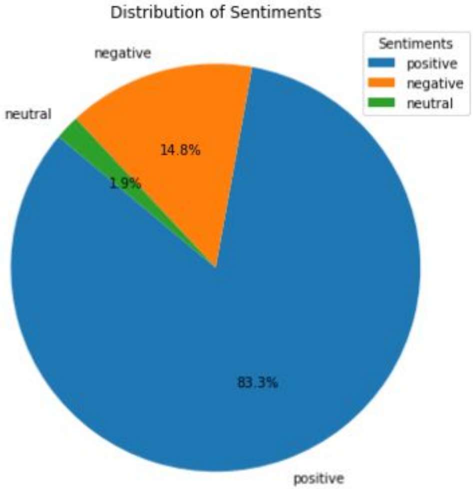


Rose Amy is the author with greatest count. Besides the top 6 authors, the count for each author becomes similar

2. Insights from Data Exploration

Rating Sentiment Bar Plot & Pie Chart



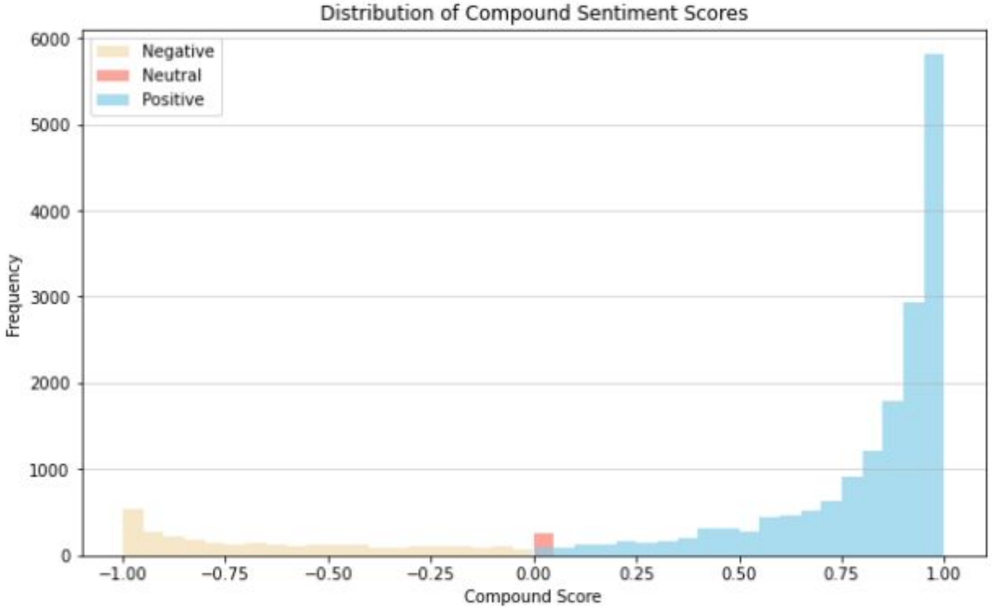


It shows that most of the texts are positive and only a few of the texts is neutral or negative.

The sentiment scores are calculated using the vaderSentiment package

2. Insights from Data Exploration

Rating Sentiment Histogram

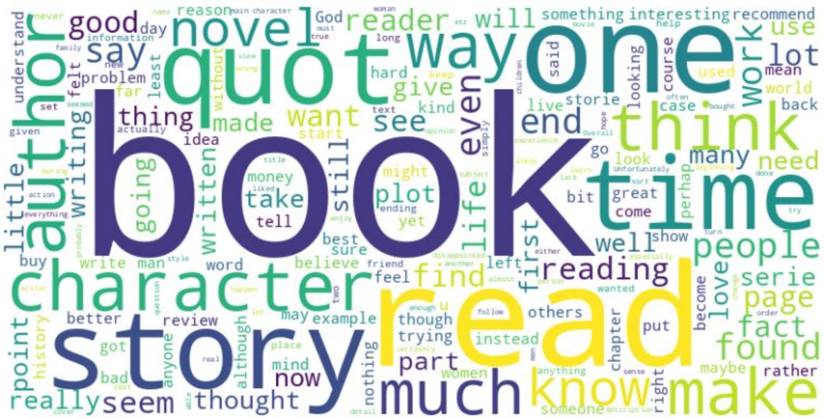


The histogram further shows that most of the texts is positive, and many of them has high positive score. Notice that there is only a few neutral texts because the compound score can hardly be close to 0.

