Assignment 6

CS 532: Introduction to Web Science

Spring 2018

Chandrasekhar Reddy Muthyala

Finished on March 27, 2018

1

Question

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".

Answer

Instead of manually picking out users from the data file provided, I wrote a script called **substituteMe.py** , shown in Listing 1, which filters all users by the gender Male "M", the occupation of "engineer" and the age **24**. This script found multiple users with age 24. This left me with 3 users with the ids: 105, 268 and 69.

To find their favorite and least favorite movies I added a function called *findMoviesMerge*, which matched each user's review to their movie names and return the bottom and top movies sorted by their ratings, there were of course other movies with rating 5 but I simply took the top and bottom 3 provided. Their tables for top 3 films and bottom 3 films are shown in Table 1, Table 2 and Table 3 respectively.

Rank	**Top Favorite Movies**	**Rating**	**Least Favorite Movie**	**Rating**
1	Gattaca (1997)	5.0	Devil's Advocate, The (1997)	2.0
2	Titanic (1997))	5.0	Saint, The (1997)	2.0
3	L.A. Confidential (1997)	5.0	Mimic (1997)	2.0

Table 1: User 105's favorite and least favorite movies

Rank	**Top Favorite Movies**	**Rating**	**Least Favorite Movie**	**Rating**
1	Aliens (1986)	5.0	Lawnmower Man, The (1992)	1.0
2	Empire Strikes Back, The (1980)	5.0	Forget Paris (1995)	1.0
3	Close Shave, A (1995)	5.0	Santa Clause, The (1994)	1.0

Table 2: User 268's favorite and least favorite movies

Rank	Top Favorite Movies	Rating	Least Favorite Movie	Rating
1	Graduate, The (1967)	5.0	Devil's Own, The (1997)	1.0
2	Scream (1996)	5.0	Peacemaker, The (1997)	1.0
3	Empire Strikes Back, The (1980)	5.0	Saint, The (1997)	2.0

Table 3: User 69's favorite and least favorite movies

```
import csv
from pprint import pprint as pp
def substituteUsers():

        users = []
        with open("dataset/u.user") as csvFile:
                lines = csv.reader(csvFile, delimiter='|')
                for line in lines:
                        age = 24
                        gender = 'M'
                        occupation = 'engineer'
                        if age == int(line[1]) and gender ==
                            line[2] and occupation == line[3] and
                             len(users)<3:
                                users.append(line)


        return users
def getMoviereviews(userIds):
        reviews = {}
        for i in userIds:
                reviews[i] = []
                with open("dataset/u.data") as csvFile:
                        lines = csv.reader(csvFile, delimiter =
                            "       ")
                        for line in lines:
                                if i == line[0]:
                                        reviews[i].append(line)
        return reviews
def findMovie(movieId):
    with open("dataset/u.item", 'r') as f:
        reader = csv.reader(f, delimiter='|')
        for i in reader:
            itemId = i[0]
            if movieId == itemId:
                # id, name, URI
                return (i[0], i[1], i[4])


def findMoviesMerge(reviewDict):
    userMovieDict = {}
```

```
39      for userId, reviews in reviewDict.items():
40
41          userMovieDict[userId] = {}
42          moviesReviewed = []
43          botMovies = []
44          topMovies = []
45          for r in reviews:
46              movieId = r[1]
47              rating = r[2]
48
49              movie = findMovie(movieId)
50              movie = tuple(rating) + movie
51              moviesReviewed.append(movie)
52
53          # botMovies.sort(key=lambda tup: tup[0])
54          moviesReviewed.sort(key=lambda tup: tup[0])
55          botMovies = moviesReviewed[:3]
56          topMovies = moviesReviewed[-3:]
57          userMovieDict[userId]["bottomMovies"] = botMovies
58          userMovieDict[userId]["topMovies"] = topMovies
59
60      return userMovieDict
61
62
63
64
65  if __name__ == '__main__':
66          selectedUsers = substituteUsers()
67          userIds = []
68          for user in selectedUsers:
69                  userIds.append(user[0])
70          userMovieReviews = getMoviereviews(userIds)
71          with open("dataset/closestUsers.txt", 'w') as f:
72                  pp(findMoviesMerge(userMovieReviews), stream=f)
```

