Jonathan Nusantara jan 26th

Problem 1 (30 points). Consider the following movie rating matrix with five users.

	LOTR	HPATPOZ	Snatch	LSATSB	The Gentlemen	The Hobbit
A	5	?	1	2	3	4
B	5	4	2	2	2	5
C	1	2	4	?	4	3
D	?	2	3	5	?	?
E	?	3	5	4	5	1

1. Compute the user-user similarity for all the 10 pairs of users and 15 pairs of movies using Pearson's similarity. (Ignore missing values when computing similarity, i.e., you only need to consider the *commonly rated* entries, which can be a lower-dimensional vector).

2. Let $k = 3$. Fill the missing entries of the matrix above using k-NN user-user CF and Pearson's similarity. When predicting, use the following formula:

$$\hat{r}_{u,i} = \bar{r}_u + \frac{\sum_{j \in K_u} S(u,u_j)(r_{u_j,i} - \bar{r}_{u_j})}{\sum_{j \in K_u} |S(u,u_j)|},$$

where $\bar{r}_u$ is the average rating of user u (on the items that they actually have rated) and K_u is the top neighbours of u who also rated item i. **Note here we rank user similarities by the absolute values of** $S(u,u_j)$ since $S(u,u_j)$ can be negative for pearson similarity. (If I always like things you hate, then your rating is also very useful to me.)

(1) Pearson's similarity $s_p(u,v) = \frac{(\bar{r}_u - \hat{r}_u) \cdot (\bar{r}_v - \hat{r}_v)}{\| \bar{r}_u - \hat{r}_u \| \| \bar{r}_v - \hat{r}_v \|}$

The ave here is from common entries. Yuhan said this is okay because I've done it before announcement

AB: $\bar{r}_A = (5, 1, 2, 3, 4)$ $\quad \bar{r}_B = (5, 2, 2, 2, 5)$

$\hat{r}_A = (3, 3, 3, 3, 3)$ $\quad \hat{r}_B = (3.2, 3.2, 3.2, 3.2, 3.2)$

$\bar{r}_A - \hat{r}_A = (2, -2, -1, 0, 1)$ $\quad \bar{r}_B - \hat{r}_B = (1.8, -1.2, -1.2, -1.2, 1.8)$

in all entries in this section, the mean is calculated only from commonly rated entries between the user or item we are comparing.

Calculate using python

```
a = np.array([5,1,2,3,4])
b = np.array([5,2,2,2,5])

mean_a = np.mean(a)
mean_b = np.mean(b)

print(mean_a, mean_b)

a = a - mean_a
b = b - mean_b

print(a,b)

c = np.dot(a,b)
d = np.linalg.norm(a) * np.linalg.norm(b)
print(c,d)

similarity = c / d
print(similarity)
```

```
3.0 3.2
[ 2. -2. -1.  0.  1.] [ 1.8 -1.2 -1.2 -1.2  1.8]
9.0 10.392304845413266
0.8660254037844385
```

$s_p(A,B) =$

$$\frac{(2,-2,-1,0,1) \cdot (1.8,-1.2,-1.2,-1.2,1.8)}{\|(2,-2,-1,0,1)\| \|(1.8,-1.2,-1.2,-1.2,1.8)\|}$$

$$= \frac{2\times1.8 + -2\times-1.2 + -1\times-1.2 + 0\times-1.2 + 1\times1.8}{\sqrt{2^2+(-2)^2+(-1)^2+0^2+1^2} \times \sqrt{1.8^2+(-1.2)^2+(-1.2)^2+(-1.2)^2+(1.8)^2}}$$