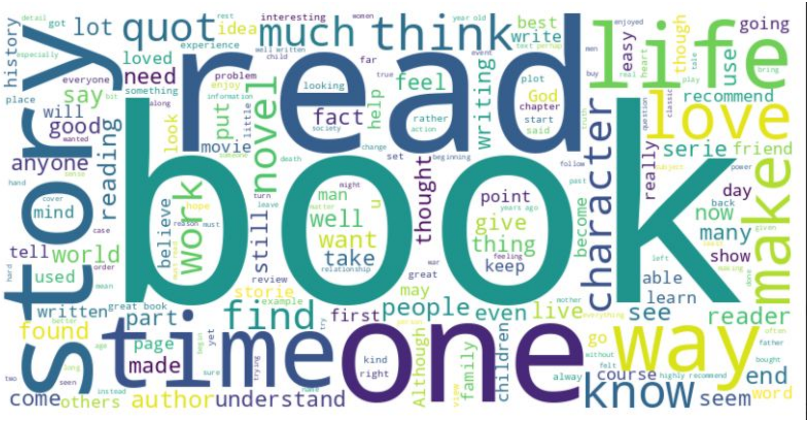
2. Insights from Data Exploration

Word Cloud for text with different reviews scores.

Reviews with score <= 3



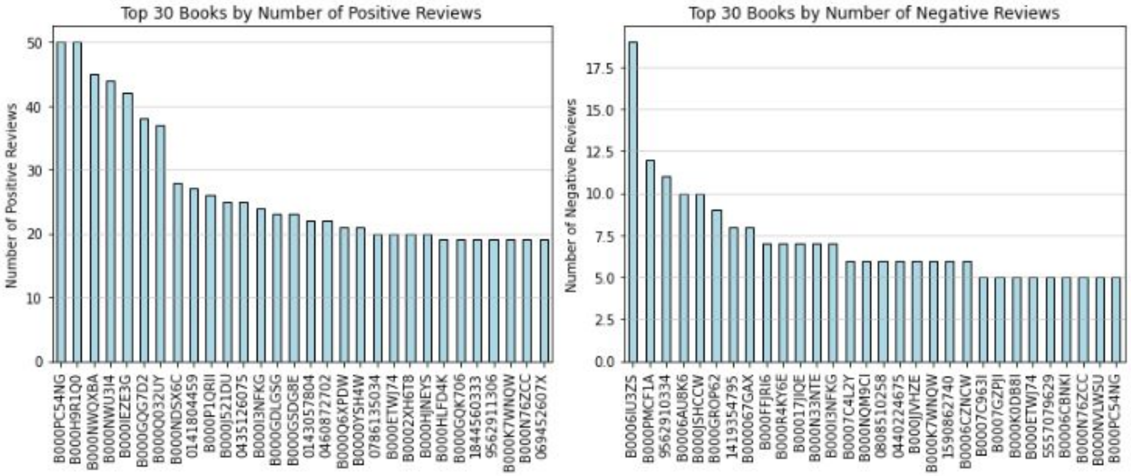
Reviews with score > 3



It seems like there is no big difference between words frequency for review score less or equal to 3 and greater than 3.

