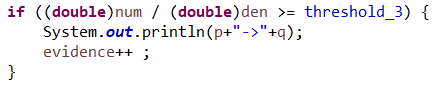
실전 프로젝트1 – Big Data Analisys

21600193 Kim Hyo Rim

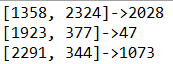
21600277 Park Ye Gyeom

1. Task2 - Find at least 3 rules that have high confidence values





Confidence 값이 0.8이상인 Set들을 프린트했을 때(일부)

[Sling Blade, Life is Beautiful] →Saving Private Ryan

[There’s Something About Mary, Speed] → Seven(a.k.a Se7en)

[Edward Scissor hands, Ace Ventura : Pet Detective] → Willy Wonka & the Chocolate Factory

이러한 결과로 인해

Find at least 3 rules that have high confidence values에 대한 3가지 rule은

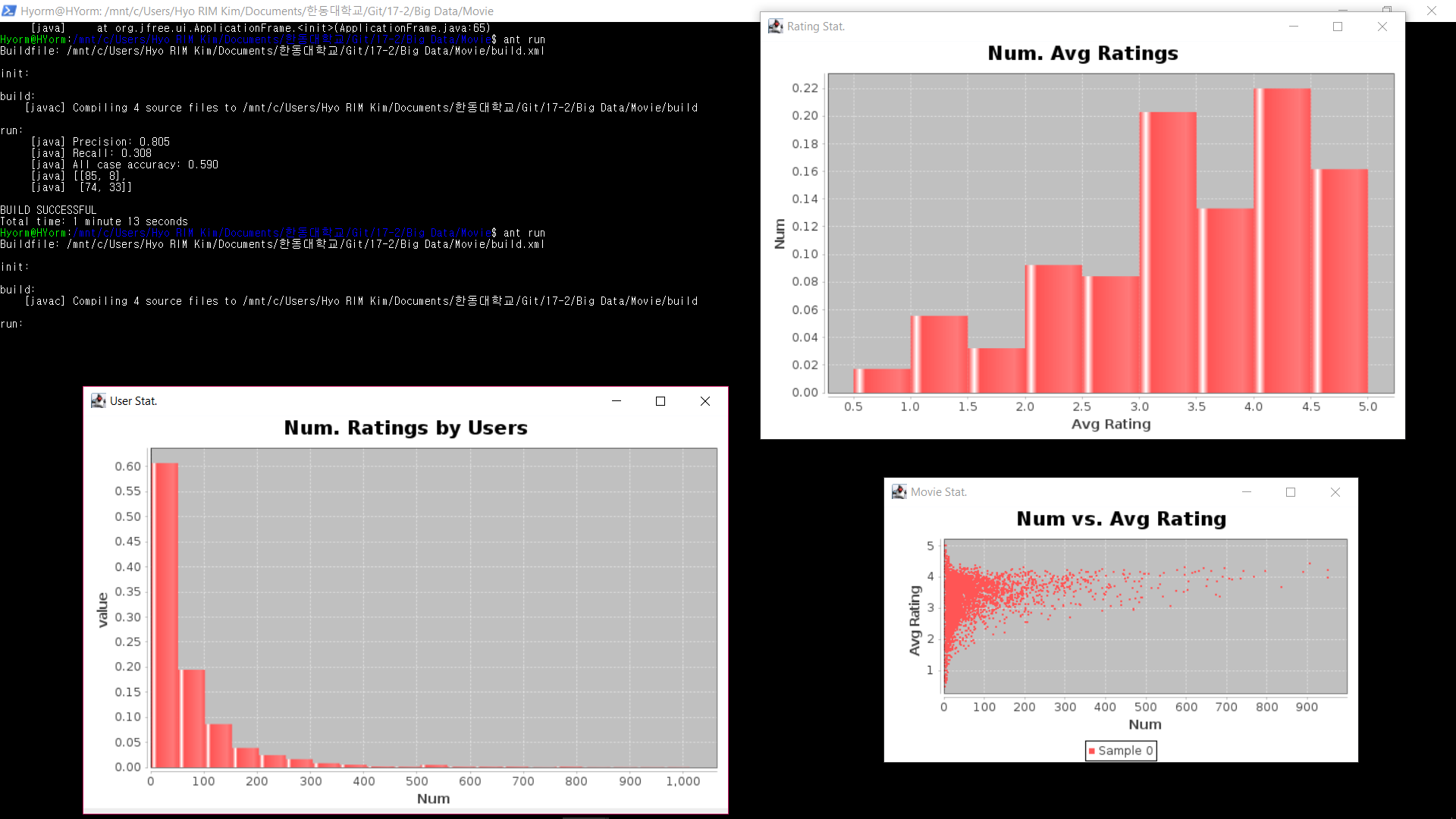
1. user의 평가 data가 많을수록 높은 confidence value를 갖는 조합을 찾을 수 있다.
2. 첫번째 조합인 [Sling Blade, Life is Beautiful] → Saving Private Ryan을 보면 세 영화 모두 장르가 drama인 것을 알 수 있다. 따라서 장르가 비슷한 영화들끼리의 confidence가 높은 값이 출력될 것이라고 기대할 수 있다.
3. 적절한 threshold 값을 사용하면 적당히 높은 confidence를 갖는 조합을 출력할 수 있다.