$= 0.866$

AC : $\bar{r}_A = (5, 1, 3, 4)$ $\bar{r}_C = (1, 4, 4, 3)$

$\hat{r}_A = (3.25, 3.25, 3.25, 3.25)$ $\hat{r}_C = (3, 3, 3, 3)$

$\bar{r}_A - \hat{r}_A = (1.75, -2.25, -0.25, 0.75)$ $\bar{r}_C - \hat{r}_C = (-2, 1, 1, 0)$

$$s_p(A,C) = \frac{(1.75, -2.25, -0.25, 0.75) \cdot (-2, 1, 1, 0)}{\|(1.75, -2.25, -0.25, 0.75)\| \|(-2, 1, 1, 0)\|}$$

$$= -0.828$$

AD : $\bar{r}_A = (1, 2)$ $\bar{r}_D = (3, 5)$

$\hat{r}_A = (1.5, 1.5)$ $\hat{r}_D = (4, 4)$

$\bar{r}_A - \hat{r}_A = (-0.5, 0.5)$ $\bar{r}_D - \hat{r}_D = (-1, 1)$

$$s_p(A,D) = \frac{(-0.5, 0.5) \cdot (-1, 1)}{\|(-0.5, 0.5)\| \|(-1, 1)\|}$$

$$= 0.999$$

AE : $\bar{r}_A = (1, 2, 3, 4)$ $\bar{r}_E = (5, 4, 5, 1)$

$\hat{r}_A = (2.5, 2.5, 2.5, 2.5)$ $\hat{r}_E = (3.75, 3.75, 3.75, 3.75)$

$\bar{r}_A - \hat{r}_A = (-1.5, -0.5, 0.5, 1.5)$ $\bar{r}_E - \hat{r}_E = (1.25, 0.25, 1.25, -2.75)$

$$s_p(A,E) = \frac{(-1.5, -0.5, 0.5, 1.5) \cdot (1.25, 0.25, 1.25, -2.75)}{\|(-1.5, -0.5, 0.5, 1.5)\| \|(1.25, 0.25, 1.25, -2.75)\|}$$

$$= -0.750$$

BC : $\bar{r}_B = (5, 4, 2, 2, 5)$ $\bar{r}_C = (1, 2, 4, 4, 3)$

$\hat{r}_B = (3.6, 3.6, 3.6, 3.6, 3.6)$ $\hat{r}_C = (2.8, 2.8, 2.8, 2.8, 2.8)$

$\bar{r}_B - \hat{r}_B = (1.4, 0.4, -1.6, -1.6, 1.4)$ $\bar{r}_C - \hat{r}_C = (-1.8, -0.8, 1.2, 1.2, 0.2)$

$$s_p(B,C) = \frac{(1.4, 0.4, -1.6, -1.6, 1.4) \cdot (-1.8, -0.8, 1.2, 1.2, 0.2)}{\|(1.4, 0.4, -1.6, -1.6, 1.4)\| \|(-1.8, -0.8, 1.2, 1.2, 0.2)\|}$$

$$= \underline{-0.809}$$

BD: $\bar{r}_B = (4, 2, 2)$ $\quad \bar{r}_D = (2, 3, 5)$

$\hat{r}_B = (2.67, 2.67, 2.67)$ $\quad \hat{r}_D = (3.33, 3.33, 3.33)$

$\bar{r}_B - \hat{r}_B = (1.33, -0.67, -0.67)$ $\quad \bar{r}_D - \hat{r}_D = (-1.33, -0.33, 1.67)$

$$s_p(B,D) = \frac{(1.33, -0.67, -0.67) \cdot (-1.33, -0.33, 1.67)}{\|(1.33, -0.67, -0.67)\| \|(-1.33, -0.33, 1.67)\|}$$

$$= \underline{-0.756}$$

BE: $\bar{r}_B = (4, 2, 2, 2, 5)$ $\quad \bar{r}_E = (3, 5, 4, 5, 1)$

$\hat{r}_B = (3, 3, 3, 3, 3)$ $\quad \hat{r}_E = (3.6, 3.6, 3.6, 3.6, 3.6)$

$\bar{r}_B - \hat{r}_B = (1, -1, -1, -1, 2)$ $\quad \bar{r}_E - \hat{r}_E = (-0.6, 1.4, 0.4, 1.4, -2.6)$

