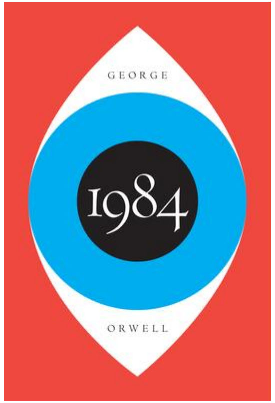
BOOK RECOMMENDATION SYSTEM

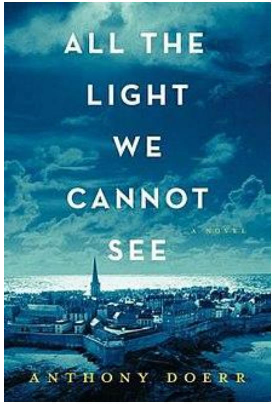
NATURAL LANGUAGE PROCESSING CLASSIFICATION

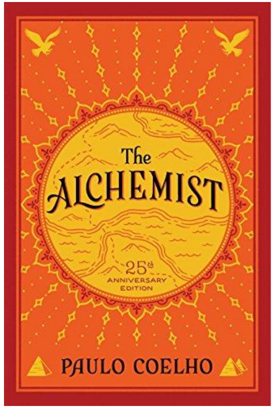
Lisa Liang | DSI Capstone

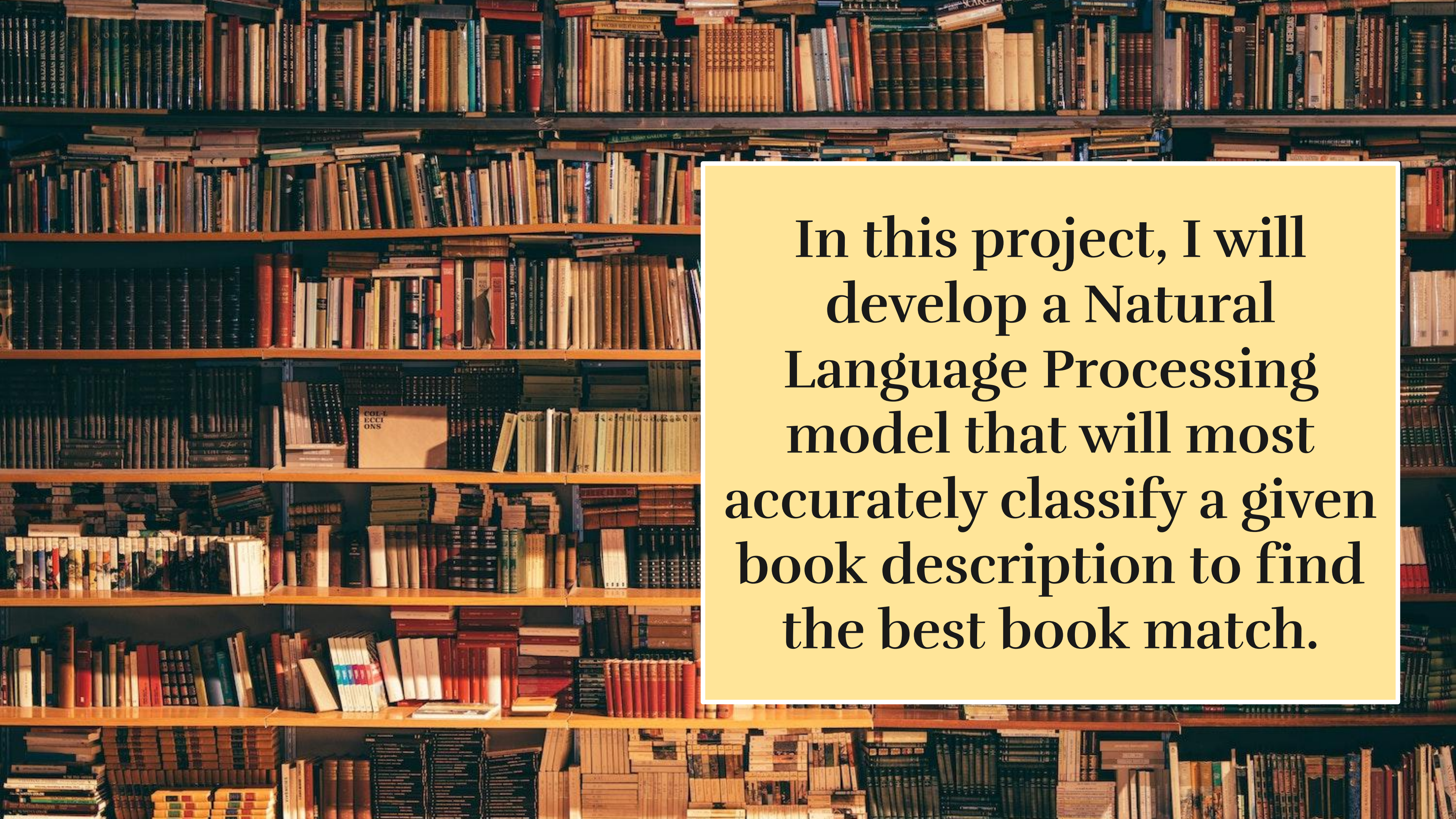
Do You Enjoy Reading?

