Predicting Content Category

September 2016

DRAFT

What Is The Problem?

Currently there is no method of detecting whether a live broadcast contains community content or non-community content. Non-community content is defined by any esports or major event which is not broadcasted by an individual streamer.

This will be a classification problem.

Potential Solution?

By training a predictive model to test on the following criterion, we can attempt to tag future broadcasts as community or non-community content to a certain degree of confidence:

- ➔ Day of the Week
- ➔ Game
- ➔ Words in Broadcast Title
- ➔ Average CCUs (Concurrent Users / Viewers)

I hypothesize that of the features listed above, the words in a broadcast title and average CCUs should be the most effective in predicting whether a broadcast is community content or not.

Data

- Dataset is achieved by using a SQL query of our internal data that sits in AWS (Redshift)
- In later iterations, I hope to connect to AWS directly with Python rather than using the frontend tool Mode Analytics and generating a csv with SQL
- Current dataset has 1,000 datapoints
    - Limitation here is due to the fact that we have to go through the dataset manually to tag historical broadcasts as community or non-community

Next Steps

Clean data and decide on which features and predictive models to use. Based on models learnt so far, the potential models that can be applied here are:

- ➔ KNN
- ➔ Naive-Bayes
- ➔ Decision Tree