Nisa Gurung & Kristel Tan
CS591 Data Science with Python
Submission Date: Friday, November 4

Project Proposal

1. *Discuss the dataset and their nature. How you got them, how they have been collected, what are their main characteristics, if you need to pre-process them, etc.*

   Our final dataset is a CSV file containing attributes of 5000+ movies from IMDB. We found our dataset from Kaggle, which was created by data scientist, Chuan Sun, from New York. In a step by step instruction of how he gathered the data, we learned that Sun used the scrapy Python library to manually scrape movie information from IMDB's website and aggregated it into a publicly available CSV.

   This file contains information for each movie such as the title, director, movie release date, featured actors, gross profits, IMDB rating, movie duration, keywords describing the plot, and Facebook likes. We will not need to pre-process this information because they are already in a format that can be easily analyzed.

2. *Expected analysis on the dataset (this of course is something that may change during the project.) What kind of techniques you plan to use.*

   We plan to use a decision tree based classification technique, Hunt's Algorithm, to analyze the dataset. The class to predict is whether or not a movie will be successful based on its given attributes. We will perform this algorithm on the top three released movie genres in the US. Three indicators of success are the gross profit of the movie, its IMDB rating, and movie Facebook likes.

3. *Application. Which application is associated with the dataset and how it will benefit from your analysis.*

   We believe there are two main parties that will benefit from this analysis. They are the film industry and the advertising industry. Based on our analysis, filmmakers, directors and actors can determine whether or not they would like to pursue a movie idea as well as how to present it most effectively to their target audience. These factors will also influence how the movie is marketed to the general public.

4. *Expected results (that also can change after you do the actual analysis).*

We expect that the attributes that contribute to the success of a movie release will differ based upon the genre it is in. For example, we predict that a romantic comedy that is long in duration may not do as well as a short one because movies of this genre do not typically have a complicated plot. Another attribute that we think will affect a movie's success is the number of faces in its movie poster. We predict that an action movie, for instance, will do better with more faces on the poster because action movies often have many characters involved. A more general hypothesis is that the popularity of the actors starring in a movie and a high film budget will also inevitably boost its success.