- Between 2016-2021, there was a decrease in average books read (approximately 15.6 to 12.6 books)
- Bigger decline in college graduates, dropping from 21.1 books to 14.6



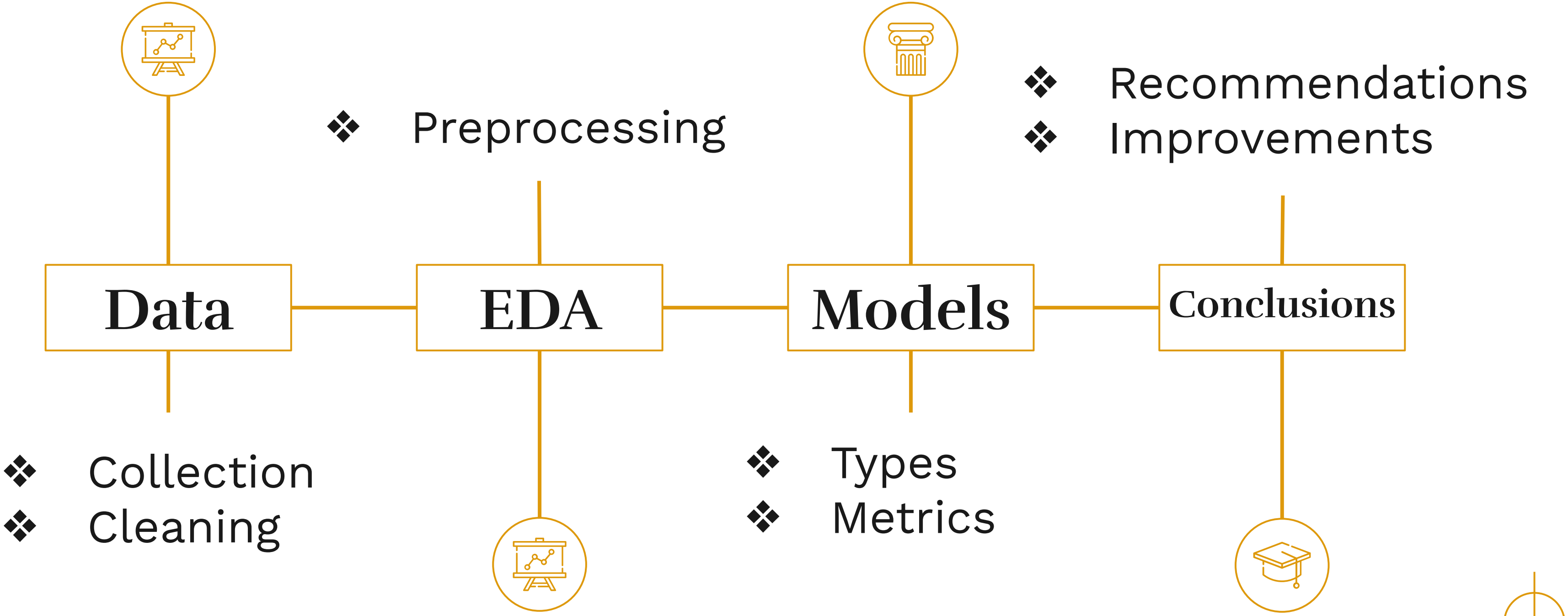






In this project, I will develop a Natural Language Processing model that will most accurately classify a given book description to find the best book match.

PROCESS



Data

- ❖ Collection
- ❖ Cleaning

EDA

- ❖ Preprocessing

Models

- ❖ Types
- ❖ Metrics

Conclusions

- ❖ Recommendations
- ❖ Improvements

DATASET DESCRIPTION

Amazon Book Reviews

- 3M rows
- Reviews posted between May 1996 - July 2014
- Columns: Text review, scores

Google Book Details

- 200K rows
- Corresponding to unique books in Amazon Book Reviews
- Columns: Description, authors, images

