|  |
| --- |
|  |
| CS432 Spring 2018 |
| Assignment 6 |
|  |
| **Jonathan Kruszewski** |
| **31/03/2018** |

|  |
| --- |
|  |

The goal of this project is to use the basic recommendation principles

we have learned for user-collected data. You will modify the code (https://github.com/arthur-e/Programming-Collective-Intelligence/blob/master/chapter2/recommendations.py)

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 22nd, 1998. We are using the

"100k dataset"; available for download from:

http://grouplens.org/datasets/movielens/100k/

There are three files which we will use:

1. 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 | timestamp

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

Example:

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

2. 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 | 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 data set.

Example:

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|0|0|0|0|0|1|0|0|0|0

162|On Golden Pond (1981)|01-Jan-1981||http://us.imdb.com/M/title-exact?On%20Golden%20Pond%20(1981)|0|0|0|0|0|0|0|0|1|0|0|0|0|0|0|0|0|0|0

163|Return of the Pink Panther, The (1974)|01-Jan-1974||http://us.imdb.com/M/title-exact?Return%20of%20the%20Pink%20Panther,%20The%20(1974)|0|0|0|0|0|1|0|0|0|0|0|0|0|0| 0|0|0|0|0

3. u.user: Demographic information about the users. This is a tab

separated list of:

user id | age | gender | occupation | zip code

The user ids are the ones used in the u.data data set.

Example:

1|24|M|technician|85711

2|53|F|other|94043

3|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. Feel free to modify the PCI code to answer the

following questions.

Questions (10 points).

1. Find 3 users who are closest to you in terms of age,

gender, and occupation. For each of those 3 users:

- what are their top 3 favorite films?

- bottom 3 least favorite films?

Based on the movie values in those 6 tables (3 users X (favorite +

least)), choose a user that you feel is most like you. Feel

free to note any outliers (e.g., "I mostly identify with user 123,

except I did not like ``Ghost'' at all").

This user is the "substitute you".

2. Which 5 users are most correlated to the substitute you? Which

5 users are least correlated (i.e., negative correlation)?

3. Compute ratings for all the films that the substitute you

have not seen. Provide a list of the top 5 recommendations for films

that the substitute you should see. Provide a list of the bottom

5 recommendations (i.e., films the substitute you is almost certain

to hate).

4. Choose your (the real you, not the substitute you) favorite and

least favorite film from the data. For each film, generate a list

of the top 5 most correlated and bottom 5 least correlated films.

Based on your knowledge of the resulting films, do you agree with

the results? In other words, do you personally like / dislike

the resulting films?

**Part 1:**

1. Find 3 users who are closest to you in terms of age,

gender, and occupation. For each of those 3 users:

- what are their top 3 favorite films?

- bottom 3 least favorite films?

Based on the movie values in those 6 tables (3 users X (favorite +

least)), choose a user that you feel is most like you. Feel

free to note any outliers (e.g., "I mostly identify with user 123,

except I did not like ``Ghost'' at all").

This user is the "substitute you".

This task is accomplished by the use of the following python programs: "FindSurragetes.py", "FindSurrageteRatings.py", "FindTopRatings.py", "FindBotRatings.py", "FindMovieTitles.py", "MergeMovieRatings.py", and "MakeSurrageteTable.py". First "FindSurragetes.py" is executed to find any matches to my stats (age: 26, Gender: M, Occupation: Student), it then output a text file “SurrageteList.txt” which contains the id number for all matches, each matching user’s data is output to a separate file “Sur<userID>.txt”. Next "FindSurrageteRatings.py" is executed, which reads the “SurrageteList.txt” file and searches “U.data” for all entries that correspond to one of the surrogates, then a separate file is output for each surrogate containing their ratings “Sur<userID>Ratings.txt”. "FindTopRatings.py" and "FindBotRatings.py" are then executed, each reading the “SurrageteList.txt” file before locating and reading each individual “Sur<userID>Ratings.txt”, they then make a list of any ratings of 5 or 1, respectively, before outputting them to separate text files for each surrogate titled “Sur<userID>TopRatings.txt” and “Sur<userID>BotRatings.txt”. "FindMovieTitles.py" is then executed which reads in the surrogate list, and located the top and bot ratings files before associating each rating with the corresponding movie title from “U.Item”, to top and bottom movies are then output to text files titles “Sur<userID>TopMovies.txt” and “Sur<userID>BotMovies.txt”. "MergeMovieRatings.py" is then used to create a file, “Sur<userID>TopBottomMovies.txt for each surrogate consisting on their user id, movie title, and their rating. Lastly "MakeSurrageteTable.py" is used to output a single text file, “SurragetesTopBottom3.txt” consisting of at most 3 movies rated 5 and 3 movies rated 1 for each surrogate. This text file was used to generate the following table of users ratings.



I have selected 771 as my surrogate me, based simply on the top ranking of toy story. Of all 5 I have seen a very few of the movies that were rated.

**Part 2:**

2. Which 5 users are most correlated to the substitute you? Which

5 users are least correlated (i.e., negative correlation)?

Accomplished through use of sample code from <https://github.com/ferronrsmith/programming-collective-intelligence-code/blob/master/chapter2/recommendations.py> file “Reccomendations.py” was used to calculate correlation, 12 users shared the same negative correlation with my surrogate and were all the least correlated, can be viewed on table to left, the 5 users below in the table to the right were the most correlated to my surrogate, 4 of the 5 sharing the same correlation. All other correlations can be found in the text file “Correlation.txt” or sorted in the “Correlation.csv” file



**Part 3**

3. Compute ratings for all the films that the substitute you

have not seen. Provide a list of the top 5 recommendations for films

that the substitute you should see. Provide a list of the bottom

5 recommendations (i.e., films the substitute you is almost certain

to hate).

Accomplished through use of sample code from <https://github.com/ferronrsmith/programming-collective-intelligence-code/blob/master/chapter2/recommendations.py> file “Reccomendations.py” was used to calculate the ratings for each movie based on the correlations, a complete list of the results can be found in “Reccomendation.txt” or “Reccomendation .csv” files. Top 5 recommendations: 5.0 Visitors, The (Visiteurs, Les) (1993), 5.0 Two or Three Things I Know About Her (1966), 5.0 They Made Me a Criminal (1939), 5.0 Someone Else's America (1995), 5.0 Sliding Doors (1998). Bottom 5 recommendations: 1.0 Amityville: Dollhouse (1996), 1.0 Amityville: A New Generation (1993), 1.0 Amityville 1992: It's About Time (1992), 1.0 American Strays (1996), 1.0 3 Ninjas: High Noon At Mega Mountain (1998).

**Part 4**

4. Choose your (the real you, not the substitute you) favorite and

least favorite film from the data. For each film, generate a list

of the top 5 most correlated and bottom 5 least correlated films.

Based on your knowledge of the resulting films, do you agree with

the results? In other words, do you personally like / dislike

the resulting films?