Machine Learning
Assignment 1
CLO3

Ida Bagus Dwi Satria Kusuma
1301140297

September 16, 2017

1. In this problem we will consider similarity measures for movies on the Movielens dataset. Download the Movielens data that we sent to you through email. In addition to the data, the file also contains some functions for easily loading the data into Matlab/Octave/R and some example code that you can use if you wish. See the README files for details.

   (a) We will now construct a similarity measure over the movies. For simplicity, let us first consider a simple measure that does not use the explicit (numerical) ratings given by the users, nor the time stamps of the ratings, but only whether or not a given movie was rated by a given user.

      i. **(20 points)** Create a function that, given two different movie IDs as input, outputs the Jaccard coeffcient: the number of users who rated both movies divided by the number of users who rated at least one of the movies. For example, for the movies 'Toy Story' and 'GoldenEye' the coefficient should be 0.217.
      Answer Gunakan jaccard_coeff.m untuk mendapatkan Jaccard Coefficient. Parameternya adalah dua id film yang ingin diukur.

```
>> jaccard_coeff(1,2)
Jumlah user yang merate kedua film :     104
Jumlah user yang merate film (1) Toy Story (1995) :      348
Jumlah user yang merate film (2) GoldenEye (1995) :      27
Jaccard coefficient : 0.217
```

ii. **(5 points)** What is the Jaccard coefficient between 'Three Colors: Red' and 'Three Colors: Blue'?
Answer Gunakan searchMovie.m untuk mendapatkan ID film dan gunakan jaccard_coeff.m untuk mendapatkan Jaccard Coefficient.

```
>> searchMovie('Three Colors: Red')
Indeks of Three Colors: Red is : 59
>> searchMovie('Three Colors: Blue')
Indeks of Three Colors: Blue is : 60
>> jaccard_coeff(59,60)
Jumlah user yang merate kedua film :     55
Jumlah user yang merate film (59) Three Colors: Red (1994) :     28
Jumlah user yang merate film (60) Three Colors: Blue (1993) :    9
Jaccard coefficient : 0.598
```

iii. **(10 points)** What are the 5 movies with highest Jaccard coefficient to 'Taxi Driver'?
Answer Gunakan searchMovie.m untuk mendapatkan id film 'Taxi Driver' dan gunakan jaccard_coeff_n.m untuk mendapatkan film-film dengan Jaccard Coefficient sebanyak n tertinggi. dalam kasus ini, n = 5.

```
Showing movies that related to Taxi Driver (1976)
ID : 182         | GoodFellas (1990)     | Jaccard Coeff : 0.417
ID : 187         | Godfather: Part II, The (1974)        | Jaccard Coeff : 0.417
ID : 179         | Clockwork Orange, A (1971)    | Jaccard Coeff : 0.404
ID : 134         | Citizen Kane (1941)   | Jaccard Coeff : 0.397
ID : 654         | Chinatown (1974)      | Jaccard Coeff : 0.394
```

iv. **(10 points)** Select a movie of your own choosing (which you are familiar with), what are the 5 movies with highest Jaccard coefficient to that movie? Do they make sense?
Answer Gunakan searchMovie.m untuk mendapatkan id film 'Star Wars' dan gunakan jaccard_coeff_n.m untuk mendapatkan film-film dengan Jaccard Coefficient sebanyak n tertinggi. dalam kasus ini, n = 5.

```
Showing movies that related to Star Wars (1977)
ID : 181         | Return of the Jedi (1983)     | Jaccard Coeff : 0.787
ID : 174         | Raiders of the Lost Ark (1981)        | Jaccard Coeff : 0.610
ID : 1   | Toy Story (1995)      | Jaccard Coeff : 0.583
ID : 172         | Empire Strikes Back, The (1980)       | Jaccard Coeff : 0.570
ID : 100         | Fargo (1996)          | Jaccard Coeff : 0.565
```

(b) Now lets try a similarity measure that uses the explicit ratings.

i. **(20 points)** Create a second function that, given two different

movie IDs as input, outputs the correlation coefficient of the ratings given to those two movies by all users which have rated both movies. (Note, the function may need to return 0 when the number of users who have rated both is so low that one cannot compute a correlation coefficient.)