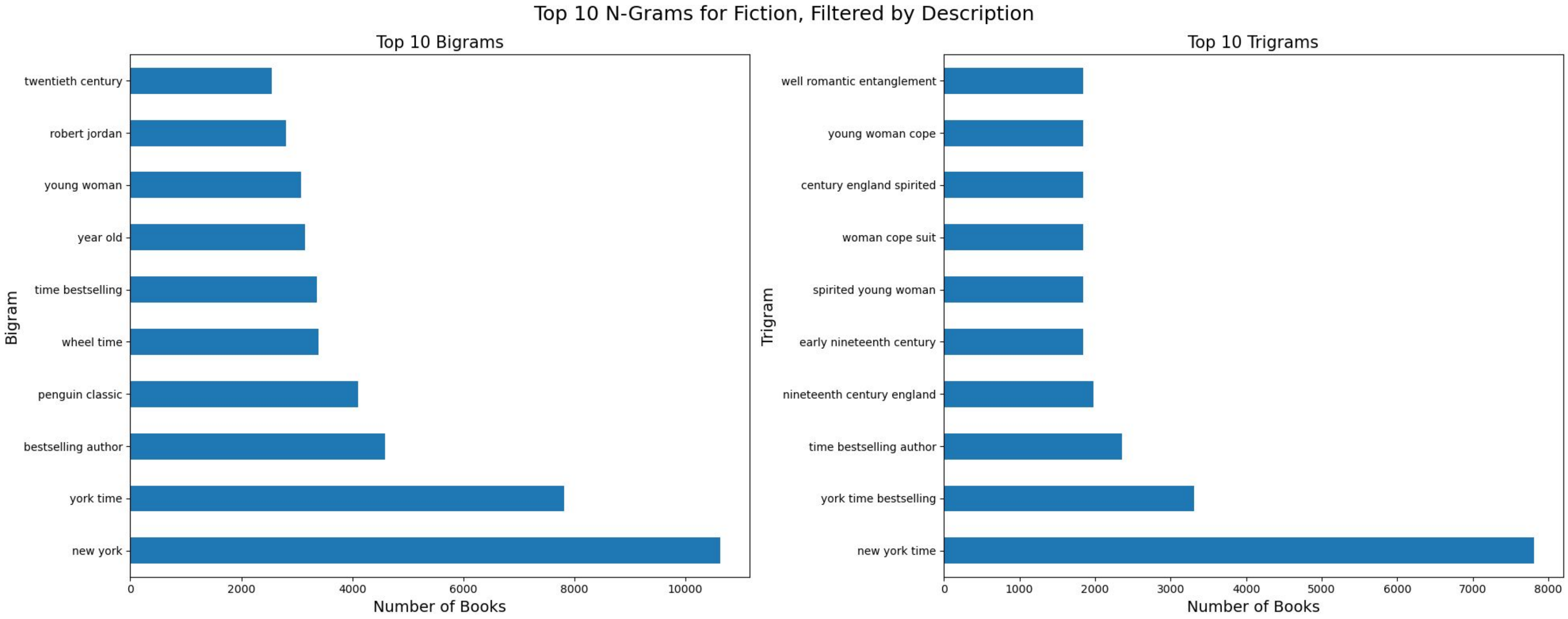
DATA CLEANING

Cleaning Process

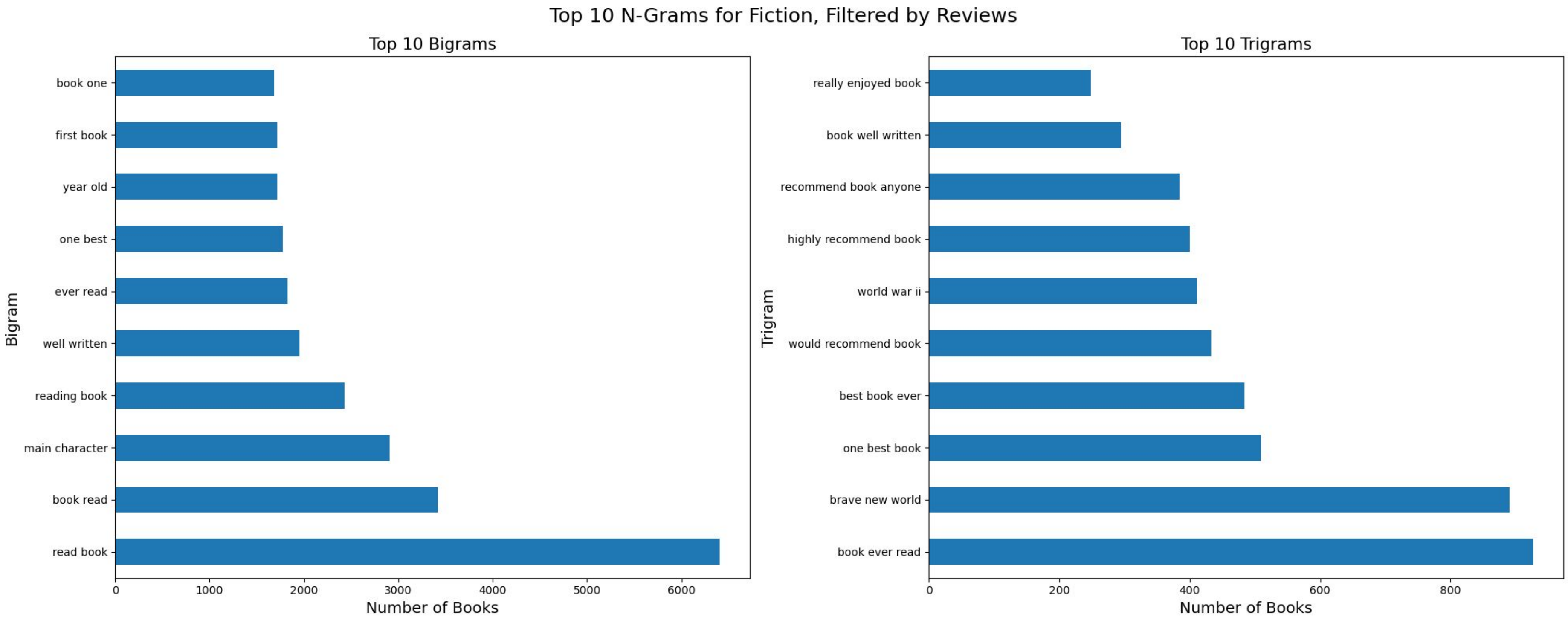
- ❖ Merged datasets, dropped rows without titles
- ❖ Dropping null values
- ❖ Removing punctuation marks in columns (e.g. encased with [‘Title’])

