Network Tour of Data Science - Project proposal
What Impacts the Success of a Movie?

Célia RAPOSO, Valentin KINDSCHI, Yong Joon THOO

November 27, 2017

Goal of the project

From the start of the declaration of a new film, film companies try to build up hype for their future movie to gain attention. However, gaining much hype or having many page views does not necessarily give a good indication of the success of a film. Indeed, in recent years, multiple movies have crashed at the box office despite having reasonably well known actors and a big budget[1]. The goal of this project is to be able to provide a tool enabling people to try to help evaluate the eventual success of an upcoming film based on multiple features such as its cast or its genre by comparing them to the features of past successful or non successful films in IMDB[2] (Internet Movie Data Base).

Data Acquisition

Most of the data will be provided by a dataset[3] obtained from kaggle.com which contains numerous features about each movie e.g. cast, budget, genre, IMDB's rating, writers, popularity, gross, etc. It also provides a movie's id which can be used with the IMDB API[4] to find additional information. The YouTube Analytics API[5] will also be used to collect the number of views of the trailers before the release date of a movie. The data will then be cleaned by selecting the movies and the features that are the most relevant.

Data Exploitation

To explore the data we will build a network where the distance between nodes would give an indication of the similarity (based on the features discussed above) between the collected movies. The relevant features will then be projected onto lower dimensions to observe the impact/importance of each of these features when it comes to the success of a film. For this idea to work, there will be a good proportion between successful and unsuccessful movies in our dataset to observe trends and eventual clusters in our data. Once this has been accomplished, attempts will be done on a selected testing set to see how robust our algorithm is when it comes to predicting their outcome.

[1] http://ew.com/movies/2017/09/04/summer-box-office-fail/

[2] http://www.imdb.com/

[3] https://www.kaggle.com/rounakbanik/the-movies-dataset/

[4] https://www.theimdbapi.org/

[5] https://developers.google.com/youtube/analytics/