Exercise on collaborative filtering (RecommenderSystems.pdf, slide 43)

Given the ratings shown in the following matrix, predict Eric's rating for Titanic.

	The Matrix	Titanic	Die Hard	Forrest Gump	Wall•E
John	5	1		2	2
Lucy	1	5	2	5	5
Eric	2	**?**	3	5	4
Diane	4	3	5	3	

For short, let's rename users and movies:

John: J
Lucy: L
Eric: E
Diane: D

The Matrix: m
Titanic: t
Die Hard: d
Forrest Gump: f
Wall•E: w

Let's write the ratings matrix using the name and movie abbreviations as above and let's calculate the mean rating μ for each user:

	m	t	d	f	w	μ
J	5	1		2	2	= (5+1+2+2)/4 = 2.5
L	1	5	2	5	5	= (1+5+2+5+5)/5 = 3.6
E	2	**?**	3	5	4	= (2+3+5+4)/4 = 3.5
D	4	3	5	3		= (4+3+5+3)/4 = 3.75

Let's normalize the ratings, by subtracting from each rating the average rating for that user:

	m	T	D	f	W
J	5-2.5=2.5	1-2.5=-1.5		2-2.5=-0.5	2-2.5=-0.5
L	1-3.6=-2.6	5-3.6=1.4	2-3.6=-1.6	5-3.6=1.4	5-3.6=1.4
E	2-3.5=-1.5		3-3.5=-0.5	5-3.5=1.5	4-3.5=0.5
D	4-3.75=0.25	3-3.75=-0.75	5-3.75=1.25	3-3.75=-0.75	

That is, the normalized ratings matrix is:

	m	T	d	f	W
J	2.5	-1.5		-0.5	-0.5
L	-2.6	1.4	-1.6	1.4	1.4
E	-1.5	**?**	-0.5	1.5	0.5
D	0.25	-0.75	1.25	-0.75	

Now, considering the normalized ratings matrix above, let us calculate the cosine similarity between Eric and each other user. Since we are interested in predicting the rating for Eric, we need to calculate the cosine similarity only for those pairs that include Eric.
Recall that the formula for the cosine similarity is the following, where A and B are two generic users, and A_i and B_i are the normalized ratings that A and B have assigned to movie *i*:

$$cosine\ similarity = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

If the rating is missing, we consider it as being equal to 0 when computing the similarity.

To calculate the cosine similarity between John and Eric, let's consider their corresponding rows in the normalized ratings matrix.
J = (2.5, -1.5, 0, -0.5, -0.5)
E = (-1.5, 0, -0.5, 1.5, 0.5)
Applying the above formula, the cosine similarity between John and Eric is:

$$s_{JE} = \frac{2.5*(-1.5)+(-0.5)*1.5+(-0.5)*0.5}{\sqrt{2.5^2+(-1.5)^2+(-0.5)^2+(-0.5)^2}\sqrt{(-1.5)^2+(-0.5)^2+1.5^2+0.5^2}} = \frac{-4.75}{3*2.236} = -0.708$$

To calculate the cosine similarity between Lucy and Eric, let's consider their corresponding rows in the normalized ratings matrix and apply the formula:
L = (-2.6, 1.4, -1.6, 1.4, 1.4)
E = (-1.5, 0, -0.5, 1.5, 0.5)

$$s_{LE} = \frac{(-2.6)*(-1.5)+(-1.6)*(-0.5)+1.4*1.5+1.4*0.5}{\sqrt{(-2.6)^2+1.4^2+(-1.6)^2+(1.4)^2+(1.4)^2}\sqrt{(-1.5)^2+(-0.5)^2+1.5^2+0.5^2}} = \frac{7.5}{3.899*2.236} = 0.86$$

To calculate the cosine similarity between Eric and Diane, let's consider their corresponding rows in the normalized ratings matrix and apply the formula:
E = (-1.5, 0, -0.5, 1.5, 0.5)
D = (0.25, -0.75, 1.25, -0.75, 0)

$$s_{ED} = \frac{(-1.5)*(0.25)+(-0.5)*1.25+1.5*(-0.75)}{\sqrt{(-1.5)^2+(-0.5)^2+1.5^2+0.5^2}\sqrt{(0.25)^2+(-0.75)^2+1.25^2+(-0.75)^2}} = \frac{-2.125}{2.236*1.658} = -0.573$$

We can now calculate Eric's predicted rating for Titanic:

$$\hat{r}_{Eric,Titanic} = 3.5 + \frac{[-0.708*(1-2.5)+0.86*(5-3.6)-0.573*(3-3.75)]}{(-0.708+0.86-0.573)} = -2.91$$

Since the predicted rating is below the minimum value of the rating scale (that we assume equal to 1), we set it to the minimum value, that is 1.

For sake of completeness, the cosine similarity values calculated for all the pairs of users are shown in the following matrix:

	J	L	E	D
J	1	-0.854	-0.708	0.427
L	-0.854	1	0.86	-0.735
E	-0.708	0.86	1	-0.573
D	0.427	-0.735	-0.573	1

Recall that when the user is the same one (e.g., Eric and Eric) the cosine similarity is equal to 1, and that the cosine similarity matrix is symmetric, that is the similarity between A and B is equal to the similarity between B and A (e.g., the similarity between Eric and John is equal to the similarity between John and Eric, -0.708 in our case).

Using this similarity matrix, you can also predict the remaining missing ratings that is, John's rating for Die Hard and Diane's rating for Wall•E.