Listing 1: Python script for determining closest 3 users

2

Question

`2. Which 5 users are most correlated to the substitute you? Which 5 users are least correlated (i.e., negative correlation)?`

Answer

This question's answer relied heavily off the source code provided by the Programming Collective Intelligence book [1]. I created a script called **pearsonCorrelation.py**, shown in Listing 4, which is also used later in questions 3 and 4. This question used the *loadMovieLens*, which generated the user preferences, and *findCorrelations* methods which finds the Sim Pearson correlation coefficient between "substitute me," user 868, and every other user based on their preferences. The results of this are shown in Tables 4 and 5 and were saved to **correlatedUsers.txt** in my Github repository [**?**].

User ID	Correlation
853	+1.0
857	+1.0
898	+1.0
626	+1.0
724	+1.0

Table 4: Most correlated users

User ID	Correlation
36	-1.0
404	-1.0
599	-1.0
628	-1.0
736	-0.9045340337332909

Table 5: Least correlated users

```
1
2  import csv
3  from math import sqrt
4  from pprint import pprint as pp
5
6
7  def sim_pearson(prefs, p1, p2):
8      '''
9      Returns the Pearson correlation coefficient for p1 and p2.
10     '''
11
12     # Get the list of mutually rated items
13     si = {}
14     for item in prefs[p1]:
15         if item in prefs[p2]:
16             si[item] = 1
17     # If they are no ratings in common, return 0
18     if len(si) == 0:
19         return 0
20     # Sum calculations
21     n = len(si)
22     # Sums of all the preferences
23     sum1 = sum([prefs[p1][it] for it in si])
24     sum2 = sum([prefs[p2][it] for it in si])
25     # Sums of the squares
26     sum1Sq = sum([pow(prefs[p1][it], 2) for it in si])
27     sum2Sq = sum([pow(prefs[p2][it], 2) for it in si])
28     # Sum of the products
29     pSum = sum([prefs[p1][it] * prefs[p2][it] for it in si])
30     # Calculate r (Pearson score)
31     num = pSum - sum1 * sum2 / n
32     den = sqrt((sum1Sq - pow(sum1, 2) / n) * (sum2Sq - pow(sum2,
           2) / n))
33     if den == 0:
34         return 0
35     r = num / den
36     return r
37
38
39 def findCorrelations(prefs):
40     most_correlated = []
41     least_correlated = []
42     correlations = {}
43     substituteMe = str(868)
44
45     users = {}
46     for line in open('dataset/u.user'):
47         (user, age, gender, job, zipcode) = line.split('|')
```

```
48          users.setdefault(user, {})
49          users[user] = {'age': age, 'gender': gender,
50                         'job': job, 'zipcode': zipcode}
51
52      for user, rest in users.items():
53          if substituteMe == user:
54              pass
55          else:
56              r = sim_pearson(prefs, substituteMe, user)
57              correlations[int(user)] = r
58
59      correlations = sorted(correlations.items(), key=lambda x: x
            [1])
60      pp(correlations)
61      least_correlated = correlations[:5]
62      most_correlated = correlations[-5:]
63
64      with open("dataset/correlatedUsers.txt", 'w') as f:
65          #print("Most Correlated:", file=f)
66          pp(most_correlated, stream=f)
67          #rint("Least Correlated:", file=f)
68          pp(least_correlated, stream=f)
69
70
71  def transformPrefs(prefs):
72      '''
73      Transform the recommendations into a mapping where persons
            are described
74      with interest scores for a given title e.g. {title: person}
            instead of
75      {person: title}.
76      '''
77
78      result = {}
79      for person in prefs:
80          for item in prefs[person]:
81              result.setdefault(item, {})
82              # Flip item and person
83              result[item][person] = prefs[person][item]
84      return result
85
86
87  def getRecommendations(prefs, person, similarity=sim_pearson):
88      '''
89      Gets recommendations for a person by using a weighted
            average
90      of every other user's rankings
91      '''
92
```

```
93      totals = {}
94      simSums = {}
95      for other in prefs:
96          # Don't compare me to myself
97          if other == person:
98              continue
