Project Proposal

Rebeka Mukherjee, Archit Rathore, Yash Gangrade

1. How we obtained the data:

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

2. How large is the data:

   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.

3. In what format are we storing the data:

   The original dataset is represented in JSON. We want to convert this into a matrix.

4. Did we need to process the original data to get it into an easier, more compressed format:

5. How we would simulate similar data: