Predicting Reddit Submission Category

Roxana Ruvalcaba

Agenda

- Problem Statement
- Data Gathering
- EDA
- Models
- Results & Next Steps

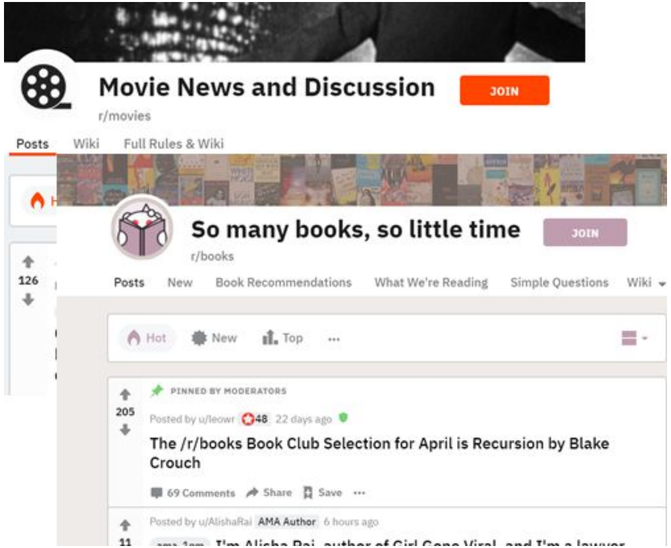
Data Science Problem

Are subreddit titles predictive of subreddit categories movies and books

- Pushshift API
- NLP to train binary classifier

Gathering Data

- Choosing subreddit: Movies & Books
- Leveraged query_upshift function
  - Requests library
  - Size = 500 per pull
  - Choose subfields
  - Dropped Duplicates



Movie News and Discussion

r/movies

So many books, so little time

r/books

PINNED BY MODERATORS

Posted by u/leowr 48 22 days ago

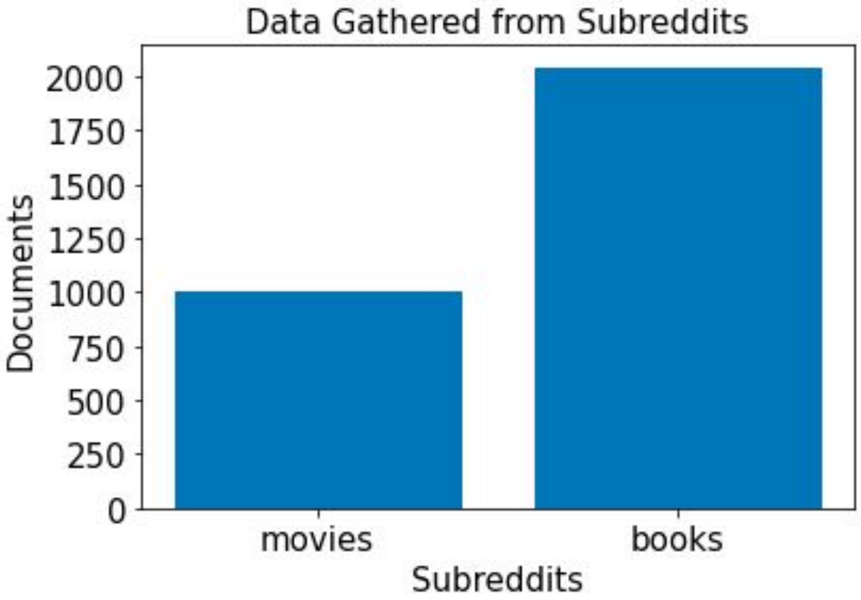
The /r/books Book Club Selection for April is Recursion by Blake Crouch

