Classifying Subreddit Posts

July 12, 2019
Jeremy Ondov

Oh No, Our Subreddits!

Our Problem - Somebody (to remain nameless) accidentally stripped the subreddit details of all the posts in r/gaming and r/pcgaming. We need a model to get them in the right place.

Luckily, we had recently pulled 1000 posts from each subreddit right before the incident happened.

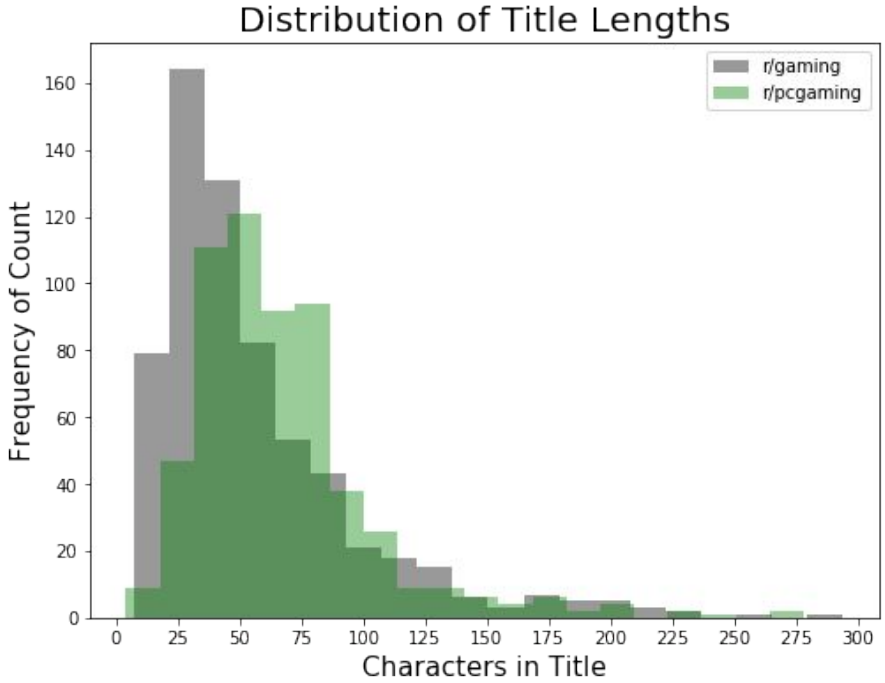
Unluckily, many of the pulled posts were repeated. Our final dataset:

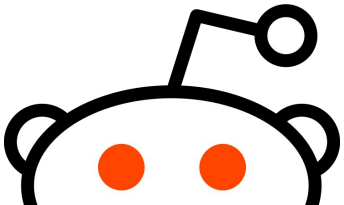
	r/gaming	r/pcgaming
Unique Posts	645	583

What Do Our Posts Look Like?

r/gaming has simpler posts, with a lot more memes/images.

r/pcgaming has more complex titles, typically aimed at creating discussions (and has its share of memes as well).

