Project 3: Reddit NLP PRESENTATION

By: Tanya Seegmiller

```
<p> Fantasy subreddit Vs. Horror
subreddit. </p>
```

Problem Statement

Book House Publishing is a children's book publisher looking to rebrand their image and publish books that appeal to a more mature and diverse audience. Book House has a short list of authors in Horror and Fantasy Literature to vet.

To better understand the dynamics and preferences of these communities, Book House wants NLP done.

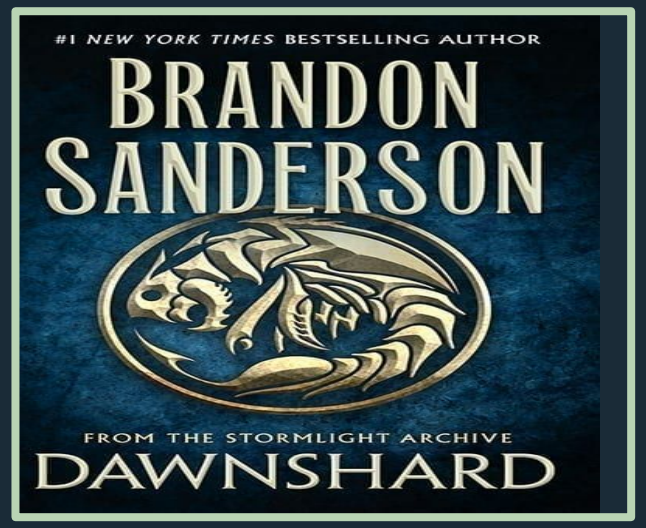


TABLE OF CONTENTS.

Data Collection: Subreddits

r/Fantasy: "The internet's largest forum dedicated to speculative fiction in literature, games, film and the wider world."

- **1973 comments scraped w/PRAW**

r/Horrorlit: "An inclusive community dedicated to the discussion, elevation, and expansion of the horror literary genre."

- **1955 comments scraped w/PRAW**

Data Cleaning:

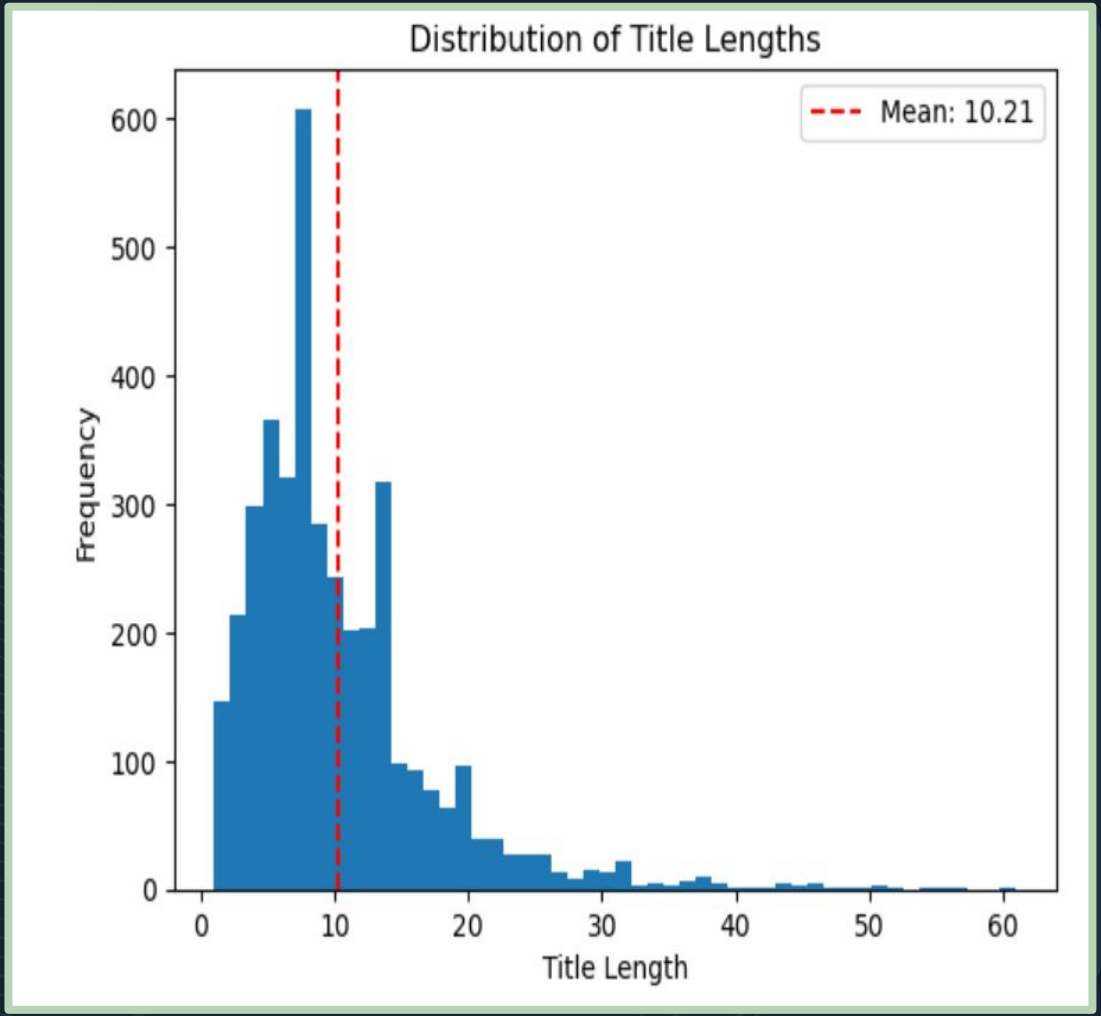
Drop it like it's…duplicated?

