Film Analytics

Data Dawgs

(Siddhant Sutar, Dalton Childers, Neil Matin, Abhimanyu Tomar)

PROGRESS REPORT 2

Achievements

- Previously, we fetched movie data for 1.1 million movies from OMDb and wrote it into a csv file (approx 500 MB size).
- Since then, we have tried to implement the multiple linear regression model for building our movie ratings predictor.
- We consider independent variables like Rating (for age, like PG-13, R, etc.), Genre(s), Director(s), Actor(s), Writer(s).
- The dependent variable is the IMDb rating, which is what we are trying to predict.
- Initially we have a Pandas dataframe which contains feature vectors for the independent variables; then, we use get_dummies to create a dummy feature vector, which is then concatenated into the original dataframe.
- Next, we used the linear regression model using the scikit-learn module, which "fits" the x and y dataframes corresponding to the feature columns and the prediction column; here the dataframe x would contain only those columns from the original dataset, and y would correspond to the dataframe that contains the dependant variable (IMDb rating).
- We have our new predictor model using multiple linear regression.
-

Future Plans

- The rating with certain predictors crosses the 10 point mark - needs to be fixed.
- To consider different kinds of machine learning models and see which is more effective and gives a lesser RMSE error.