Project 4 Amazon Movie Reviews

STATW4249_Project4_Team1_2016

Wednesday, April 13, 2016

Contents

Introduction

In this project, we used the Stanford dataset consisting of Amazon movie from Aug 1997 to Oct 2012 to build a movie recommendation system. We filtered the dataset to a smaller one by choosing users with more than 100 reviews and movies that has more than 100 reviews.

Example of movie reviews

- product/productId: B003AI2VGA
- review/userId: A141HP4LYPWMSR
- review/profileName: Brian E. Erland "Rainbow Sphinx"
- review/helpfulness: 7/7
- review/score: 3.0
- review/time: 1182729600
- review/summary: "There Is So Much Darkness Now ~ Come For The Miracle"
- review/text: Synopsis: On the daily trek from ...

Recommendation System

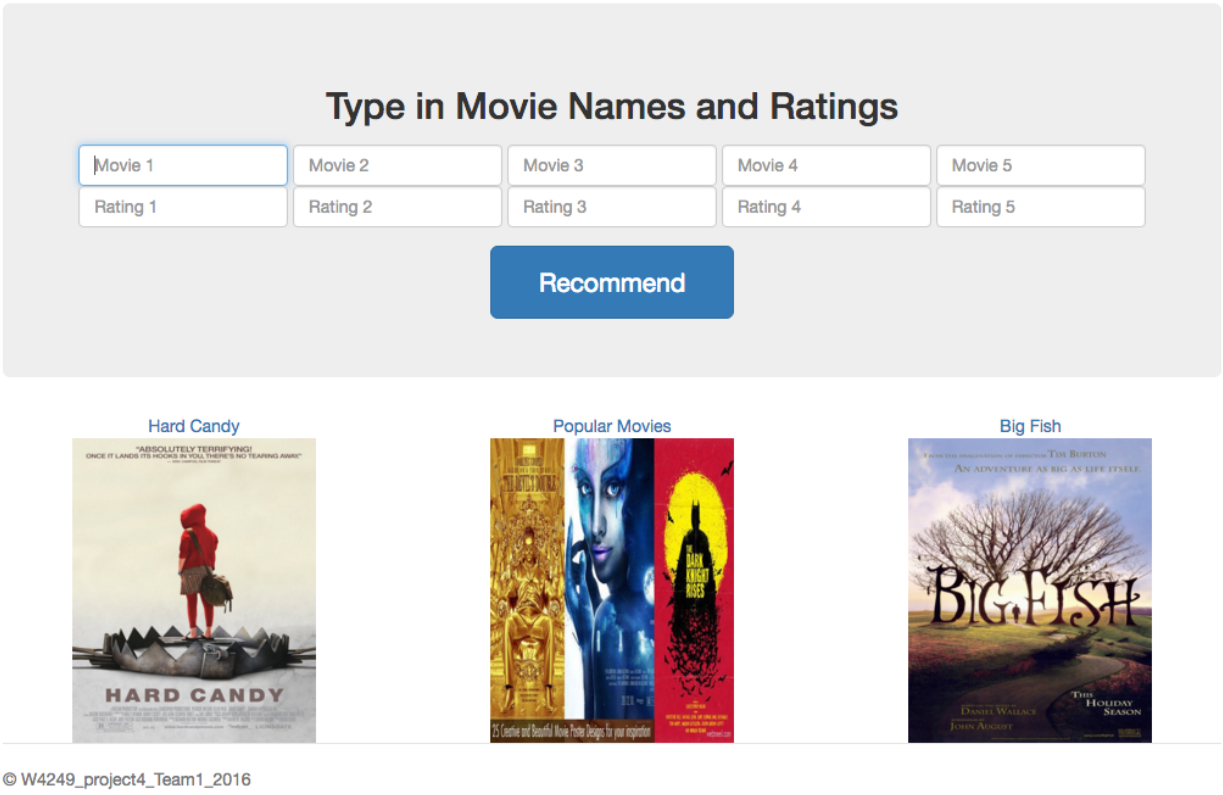
Algorithm

The recommendation system is `users-based` and it recommends the user three movies based on the movies feed to the system and their ratings entered.

When new user enters the movie names and the corresponding ratings, the recommender will look for the most similar users in the existing dataset by `cosine` similarity. This could be a collection of more than one similar existing users. If this is the case, the system will pick the top three movies from all the movies rated by this group of similar users. During the process, we used Amazon API to transform movie name to ASIN. At last, three movies will be returned by the system that are thought to suit the user's taste well.

Web app

Here is the screenshot of our movie recommendation web app.