- Then treated missing self_text values

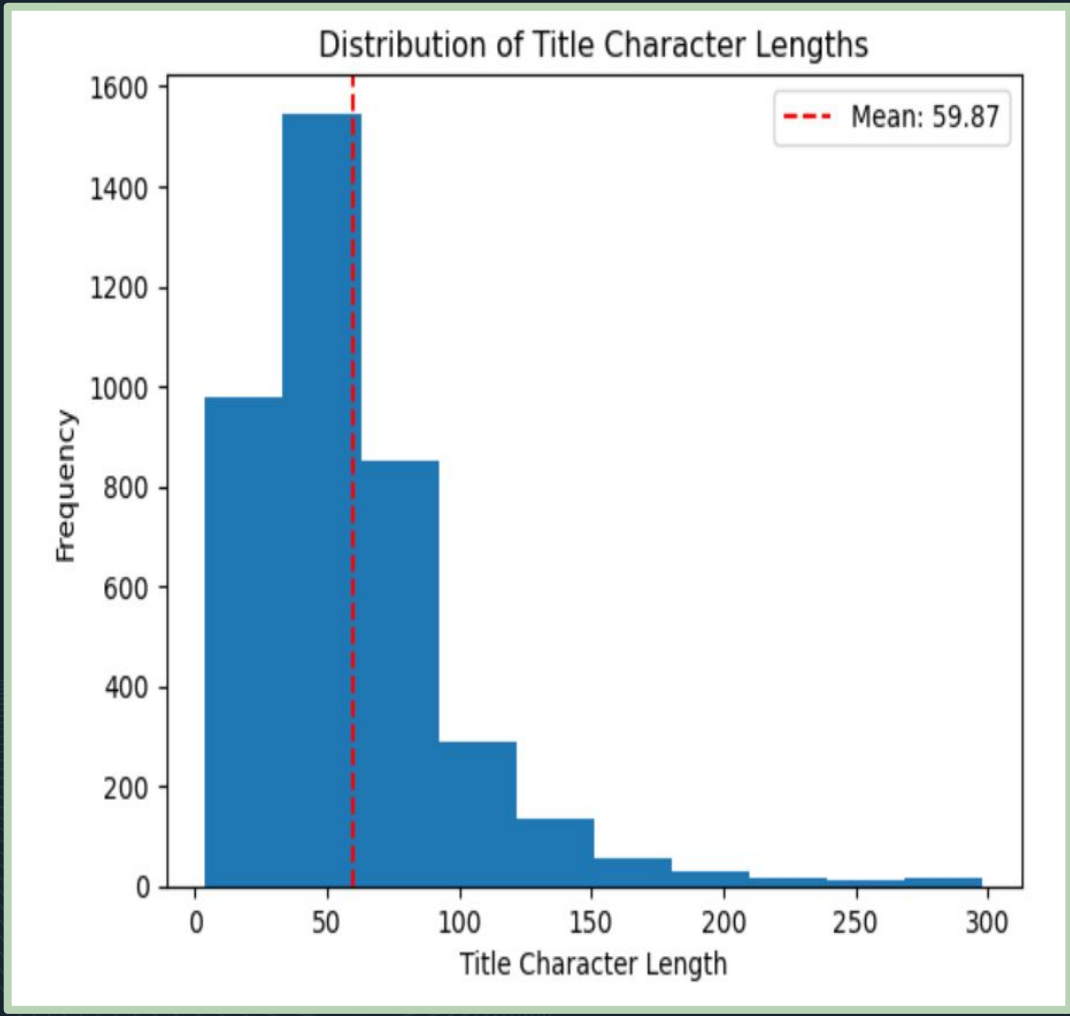
f**2 + h**2 = subreddit.csv

- (geometry joke totally intended, hehe!)
- Concatenated all 4 csv's into 1 subreddit.csv
- Mapped 'reddits' column to: {'Fantasy': 0, 'horrorlit': 1}