Datasets

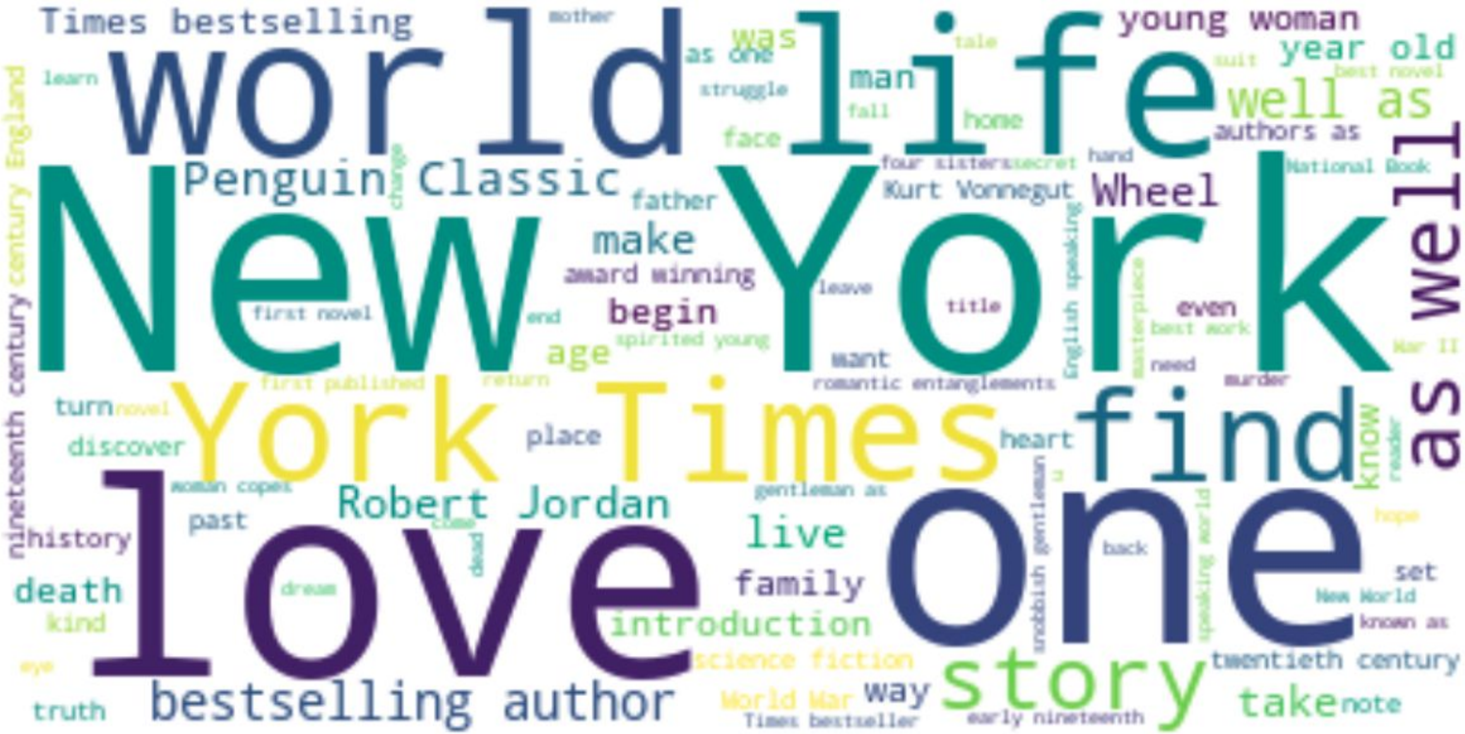
- ❖ Genres (Fiction, Juvenile Fiction, Biography/Autobiography)
- ❖ Overall
- ❖ Books with over 2000 reviews



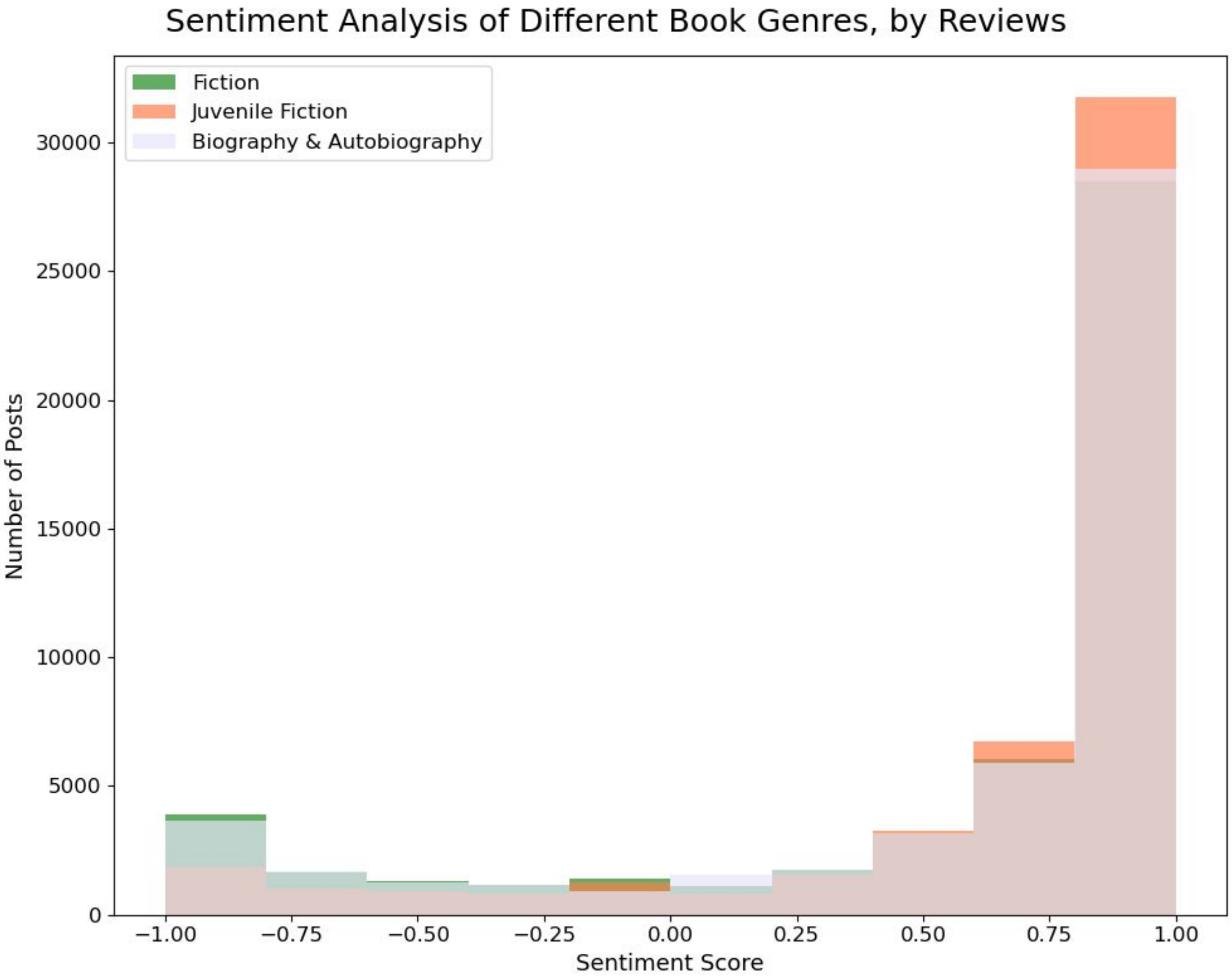
Description: Bigrams & Trigrams of Fiction Book Descriptions