Network Analysis and Visualization

Network analysis is the process of investigating social structures through the use of network and graph theories. Based on the similarity algorithm in recommendation system, we built the network of users and movies.

Network of users

```
## Warning: package 'arules' was built under R version 3.2.4
```

```
## Warning: package 'statnet' was built under R version 3.2.4
```

```
## Warning: package 'tergm' was built under R version 3.2.4
```

```
## Warning: package 'ergm' was built under R version 3.2.4
```

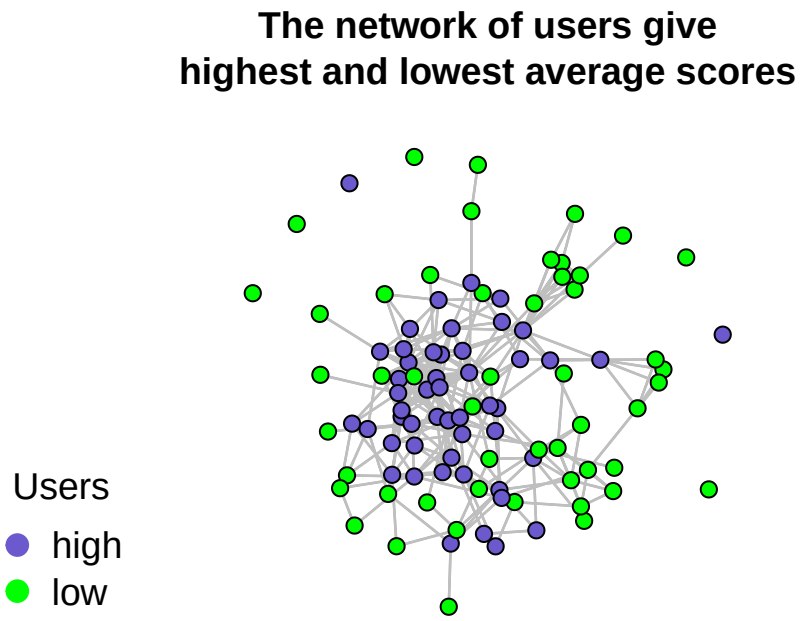
```
## Warning: package 'ergm.count' was built under R version 3.2.4
```

```
## Warning: replacing previous import by 'sna::%c%' when loading 'statnet'
```

Select 50 users who give highest average score and 50 users who give lowest average score.

Combine the two groups of people who give highest and lowest scores. Calculate the similarity matrix and build network. The vertexes of network are users. If the similarity of two users is greater than 0.05, there will be an edge between the two vertexes. Otherwise, there will be no edge.

Set `high` and `low` as vertex attributes and plot the network of users.



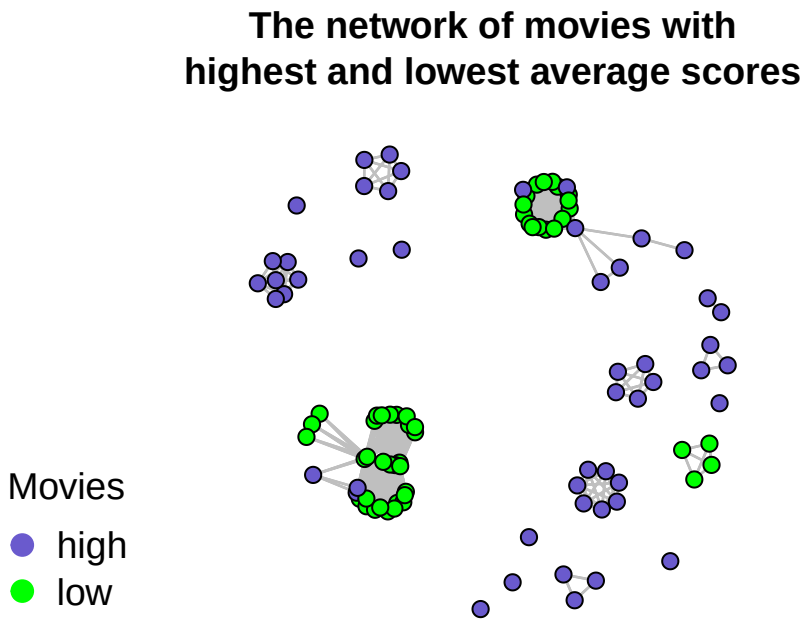
It is obvious that users who give highest average scores are highly related, while many users who give lowest scores are less similar to others.

Network of movies

Sort the movies table by average score, then select 50 movies with highest average scores and 50 movies with lowest average scores.

Combine the two groups of movies which have highest and lowest average scores. Calculate the similarity matrix and build network. The vertexes of network are movies. If the similarity of two movies is greater than 0.01, there will be an edge between the two vertexes. Otherwise, there will be no edge.

