Movie and TV Show Dataset  
Project Scope

Magnus Hjorleifsson and RamKishore Karuppiah Nadar Venkateswaran

1. **Background:**
   1. **Magnus Hjorleifsson (mrh369)**

I have a background in computer science, finishing my computer science undergraduate degree at Drexel University earlier this year. I have worked with several APIs in the past, both in past courses and co-ops over the last couple of years. However, I did not do too much complex data pre-processing as typically I was only pulling a couple fields of information and instantly working with the data. I have completed several software engineering projects and several projects involving data structures, all with varying aspects within Python.

In this project, I am aiming to grow my data structure familiarity with big data specifically. Any time I have used data, it always seemed to be in small sets which are very easy to debug. Working with larger data sets presents bigger issues and more edge cases to account for. In addition, I hope to increase my familiarity with APIs, specifically in python (I used other languages with the APIs in the past) and look forward to using multiple sources of data to combine into a single data set. Finally, I hope to expand my experience with web scraping as I have never done it before, and hope that will provide useful information to the dataset.

* 1. **Ram Kishore Karuppiah Nadar Venkateswaran**

I have completed my bachelor's degree in engineering with a specialization in electronics and instrumentation. I have worked in industrial automation companies and as a front end web developer in the past. During my undergraduate degree, I learned basic python and data structures. However, I have only a little experience in working with data.

In this project, I am aiming to expand my knowledge in data collection methods, APIs and data processing. With this project, I hope to learn about web scraping, data integration and grow my familiarity in data structures and APIs.

1. **Goals**

This data will be used primarily to pull information from all different movies and television shows including ratings, genre, starring actors, titles, revenue, budget etc. From this dataset, we hope to provide insightful information to movie and TV show watchers and creators (screenwriters, directors, actors, etc.) as to what the perfect movie or TV show would entail based on the genre, actors, budget, ratings and everything else in between. In addition, we can also generate which features of a movie or television show makes it most financially profitable by finding the piece with the highest budget to revenue ratio. Finally, we can determine who the best director or the best movie stars are under specific genres. Laterally shifting, we can find the best actors and the most successful genre is for TV shows. Our goal is to have all of the information required to answer these questions in one, full, comprehensive dataset.

1. **Audience**

While this dataset can be used for a plethora of different reasons, my primary intended use for this dataset is for movie and TV show screenwriters directors to have enough information to create their movie or TV show based on parameters that are most enjoyable to the average viewer.

1. **Data**
   1. **Types**

We plan to use two different primary strategies to pull information to create this dataset. For a primary API source, we plan to use the movie database. The movie database (<https://www.themoviedb.org/>) is a massive API with information on millions of movies and tv shows. Information ranges from the title, ratings, and genre, to directors and the starring cast. While the quantity and detail of this API seems to be remarkable, we would like to accompany this API with a web scraping technique as well, pulling from IMDB, Wikipedia, and other movie sources for additional information such as the program's revenue and budget. We can also use web scraping to ensure the accuracy of the data, as we know one data source may be skewed, but having an extra one or two sources to corroborate the data may be the best way to get the most accurate data.

* 1. **Examples**

Data gathered from the API will be in a JSON format. JSON as we know will take a dictionary structure, and can be searchable through keys. Each feature of the movie will be a key in each element of the dictionary, while we can perform operations on the values. Although I have never done web scraping before, I assume we will need to parse HTML which will give us string or integer return value.

* 1. **Access Rights**

In order to access this data and create valid OAuth credentials, an account must be made on the movie database website. The account requires an email address, username, password, while the API requires some additional information including information about the app we are developing, an address, and a phone number. Depending on the additional websites we decide to use in conjunction with IMDB and Wikipedia for web scraping, we may need accounts on several news sites to obtain more information from web scraping.

* 1. **Limitations**

We might be limited on what data the API is able to cover, and may need to supplement the missing pieces of this data with web scraping. We may also be limited to the number of API calls per day, but I have not encountered this information on the movie database website above. All I found is that we are unable to make more than 40 calls every 10 seconds.