CS595 Intro to Web Science, Assignment #8

Valentina Neblitt-Jones

November 14, 2013

Instructions

The goal of this project it is to use the basic recommendation principles we have learned for user-collected data. You will modify the code given to you which performs movie recommendations from the MovieLense data sets.

The MovieLense data sets were collected by the GroupLens Research Project at the University of Minnesota during the seven-month period from September 19th, 1997 through April 22, 1998. It is available for download from `http://www.grouplens.org/node/73`

There are three files which we will use:

u.data

u.data: 100,000 ratings by 943 users on 1,682 movies. Each user has rated at least 20 movies. Users and items are numbered consecutively from 1. The data is randomly ordered. This is a tab-separated list of user id, item id, rating, and timestamp.

The time stamps are unix seconds since 1/1/1970 UTC.

Example:

user id	item id	rating	timestamp
196	242	3	881250949
186	302	3	891717742
22	377	1	878887116
244	51	2	880606923
166	346	1	886397596
298	474	4	884182806
115	265	2	881171488

u.item

u.item: Information about the 1,682 movies. This is a tab separated list of movie id, movie title, release date, video release date, IMDb URL, unknown, Action, Adventure, Animation, Children's, Comedy, Crime, Documentary, Drama, Fantasy, Film-Noir, Horror, Musical, Mystery, Romance, Sci-Fi, Thriller, War, and Western.

The last 19 fields are the genres, a 1 indicates the movie is of that genre, a 0 indicates it is not; movies can be in several genres at once. The movie ids are the ones used in the u.data set.

Example:

movie id	movie title	release date	video release date	IMDb URL
161	Top Gun (1986)	01-Jan-1986		http://us.imdb.com/M/title-exact?Top%20Gun%20(19

```
161|Top Gun (1986)|01-Jan-1986||http://us.imdb.com/M/title-exact?Top\%20Gun\%20(1986)|0|1|0|0|0|0|0|0|0
162|On Golden Pond (1981)|01-Jan-1981||http://us.imdb.com/M/title-exact?On\%20Golden\%20Pond\%20(1981)|(
163|Return of the Pink Panther, The (1974)|01-Jan-1974||http://us.imdb.com/M/title-exact?Return\%20of\%2
```

u.user

u.user: Demographic information about the users. This is a tab-separate list of user id, age, gender, occupation, and zip code.

Example:

user id	age	gender	occupation	zip code
1	24	M	technician	85711
2	53	F	other	94043
2	23	M	writer	32067
4	24	M	technician	43537
5	33	F	other	15213

The code for reading from the u.data and u.item files and creating recommendations is described in the book Programming Collective Intelligence (check mail for more details). You are to modify recommendations.py to answer the following questions. Each question your program answers correctly will award you 10 points. You must have the question answered completely correct; partial credit will be only awarded if your answer is very close to the correct one. Your output should clearly indicate the answers from the question you answered. Provide any relevant discussion.

Answers

1. What 5 movies have the highest average ratings? Show the movies and their ratings sorted by their average ratings.

2. What 5 movies have received the most ratings? Show the movies and the number of ratings sorted by number of ratings.

3. What 5 movies were rating the highest on average by women? Show the movies and their ratings sorted by ratings.

4. What 5 movies were rating the highest on average by men? Show the movies and their ratings sorted by ratings.

5. What movie received ratings most like Top Gun? Which movie received ratings that were least like Top Gun (negative correlation)?

6. Which 5 raters rated the most films? Show the raters' IDs and the number of films each rated.

7. Which 5 raters most agreed with each other? Show the raters' IDs and Pearson's r, sorted by r.

8. Which 5 raters most disagreed with each other (negative correlation)? Show the raters' IDs and Pearson's r, sorted by r.

9. What movie was rated highest on average by men over 40? By men under 40?

10. What movie was rated highest on average by women over 40? By women under 40?

Resources