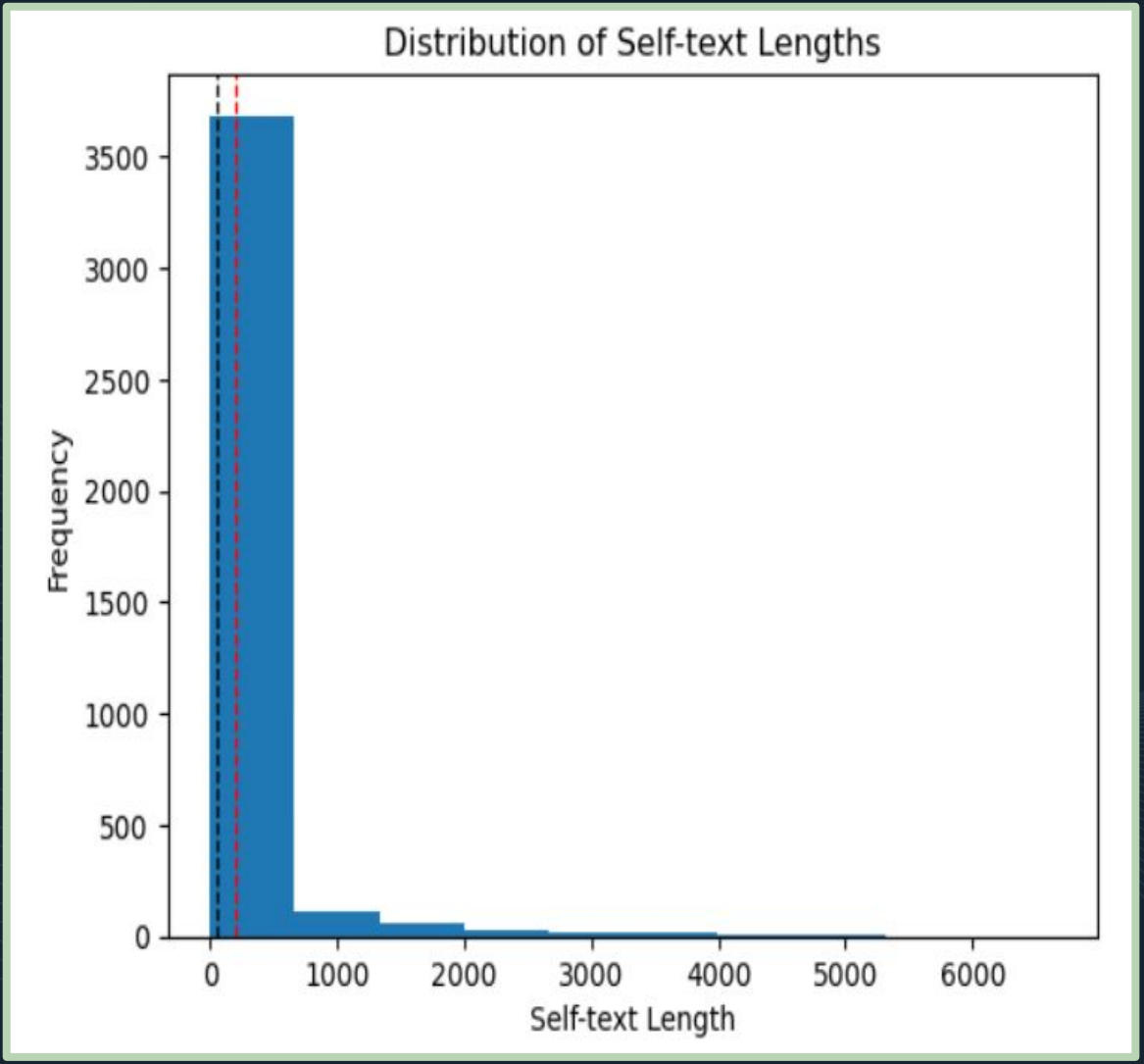
Avg. Title Word Length: 10

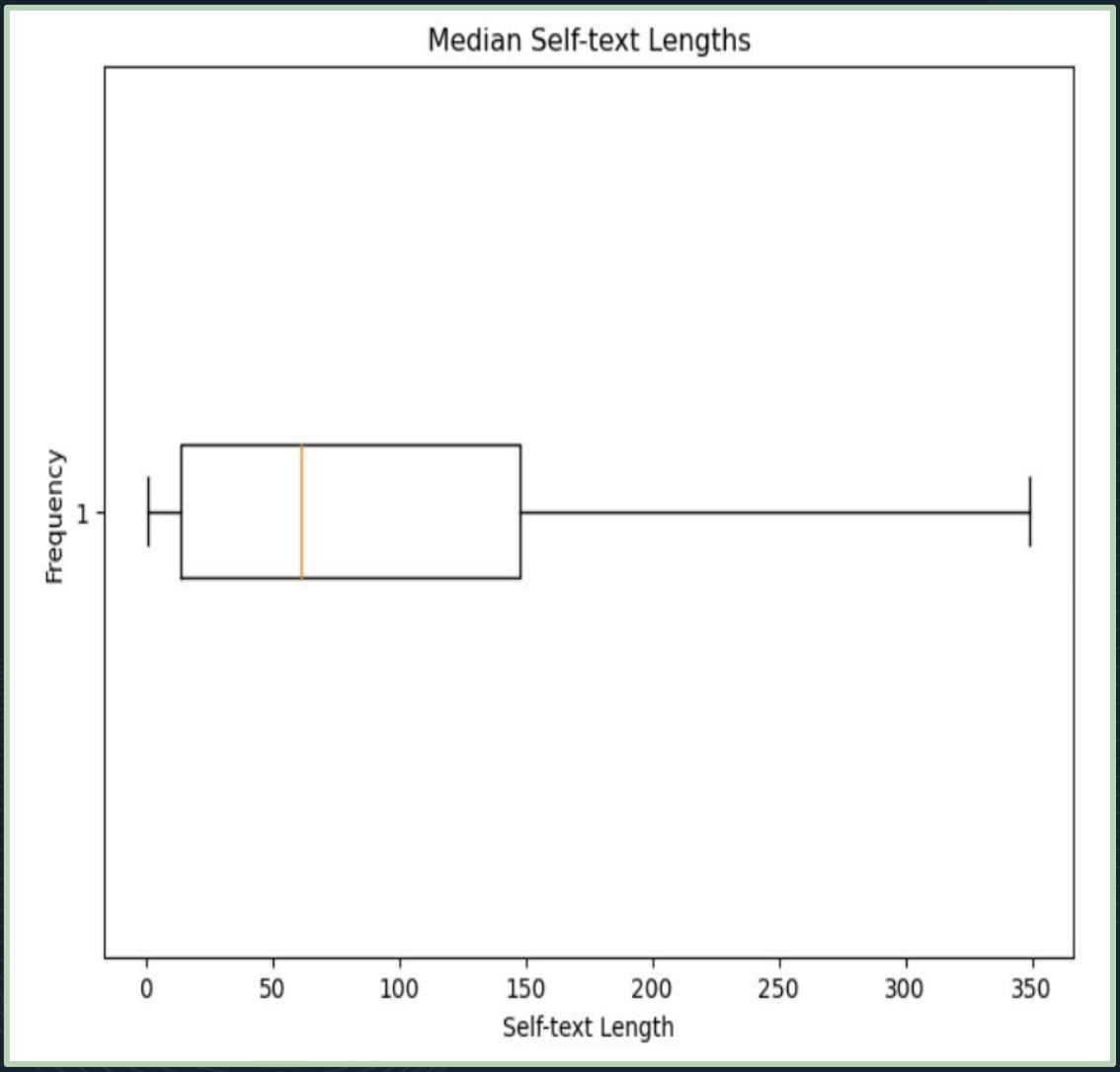


Avg. Title Char Length: 60



Median Self_text Length: 61
Mean self_text length: 199

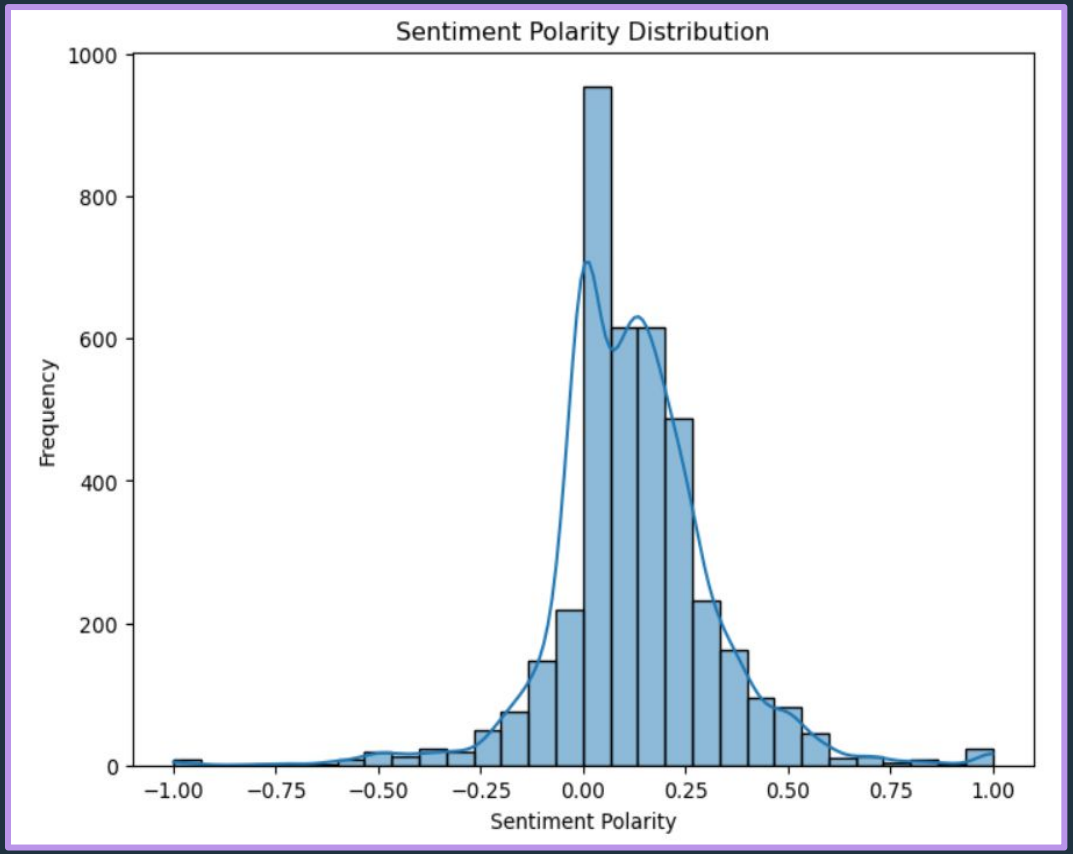


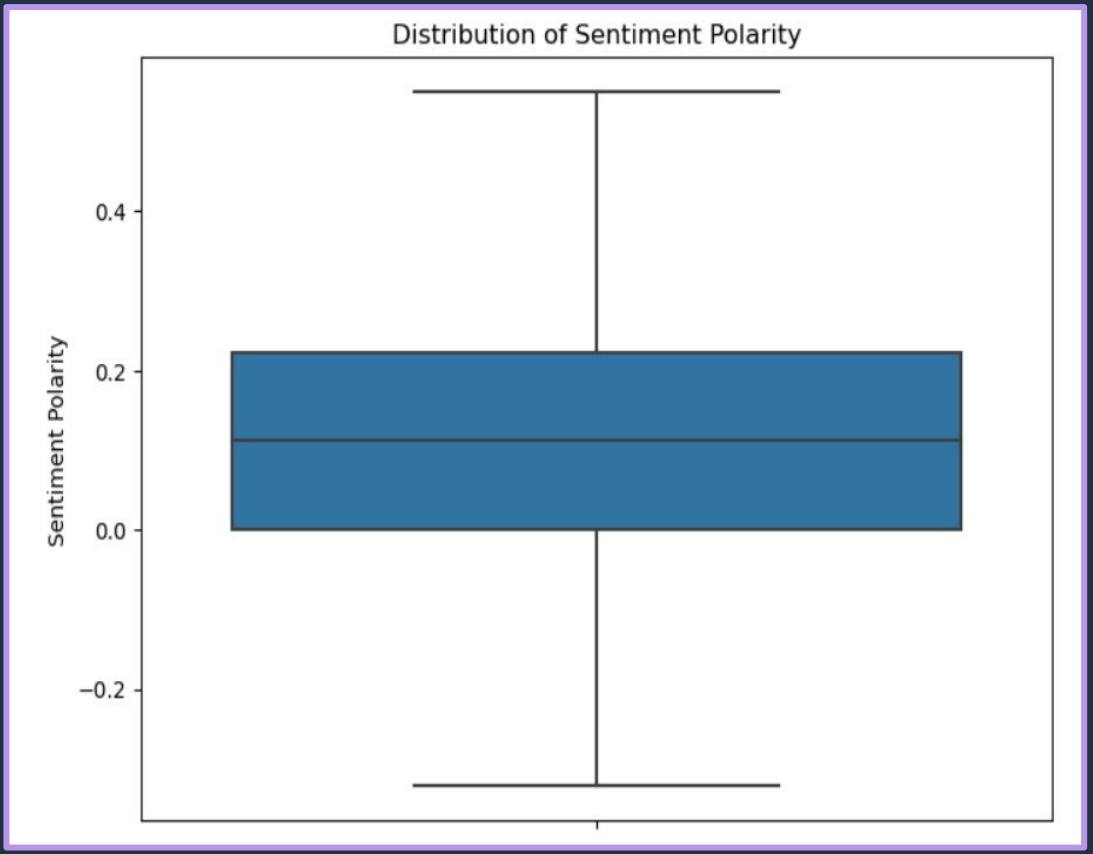


Sentiment Polarity

