**Movie Dataset Analysis**

**Problem Statement** : Movie dataset analysis

**Team size** : 3

**Team members** :Ramya Ramanathan, Aakanksha Ashok, Malika Makker

**Problem statement defined in simple terms** : To draw intelligence and trends from the Bollywood Dataset consisting of poster, trailer, wikipedia and script data.

**Technologies/Platforms/APIs planned to use** : Python 3.5, IBM Watson, nltk, wordcloud, scikit, matplotlib, shapely etc.

**Role of each team member** :

Malika Makker – Word cloud, get frequent characters, and graph based trend analysis

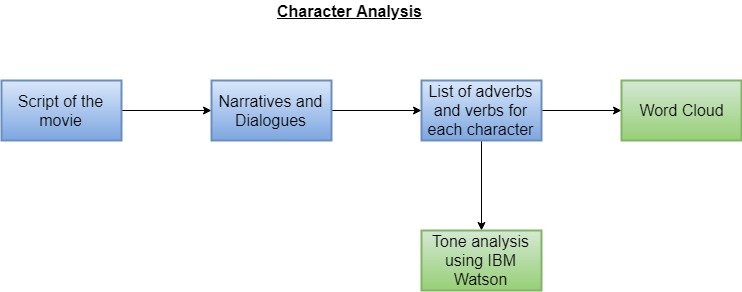
Ramya Ramanathan – UI, Clustering and graph based trend analysis

Aakanksha Ashok – System integration and graph plotting

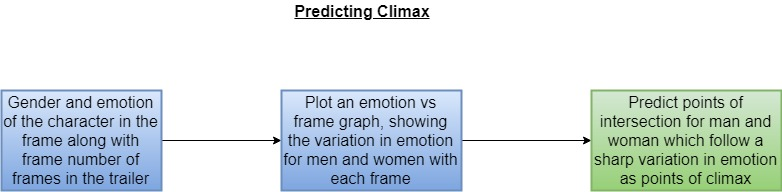
**Scope of work in the opted problem statement** :

For the dataset provided, we decided to do the following analysis:

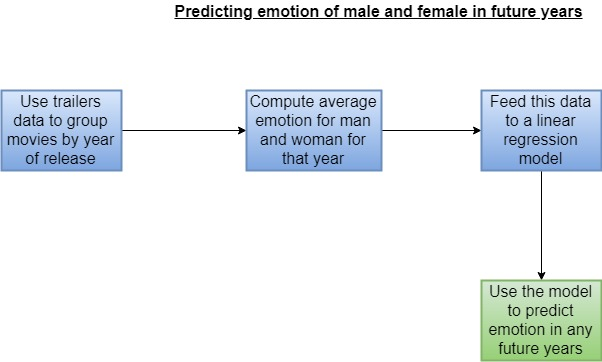
1. Character analysis of various actors in a movie using the script of the movie and displaying the result in the form a word cloud
   * The script of the movie is tokenized and divided into dialogues and narratives
   * The narrative is analysed to pick the top three main actors in the movie
   * The verbs and the adverbs used for a particular character are put in respective lists
   * Those lists are used to create word clouds of different actors, giving an overview of the character being portrayed in the movie
   * Further, IBM Tone Analyzer API is used to find the tone of the ending of the movie as well as that of the character.



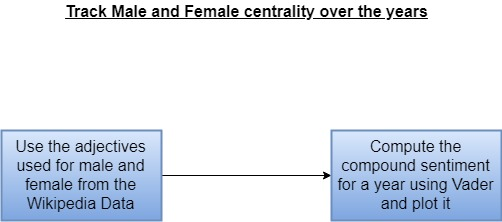
1. Predicting climax of a movie
   * To predict the climax of the movie, we used the trailer data of individual movies present in the *individual-trailers-data* folder. The data consisted of frame number, emotion of the frame and the gender of the actor present in the frame.
   * We made a range of various emotions varying from negative to positive. Angry being 0 and happy being 6
   * The graph was plotted with frame number being x-axis and emotion being y-axis
   * The graph had a line for the emotion variance of both man and woman
   * The points where the emotion of both man and woman intersected and showed a great variation after the point of intersection was predicted as the frame of climax



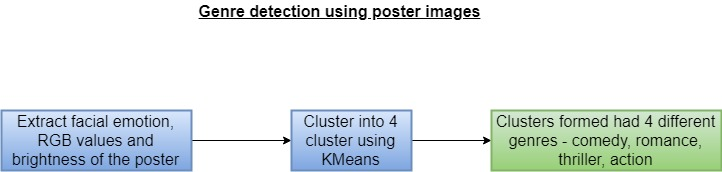
1. Predicting emotion of male and female in movies in future years
   * To predict emotion, we used complete data dataset available in trailers datawhich contains male and female emotion information over the years
   * After grouping by the year of release, the emotion was scaled using above mentioned scale having 0 for anger and 6 for happiness and plotted.
   * The trend was then extended to year 2020, by using the Linear Regression Model.



1. Tracking male and female centrality using Sentiment Analysis
   * To predict emotion, we used commonly used adjectives of male and female over the years, present in Wikipedia data
   * To get the value of emotion, Vader sentiment of the nltk package was used on adjectives..
   * The compound score of sentiment was plotted for men and women.



1. Genre detection using information extracted from poster images
   * From the images of poster data, we extracted image features such as brightness, and average RGB values
   * The most prevalent emotion of the people in the poster is also calculated.
   * The data is then clustered using K-means algorithm to identify thriller, action, comedy and romance as the respective classes.
   * These classes were selected because of their significance in Bollywood industry



6. Songs DB male vs female plotting

* SongsDB.csv file was used for the analysis.
* The movies were grouped by the year of release.
* The male and female singers count for each of those movies was summed and plotted for that particular year.

Business Value:

* Through the male centrality analysis, we were able to predict the average emotion that would be seen in the future years according to the trend, which might eventually lead to a change of approach in script writing.
* The genre classification module helps identify the image characteristics that are typically expected by users from each type of genre
* The climax detection module helps to estimate the positions at which climax is generally found, and hence identify the best impact of the same in trailers.
* Character Analysis helps to better the script and gain deep understanding of a character and their behavior/characteristics.
* A change in adjectives used for women and men used over the years, is used to identify the trends of similarities and dissimilarities in the sentimentality used for each gender across the decades.