THE UNIVERSITY OF SYDNEY

School of Information Technologies
Dr. Ying Zhou & A/P Uwe Roehm

COMP5349: Cloud Computing **Sem. 1/2016**

Week10: Assignment 2 Additional Information

12.05.2016

FAQ

1. **Question:** What is the expected output format of workload one?
   Answer: There is no required format. It is hard to compute all results within one job. You may `collect` partial results as local variables from various RDDs and merge them as the final result to be kept in a single file. Or you may leave the raw output from different RDDs in different files. However, in your report's result section, you need to produce the aggregated result in a list or table format to make it easy to find all information about a particular genre's top users.

2. **Question:** How do we set the personal ratings for a list of 15 movies in workload 2
   Answer: This is up to you. You may hard code the rating or read that from a text file as you did in week 9 tutorial question 2. We may supply a list of ratings for you to use in the demo. Make sure you can easily switch between different lists.

Workload 2 worked example

A few students have trouble understanding the cosine similarity formula and the rate prediction formula. Below is a worked example to help you implement the calculation.

Table **??** shows a sample rating data of 6 users and 6 movies. Each user's average rating is listed in the last column.

Table 1: Rating Data

user/movie	1	2	3	4	5	6	user average rating($\bar{R_u}$)
1	5		4	3			4
2	4	4.5		5	2	3	3.7
3	5		3.5		2		3.5
4	4		3	5		4.5	4.125
5	4	4.5	5	4.5	3		4.2
6	4	4	4.5		2.5	5	4

Suppose we want to make recommendation for user 1 who has rated movie 1, 3 and 4. We need to predict user 1's ratings for movie 2,5 and 6 to work out which one might have the highest value.

We show a step by step workout to predict user 1's rating of movie 2.

First, we need to compute the similarities between movie 2 and the three movies user 1 has rated. There are three of them, denoted as $sim(2,1)$, $sim(2,3)$ and $sim(2,4)$.

To compute $sim(2,1)$, we first find out the users who co-rated these two movies. There are three of them: user 2, 5 and 6 (see table **??**). We will apply the adjusted cosine similarity formula to compute $sim(2,1)$.

$$sim(i,j) = \frac{\sum_{u \in U}(R_{u,i} - \bar{R_u})(R_{u,j} - \bar{R_u})}{\sqrt{\sum_{u \in U}(R_{u,i} - \bar{R_u})^2}\sqrt{\sum_{u \in U}(R_{u,j} - \bar{R_u})^2}} \quad (1)$$

Our $U = \{2,5,6\}$. Movie 1 and 2 are represented as two vectors: $\begin{bmatrix}4\\4\\4\end{bmatrix}$ and $\begin{bmatrix}4.5\\4.5\\4\end{bmatrix}$.

For user 2: his average rating is 3.7; the rating for movie 1 is 4, the rating for movie 2 is 4.5. So we have our first $(R_{2,1} - \bar{R_2})(R_{2,2} - \bar{R_2})$ as (4-3.7) * (4.5-3.7). We do the same thing for user 5 and user 6. $sim(2,1)$ is calculated as follows:

$$sim(2,1) = \frac{(4-3.7)*(4.5-3.7)+(4-4.2)*(4.5-4.2)+(4-4)*(4-4)}{\sqrt{(4-3.7)^2+(4-4.2)^2+(4-4)^2}*\sqrt{(4.5-3.7)^2+(4.5-4.2)^2+(4-4)^2}} = 0.58$$

Table 2: Users co-rated movie 1 and 2

user/movie	1	2	3	4	5	6	user average rating ($(\bar{R_u})$)
1	5		4	3			4
2	4	4.5		5	2	3	3.7
3	5		3.5		2		3.5
4	4		3	5		4.5	4.125
5	4	4.5	5	4.5	3		4.2
6	4	4	4.5		2.5	5	4

Two users (user 5 and user 6) co-rated movie 2 and 3. So we compute $sim(2,3)$ as follows:

$$sim(2,3) = \frac{(4.5-4.2)*(5-4.2)+(4-4)*(4.5-4)}{\sqrt{(4.5-4.2)^2+(4-4)^2}*\sqrt{(5-4.2)^2+(4.5-4)^2}} = 0.85$$

We can also work out that $sim(2,4) = 0.99$.

Suppose we only want to use the two most similar neighbours to predict a user's rating. The two most similar neighbours of movie 2 are: movie 4 and movie 3. Using the weighted sum predication generation formula:

$$P_{u,i} = \frac{\sum_{n \in N}(s_{i,n} * R_{u,n})}{\sum_{n \in N}(\mid s_{i,n} \mid)} \tag{2}$$

We have $sim(2,4) = 0.99$, $R_{1,4} = 3$, $sim(2,3) = 0.85$, $R_{1,3} = 4$. We can predict user 1's rating for movie 2 as:

$p_{1,2} = \frac{0.99*3+0.85*4}{|0.99|+|0.85|} = 3.46$

[**Note:**]The adjusted cosine similar formula computes similarity value in the range of [-1, 1]. Larger value means higher similarity. A negative similarity value means the two movies are kind of opposite (different) to each other. A value -1 would mean direct opposite. For instance, if you compute the the similarity between movie 5 and 1, you will end up with a negative value. It is not hard to see from the rating data that movie 1 and 5 receive very different ratings from the user group. For any user who has co-rated movie 1 and movie 5, he rated 1 above his average and 5 below his average.

With negative similarity value, it is possible to compute negative predicted rating value using the weighted sum formula. This will not affect the overall recommendation result as negative value indicates not similar. If a movie is quite different to most of the movies a user has rated, it should not appear in the recommendation list. You will find slightly different rate generation formula in other research papers trying to avoid the negative rating value issue. But we stick to the formula used in the same paper.