Sentiment Polarity: The measure of sentiment expressed in a piece of text where

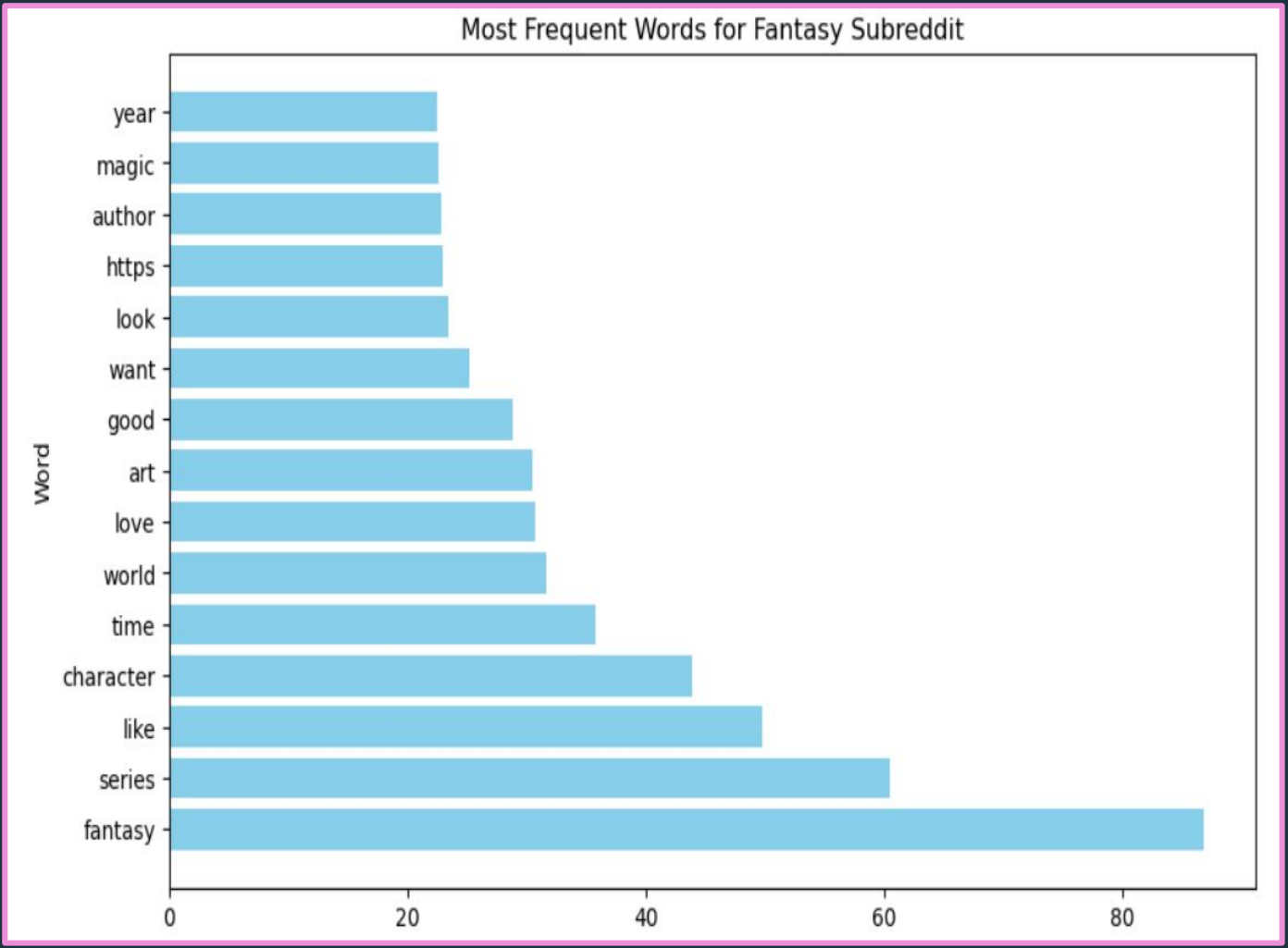
-1 = 0= 1=



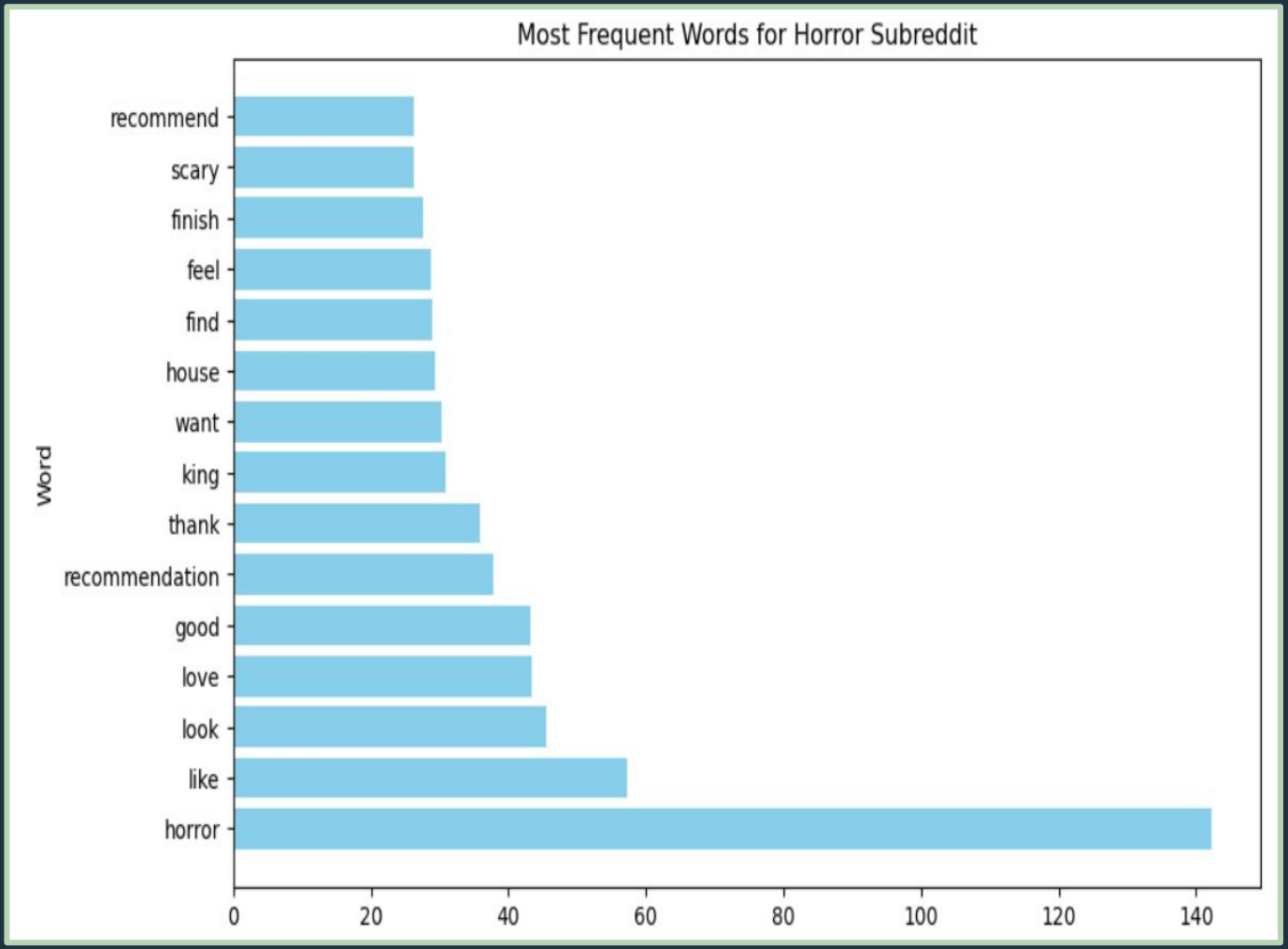


Most Frequent Words: Fantasy

```
Most frequent words for subreddit 0:
fantasy      86.792918
series       60.458241
like         49.748452
