Project Proposal

Rebeka Mukherjee, Archit Rathore, Yash Gangrade

1. What data we plan to use and where we plan to get it from:

   For our project, we want to study gender equality in movie roles. We plan to use the data from `https://www.kaggle.com/rounakbanik/the-movies-dataset`.

   The dataset contains metadata for 45,000 movies that were released on or before July 2017. The data points include cast, crew, plot keywords, budget, revenue, posters, release dates, languages, production companies, countries, TMDB vote counts and vote averages.

2. What structure we want to mine from the data:

   The original dataset is represented in JSON. We want to convert this into a matrix. Then we want to cluster the data by movie roles and look at the distributions of the genders. We also want to find similarities between genres by comparing the distribution of gender in movie roles.

3. Why this problem is interesting:

   The problem is interesting because we think that representation of the genders on the screens reflects our perception of the genders in real life. This project could be useful in studying the socio-economic aspect of gender equality.

4. What is new, or what the intructor will learn:

   The topic of gender equality is very relevant today. Mining data from movies and using proper visual analysis techniques to represent the data is an interesting way to explore the topic.