$$s_p(B,E) = \frac{(1, -1, -1, -1, 2) \cdot (-0.6, 1.4, 0.4, 1.4, -2.6)}{\|(1, -1, -1, -1, 2)\| \|(-0.6, 1.4, 0.4, 1.4, -2.6)\|}$$

$$= \underline{-0.951}$$

CD: $\bar{r}_C = (2, 4)$ $\quad \bar{r}_D = (2, 3)$

$\hat{r}_C = (3, 3)$ $\quad \hat{r}_D = (2.5, 2.5)$

$\bar{r}_C - \hat{r}_C = (-1, 1)$ $\quad \bar{r}_D - \hat{r}_D = (-0.5, 0.5)$

$$Sp(C,D) = \frac{(-1,1)\cdot(-0.5,0.5)}{\|(-1,1)\|\ \|(-0.5,0.5)\|}$$

$$= 0.999$$

CE : $\bar{r}_C = (2,4,4,3)$ $\quad \bar{r}_E = (3,5,5,1)$

$\hat{r}_C = (3.25, 3.25, 3.25, 3.25)$ $\quad \hat{r}_E = (3.5, 3.5, 3.5, 3.5)$

$\bar{r}_C - \hat{r}_C = (-1.25, 0.75, 0.75, -0.25)$ $\quad \bar{r}_E - \hat{r}_E = (-0.5, 1.5, 1.5, -2.5)$

$$Sp(C,E) = \frac{(-1.25, 0.75, 0.75, -0.25)\cdot(-0.5, 1.5, 1.5, -2.5)}{\|(-1.25, 0.75, 0.75, -0.25)\|\ \|(-0.5, 1.5, 1.5, -2.5)\|}$$

$$= 0.636$$

DE : $\bar{r}_D = (2,3,5)$ $\quad \bar{r}_E = (3,5,4)$

$\hat{r}_D = (3.33, 3.33, 3.33)$ $\quad \hat{r}_E = (4,4,4)$

$\bar{r}_D - \hat{r}_D = (-1.33, -0.33, 1.67)$ $\quad \bar{r}_E - \hat{r}_E = (-1,1,0)$

$$Sp(D,E) = \frac{(-1.33, -0.33, 1.67)\cdot(-1,1,0)}{\|(-1.33, -0.33, 1.67)\|\ \|(-1,1,0)\|}$$

$$= 0.327$$

Next, for simplicity, in calculating the pearson's similarity for movie pairs, I will be showing the vector values, mean, and similarity.

We will represent movies with number for simplicity too. Calculated in python.