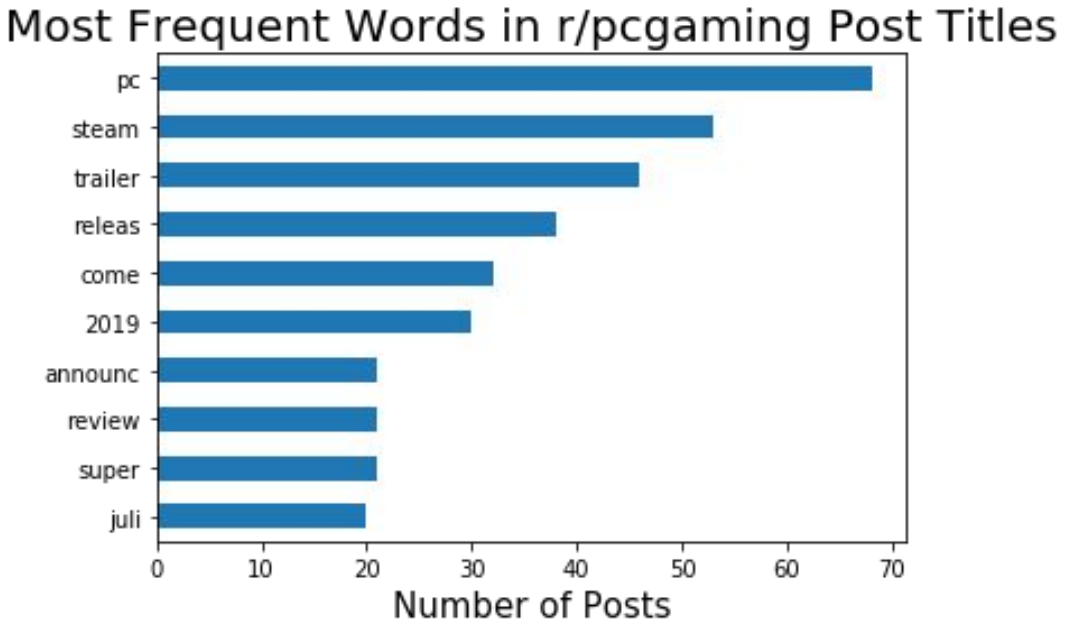




PC Gamers Use Specific Language

We removed “stop words”, and stemmed the rest down to their roots.

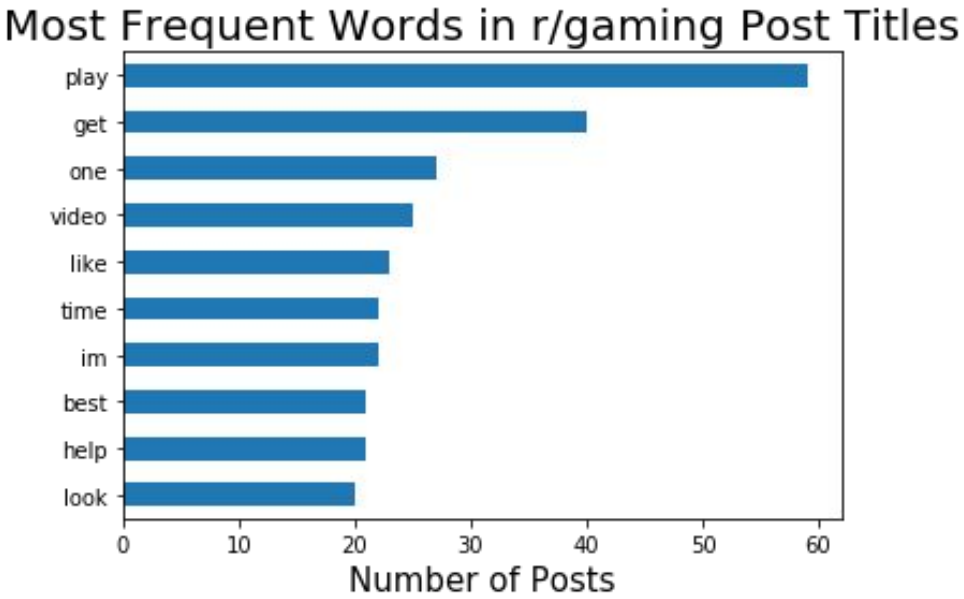
PC gamers seem to talk about PCs and Steam (a PC-based client), among other relevant gaming topics.



Gamers are More Vague

The top words in r/gaming have looser connections, but still can be relevant

The difference in typical posting styles should help the model with classification



How Do We Properly Model?

1. The data was split into a training set and a testing set
2. A wide range of modeling techniques were used: 2 vectorizers, and many classifiers including logistic regression, k-nearest neighbors, and multiple decision tree-based models
3. A thorough evaluation of each model was performed, using a grid search and cross-validation.