2. Task3 - Write at least one idea of improving the recommendation system

1) Analysis

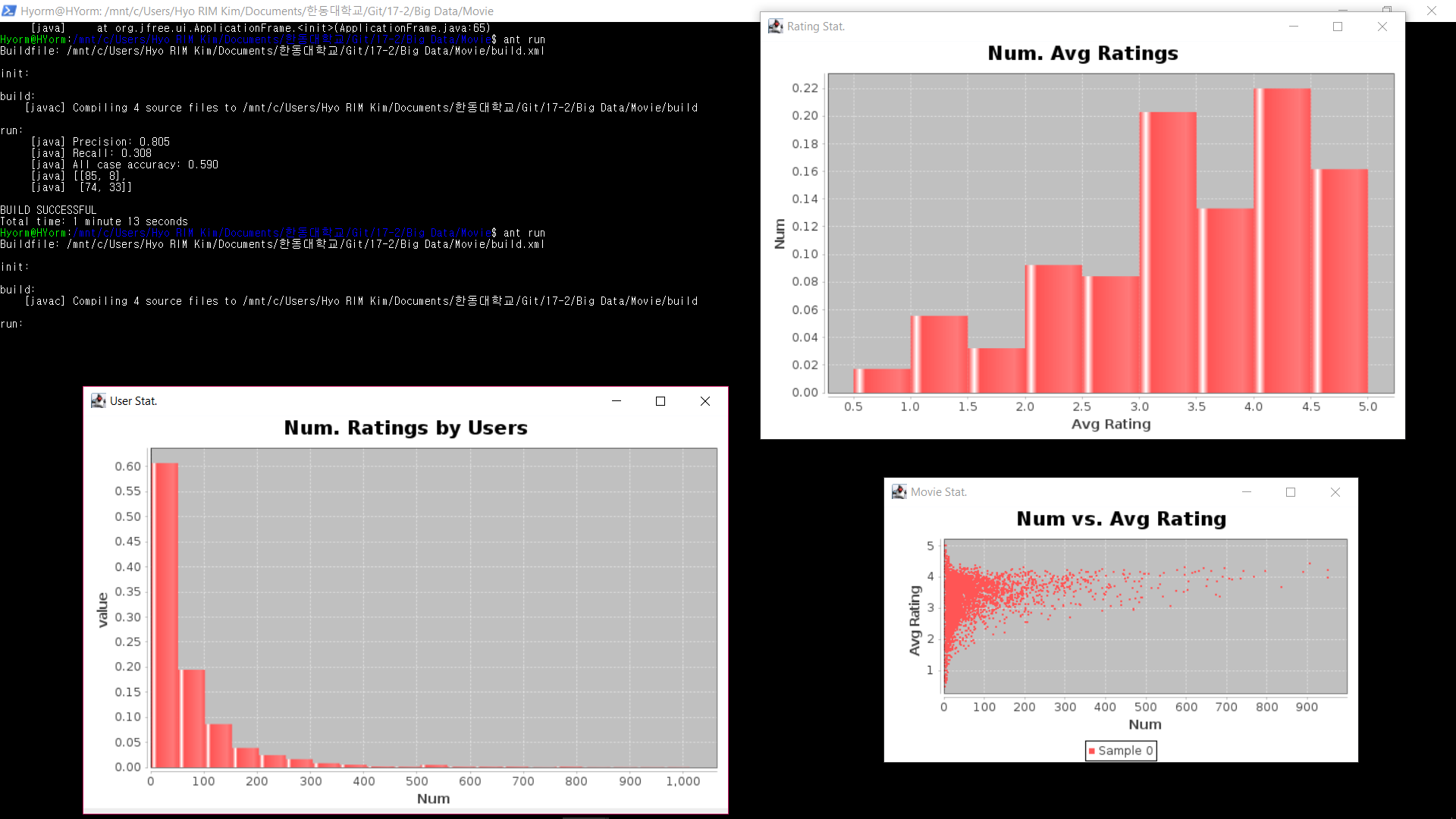
Add <arg value = "- d" /> to build.xml and run ‘ant run’, draw three gui charts for movie recommendation data.

(1) Movie Stat.

This graph shows the average rating of people for movies. Each point represents a movie.

The Avg Rating is divided into 0 ~ 5.0 units. If we look at the right of the xy plot, we can see that a lot of people have watched a certain movie and gave it a high score. These data are highly reliable. However, the left side of the xy plot is a rating that few people have seen. These data should consider personal preferences more then the xy plot on the left.

(2) User Stat

This graph shows the number of movie ratings each person has written.

Each unit is divided into 50 units. In the market-basket model, people who write too many ratings, or those who write very positive ratings, should do outlier processing. In the graph, about 80% people wrote 0-100 ratings. 20% of the people to the right of the graph have more ratings than 80% of the people, but they are outliers because they are too small. At this time, each value of the right group did not have a value of 10% which is half of 20%.

(3) Rating Stat

This graph shows where each movie's rating belongs and how many percentages each rating has.

Each unit is divided into 0.5 units. When the total movie set was 100%, the ratio of 4.0 to 4.5 was the highest. There are 4 groups with a rating of 3.0 or higher, and they account for about 70% of the total.

2) Conclusion

It is important to add other conditions to improve the recommendation system, but it is important to adjust the config.properties according to the scale of the data. The config values that affect the result are data.like\_threshold, data.outlier\_threshold, prediction.threshold\_2, prediction.threshold\_3, and prediction.min\_evidence\_3. At this point, the two parts are easy to recognize on the graph. The following is a way to automatically generate these values through the data and graph.

|  |  |
| --- | --- |
| Config Value | How to change |
| data.like\_threshold | It is calculated based on half of the average rating of upper movie rating set. |
| data.outlier\_threshold | Divide the user group into 50 user ratings. And subtract the percentage of the first group out of the total percentage. Divide the extracted ratio in half. If the ratio of the next group is less than half, the group other than the first group is treated as outlier. If more than half, repeat the previous step for the next group. |