4502 Group 4 (in progress)

Title: Box Office Revenue Prediction

Team members:

- Melissa Pollich
- Nhan Nguyen
- Benjamin Breyfogle
- Heather Besch

Description:

- In a world, where movies made an estimate of $41.7 billion in 2018, the film industry is more popular than ever. But what movies make the most money at the box office? How much does a director matter? Or the budget? For some movies, it's "You had me at 'Hello. In this competition, you're presented with metadata on over 7,000 past films from The Movie Database to try and predict their overall worldwide box office revenue. Also, the data points provided include cast, crew, plot keywords, budget, posters, release dates, languages, production companies, and countries. Furthermore, you can collect other publicly available data to use in your model predictions. These will be used to predict a movie's worldwide box office revenue.

Prior Work:

- There have been several studies that have been performed using data mining techniques to predict movie success
  - https://www.researchgate.net/publication/282133920_Predicting_Movie_Success_Based_on_IMDB_Data
  - https://pdfs.semanticscholar.org/6d4f/1003fd164ffe30e2e45dd252715efecf9e61.pdf

Dataset: (in progress)

- https://www.kaggle.com/stephanerappeneau/350-000-movies-from-themoviedborg
- https://www.kaggle.com/eliasdabbas/boxofficemojo-alltime-domestic-data#boxoffice.csv
- Looking for revenue, month release, budget, actors…

Proposed work:

- Data cleaning: Remove attributes that will not be used in mining
- Data preprocessing: Normalize data and join across datasets for one complete set with which to work
- Data integration: possibly include other data sources (TBD) (IMDB, Box office Mojo?)

List of tool(s)

- Github
- Python w/ packages for statistical analysis (dataframe, pandas, numpy, sklearn)
- Logistic Regression, Linear Regression, SVM Regression
- Neural network prediction

Evaluation:

- We will prepare a training set and a testing data set for our analyses. This will allow us to determine our models' accuracies
- We will be able to evaluate the relative contributions of different factors affecting movie revenue at the box office.