character    43.799830
time         35.740716
world        31.574656
love         30.735291
art          30.507070
good         28.803490
want         25.185668
look         23.415675
https        22.927411
author       22.756687
magic        22.535572
year         22.476298
```



Most Frequent Words: Horror

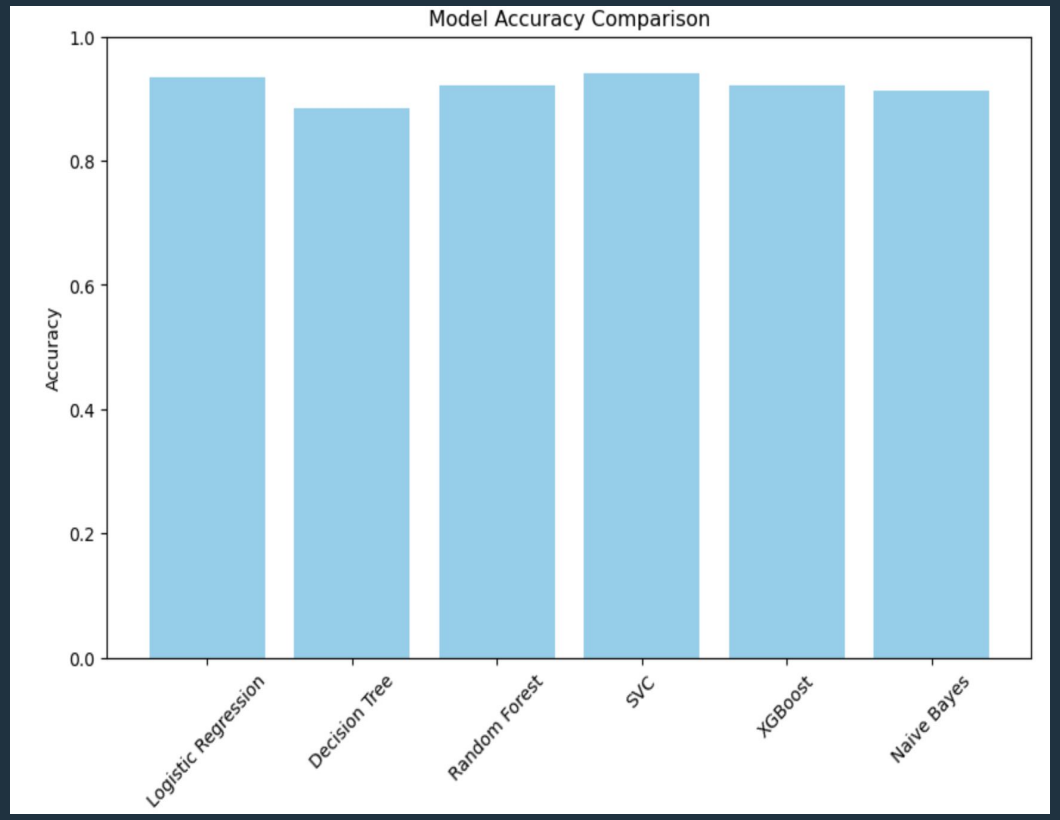



```
Most frequent words for subreddit 1:
horror            142.204028
like               57.279154
look               45.505959
love               43.391293
good               43.262734
recommendation     37.832223
thank              35.815091
king               30.778066
want               30.267899
house              29.249897
find               28.924760
feel               28.758687
finish             27.526748
scary              26.314606
recommend          26.245879
```

Assumptions and Initial Model Performance

* Assumptions: Simple Lemmatization and TfidfVectorizer would return best results

