CS105 Project Proposal

• Team members names and Github accounts

Tongyuan He Github: the1323, Link: https://github.com/the1323

Boning Li Github: BBBonnie, Link: https://github.com/BBBonnie

• Which datasets (minimum is 2) you plan to use and how you will obtain them (crawling, API, download).

1. IMDb:
   a. Data Files Downloads: https://datasets.imdbws.com/
      description of the dataset attributes is located on the next page and also at https://www.imdb.com/interfaces/
   b. Web scraping API (Movie Database (IMDB Alternative)) : https://rapidapi.com/rapidapi/api/movie-database-imdb-alternative/
      1000 Free requests per day, This API only use as a helper to gather information on specific titles.
2. Box Office web
   a. Web scraping: https://www.boxofficemojo.com/chart/ww_top_lifetime_gross/?area=XWW&ref_=bo_cso_ac
   b. Box Office Mojo API https://github.com/skozilla/BoxOfficeMojo
      This API uses python for web scraping movie information from www.boxofficemojo.com. It uses movie titles for searching and provides responses in JSON format of webpage information.

• Description of how the datasets are correlated, what information they provide, and the type of analysis you plan to perform.

We will be using portions of name.basics.tsv.gz, title.akas.tsv.gz and title.ratings.tsv.gz files downloaded from IMDb Datasets to gather the movie info (title, language, votes, rating, etc.):
(Bolded words are data variables that might be useful)

title.akas.tsv.gz - Contains the following information for titles:

- **titleId (string)**, , **title (string), language**, ordering (integer) **,** region (string), types (array), attributes (array), isOriginalTitle (boolean) original title

title.basics.tsv.gz - Contains the following information for titles:

- **tconst (string)**, **titleType (string)**, primaryTitle (string), **originalTitle (string)**, **isAdult (boolean)**, startYear (YYYY), endYear (YYYY), runtimeMinutes, genres (string array)

title.ratings.tsv.gz – Contains the IMDb rating and votes information for titles

- tconst (string), averageRating, **ratings, numVotes**

Additionally, we will use a Web scraping API (Movie Database (IMDB Alternative)) to gather movie information of specific titles that may not be practical to search in the data files that's over 2 Gigabytes. We will also scrape movie titles, year, their Worldwide Lifetime Gross, Domestic Lifetime Gross, and Foreign Lifetime Gross from Box Office Mojo (https://www.boxofficemojo.com/chart/ww_top_lifetime_gross/?area=XWW&ref_=bo_cso_ac) to compare with the rating data, voting data, language data, etc. to determine how these data affect the Lifetime Gross of a movie Domestically and foreignly.

We plan to perform Exploratory Data Analysis from the data visualizations (tables and charts) we generate from the datasets. We will make hypotheses throughout the first two phases of the project while discovering interesting facts and exploring in phase three.
In the end, we will perform hypothesis testing to provide a better understanding of data set variables and the relationships between them.

Some interesting analysis:

Which movie category is most popular in different year cohorts?
Which year cohort makes the best worst movies?
The best and worst movies people rated in different regions of the world.
Movies that have higher ratings also have higher Lifetime Gross?