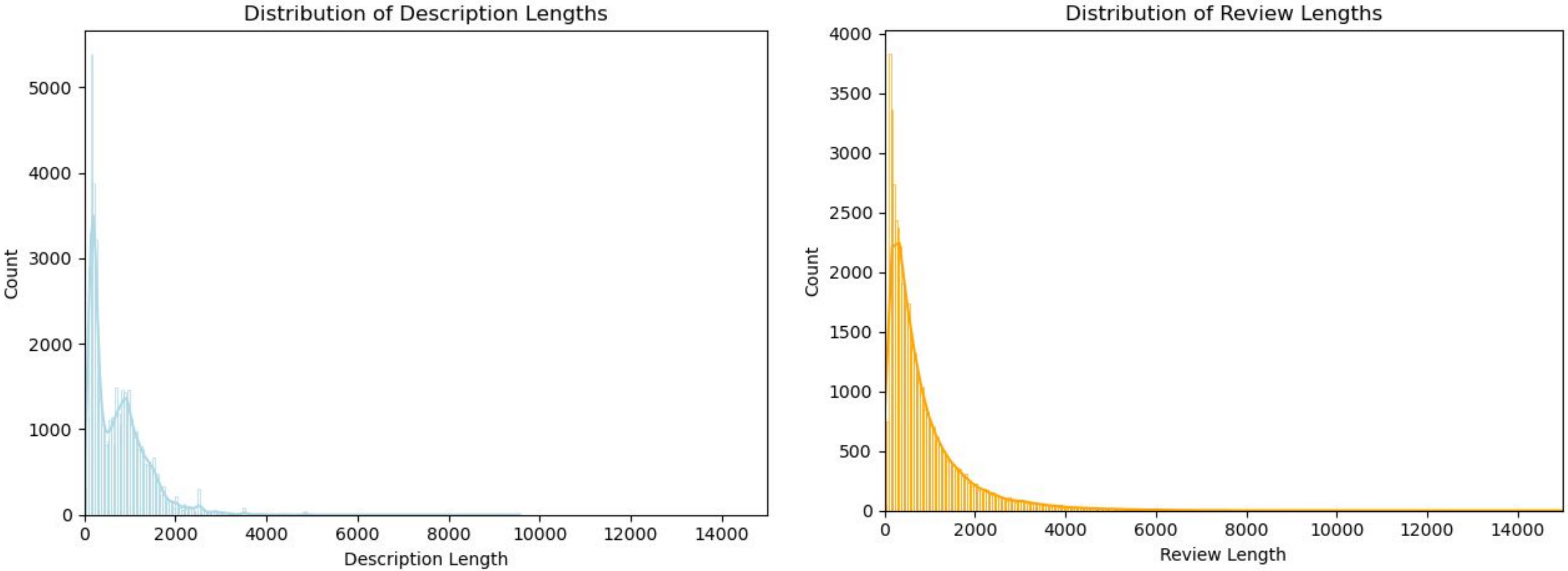


Description: Bigrams & Trigrams of Fiction Book Reviews

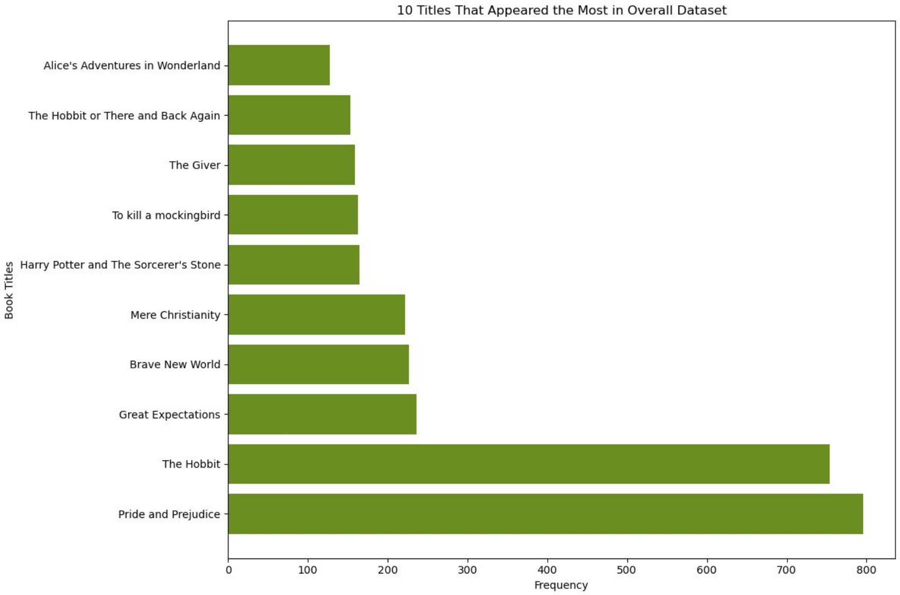


Description: Word Cloud for Fiction Book Descriptions

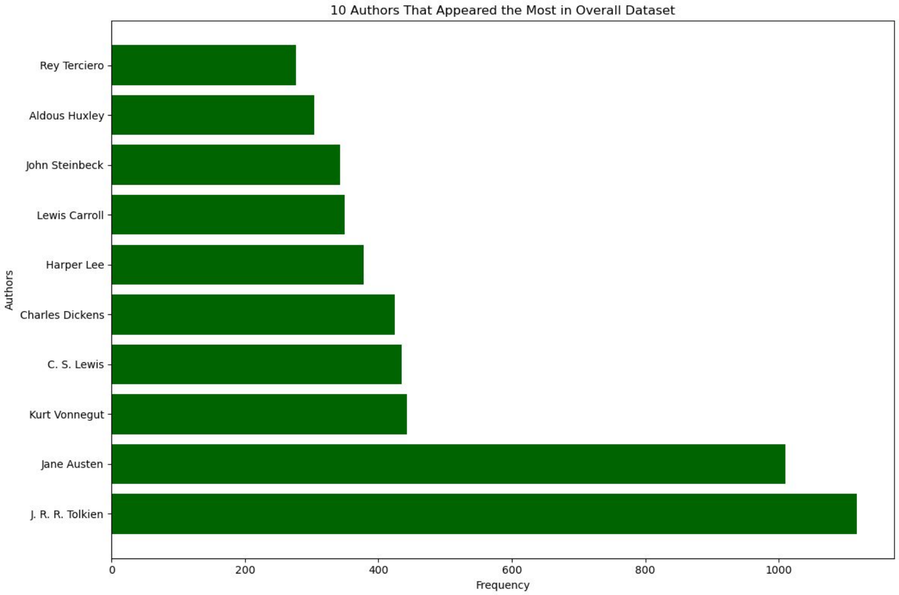

Sentiment Analysis of Different Book Genres, by Reviews
🟩 Fiction
🟧 Juvenile Fiction
⬜ Biography & Autobiography
Number of Posts
30000
25000
20000
15000
10000
5000
0
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Sentiment Score