LOTR => 1 $\quad$ Snatch => 3 $\quad$ The gentleman => 5

HP/HTPOZ => 2 $\quad$ LSATSB => 4 $\quad$ The Hobbit => 6

1-2 : $\bar{r}_1 = (5, 1)$ $\bar{r}_2 = (4, 2)$

$\hat{r}_1 = 3$ $\hat{r}_2 = 3$

$S_p(1,2) = 0.999$

1-3 : $\bar{r}_1 = (5, 5, 1)$ $\bar{r}_3 = (1, 2, 4)$

$\hat{r}_1 = 3.666$ $\hat{r}_3 = 2.333$

$S_p(1,3) = -0.945$

1-4 $\bar{r}_1 = (5, 5)$ $\tilde{r}_4 = (2, 2)$

$\hat{r}_1 = 5$ $\hat{r}_4 = 2$

$S_p(1,4) = 0$

1-5 $\bar{r}_1 = (5, 5, 1)$ $\bar{r}_5 = (3, 2, 4)$

$\hat{r}_1 = 3.666$ $\hat{r}_5 = 3$

$S_p(1,5) = -0.866$

1-6 $\bar{r}_1 = (5, 5, 1)$ $\bar{r}_6 = (4, 5, 3)$

$\hat{r}_2 = 3.666$ $\hat{r}_6 = 4$

$S_p(1,6) = 0.866$

2-3

$\bar{r}_2 = (4, 2, 2, 3)$

$\hat{r}_2 = 2.75$

$\bar{r}_3 = (2, 4, 3, 5)$

$\hat{r}_3 = 3.5$

$sp(2,3) = -0.405$

2-4

$\bar{r}_2 = (4, 2, 3)$

$\hat{r}_2 = 3$

$\bar{r}_4 = (2, 5, 4)$

$\hat{r}_4 = 3.666$

$sp(2,4) = -0.982$

2-5

$\bar{r}_2 = (4, 2, 3)$

$\hat{r}_2 = 3$

$\bar{r}_5 = (2, 4, 5)$

$\hat{r}_5 = 3.666$

$sp(2,5) = -0.654$

2-6

$\bar{r}_2 = (4, 2, 3)$

$\hat{r}_2 = 3$

$\bar{r}_6 = (5, 3, 1)$

$\hat{r}_6 = 3$

$sp(2,6) = 0.499$

3-4

$\bar{r}_3 = (1, 2, 3, 5)$

$\hat{r}_3 = 2.75$

$\bar{r}_4 = (2, 2, 5, 4)$

$\hat{r}_4 = 3.25$

$sp(3,4) = 0.683$

3-5

$\bar{r}_3 = (1, 2, 4, 5)$

$\hat{r}_3 = 3$

$sp(3,5) = 0.849$

$\bar{r}_5 = (3, 2, 4, 5)$

$\hat{r}_5 = 3.5$

3-6

$\bar{r}_3 = (1, 2, 4, 5)$

$\hat{r}_3 = 3$

$sp(3,6) = -0.855$

$\bar{r}_6 = (4, 5, 3, 1)$

$\hat{r}_6 = 3.25$

4-5

$\bar{r}_4 = (2, 2, 4)$

$\hat{r}_4 = 2.667$

$sp(4,5) = 0.945$

$\bar{r}_5 = (3, 2, 5)$

$\hat{r}_5 = 3.333$

4-6

$\bar{r}_4 = (2, 2, 4)$

$\hat{r}_4 = 2.666$

$sp(4,6) = -0.971$

$\bar{r}_6 = (4, 5, 1)$

$\hat{r}_6 = 3.333$

5-6

$\bar{r}_5 = (3, 2, 4, 5)$

$\hat{r}_5 = 3.5$

$sp(5,6) = -0.983$

$\bar{r}_6 = (4, 5, 3, 1)$

$\hat{r}_6 = 3.25$

The ave used here is from common entries. Yuhan said this is okay because I've done it before announcement, and it actually makes more sense. I am putting a note here that I am using a different mean as instructed by YuHan.

(2)

$$\hat{r}_{u,i} = \bar{r}_u + \frac{\sum_{j \in K_u} S(u,u_j)(r_{u_j,i} - \bar{r}_{u_j})}{\sum_{j \in K_u} |S(u,u_j)|},$$

$\in$ in this formula, $\hat{r}$ is the grade and $\bar{r}$ is average.

I will use the other way around (like in the lecture)

$|K| = 3$

A,2: closest $\Rightarrow k \in \{B, C, D\}$

$A-B = 0.866$, $A-C = -0.828$, $A-D = 0.999$

$$\bar{r}_{A,2} = 3 + \frac{0.866\times(4-3.33) + (-0.828)(2-2.8) + (0.999)(2-3.33)}{|0.866| + |-0.828| + |0.999|}$$

$$= 2.966$$

In this section, the mean is from all entries.
But the "similarity" value uses mean from commonly rated entry.
Yuhan verified that this is okay as long as I'm consistent.

