Final Capstone Project

Proposal

**Box Office Movie Prediction – Round 3**

**Problem:**

How much money will a movie make opening weekend? How much money will a film make in total after it’s opening weekend? I plan on doing two separate

**Solution:**

Calculate a film’s digital footprint prior to release, i.e. how many consumers have searched, reacted, engaged, etc. with the film prior to release.

**Data sources / access:**

1. Wikipedia
2. Wikipedia pageview API
3. RottenTomatoes
4. Metacritic
5. Box Office Mojo
6. Youtube trailer views
7. TheMovieDB
8. Facebook

**Techniques:**

1. Scraping
2. Clustering
3. PCA
4. Linear regression
5. CNN

**Challenges:**

Most of the sources will be fairly easy, mainly because I’ve done the work on scraping and cleaning those sources in prior projects. Also, finding TheMovieDB API is a lifesaver to make getting the trailer views much more manageable. However, getting the right movies from TheMovieDB to match will require me to have the IMDB link, so I need to make sure I’m grabbing enough of those from Wikipedia articles. Facebook scraping might be a bit of an issue, just because I’m unfamiliar with it.

The Wikipedia pageview data will be a challenge. In a prior project I didn’t find the information to be highly correlated to box office success. Plus it limits my dataset to only films after 2015 when the API was made available. A few things I’m going to try differently; 1) is PCA on all of the search metrics I come up with, and 2) try some clustering techniques to identify similarities in pageviews that might help box office correlation. Also, I had tried PCA in the unit 3 capstone, but I only pulled out the most highly correlated variable, which I’ve since learned doesn’t make as much sense to keep all of the output PCA features.

Finally, related to techniques, I’m going to compare my linear regression model to a neural network model for box office prediction. If the neural network outperforms the linear regression, that will be great from an output perspective, but it will be very difficult to determine what input variables are important to the CNN model. I will probably try to run the linear regression a few different ways, leaving out different input variables to try and back into feature importance, but that will be a challenge in the model evaluation phase.

My theory is that the additional data I am bringing into this challenge (trailer views, twitter mentions, using both RT audience reviews and Wikipedia pageviews together), will be more accurate than prior iterations. I think both the CNN and linear regression with clustering techniques will be able to account for the different release patterns (greater than or less than 1k theater releases), but I need to reduce the errors (predicted versus actual box office) I saw in unit 3 and unit 4 capstones. This is also the first time I am going to

**Additional detail**:

* RT reviews – scores calculated from top critics
* RT audience reviews – text from first 1000 audience reviews
  + Clustering technique applied to reviews to group similar films
* Metacritic scores – scores calculated
* Director, actor, distributor clustering from BOMojo
* Number of theaters information from BOMojo
* Output variables – total box office and opening weekend box office from BOMojo
* Film trailer views from YouTube
  + Get as many as possible: calculate min, max, average
* Wikipedia pageview data – use metrics from unit 3 capstone to identify time before release
  + Clustering? PCA? Try both techniques to come up with higher correlated variable to box office
* Crawl Facebook for posts and followers
  + Film-specific hashtag from theMovieDB API
  + How to get historical or time-specific mentions from hashtag through API

**Example of TheMovieDB API /movies request:**

