Homework 4

Recitation Problems

**Problem 9.2.1**: Three computers A, B, C have numerical features listed below

|  |  |  |  |
| --- | --- | --- | --- |
| Feature | A | B | C |
| Processor Speed | 3.06 | 2.68 | 2.92 |
| Disk Size | 500 | 320 | 640 |
| Main-Memory Size | 6 | 4 | 6 |

We may imagine these values as defining a vector for each computer; for instance,

A’s vector is [3.06, 500, 6]. We can compute the cosine distance between

any two of the vectors, but if we do not scale the components, then the disk

size will dominate and make differences in the other components essentially invisible.

Let us use 1 as the scale factor for processor speed, α for the disk size and β for the main memory size.

(a) In terms of α and β, compute the cosines of the angles between the vectors

for each pair of the three computers.

(b)What are the angles between the vectors if α = β = 1?

(c) What are the angles between the vectors if α = 0.01 and β = 0.5?

**Answer:**

Cosine of the angle between two vector is represented as

= ∑ x.y / (|x|. |y|)

|x|= length of vector x =√ (∑ (x2))

|y|= length of vector y =√ (∑ (y2))

a) After applying Scaling factor

|  |  |  |  |
| --- | --- | --- | --- |
| Feature | A | B | C |
| Processor Speed | 3.06(1) | 2.68(1) | 2.92(1) |
| Disk Size | 500α | 320α | 640α |
| Main Memory size | 6β | 4β | 6β |

Cosine of the angle between A[3.06 500α 6β]and B[2.68 320α 4β]

= 3.06\*2.68+ 500α\* 320α+ 6β\* 4β/ (√ (3.062+500α2+6β2 ) \* √ (2.682+320α2+4β2))

= 8.2008 + 160000α2 +24β2/ √9.3636 +250000α2+36β2 \* √ 7.1824 +102400α2 +16β2

Cosine of the angle between A [3.06 500α 6β] and C [ 2.92 640α 6β]

= 8.9352+ 320000α2+ 36β2/√9.3636 +250000α2+36β2 \* √8.5264+409600α2+36β2

Cosine of the angle between B [2.68 320α 4β] and C [ 2.92 640α 6β]

= 7.8256+204800α2+24β2/ √ 7.1824 +102400α2 +16β2 \* √8.5264+409600α2+36β2

b) the angles between the vectors if α = β = 1

Cosine of the angle between A [3.06 500α 6β]and B[2.68 320α 4β]

= 8.2008 + 160000α2 +24β2/ √9.3636 +250000α2+36β2 \* √ 7.1824 +102400α2 +16β2

=8.2008 + 160000\*12 +24\*12/ √9.3636 +250000\*12+36\*12 \* √ 7.1824 +102400\*12 +16\*12

= 0.99999733

theta =cos-1(0.99999733)

= 0.132

Cosine of the angle between A[3.06 500α 6β] and C [ 2.92 640α 6β]

= 8.9352+ 320000α2+ 36β2/√9.3636 +250000α2+36β2 \* √8.5264+409600α2+36β2

= 8.9352+ 320000\*12+ 36\*12/√9.3636 +250000\*12+36\*12 \* √8.5264+409600\*12+36\*12

theta =cos-1 (0.99999534)

= 0.175

Cosine of the angle between B[2.68 320α 4β] and C [ 2.92 640α 6β]

= 7.8256+204800α2+24β2/ √ 7.1824 +102400α2 +16β2 \* √8.5264+409600α2+36β2

= 7.8256+204800\*12+24\*12/ √ 7.1824 +102400\*12 +16\*12 \* √8.5264+409600\*12+36\*12

theta =cos-1 (0.99998785)

= 0.282

c)the angles between the vectors if α = 0.01 and β = 0.5

Cosine of the angle between A [3.06 500α 6β]and B[2.68 320α 4β]

= 3.06\*2.68+ 500α\* 320α+ 6β\* 4β/ √3.062+500α2+6β2 \* √ 2.682+320α2+4β2

= 8.2008 + 160000\*0.012 +24\*0.52/ √9.3636 +250000\*0.012+36\*0.52 \* √ 7.1824 +102400\*0.012 +16\*0.52

theta =cos-1 (0.9908815)

=7.74

Cosine of the angle between A[3.06 500α 6β] and C [ 2.92 640α 6β]

= 8.9352+ 320000\*α2+ 36β2/√9.3636 +250000α2+36β2 \* √8.5264+409600α2+36β2

= 8.9352+ 320000\*0.012+ 36\*0.52/√9.3636 +250000\*0.012+36\*0.52 \* √8.5264+409600\*0.012+36\*0.52

theta =cos-1 ( 0.99155471)

= 7.45

Cosine of the angle between B[2.68 320α 4β] and C [ 2.92 640α 6β]

= 7.8256+204800α2+24β2/ √ 7.1824 +102400α2 +16β2 \* √8.5264+409600α2+36β2

= 7.8256+204800\*0.012+24\*0.52/ √ 7.1824 +102400\*0.012 +16\*0.52 \* √8.5264+409600\*0.012+36\*0.52

theta = cos-1 (0.96917792)

= 14.26

(d) One fair way of selecting scale factors is to make each inversely proportional to the average value in its component. What would be the values of α and β, and what would be the angles between the vectors?