Description: Distribution of Description & Review Lengths

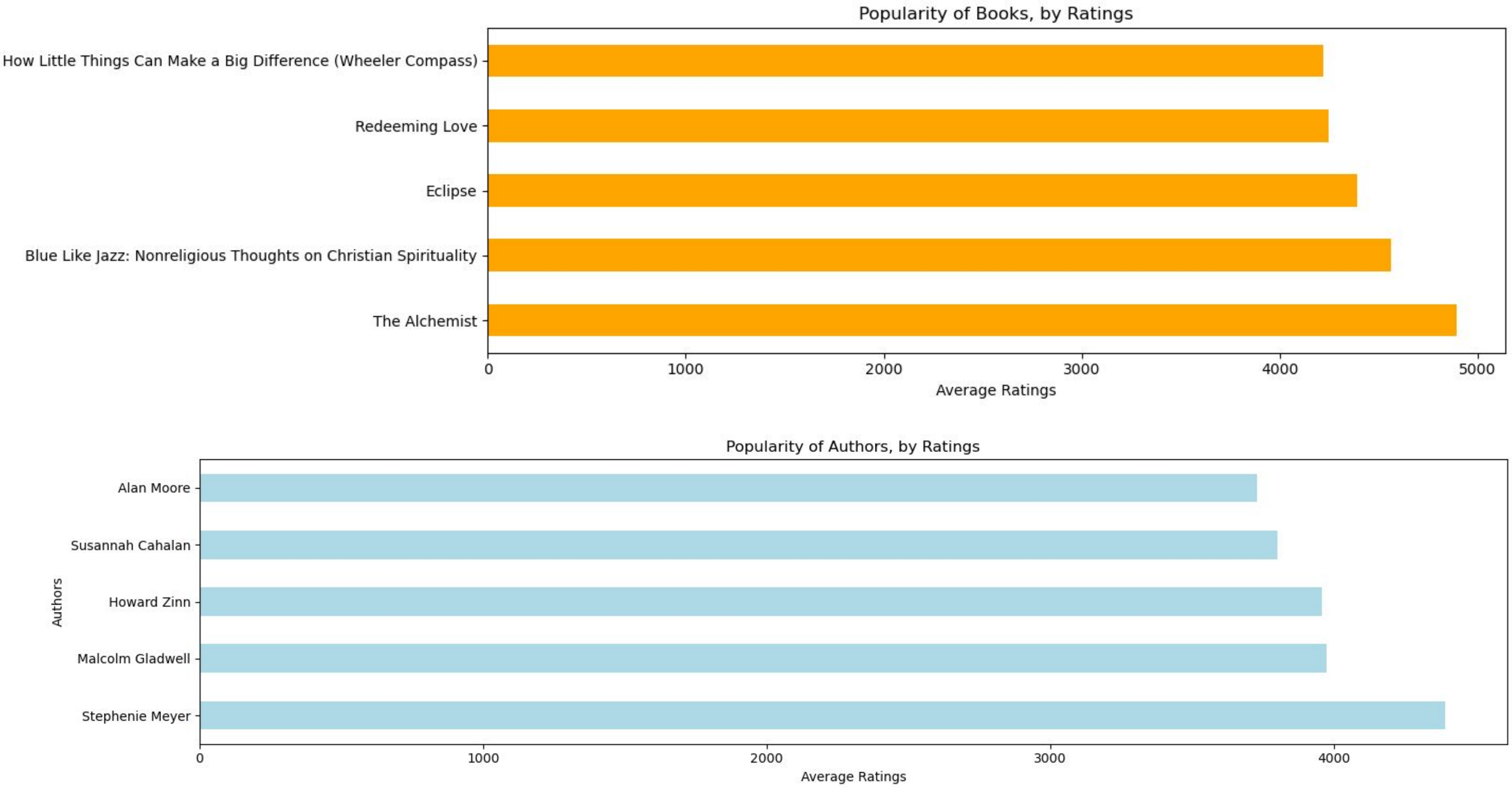


10 Titles That Appeared the Most in Overall Dataset

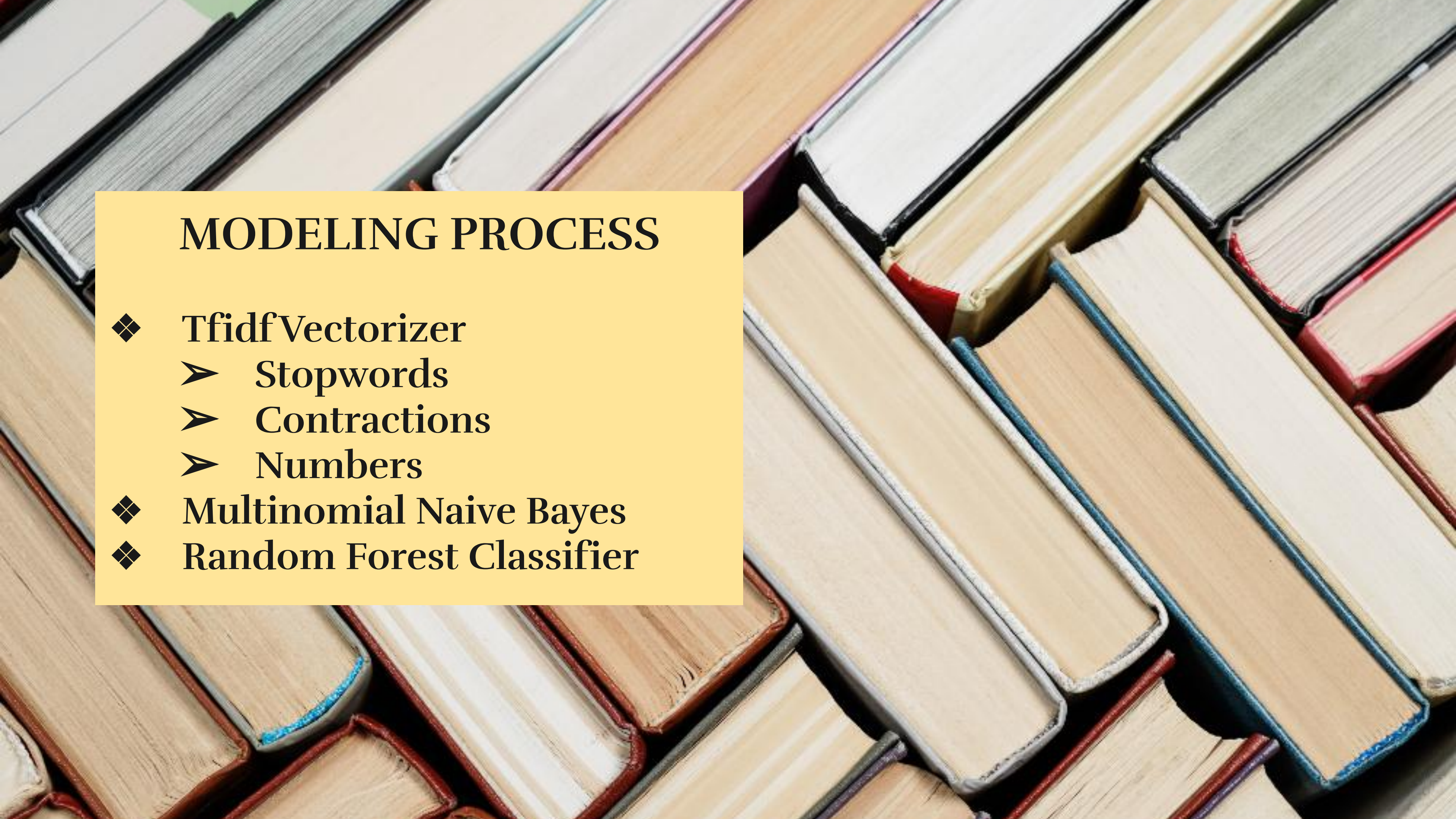


10 Authors That Appeared the Most in Overall Dataset

Description: Top 10 Authors & Titles



Description: Popular Authors & Titles, by Review Count



MODELING PROCESS

- TfidfVectorizer
  - Stopwords
  - Contractions
  - Numbers
- Multinomial Naive Bayes
- Random Forest Classifier

