VISUALIZATION FINAL PROJECT PROPOSAL
TOPIC (DATASET) : **IMDB 5000 MOVIE DATASET**

BY: Namita Mhatre (110929172), Pratik Gaikar (110937719)

There are so many things we could do with the data we obtain from existing movies. For example, let's consider the Netflix data. Let's say we are trying to visualize something like: favorite/most watched movies of each region/state. We would get some really cool visualization as follows:

Similarly, how can we gauge the success of a movie before its release? We usually rely on critics or our own instincts which are both not definitive and are based on a personal opinion. So, we decided to work on and find some amazing results from the "5000 IMDB movies dataset" which is available on Kaggle. IMBD is one of the world's most popular source for movie, TV and celebrity content. This dataset has information about approximately 5000 movies, all extracted from IMDB. And, we have 28 attributes to describe each movie. Following are the attributes; Name of the movie, Color, Director, actors, number of critic reviews, duration, Facebook likes, director Facebook likes, actor Facebook likes, gross income, genre, plot keywords, language, country and many more. We can notice that some attributes (like actors, genre etc.) play an important role in a movie's popularity, whereas attributes like color, don't contribute much. So, we would like to work on this data to come up with some interesting results.

So, our goal is to try to find correlations between attributes that can help us predict about a movie's success, or to try to find the top most important variables playing part in a movie's popularity. In this process, the measure of success and popularity would be the positive reviews by critics, likes by people, etc. So basically we are trying to come up with interesting facts about how a movie's success can be measured.

Problems:

This is a very huge dataset with a large number or attributes. There are many hypotheses that can be formed. As of now, we have found of the following problems that we would like to work on.

1. "Given that thousands of movies were produced each year, is there a better way for us to tell the greatness of movie without relying on critics or our own instincts?" This is the most important question we would like to answer. Can a movie be judged without human inputs like critic and user reviews, but solely on facts?

2. This is an interesting problem. "Will the number of human faces in movie poster correlate with the movie rating?" So normally it seems like these two attributes won't be correlated at all. So let's try finding that out.

3. "How is the curve for a director when we try to plot it with the gross income for every year?"

4. "Which genres are the most famous and do they depend on the country?"

5. "What drives gross?" I bet everyone in the movie industry would be trying to figure this!

6. "What is predictive model for duration?"

7. "Do movie posters really matter?"

These are some of the base questions we want to answer. We would like to keep on adding more interesting hypotheses along the timeline.

Approach:

1. **Data**

   We have 28 variables for 5043 movies, spanning across 100 years in 66 countries. There are 2399 unique director names, and thousands of actors/actresses. This was scraped from IMDB and made available on Kaggle.

The data is in the .csv file format and since it has 5000 data points with total 28 variables we feel that it would be interesting dataset to analyze.

1. Movie title, release date and duration
2. Country of movie, language and aspect ratio
3. Genre of the movie (Action/Comedy/Sci-Fi)
4. Content rating (PG/PG-13/R)
5. Plot keywords
6. Gross income of the movie (in respective currency)
7. Budget of the movie
8. IMDB rating
9. Names of the three actors in the movie
10. Total Facebook likes for all the actors
11. Total number of users who rated the movie
12. Actual number of critics who rated the movie

Preprocessing:

The data would need some preprocessing for filling in the missing values. Not all the directors or actors have Facebook pages and accounts and so 'Facebook' column has some missing values or default values of 0 in that column. We plan to use the Facebook API's for collecting such missing Facebook values. Also, people have some variation in their names on Facebook pages so we plan to try some combinations of that also.

We need to normalize the gross income and budget for movies using web scraping libraries

2. Analysis

We plan on first finding the most important attributes (Dimensionality reduction) using PCA (Principle component Analysis) or MDS. Also, for visualizing the data, we would like to sample it so that the visualization does not look messy. We would also be using some machine learning techniques like clustering to try to prove some hypotheses.

3. Visualization

We would like to make the best use of resources (d3) to visualize our data. We are planning to visualize data with different types of graphs and charts. We are planning to show this entirely in a single web application with an aesthetically pleasant User Interface.

1. Plotting the correlation matrix, PCA components and clusters
2. Building the histogram/bar chart of different attributes like budget, gross income.
3. Building a word cloud of the plot keywords and see the most frequently used plot keywords
4. Dendrogram for directors, actors and movies.
5. Bubble chart using Genre or Language
6. Increased in movie budget on time scale using histogram.

Technologies involved:

We will be using python for the preprocessing and analysis part, the d3 for visualization part and web technologies (HTML, CSS, XAMPP) for the UI part. For the database part, we plan to use .csv format files as it is much easier to process.