Data Mining Task:

I decided on the topic similar to the previous Programming Assignment, in where I create a recommendation system using hybrid collaborative filtering, but for animes instead of movies (Japanese animated series). This is because as nowadays, despite anime becoming more mainstream, many new to the community prefer to stick to their interests and miss the opportunity to watch many other great shows or hidden gems, new or old.

Dataset:

The dataset used will be one I acquired from Kaggle, which was mined from MyAnimeList, a site similar to Imdb, where users rate and review shows they liked or disliked, on May 2018. This particular dataset was mined by Kaggle user, Azathoth:

https://www.kaggle.com/datasets/azathoth42/myanimelist

Methodology:

The data is going to be analyzed similarly to a content-based filtering system, and a collaborative-filtering system, in other words, a hybrid collaborative filtering recommendation system. It works by calculating the weighted average of the average ratings of each show, then summing it with the popularity score and number of users who favourited of the respective shows, using 33.3% priority to the three variables to determine a final recommendation score, which would then be sorted from highest to lowest with all relevant related information such as episode number or genres are listed among the recommendations, for each year. The formula used to determine weighted average is as such:

$$W = \frac{Rv + Cm}{v + m}$$

where:

W = Weighted Rating
R = average for the movie as a number from 0 to 10 (mean) = (Rating)
v = number of votes for the movie = (votes)
m = minimum votes required to be listed in the Top 250 (currently 3000)
C = the mean vote across the whole report (currently 6.9)

Sources: https://medium.com/@developeraritro/building-a-recommendation-system-using-weighted-hybrid-technique-75598b6be8ed | http://trailerpark.weebly.com/imdb-rating.html?source=post_page

Which is the weighted average formula currently used by Imdb ratings. The project will be written in Python, and using the pandas and numpy libraries for data frame editing.

Final Product:

The final product is expected to be a function that will, as long as the data is up to date, reliably recommend the top-rated shows for each year between the earliest and latest in record. To see if it succeeds, I can cross-reference it with the general consensus of the shows found online, or another rating site using different methods. This project would be a stepping stone for me to understand more of how more complex recommendation systems may function.