2017

Assignment 7

ROSS REELACHART

PROFESSOR MICHAEL L. NELSON

Question 1:

Before I even began to find the user who would represent me in the following questions, I needed to extract and arrange the data from the raw data provided in *ml-100k*[1], which had been extracted to a directory called *movieData*. Specifically, I used a trio of python programs to extract the movie data, user data, and rating data. (Like previous assignments, I preferred to split-up my functions in case something went wrong.) *getMovieData.py* extracted the movie names and IDs from *u.item*. *getUserData.py* extracted all the user details and their movie ratings from *u.user*. *getRatingData.py* extracted the individual ratings for movies from each user.

Now that I had all the data arranged in a useable JSON format, I could begin searching for my substitute user. With over 900 users to choose from, I needed to whittle down the selection to something that was both manageable and only included those that were similar to me. So I made another trio of python programs, with each one narrowing down the list further. First, *getMaleUsers.py* extracted only the users who had their genders listed as Male. This took the list from 943 users to just the 670 male users, stored in the *maleUsers* file.

```
rreelac@atria:~/assignment 7/Question 1$ python getMaleUsers.py
Total users:943
Number of male users:670
```

Figure 1 - Output of getMaleUsers.py

Then I took the *maleUsers* file and ran it through the *getAgeUsers.py* program, which extracted only the users who were exactly 27-years-old. This reduced the list from 670 male users to 25 male users that were 27-years-old. This list was stored in the *maleUsersWithAge* file.

```
rreelac@atria:~/assignment 7/Question 1$ python getAgeUsers.py
Number of male users:670
Number of male users who are age 27:25
```

Figure 2 - Output for getAgeUsers.py

Finally, *maleUsersWithAge* was run through the *getOccupationUsers.py* program, which extracted only the users who had listed their occupation as "student." This final list of 6 possible users was stored in the file *maleUsersWithAgeAndOccup*.

```
rreelac@atria:~/assignment 7/Question 1$ python getOccupationUsers.py
Number of male users who are age 27:25
Number of male users who are age 27 and a Student:6
```

Figure 3 - Output of getOccupationUsers.py

Continued on Next Page

So with a mere 6 users to choose from, I needed to select the user that I felt best represented me. This required me to do some manual selection. So I started with one criteria: Who listed "The Empire Strikes Back," "The Terminator," and "Raiders of the Lost Ark" as some of their top-rated movies? Given those three movies, only User 429 would fit as my substitute.

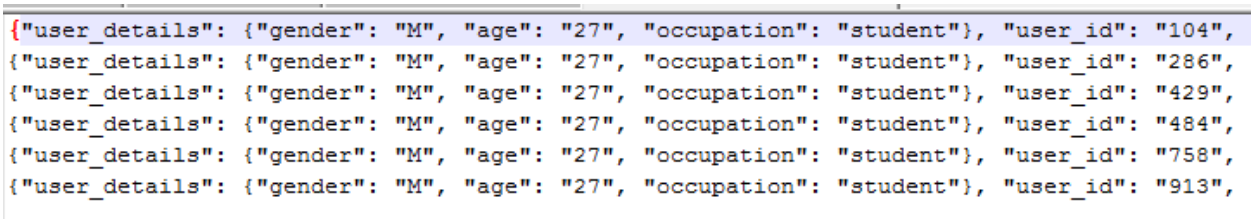

Figure 4 - The final selection of 6 users who matched my in gender, age, and occuption

The top three possible candidates, including User 429, are listed below with their Top and Bottom films.

User 429	*Most Favorite Films*	*Least Favorite Films*
	The Empire Strikes Back (1980)	The Jackal (1997)
	Highlander (1986)	If Lucy Fell (1996)
	Schindler's List (1993)	The Fighteners (1996)

User 484	*Most Favorite Films*	*Least Favorite Films*
	Raiders of the Lost Ark (1981)	Mrs. Doubtfire (1993)
	Apollo 13 (1995)	The Indian in the Cupboard (1995)
	E.T. the Extra-Terrestrial (1982)	The Lion King (1994)

User 758	*Most Favorite Films*	*Least Favorite Films*
	Naked Gun 33 1/3: The Final Insult (1994)	Conspiracy Theory (1997)
	Monty Python and the Holy Grail (1974)	Reality Bites (1994)
	The Terminator (1984)	The Saint (1997)

Continued on Next Page

Question 2:

To get the 5 most and least correlated users I needed to calculate their Pearson score[2]. Using the *userData.json* file and the user ID of my substitute (429), I calculated the Pearson score and found the top and bottom five using *getCorrelation.py* and piping the output to the text file *CorrelatedUsers.txt.*

The Top and Bottom five users are listed below their calculated Pearson scores. The most correlated users have scores closer to "1.0," while least correlated users have scores closer to "-1.0."