99          sim = similarity(prefs, person, other)
100         # Ignore scores of zero or lower
101         if sim <= 0:
102             continue
103         for item in prefs[other]:
104             # Only score movies I haven't seen yet
105             if item not in prefs[person] or prefs[person][item]
                    == 0:
106                 # Similarity * Score
107                 totals.setdefault(item, 0)
108                 # The final score is calculated by multiplying
                        each item by the
109                 #   similarity and adding these products
                        together
110                 totals[item] += prefs[other][item] * sim
111                 # Sum of similarities
112                 simSums.setdefault(item, 0)
113                 simSums[item] += sim
114     # Create the normalized list
115     rankings = [(total / simSums[item], item) for (item, total)
            in
116                 totals.items()]
117     # Return the sorted list
118     rankings.sort()
119
120     lowestRankings = rankings[:5]
121     topRankings = rankings[-5:]
122
123     return (lowestRankings, topRankings)
124
125
126 def topMatches(
127     prefs,
128     person,
129     n=5,
130     similarity=sim_pearson,
131 ):
132     '''
133     Returns the best matches for person from the prefs
            dictionary.
134     Number of results and similarity function are optional
            params.
135     '''
```

```
136
137     scores = [(similarity(prefs, person, other), other) for
            other in prefs
138               if other != person]
139     scores.sort()
140     # scores.reverse()
141     lowestScores = scores[:n]
142     highestScores = scores[-n:]
143     return (lowestScores, highestScores)
144
145
146 def loadMovieLens(path='./'):
147     # Get movie titles
148     movies = {}
149     for line in open(path + 'dataset/u.item'):
150         (id, title) = line.split('|')[0:2]
151         movies[id] = title
152
153     # Load data
154     prefs = {}
155     for line in open(path + 'dataset/u.data'):
156         (user, movieid, rating, ts) = line.split('\t')
157         prefs.setdefault(user, {})
158         prefs[user][movies[movieid]] = float(rating)
159     return prefs
160
161
162 def saveScores(filename, lowScores, highScores):
163     with open(filename, 'w') as f:
164         #print("Lowest Scores:", file=f)
165         pp(lowScores, stream=f)
166         #print("Highest Scores:", file=f)
167         pp(highScores, stream=f)
168
169
170 if __name__ == "__main__":
171     # q2
172     prefs = loadMovieLens()
173     findCorrelations(prefs)
174     # 868 is substiteMe
175     # q3
176     getRecommendations(prefs, '868')
177     # q4
178     prefs = transformPrefs(prefs)
179     (lowestScore, highestScore) = topMatches(prefs, 'Citizen
            Kane (1941)')
180     saveScores("dataset/favoriteFilmCorrelation.txt",
            lowestScore, highestScore)
181     (lowestScore, highestScore) = topMatches(prefs, 'Mars
```

```
        Attacks! (1996)')
182     saveScores("dataset/worstFilmCorrelation.txt", lowestScore,
            highestScore)
```

Listing 2: Python script utilizing Programming Collective Intelligence's code

3

Question

```
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).
```

Answer

The answer for this question again relied heavily upon Programming Collective Intelligence's code since it provided the main method to solve this problem, the *getRecommendations* method, and is shown in Listing 4. This again utilized the Sim Pearson correlation coefficient between users and found all the movies my substitute user, user 868, hasn't seen. After a final score was calculated these scores were normalized on a 1 to 5 scale and sorted in order. I took the lowest 5 movies and the top 5 movies, again with some being the same weight I simply took the top 5 it provided. These are shown in Table 6 and 7.

I was taken aback when substitute me's second most recommended movie was "Santa with Muscles (1996)." He also apparently hates any kind of Amityville horror movie. I'll have to look into Saint of Fort Washington.