Answer Fungsi Correlation Coefficient yang digunakan adalah Pearson Correlation Coefficient yang didapatkan dari persamaan :

$$r = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{\sqrt{\left[n\left(\sum x^2\right) - \left(\sum x\right)^2\right]\left[n\left(\sum y^2\right) - \left(\sum y\right)^2\right]}} \quad (1)$$

Fungsi ini dapat dilihat di pearson_coeff.m

ii. **(5 points)** What is now the similarity between 'Toy Story' and 'GoldenEye'?

Answer

```
E(x)    : 402.000
E(y)    : 333.000
E(xy)   : 1307.000
E(x^2)  : 1646.000
E(y^2)  : 1153.000
(E(x))^2        : 161604.000
(E(y))^2        : 110889.000
Movie 1 : (1) Toy Story (1995)
Movie 2 : (2) GoldenEye (1995)
 Pearson Correlation Coefficient        : 0.2218
```

iii. **(5 points)** How about 'Three Colors: Red' and 'Three Colors: Blue'?

Answer

Gunakan searchMovie.m untuk mendapatkan ID film dan gunakan pearson_coeff.m untuk mendapatkan Pearson Correlation Coefficient.

```
>> searchMovie('Three Colors: Red')
Indeks of Three Colors: Red is : 59
>> searchMovie('Three Colors: Blue')
Indeks of Three Colors: Blue is : 60
>> pearson_corrcoeff(59,60)
E(x)    : 229.000
E(y)    : 221.000
E(xy)   : 958.000
E(x^2)  : 997.000
E(y^2)  : 945.000
(E(x))^2        : 52441.000
(E(y))^2        : 48841.000
Movie 1 : (59) Three Colors: Red (1994)
Movie 2 : (60) Three Colors: Blue (1993)
 Pearson Correlation Coefficient        : 0.7597
```

iv. **(10 points)** What are the 5 movies with highest similarity to 'Taxi Driver'?

Answer

Gunakan searchMovie.m untuk mendapatkan ID film dan gunakan pearson_coeff_n.m untuk mendapatkan film-film dengan Pearson Correlation Coefficient sebanyak n tertinggi. Dalam kasus ini, n = 5.

```
Showing movies that related to Taxi Driver (1976)
ID : 35         | Free Willy 2: The Adventure Home (1995)       | PCC : 1.000
ID : 766        | Man of the Year (1995)        | PCC : 1.000
ID : 909        | Dangerous Beauty (1998)       | PCC : 1.000
ID : 918        | City of Angels (1998)         | PCC : 1.000
ID : 927        | Flower of My Secret, The (Flor de mi secreto, La) (1995)      | PCC : 1.000
```

v. **(10 points)** Again, select a movie of your own choosing and list the 5 movies with highest similarity.

Answer

```
Showing movies that related to Star Wars (1977)
ID : 119        | Maya Lin: A Strong Clear Vision (1994)        | PCC : 1.000
ID : 766        | Man of the Year (1995)        | PCC : 1.000
ID : 1096       | Commandments (1997)   | PCC : 1.000
ID : 1123       | Last Time I Saw Paris, The (1954)     | PCC : 1.000
ID : 1237       | Twisted (1996)        | PCC : 1.000
```

(c) **(5 points)** Provide some brief thoughts on which similarity measure seems to work better, in the sense that the computed similarity matches your intuitive sense of similarity. Why do you think this is? Explain.

Answer

Berdasarkan hasil pengamatan, fungsi Jaccard Coefficient memiliki kemiripan yang lebih baik dari segi judul. Namun, untuk seberapa besar sebuah ID memberikan rating terhadap kedua film

tersebut dapat ditunjukkan oleh Pearson Correlation Coefficient.