69 Comments Share Save

Posted by u/AlishaRai AMA Author 6 hours ago

I'm Alisha Rai, author of Girl Gone Viral, and I'm a lawyer

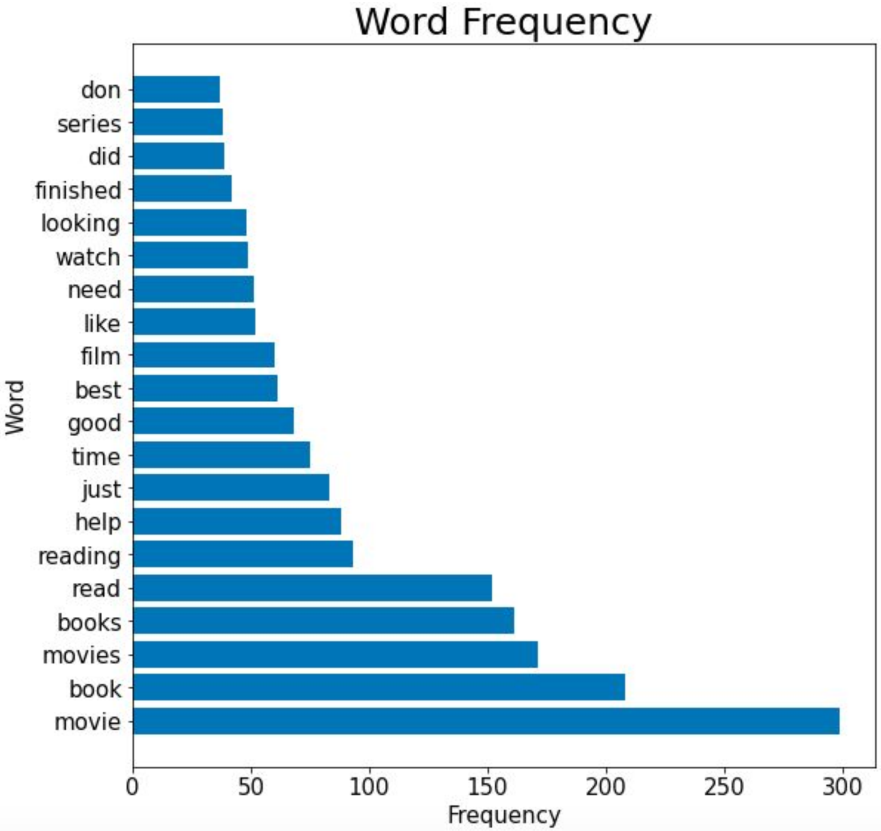
Exploring Data



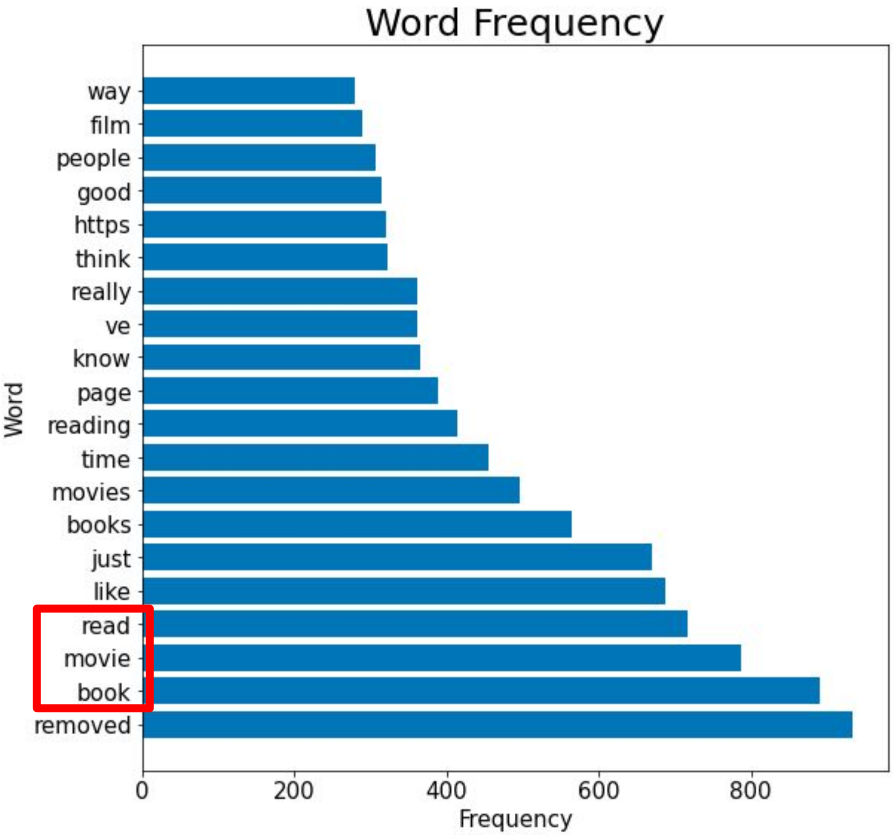
- Removed null values
  - 3% of movies selftext null
  - 6% of books selftext null
- Subreddit column = 1 or 0
- Took ~900 samples from each