Rank	**Movie**	**Rating**
1	Saint of Fort Washington, The (1993)	5.0
2	Santa with Muscles (1996)	5.0
3	Someone Else's America (1995)	5.0
4	The Deadly Cure (1996)	5.0
5	They Made Me a Criminal (1939)	5.0

Table 6: Most recommended movies

Rank	Movie	Rating
1	3 Ninjas: High Noon At Mega Mountain (1998)	1.0
2	Amityville 1992: It's About Time (1992)	1.0
3	Amityville: A New Generation (1993)	1.0
4	Amityville: Dollhouse (1996)	1.0
5	Babyfever (1994)	1.0

Table 7: Least recommended movies

4

Question

```
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?
```

Answer

When looked through the movie data I saw one of my favorite movies of all time, Citizen Kane (1941) because of the amazing cinematography for its time, I chose this film as my favorite. For me least favorite film I chose Mars Attacks! (1996). This question used the *getRecommendations* method like in question 3, but this time I had to transform the preferences into a dictionary of movie as key and users as values. The method to do this was *transformPrefs*, which was also provided by the Programming Collective Intelligence book [1]. I then used the *topMatches* method to get the bottom and top most correlated films. The results are shown in Tables 8, 9, 10 and 11 with the data saved to a files named **favoriteFilmCorrelation.txt** and **worstFilmCorrelation.txt** on my Github repository [**?**].

Its difficult to comment on the results of this problem as I have never seen any of the films in the tables below, aside from Free Willy, but not even that sequel number. I imagine that the films that least correlate with Citizen Kane also align with my dislikes after looking them up on the IMDB website. Aside from that, I really can't comment on the rest of the films.

Rank	Movie	Correlation
1	Newton Boys, The (1998)	1.0
2	Palmetto (1998)	1.0
3	Savage Nights (Nuits fauves, Les) (1992)	1.0
4	Wild America (1997)	1.0
5	Heavy (1995)	1.0

Table 8: Citizen Kane highest correlated movies

Rank	Movie	Correlation
1	Maya Lin: A Strong Clear Vision (1994)	-1.0
2	Free Willy 3: The Rescue (1997)	-1.0
3	Bad Girls (1994)	-1.0
4	Big Bully (1996)	-1.0
5	Colonel Chabert, Le (1994)	-1.0

Table 9: Citizen Kane least correlated movies

Rank	Movie	Correlation
1	Winter Guest, The (1997)	1.0
2	Wonderful, Horrible Life of Leni Riefenstahl, The (1993)	1.0
3	Wooden Man's Bride, The (Wu Kui) (1994)	1.0
4	Kaspar Hauser (1993)	1.0
5	Palmetto (1998)	1.0

Table 10: Mars Attacks! highest correlated movies

Rank	Movie	Correlation
1	Across the Sea of Time (1995)	-1.0
2	8 Heads in a Duffel Bag (1997)	-1.0
3	Swan Princess, The (1994)	-1.0
4	8 Seconds (1994)	-1.0
5	Aparajito (1956)	-1.0

Table 11: Mars Attacks! least correlated movies

5

Question

Extra credit (3 points)
5. Rank the 1,682 movies according to the 1997/1998 MovieLense data. Now rank the same 1,682 movies according to todays (March 2016) IMDB data (break ties based on # of users, for example: 7.2 with 10,000 raters > 7.2 with 9,000 raters).

Draw a graph, where each dot is a film (i.e., 1,682 dots). The x-axis is the MovieLense ranking and the y-axis is today's IMDB ranking.

What is Pearon's r for the two lists (along w/ the p-value)? Assuming the two user bases are interchangable (which might not be a good assumption), what does this say about the attitudes about the films after nearly 20 years?