Top 5 Most Correlated Users		**Bottom 5 Most Correlated Users**	
User ID	Pearson Score	User ID	Pearson Score
920	0.9	515	-0.707
675	0.948	78	-0.707
260	1.0	410	-0.707
309	1.0	50	-0.707
720	1.0	40	-0.75

Question 3:

In order to get the top and bottom most recommended films, I needed to use much of the same code from the previous question. I needed the correlation scores because I wanted to see what the users who correlated with the substitute me could possibly recommend, or not recommend, to substitute me. The rated movies of the correlated users were ordered and the top and bottom of that list were added to a list, with duplicate movies not included in the list. Then the top five and bottom five movies were printed and piped to the text file *RecommendedFilms.txt*. The top and bottom five movies, and their movie IDs, are listed below.

```
rreelac@sirius:~/assignment 7/Question 3$ python getRecommendations.py
[(5.000000000000001, u'1500'), (5.000000000000001, u'1201'), (5.0, u'814'), (5.0, u'1653'), (5.0, u'1536')]
[(1.0, u'1329'), (1.0, u'1325'), (1.0, u'1320'), (1.0, u'1309'), (1.0, u'1308')]
1500
Santa with Muscles (1996)
1201
Marlene Dietrich: Shadow and Light (1996)
814
Great Day in Harlem, A (1994)
1653
Entertaining Angels: The Dorothy Day Story (1996)
1536
Aiqing wansui (1994)
1329
Low Life, The (1994)
1325
August (1996)
1320
Homage (1995)
1309
Very Natural Thing, A (1974)
1308
Babyfever (1994)
rreelac@sirius:~/assignment 7/Question 3$
```

Figure 5 - Raw output of getRecommendations.py

Top Five Most Recommended Films for User 429	
Movie Title	Movie ID
Santa with Muscles (1996)	1500
Marlene Dietrich: Shadow and List (1996)	1201
A Great Day in Harlem (1994)	814
Entertaining Angels: The Dorothy Day Story (1996)	1653
Aiqing Wansui (1994)	1536

Bottom Five Most Recommended Films for User 429	
Movie Title	Movie ID
The Low Life (1994)	1329
August (1996)	1325
Homage (1995)	1320
A Very Natural Thing (1974)	1309
Baby Fever (1994)	1308

Question 4:

Choose a most and least favorite film among the expansive list was difficult. But I settled on a movie that I knew was my favorite from a select list: The 1997 Disney film "Hercules" is my favorite Disney movie based on the music alone. My least favorite was difficult to pick because I was always picky about which movies I saw so I didn't see too many I didn't like. I ended up picking "8 Heads in a Duffel Bag" because of what it represented. After Tarantino made "Pulp Fiction" and "Reservoir Dogs," there was a slew of crappy copycat movies that tried and failed to emulate his unique sensibilities and writing.

The movie IDs of these two films were added to the program *getRealCorrelatedFilms.py* along with code that calculated their Pearson score and printed the results to the text file *TopandBottom.txt*. The output is listed below.

I can't say much regarding the films that best correlated with "8 Heads in a Duffel Bag," except for "Twister." I remember seeing "Twister," but I don't remember anything about it. So I suppose the correlation there is whether or not these films are memorable. After looking up the other ones, I can say that they too wouldn't especially make an impression on me. Especially "Kingpin" which seems to be some kind of Farrelly brothers bowling comedy.

The top five most correlated films for "Hercules" aren't very appealing to me, except for perhaps "A Christmas Carol" because that's a classic. I suppose these results could be linked to their Pearson scores, which are all over 1.0.

Of all the correlated films, top and bottom, that appeal most to me, it's "Star Wars" for obvious reasons and "The Doom Generation[3]" which sounded like the kind of punk rock dark comedy I would like nowadays.

Top Five Correlated Films For: 8 Heads in a Duffel Bag (1997)	
Movie Title	Pearson Score
Kingpin (1996)	1.0
Othello (1995)	0.948
Twister (1996)	0.894
Rosewood (1997)	0.894
The River Wild (1994)	0.894

Bottom Five Correlated Films For: 8 Heads in a Duffel Bag (1997)	
Movie Title	Pearson Score
The English Patient (1996)	-0.471
Star Wars (1977)	-0.566
Mars Attacks! (1996)	-0.816
The Game (1997)	-0.816
Nixon (1995)	-0.948

Top Five Correlated Films For: Hercules (1997)	
Movie Title	Pearson Score
Jude (1996)	1.154
Girl 6 (1996)	1.154
The Second Jungle Book: Mowgli & Baloo (1997)	1.043
Home Alone 3 (1997)	1.032
A Christmas Carol (1983)	1.032

Bottom Five Correlated Films For: Hercules (1997)	
Movie Title	Pearson Score
The Doom Generation (1995)	-0.942
Warriors of Virtue (1997)	-0.948
Antonia's Line (1995)	-0.948
The Age of Innocence (1993)	-0.948
In the Bleak Midwinter (1995)	-1.0

References

1.) https://grouplens.org/datasets/movielens/100k/
2.) http://www.statisticshowto.com/what-is-the-pearson-correlation-coefficient/
3.) https://en.wikipedia.org/wiki/The_Doom_Generation