Project 1 Proposal

1. Summary and Motivation:

People nowadays love watching movies. I want to make use of the datasets to find out the top-rated movies in each genre, the relationship between the movie's genres, budget, casts and its reviews and etc.

2. Datasets:

1) The movies dataset that consists of the information on over 45,000 movies which released on or before July 2017. It includes posters, budget, revenue, release dates, languages, production countries and companies and so on as features.

https://www.kaggle.com/rounakbanik/the-movies-dataset#movies_metadata.csv

2) The movies dataset that contains the cast and crew information. It mainly records the character, gender, name of each person in the cast.

https://www.kaggle.com/rounakbanik/the-movies-dataset#credits.csv

3) The movies dataset that has 26 million ratings from over 270,000 users. The ratings are in the range of 0 to 5.

https://www.kaggle.com/rounakbanik/the-movies-dataset#ratings.csv

3. Data Manipulation:

After preprocessing the datasets (dealing with missing values), I will join the movies_metadata.csv, credits.csv, ratings.csv on their shared id column.

4. Large-scale Computation Tasks:

(first use Spark "flatmap" to get information regarding genres, genders...)

1) Top rated movies in each genre
2) Top high-budget movies in each genre.
3) Average female proportion/number in the cast in each genre ordered by the decreasing average female proportion.

I will use Mrjob/Spark/Sparksql to do map/reduce tasks.

5. Visualization

I will use R/Python to plot the scatterplot of rate against budget for each genre, barplot that takes genre as x and average female proportion as y, the plot that shows comparison of average rate and average budget in each genre.