C,4: closest $\Rightarrow k \in \{A, B, D\}$

$C-A = -0.828$ $C-B = -0.809$ $C-D = 0.999$

$$\bar{r}_{C,4} = 2.8 + \frac{-0.828(2-3) + (-0.809)(2-3.33) + 0.999(5-3.33)}{|-0.828| + |-0.809| + |0.999|}$$

$$= 4.155$$

D,1: closest 3 $\Rightarrow$ $k \in \{A, B, C\}$

$D-A = 0.999$ $D-B = -0.756$ $D-C = 0.999$

$$\bar{r}_{D,1} = 3.333 + \frac{0.999(5-3) + (-0.756)(5-3.33) + 0.999(1-2.8)}{|0.999| + |-0.756| + |0.999|}$$

$$= 2.947$$

D, 5 : closest 3 => $k \in \{A, B, D\}$

$D-A = 0.999$ $D-B = -0.756$ $D-C = 0.999$

$$\bar{r}_{D,5} = 3.333 + \frac{0.999(3-3) - 0.756(2-3.33) + 0.999(4-2.8)}{|0.999| + |-0.756| + |0.999|}$$

$$= 4.133$$

D, 6 : closest 3 => $k \in \{A, B, C\}$

$D-B = -0.756$ $D-C = 0.999$ $D-A = 0.999$

$$\bar{r}_{D,6} = 3.333 + \frac{0.999(4-3) - 0.756(5-3.33) + 0.999(3-2.8)}{|0.999| + |-0.756| + |0.999|}$$

$$= 3.311$$

E, 1 : closest 3 => $k \in \{A, B, C\}$

$E-A = -0.750$ $E-B = -0.951$ $E-C = 0.636$

$$\bar{r}_{E,1} = 3.6 + \frac{(-0.750)(5-3) - 0.951(5-3.33) + 0.636(1-2.8)}{|-0.750| + |-0.951| + |0.636|}$$

$$= 1.790$$

The answer may be different from the answer key because the Similarity calculated in 1.1 is using mean from commonly rated entries. However, the mean used for the formula in 1.2 uses the mean of all entries. So there are 2 types of mean involved. Please check my working instead of final answer. Thank you.

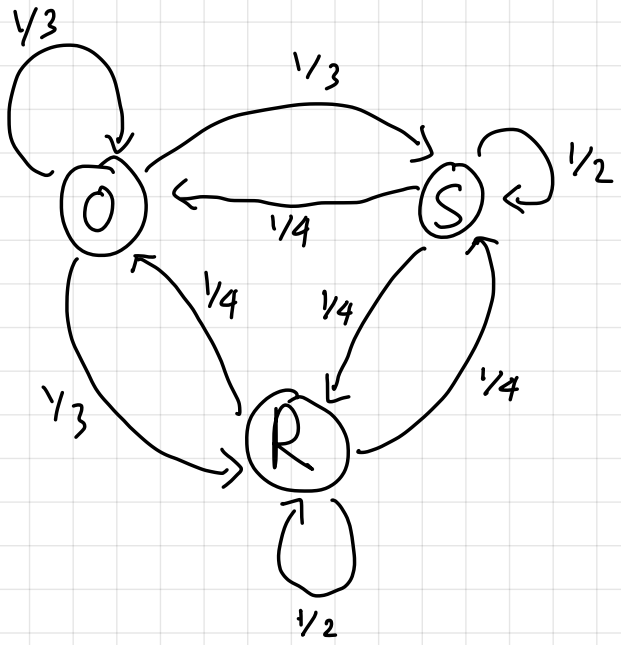
Problem 2 (30 points). Consider a Markov chain with three states, Overcast, Rain, and Sunny. The transition probabilities are given in the following table. The $(i, j)th$ entry of the matrix is the probability that the next day to be j if today is i. November 29, 2020 is Rain.

		1	2	3
		O	S	R
1	O	1/3	1/3	1/3
2	S	1/4	1/2	1/4
3	R	1/4	1/4	1/2

O -> 1
S -> 2
R -> 3

1. Draw the state transition diagram with arrows annotating the transition probabilities.
2. What is the probability that it will be Sunny on November 30th, 2020?
3. What is the probability that it will Rain on December 2nd, 2020?
4. What is the probability that it will Rain every day until December 5, 2020 (including it)?
5. Compute the probability that it will Rain on December 6, 2020?