Answer

For this problem I wrote a python program **getIMDBRating.py** to get the current IMDB rating of movie lense data through OMDB API and storing all those corresponding IMDB rating and movie lense rating into **test.csv** file to draw correlation graph i wrote a another python program **scatterplot.py**.

```
1  import json, requests
2  import yaml
3  import csv
4  # data = ''
5  count = 0
6  imdbRating = {}
7
8  movieLensDataRating = {}
9  keys = {'Rating':0.0, 'count':0}
10 for line in open('dataset/u.data'):
11         item = line.split('\t')
12         # print(item[1])
13         movieLensDataRating[item[1]] = []
14         movieLensDataRating[item[1]].append(0.0)
15         movieLensDataRating[item[1]].append(0)
16         # print(movieLensDataRating)
17
18 for line in open('dataset/u.data'):
```

```
19          item = line.split('\t')
20          print(item[1],item[2])
21          # print("before:",movieLensDataRating[item[1]])
22          movieLensDataRating[item[1]][0] = movieLensDataRating[
                item[1]][0]+float(item[2])
23          movieLensDataRating[item[1]][1] =movieLensDataRating[
                item[1]][1] + 1
24          # print("after: ",movieLensDataRating[item[1]])
25  for key,value in movieLensDataRating.items():
26          try:
27                  print("before:",movieLensDataRating[key][0],
                        movieLensDataRating[key][1])
28                  movieLensDataRating[key][0] =
                        movieLensDataRating[key][0]/
                        movieLensDataRating[key][1]
29                  print("after:",movieLensDataRating[key][0],
                        movieLensDataRating[key][1])
30                  # print(movieLensDataRating[key]['Rating'],
                        movieLensDataRating[key]['count'])
31          except:
32                  pass
33
34  for line in open('dataset/u.item'):
35          movieTitle = line.split("|")[1].split("(")[0].split(",")
                [0]
36          count +=1
37          try:
38                  url = 'http://www.omdbapi.com/?t='+movieTitle+'&
                        apikey=74052bd2'
39                  # print(line.split("|")[1].split("(")[0].split
                        (",")[0])
40                  resp = requests.get(url=url)
41                  # data = json.dump(resp.json(), file,indent=4,
                        sort_keys=True,ensure_ascii=False)
42                  data = yaml.safe_load(resp.text)
43                  imdbRating[line.split('|')[0]] = {'movieId':
                        line.split('|')[0], 'ImdbId': data['imdbID'],
                         'movieTitle':data['Title'],'Rating':data['
                        imdbRating'],'imdbVotes':data['imdbVotes']}
44                  # print(imdbRating[line.split('|')[0]])
45                  # print(data['Title'],data['imdbRating'])
46                  # writer.writerow({'movieId': line.split('|')
                        [0], 'ImdbId': data['imdbID'], 'movieTitle':
                        data['Title'],'Rating':data['imdbRating']})
47
48          except:
49                  pass
50
51
```

```
52
53 with open('dataset/test.csv', 'w') as csvfile:
54         fieldnames = ['movieId', 'ImdbId','movieTitle', 'Rating
               ','imdbVotes','movieLenseRating','movieLensUserCount
               ']
55         writer = csv.DictWriter(csvfile, fieldnames=fieldnames)
56         writer.writeheader()
57         for key,value in imdbRating.items():
58                 try:
59                         if key in movieLensDataRating:
60                                 print(movieLensDataRating[key])
61                                 writer.writerow({'movieId': key,
                                        'ImdbId': value['ImdbId'], '
                                       movieTitle':value['movieTitle
                                       '],'Rating':value['Rating'],
                                       'imdbVotes':value['imdbVotes
                                       '],'movieLenseRating':round(
                                       movieLensDataRating[key
                                       ][0],2), 'movieLensUserCount
                                       ':round(movieLensDataRating[
                                       key][1],2)})
62                 except:
63                         pass
64 # print(imdbRating)
```