EDA: Concatenated Title & Selftext

Title



Title + “ “ + Self Text



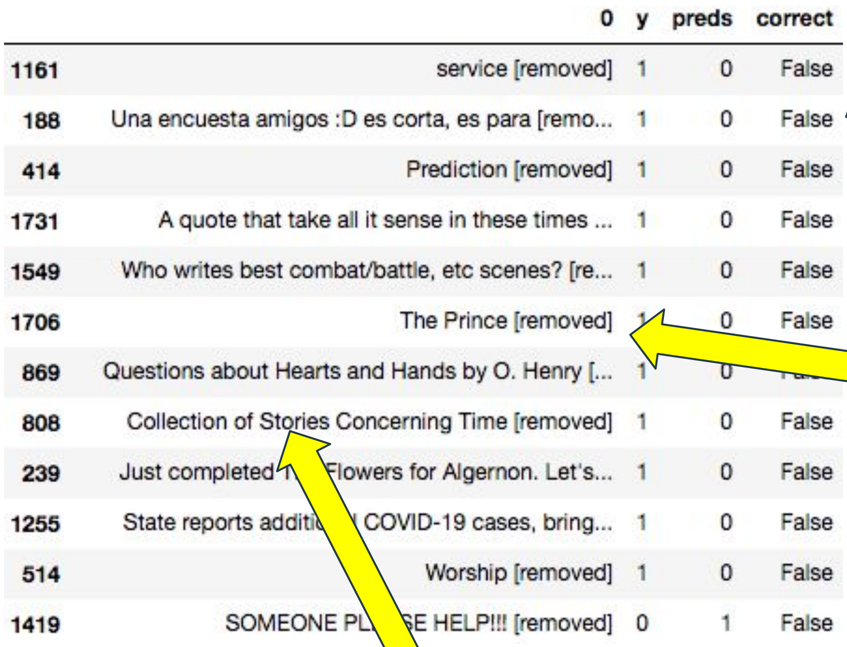
Model 1: Logistic Regression

Best Cross Val	0.9351	0.8891
Best Train Score	1.0000	0.9110
Best Test Score	0.9453	0.9037
	CountVectorizer 'model__max_iter': 1000, 'model__C': 10.0 'model__penalty': 'l2', 'vectorizer__ngram_range': (1, 2), 'vectorizer__stop_words': 'english'	TfidfVectorizer 'model__max_iter': 1000, 'model__C': 0.01, 'model__penalty': 'l2', 'vectorizer__ngram_range': (1, 1), 'vectorizer__stop_words': 'english', 'vectorizer__max_features': 400,

Model 2: Multinomial Naive Bayes

Best Cross Val	0.9285
Best Train Score	0.9832
Best Test Score	0.9431
	CountVectorizer 'vectorizer__max_df': 0.98 'vectorizer__ngram_range': (1, 1), 'vectorizer__stop_words': 'english'

What did the model predict incorrectly?



	0	y	preds	correct
1161	service [removed]	1	0	False
188	Una encuesta amigos :D es corta, es para [remo...	1	0	False
414	Prediction [removed]	1	0	False
1731	A quote that take all it sense in these times ...	1	0	False
1549	Who writes best combat/battle, etc scenes? [re...	1	0	False
1706	The Prince [removed]	1	0	False
869	Questions about Hearts and Hands by O. Henry [...	1	0	[illegible]
808	Collection of Stories Concerning Time [removed]	1	0	False
239	Just completed T[illegible] Flowers for Algernon. Let's...	1	0	False
1255	State reports additi[illegible] COVID-19 cases, bring...	1	0	False
514	Worship [removed]	1	0	False
1419	SOMEONE PL[illegible]SE HELP!!! [removed]	0	1	False

Different Language

Could be either

Stories wasn't associated with books?

Conclusion & Next Steps

Both models accurately predict if a subreddit is from the Books or Movies topics with the LR model performing slightly better.

Next Steps:

- Reduce variance without sacrificing accuracy
- Attempt other regularization parameters & models
- Attempt model with other subreddit topics