(1)



(2) $P(X_{t+1} = S \mid X_t = R) = \frac{1}{4}$ $\leftarrow$ $P_{32}^{(1)} = \frac{1}{4}$

or

Find $P_{32}^{(1)}$

$$\begin{bmatrix} 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/4 & 1/2 & 1/4 \\ 1/4 & 1/4 & 1/2 \end{bmatrix} = \begin{bmatrix} 1/4 & 1/4 & 1/2 \end{bmatrix} \Rightarrow P_{32}^{(1)} = \frac{1}{4}$$

↑ index 3 for R

↑ index 2 for S

③ 29th (R) → 30th (any) → 1st (any) → 2nd (R) ← $P_{33}^{(3)}$

$$P^3 = \begin{bmatrix} 0.27315 & 0.36343 & 0.36343 \\ 0.27257 & 0.37153 & 0.35590 \\ 0.27257 & 0.35590 & 0.37153 \end{bmatrix}$$ ← calculated in python

```
import numpy as np
from numpy.linalg import matrix_power

temp = np.array([[1/3,1/3,1/3],[1/4,1/2,1/4],[1/4,1/4,1/2]])
print(temp)
power3 = matrix_power(temp,3)
print(power3)
```

```
[[0.33333333 0.33333333 0.33333333]
 [0.25       0.5        0.25       ]
 [0.25       0.25       0.5        ]]
[[0.27314815 0.36342593 0.36342593]
 [0.27256944 0.37152778 0.35590278]
 [0.27256944 0.35590278 0.37152778]]
```

$$P_{33}^{(3)} = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 0.27315 & 0.36343 & 0.36343 \\ 0.27257 & 0.37153 & 0.35590 \\ 0.27257 & 0.35590 & 0.37153 \end{bmatrix}$$

$$= [0.27257 \quad 0.35590 \quad 0.37153] \Rightarrow \underline{0.37153}$$

3rd entry

④ dec 5 dec 4 | dec 4 dec 3 | dec 3 dec 2

$$P(X_{t+6}=R \mid X_{t+5}=R) \times P(X_{t+5}=R \mid X_{t+4}=R) \times P(X_{t+4}=R \mid X_{t+3}=R)$$

dec 2 dec 1 | dec 1 nov 30 | nov 30 nov 29

$$P(X_{t+3}=R \mid X_{t+2}=R) \times P(X_{t+2}=R \mid X_{t+1}=R) \times P(X_{t+1}=R \mid X_t=R)$$

$$= \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \underline{\frac{1}{64}}$$

⑤ $p_{33}^{(7)}$ = 3rd index of $\vec{u} \cdot P^7$ (initial state)

$$P^7 = \begin{bmatrix} 0.27273 & 0.36364 & 0.36364 \\ 0.27273 & 0.36367 & 0.36361 \\ 0.27273 & 0.36361 & 0.36367 \end{bmatrix} \leftarrow \text{calculated in python}$$

```
temp = np.array([[1/3,1/3,1/3],[1/4,1/2,1/4],[1/4,1/4,1/2]])
print(temp)
power7 = matrix_power(temp,7)
print(power7)
```

```
[[0.33333333 0.33333333 0.33333333]
 [0.25       0.5        0.25       ]
 [0.25       0.25       0.5        ]]
[[0.27272729 0.36363635 0.36363635]
 [0.27272727 0.36366689 0.36360585]
 [0.27272727 0.36360585 0.36366689]]
```

$p_{33}^{(7)} = [0\ 0\ 1] \cdot P^7 = [0.27273 \quad 0.36361 \quad 0.36367]$ (3rd entry)

$\Rightarrow$ **0.36367**