Listing 3: Python script for getting IMDB rating

```
1  import matplotlib.pyplot as plt
2  import pandas as pd
3  points = [[3,9],[4,8],[5,4]]
4
5  df  = pd.read_csv('dataset/test.csv').dropna()
6  df1 = df.copy()
7  df2 = df.copy()
8  df1 = df1.sort_values(by=['Rating', 'imdbVotes'])
9  df2 = df2.sort_values(by=['movieLenseRating', '
       movieLensUserCount'])
10 xrank ={}
11 yrank = {}
12 i=1
13 for index, row in df1.iterrows():
14         xrank[row["movieTitle"]] = i
15         i =i+1
16 # print("xrank",xrank)
17 i =1
18
19 for index, row in df2.iterrows():
20         #
21         # print(index,row['movieId'],row['Rating'],row['
```

```
                imdbVotes'])
22          yrank[row["movieTitle"]] = i
23          i +=1
24          # print(xrank[row["movieId"]],yrank[row["movieId"]])
25  # print("yrank",yrank)
26  for key,value in xrank.items():
27          # print(xrank[key],yrank[key])
28          y = int(xrank[key])
29          x = int(yrank[key])
30          plt.plot(x, y, 'bo',markersize=2)
31          plt.text(x * (1 + 0.01), y * (1 + 0.01) , key, fontsize
                =4)
32
33  plt.xlabel('MovieLense ranking', fontsize=12)
34  plt.ylabel('IMDB ranking', fontsize=12)
35  plt.show()
```

Listing 4: Python script for plotting correlation graph

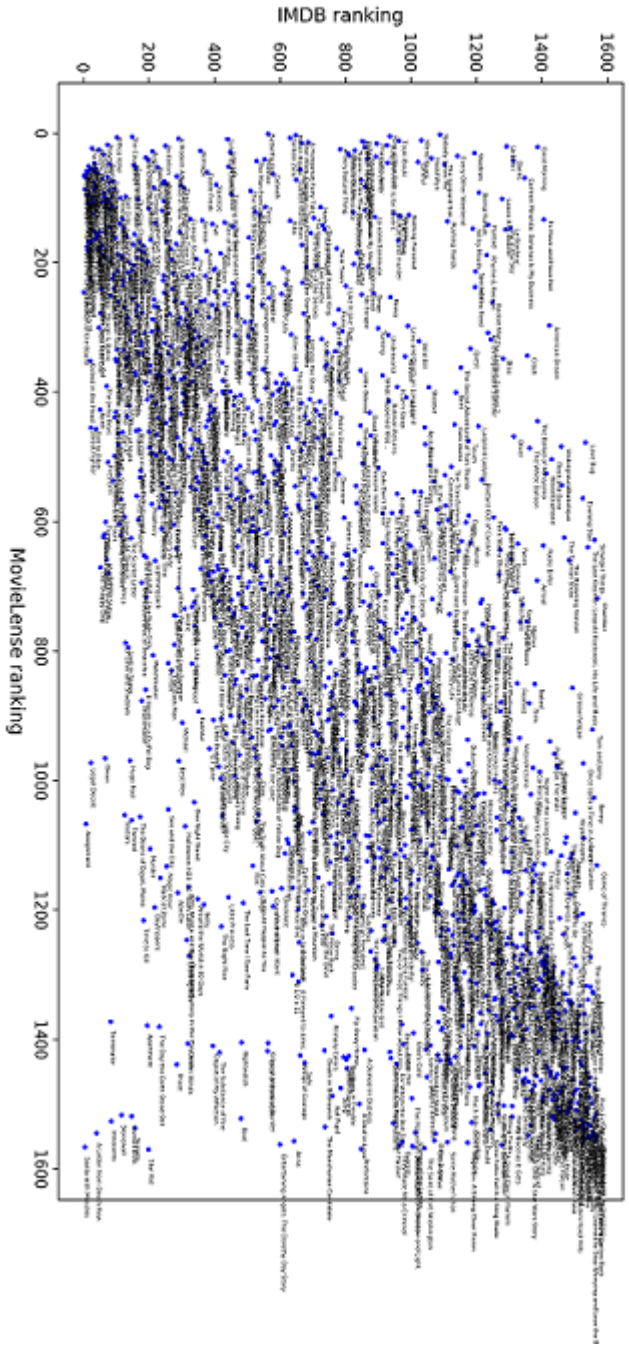

Figure 1: MovieLense ranking VS today's IMDB ranking

References

[1] Segaran, Toby. "Programming Collective Intelligence". O' Reilly, 2007. Web. 6 April 2017. http://shop.oreilly.com/product/9780596529321.do.