Week 1 – Understanding your data

Filip Piskor
12331436

This week we had to import a csv file into our program and calculate some basic statistics on the data provided.

Approach

I chose to load the ‘resutls.csv’ file into 2 HashMaps: users, items. This would allow me to load quickly and make the calculations simpler since I wouldn’t have to extract items from a HashMap after it has already been loaded.

I have decided that by using a HashMap I was able to have a mapping of userID to User object. The user object itself contains a HashMap containing a mapping of item to rating. Similarly to make it easier to calculate cumulative statistic I have created a HashMap with a mapping itemID to Item object. Each item object contains an ArrayList of every rating it received. Both User and Item objects contain methods which enable us to perform calculations to get respective statistics for each object. I could have used just HashMaps but decided that this approach will make it more readable and structured.

I have also decided to store statistics of each object in a JSON files (users.json and items.json) as it will allow for easy importation if ever need and also it allows for easy readability. Since overall mean statistics are a lot smaller I wrote them into a txt file for easy readability again. JSON in this case wouldn’t have made as much sense.

Findings

By looking at the density of the ratings matrix we can observe that the majority of the users rate a small selection of movies since only ~6% is rated overall. We can also see that the overall means of users and items differ by a small but significant amount.
When we look at the mean and median we can see that on average a user will be rating 3.5 but each item on average is rated only 3. This means that overall there is users who give very high ratings which bring the average up. This can be further seen in the total amount of ratings per rating. Here we see that a significant of ratings will be 4 and 5 which are about above the average rating which is 3.
There is a noteworthy difference between the maximum ratings given per user and per item. Here we see that users maximum rating tend to be very close to highest rating since it’s 4.98 while items maximum rating is 4.45 meaning that some items probably will have very few ratings which bring the average down. I expected both to be very close if not 5. This was a surprising but insightful result.