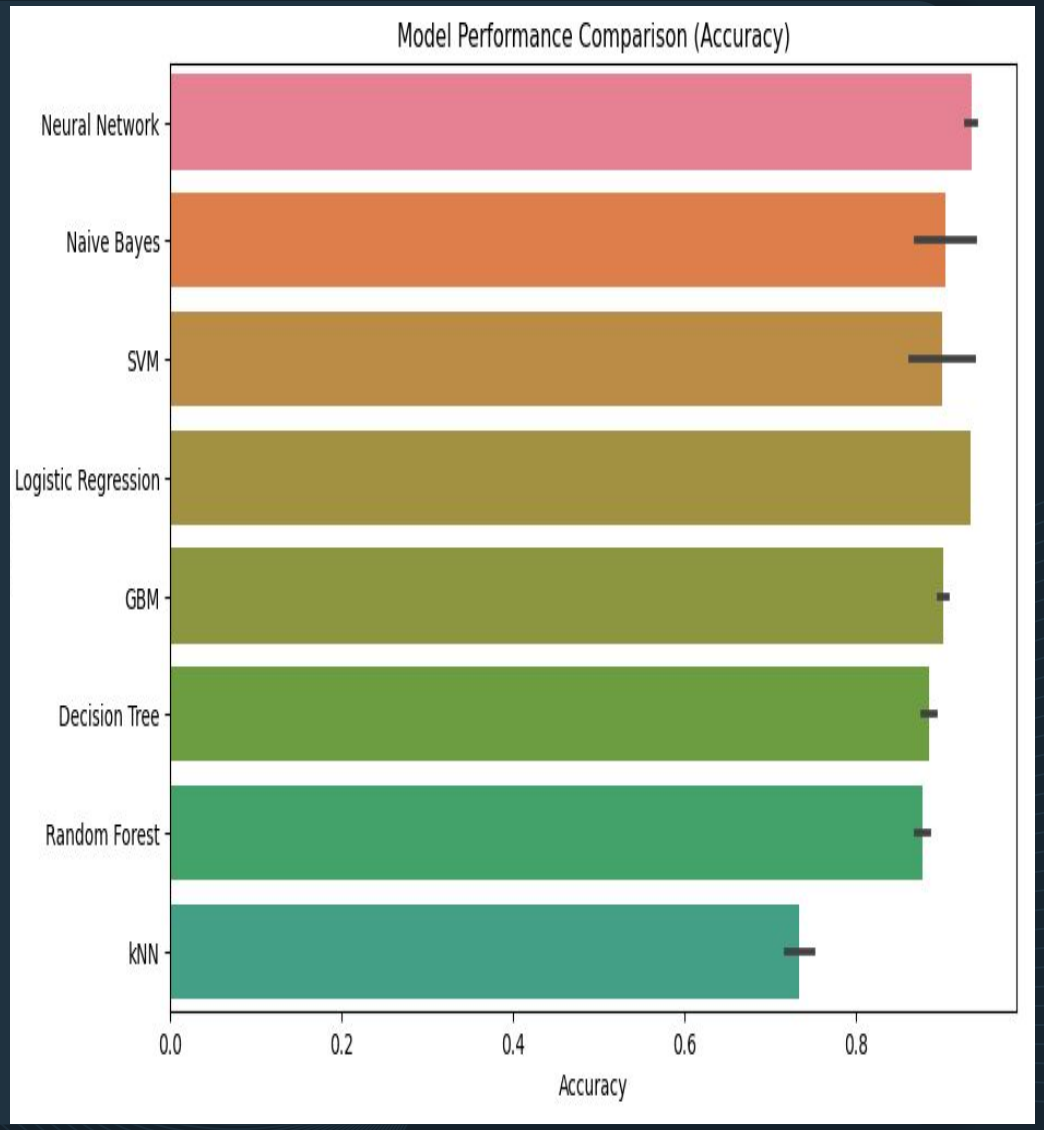
Model Type	Accuracy Score
1. Log Regression	.9351
2. Decision Tree	.8842
3. Random Forest	.9274
4. SVC	.9415
5. XGBoost	.9211
6. Naive Bayes	.9122



Did Tuning Help Model Performance: No

Hypertuned Model Performance

Model Type	Accuracy	Tuned Accuracy
1. Log Regression	.9351	.9325
2. Decision Tree	.8842	.8893
3. Random Forest	.9274	.8702
4. SVC	.9415	.9364
5. XGBoost	.9211	.9071
6. Naive Bayes	.9122	.9376
7. KNN		.7494
8. ANN		.9402



Conclusions and Recommendations

Conclusion: Both Fantasy and Horror Literature had positive sentiment with respect to their preferred genres as well as camaraderie between the genres. Positive descriptors were a key factor in every model correctly identifying comments.

Recommendations: The best best overall model as far as performance across all metrics was Naive Bayes with CountVectorizer. For any future sentiment studies conducted by Book House Publishing, I would recommend using that model to gauge genre sentiment.

THANK YOU!

Do you have any questions?

bookhouse@mail.com
800-285-4697
mydomain.com