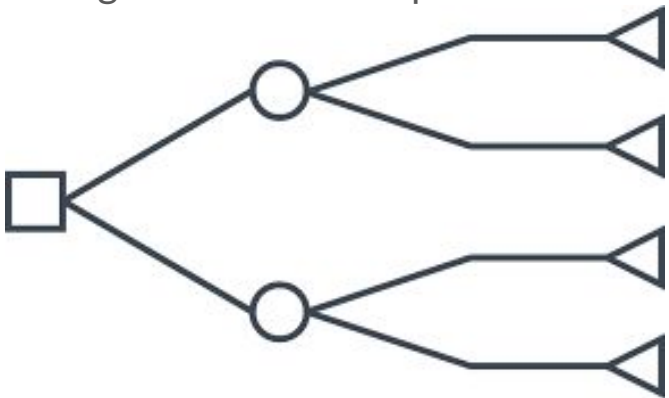
The final selected model utilized a TFIDF vectorizer with a boosted decision tree classifier.

Good Performance, with Room to Improve

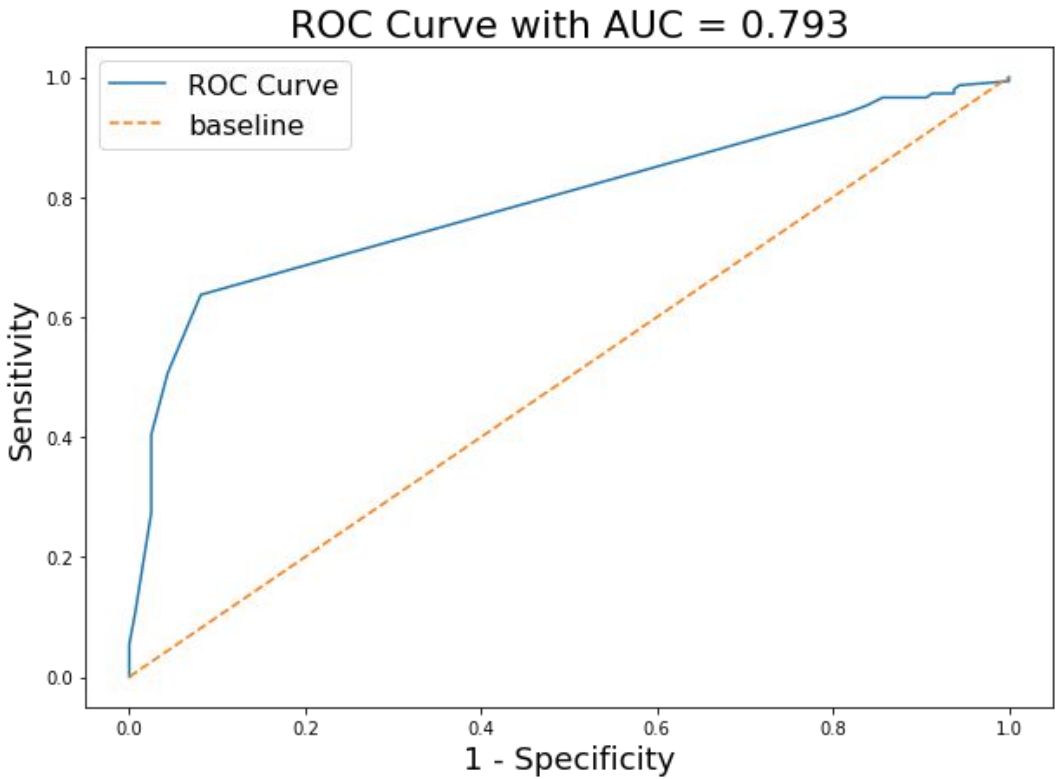
The selected model gave an accuracy score of **~82%** on the training set, and **~78%** on the testing set.

Each score was not the absolute highest out of every model, but this produced the best balance between bias and variance

Further optimizations to the stop words and decision tree hyperparameters should be analyzed - May see how title length/comments help classification



Extra Model Visualization



Thanks!