# CS 6140 - Intermediate Project Report

Archit Rathore, Rebeka Mukherjee, and Yash Gangrade March 20, 2019

## 1 Progress towards goal

Our project is more inclined towards the exploratory nature. In this context, there is no quantification number to evaluate the progress made in the project. We consider trying multiple different techniques and comparing results from them, and also the ability to answer several queries on the database as the metric of effectiveness for our project. We have tried some of the basic approaches (discussed in the next section) and suggested improvements and modifications to them in the subsequent section. We have also mentioned the experiments we have performed and ones we are planning to perform in the last section.

## 2 Basic Approaches tried

### 2.1 Storing the data

Till the data collection report submission, we performed all the cleaning of the data by removing non-essential columns and the invalid data rows from it. We then developed several classes to store the data and it's properties in a easy to use manner. For instance, the Movie class contains all the detail like movie id, genre, release data etc. Similarly, we have a Cast class which stores the information about the cast related to a particular movie. We then have multiple helper functions as well which will return the list of male and female cast separately for a movie. This data storage will be extended to a graph as well where each vertex could be a Movie class object and the edges can signify the pairwise similarities or any other comparison method between the two objects.

### 2.2 Clustering

We are using different techniques of Clustering of the data we have after the required processing and storage. Before performing clustering, we had to get all the unique 'Genres', all unique languages from the films and so on. Once, we have that, we just enumerate these unique items in form of a dictionary. We use these enumerations later in forming the numpy array. To create a numpy array, we then iterate over all the movies in the dataset (i.e. about 45000) and then get the fields like revenue, vote average, etc. in the numpy array. Then, we pass this resultant numpy array to the scikit-learn Kmeans function. We are using scikit-learn 0.20.3 version and we are using the in-built Kmeans function. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html. Similarly, we are also using the Spectral Clustering defined in the scikit-learn library. We define the number of clusters as 20 here since we have 20 different genres. Apart from that, we are using the default affinity matrix and not using the precomputed options.

## 3 Improvement over the basic approaches

- We have tried KMeans clustering and Spectral Clustering over the numpy arrays we created using the procedure defined earlier. They don't perform as expected since the data we have at hand is categorical. Genres, Languages, classes of revenue etc. are all categorical data. With the KMeans and Spectral Clustering techniques, the maximum we can say is that two points are different but we cannot quantify that difference. So if we want to quantify that distance, we would like to use some kind of comparator function to compare the two points. Essentially, we would develop a graph where each vertex can be a movie class object and the edge weights would be determined by pairwise distance metric for eg. Jaccard Distance. Now, if we apply spectral clustering or k-means clustering technique, it would work a lot better than what we have tried with the basic approach. Also, we can define our own affinity matrix and precomputed options to get better results.

- Secondly, we will fetch the scripts of the movies we have in the dataset and then convert it to a bag-of-words, topic model etc. kind of model to get semantically meaningful results. We can apply different topic modelling, NLP techniques to the word vectors to identify complicated patterns and distribution across genres, genders etc.

## 4 Experimental Analysis

- We ran our experiments using KMeans and Spectral Clustering techniques as we discussed before. The results are not up to the standard because we are using the aforementioned clustering techniques on categorical data. So we tried the basic approach to this problem but to make sure our project is robust and effective, we will implement a graph based approach as we discussed earlier.

- We would also want to perform some of the famous Topic Modelling techniques on the script of the data to gather in-depth information and answer more number of questions.

- As we discussed in the progress section, our project is an exploratory project with multiple different perspectives and angles to look at the data and gather new information. There are no exact quantification numbers we are aiming for. We want to demonstrate the effectiveness of this project by answering different queries like how does male to female ratio vary based on classes like genre, revenue etc. We can also find and analyze frequency distributions over the years and genres.