EECS 731 – Project 3 report

9/28/2020

The values that are contained within ‘ratings’ and ‘tags’ can be used by the film production industry to find what the particular popularity for a certain type of movie is. During my manipulation of the datasets, I found that “Drama” was the most popular keyword contained within all of the data. This would indicate that dramatic films are more popular than “Thriller” or “Crime” movies. Based on these findings, if I were to make a movie, I would make a movie that was dramatic.

I began the project by reading three data sets into my program. I then merged all three of the data sets into a single data set. I removed columns from the set that I did not plan on using for future calculations and manipulation of the data. I then began to manipulate the data by counting the number of occurrences for tags within the data set. After counting and totaling the number of keywords within the single data set, I then displayed the first five tags in the form of a bar graph. Please note, The first two colors are the official colors of the Kansas City Chiefs, the next two are the official colors for the University of Kansas, and the last is the official color for Kansas State University.

After displaying the keywords data, I then found the mean of all ratings for each movie and displayed the first ten mean values that had been stored. I was able to display how many movies within the data set were associated with each of the rating levels. I then set the needed values as the data points and labels for the KNeighbors Classifier to train and fit the model. I was able to then make predictions using the values that were returned after training the model. Lastly, I used KMeans clustering to form the data and displayed the data in the form of a scatter plot graph.