Spook is in the Air

❤️ 👻 👾

Omar Qusous and Brian Yee

The Question

We wanted to be able to classify the main genre of a film based on its script alone.

The Data

We got our data from imsdb.com.

Web-scraped scripts of movies from 3 different genres.

- Romance ❤️
- Horror 👻
- Fantasy 👾

These genres were chosen because we deemed them to have the least overlap.

In total, we had 116 scripts from each genre, with an average of 23221 words per script.

Data Cleaning

We created our list of stopwords by including:

- Standard english stopwords from nltk.corpus.stopwords
- Names from nltk.corpus.names
- Punctuation marks and special characters
- Digits

Lemmatized our remaining words using WordNetLemmatizer from NLTK.

- Lemmatizer needed to be given the part of speech for the words to be correctly lemmatized.

Created frequency distribution tables for the scripts.

- Stored data on all scripts in a single DataFrame.
- This DataFrame had 348 movies with 55948 unique words.

EDA and Feature Engineering

To combat sparsity, we decided to cut out words that have only appeared in a single script out of all our movie.

This cut down the number of features to 27799.

We tried being even more selective with our words by adjusting the minimum number of movies they must appear in, but this ended up removing too many words.

Number of Movies	Words Left
2+	27799
3+	12435
4+	11209
5+	10251
10+	8308
15+	6476

EDA and Feature Engineering

Then we went through each word and only kept the words that were in the top 25 most frequent of a movie.

This ended up cutting down our number of features down to 1220.

Models

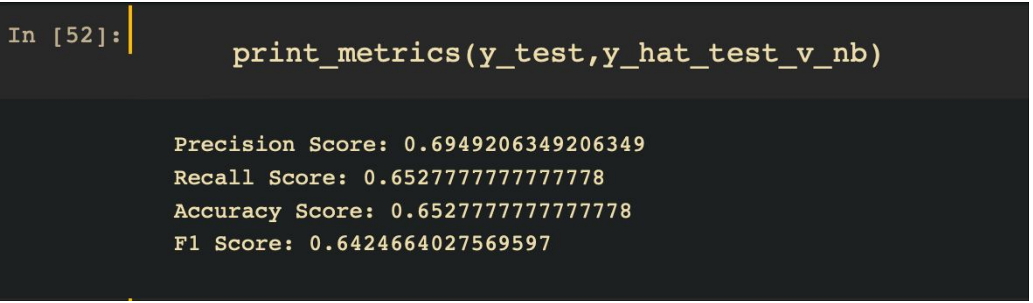
All of our models were trained on 80% of our data and verified on the remaining 20%.

All models were also tuned using GridSearchCV.

Model	Accuracy	Precision	Recall	F1
Dummy Classifier	0.32	0.32	0.32	0.32
Random Forest	0.65	0.65	0.65	0.65
Naive Bayes	**0.65**	**0.69**	**0.65**	**0.64**
XGBoost	0.64	0.63	0.64	0.63

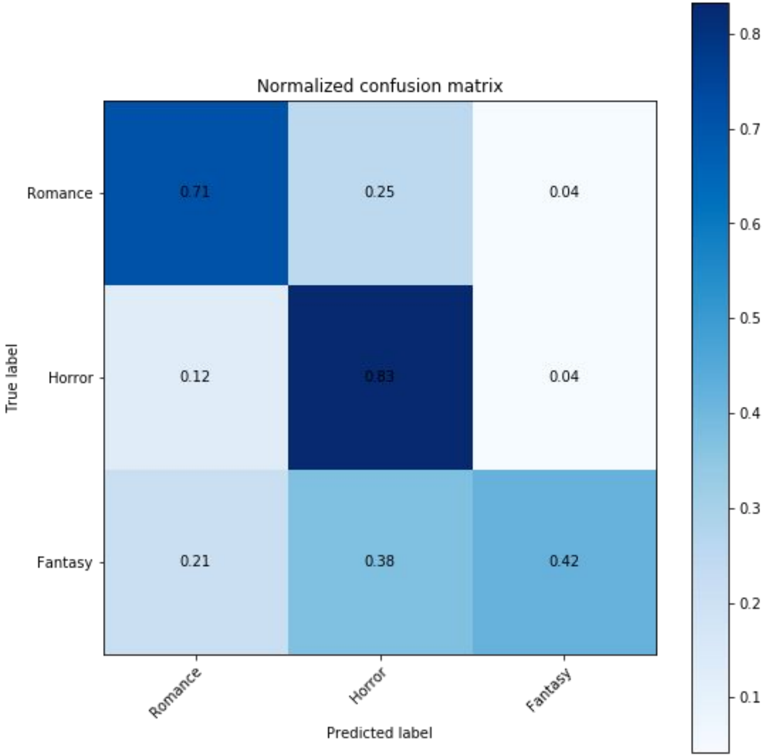
Metrics for each classification model.

Final Model - Multinomial Naive Bayes



```
In [52]:    print_metrics(y_test,y_hat_test_v_nb)

            Precision Score: 0.6949206349206349
            Recall Score: 0.6527777777777778
            Accuracy Score: 0.6527777777777778
            F1 Score: 0.6424664027569597
```



Fantasy scripts were misclassified the most because fantasy could include romantic scenes and mystical creatures.

Thank you, Questions?