MODELING RESULTS

(Biography & Autobiography Dataset)

Multinomial Naive Bayes

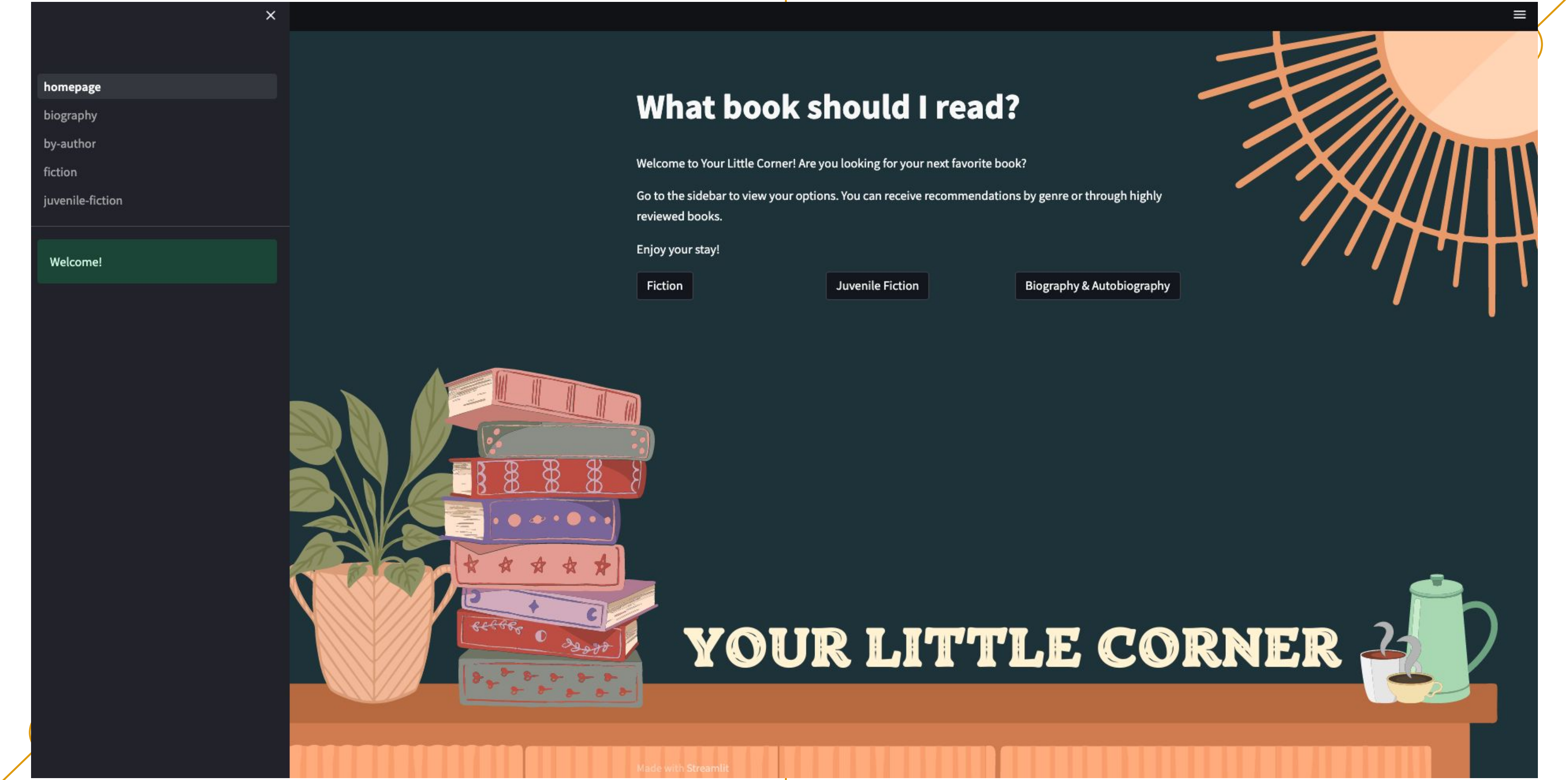
Recall: 0.8381
Precision: 0.7403
F1: 0.7794
Accuracy: 0.8381

Random Forest

Recall: 0.9149
Precision: 0.8752
F1: 0.8889
Accuracy: 0.9149

MODELING RESULTS

Dataset	Model	Recall	Precision	F1	Accuracy
Fiction	MNB	0.6714	0.5155	0.5702	0.6714
JVF	RFC	0.8477	0.7744	0.8016	0.8477
B/A	RFC	0.9149	0.8752	0.8889	0.9149
Most Reviewed	RFC	0.9958	0.9918	0.9937	0.9958



What book should I read?

Welcome to Your Little Corner! Are you looking for your next favorite book?

Go to the sidebar to view your options. You can receive recommendations by genre or through highly reviewed books.

Enjoy your stay!

CONCLUSION

For future researchers:

❖ Experimenting with different models, especially for multiclass, imbalance datasets

❖ Adding other components to modeling (e.g. user satisfaction, review data, memory storage)

❖ Explore hyperparameters within models (with computer capability)

Thank you! Any questions or comments?

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon** and infographics & images by **Freepik**