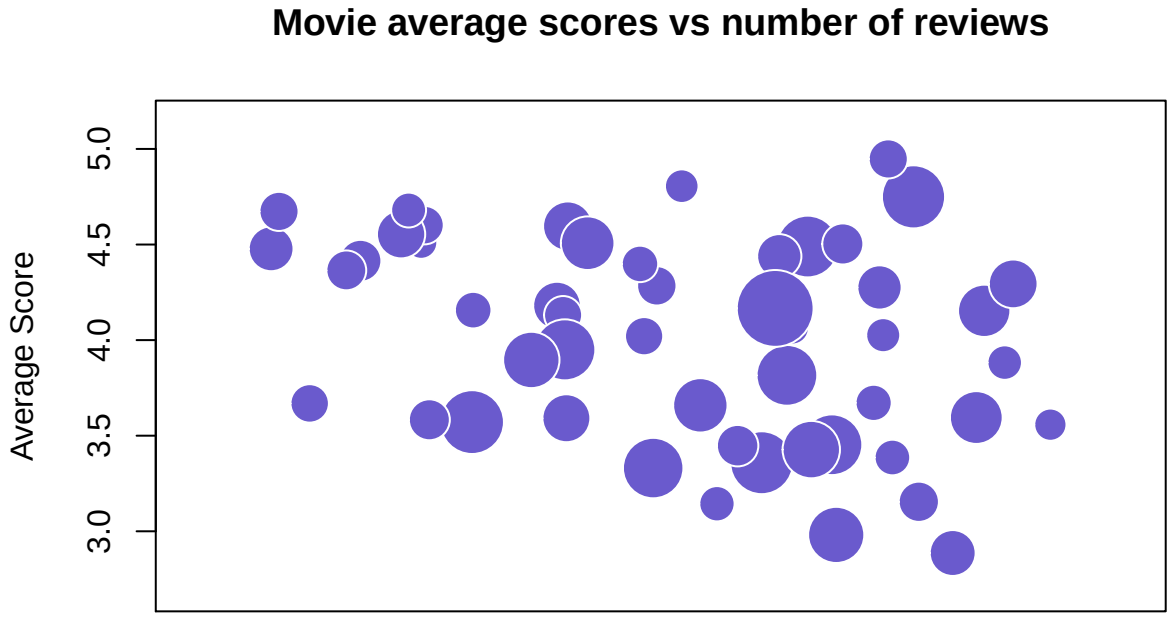
Set `high` and `low` as vertex attributes and plot the network of movies.



Movies are clustered in several groups. A few of movies with high scores are similar to movies with low scores. This phenomemon may result from the attributes of movies, like genre, director, actor, etc.

