Subreddit Classification

By Chris Shaw

```
1  scifi.title.head()
```

```
0                                   Book Suggestions
1    CBS All Access Announces Michelle Yeoh-Led 'St...
2    Terry Pratchett's "Discworld" BBC Radio adapta...
3                A theory of flight [SciFi short story]
4    What do you think about my 3D-Printed Viper MK...
```

Sci-Fi Subreddit

```
1  politics.title.head()
```

```
0                                          Cry Me a River
1    Ted Cruz defends Trump on Russia: "I don't see...
2          This is the longest shutdown in US history
3    Rhode Island Gov. Gina Raimondo to call for ma...
4    Laura Kelly sworn in as Kansas governor after ...
```

Politics Subreddit

Cleaning Done to the Data

1. Delete duplicates
2. Lower case words
3. Just letters, no punctuation or other characters

Vectorizers

Count Vectorizer: 21,606 words with no stop words
747,001 words with 1-10 n-grams

TF-IDF Vectorizer: 21,606 words with no stop words
747,001 words with 1-10 n-grams

Hashing Vectorizer: 1,048,576 words with 1-10 n-grams

Models	Train Score	Test Score
Logistic Regression with Count Vectorizer	.981	.961
Random Forest with Count Vectorizer	.847	.840
Logistic Regression with Hash Vectorizer	.936	.901
Random Forest with Hash Vectorizer	.859	.830
Logistic Regression with TF-IDF Vectorizer	**.976**	**.962**
Random Forest with TF-IDF Vectorizer	.859	.852