2. Insights from Data Exploration

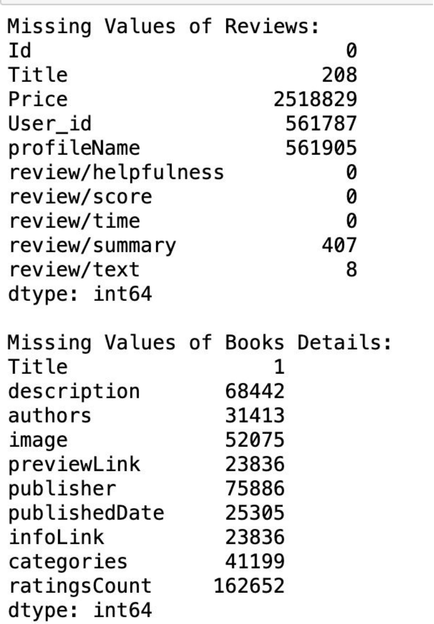
Bar plot of number of reviews of books slicing by sentiment



Horizontal axis represents the Id of different books.
B00PC54NG is the book with the greatest number of positive reviews.
B0006IU3ZS is the book with the lowest number of negative reviews.

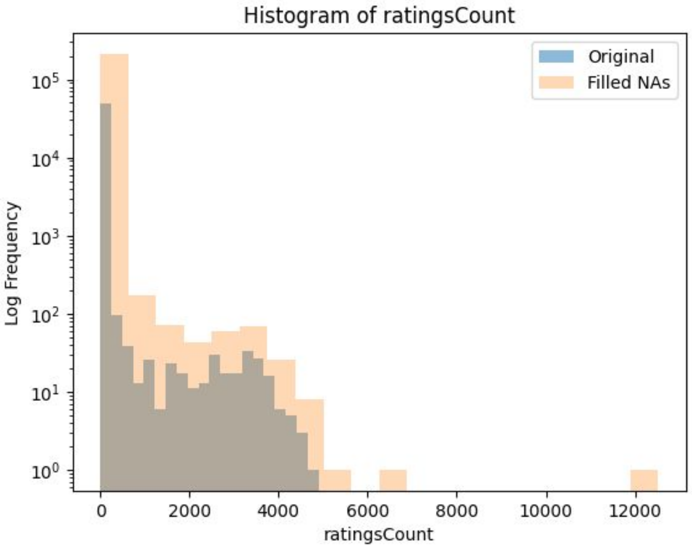
3. Cleaning & Sampling

Missing Values of two datasets



```
Missing Values of Reviews:
Id                          0
Title                     208
Price                 2518829
User_id                561787
profileName            561905
review/helpfulness          0
review/score                0
review/time                 0
review/summary            407
review/text                 8
dtype: int64

Missing Values of Books Details:
Title                 1
description       68442
authors           31413
image             52075
previewLink       23836
publisher         75886
publishedDate     25305
infoLink          23836
categories        41199
ratingsCount     162652
dtype: int64
```



76.5% of books do not have a value in the books dataset.

3. Cleaning Dataset for the Recommendation System

< Cleaning *Reviews* Dataset >

1. Drop rows with missing ‘Title’ and ‘user_id’ in *reviews* dataset

   (‘Title’ and ‘user_id’ are prerequisite for the recommendation system)

2. Drop ‘profileName’ and ‘Price’ columns that have no related information for the recommendation system
3. Fill missing review text values with NA
4. Convert ‘review/helpfulness’ from text to numeric

< Cleaning *Books Details* Dataset >

5. Drop rows with missing ‘Title’ of *books* dataset

   (‘Title’ is prerequisite for the recommendation)

6. Drop unrelated columns for the recommendation system
7. Fill missing ratingsCount using the median

Proposed ML Technique

- Content-Based Filtering: From variables such as categories, authors, and titles to recommend similar books, implement TF-IDF (Term Frequency-Inverse Document Frequency) vectorization with cosine similarity.

- Collaborative Filtering: Implement Singular Value Decomposition (SVD) or memory-based approaches like KNN (k-nearest neighbors) from user-item interaction matrices.