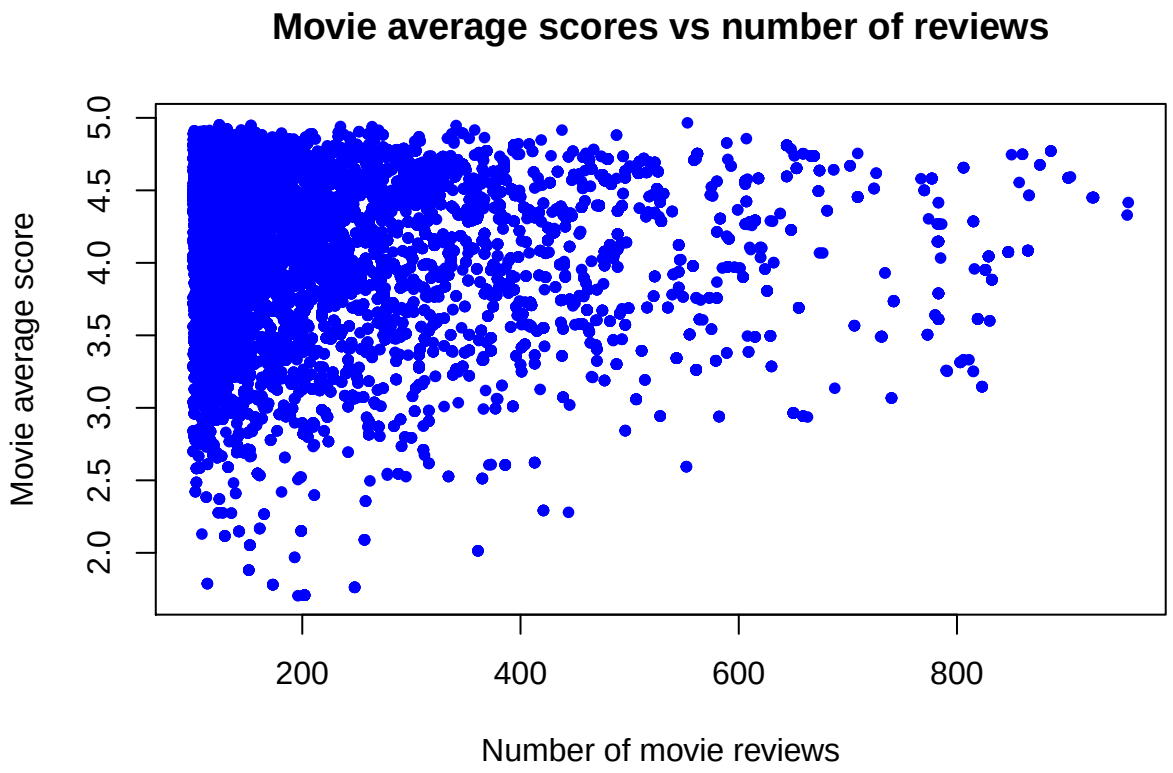
Other plots

Bubble plot of movie average scores and number of reviews



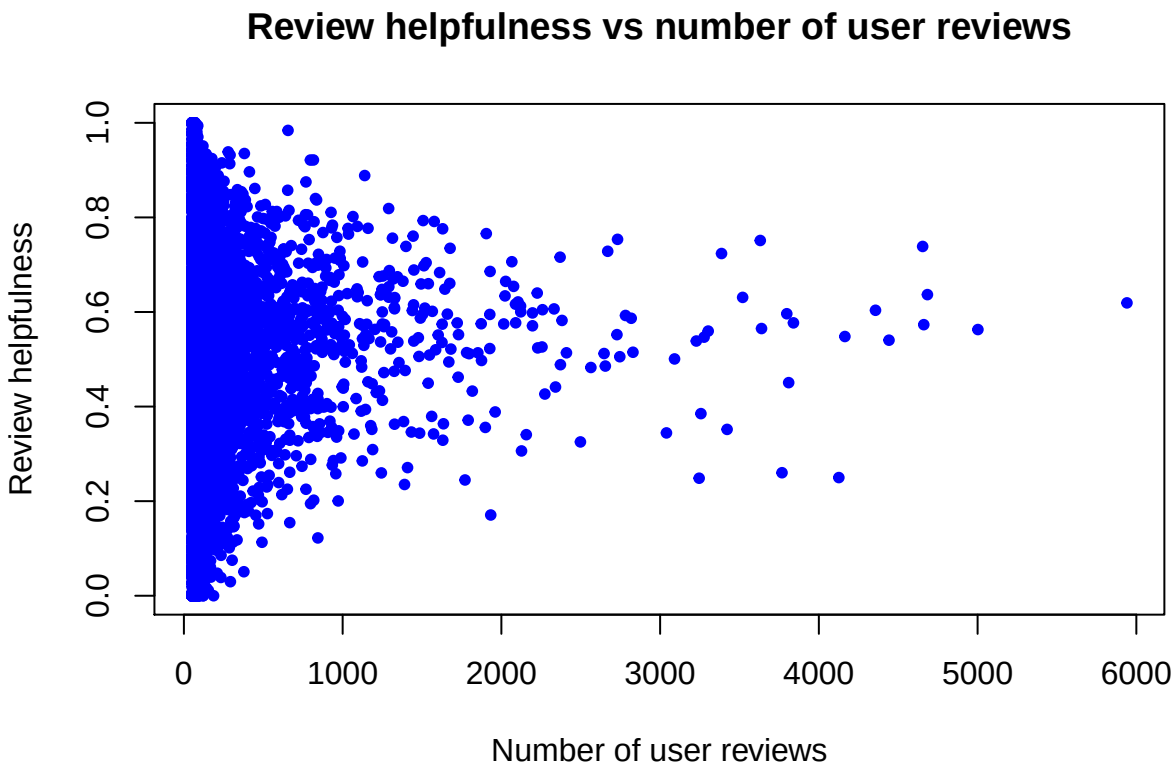
It seems that movies with more reviews tend to have higher average scores. In order to show the relationship more clearly, we plot the scatterplot of movie average score against number of movie reviews.

Movie average score against number of movie reviews



It shows that the more reviews a movie have, the less possible for the movie to have low average score. It is reasonable because better movies tend to be more popular, which leads to higher scores.

Average helpfulness of reviews against number of user reviews



As the number of user reviews, the average helpfulness of reviews converges to 0.6. In conclusion, it is less possible for users who give much reviews to have very high or low average helpfulness.