Answer-

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Feature | A | B | C | Average of Compnents |
| Processor Speed | 3.06 | 2.68 | 2.92 | 2.88 |
| Disk Size | 500 | 320 | 640 | 486.67 |
| Main Memory size | 6 | 4 | 6 | 5.33 |

Scale factor for processor speed =1/2.88 = 0.347

Scale factor for Disk size, α=1/486.67 = 0.002

Scale factor for Main memory size, β =1/5.33 = 0.187

|  |  |  |  |
| --- | --- | --- | --- |
| Feature | A | B | C |
| Processor Speed | 1.06 | 0.93 | 1.01 |
| Disk Size | 1 | 0.64 | 1.28 |
| Main Memory size | 1.12 | 0.75 | 1.12 |

Cosine of the angle between A[1.06 1 1.12]and B[0.93 0.64 0.75]

= 1.06\*0.93+ 1\*0.64+ 1.12\*0.75/ √1.062+12+1.122 \* √ 0.932+0.642+0.752

theta = cos-1 (0.9898) =8.19

Cosine of the angle between A[1.06 1 1.12] and C [1.01 1.28 1.12]

= 1.06\*1.01+ 1\*1.28+ 1.12\*1.12/√1.062+12+1.122 \* √1.012+1.282+1.122

theta =cos-1 ( 0.9915) = 7.475

Cosine of the angle between B[0.93 0.64 0.75] and C [1.01 1.28 1.12]

= 0.93\*1.01+0.64\*1.28+0.75\*1.12/√ 0.932+0.64^2+0.752 \* √1.012+1.282+1.122

theta =cos-1 (0.9692) = 14.257

**Problem 9.2.3** : A certain user has rated the three computers of Exercise 9.2.1

as follows: A: 4 stars, B: 2 stars, C: 5 stars.

(a) Normalize the ratings for this user.

(b) Compute a user profile for the user, with components for processor speed,

disk size, and main memory size, based on the data of Exercise 9.2.1.

**Answer** -

1. Normalized ratings for user X ,the average rating by a user X is 2

Average = (4+2+5)/3 = 11/3

A = 4 – (11/3) = (12-11)/3 = 1/3 = 0.333333

B = 2-(11/3) = (6-11)/3 = -5/3 = -1.666666

C = 5- (11/3) = (15-11)/3 = 4/3 = 1.33333

|  |  |  |  |
| --- | --- | --- | --- |
| Feature | A | B | C |
| User ratings | 0.34 | -1.66 | 1.34 |

b)

Processor Speed

= (3.06 \* 0.34) – (2.68\* 1.66) + (2.92\*1.34)

= (1.02) -(4.46)+(3.89)

= 0.4467

Disk Size

= (500\*0.34) – (320\*1.66) +(640\*1.34)

= 166.66-533.33+853.3

=486.667

Main Memory Size

=(6 \* 0.34) – (4\* 1.66) + (6\*1.34)

= (2.04) -(6.64)+(8.04)

= 3.44

**Exercise 9.3.1 :** Figure 9.8 is a utility matrix, representing the ratings, on a 1–5 star scale, of eight items, a through h, by three users A, B, and C. Compute the following from the data of this matrix.

(a) Treating the utility matrix as boolean, compute the Jaccard distance between each pair of users.

(b) Repeat Part (a), but use the cosine distance.

(c)Treat ratings of 3, 4, and 5 as 1 and 1, 2, and blank as 0. Compute the Jaccard distance between each pair of users.

(d) Repeat Part (c), but use the cosine distance.

(e) Normalize the matrix by subtracting from each nonblank entry the average value for its user.

(f) Using the normalized matrix from Part (e), compute the cosine distance between each pair of users

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | a | b | c | d | e | f | g | h |
| A | 4 | 5 |  | 5 | 1 |  | 3 | 2 |
| B |  | 3 | 4 | 3 | 1 | 2 | 1 |  |
| C | 2 |  | 1 | 3 |  | 4 | 5 | 3 |

Answer -

(a) Treating the utility matrix as boolean, compute the Jaccard distance between each pair of users.

Jaccard similarity between A and B = 4/8 =0.5

Jaccard similarity between A and C = 4/8 =0.5

Jaccard similarity between B and C = 4/8 =0.5

Sim(A,B) = Sim(A,C)=Sim(B,C)

(b) Repeat Part (a), but use the cosine distance.

Cosine Similarity between A and B = 0.601

Cosine Similarity between A and C = 0.615

Cosine Similarity between B and C = 0.5138

Sim(A,C)>Sim(A,B)>Sim(B,C)

(c)Treat ratings of 3, 4, and 5 as 1 and 1, 2, and blank as 0. Compute the Jaccard distance between each pair of users.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | a | b | c | d | e | f | g | h |
| A | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 |
| B | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| C | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 |

Jaccard Similarity = M11/(M01+M10+M11)

Jaccard Distance = (M10+ M01)/(M01+M10+M11) = 1-J

Jaccard Similiarity between A and B = 2/5 = 0.4

Jaccard Distance = 0.6

Jaccard Similiarity between A and C = 2/6 = 0.3333

Jaccard Distance = 0.666667

Jaccard Similiarity between B and C = 1/6 = 0.1666

Jaccard Distance = 0.833334

Sim(A,B)>Sim(A,C)>Sim(B,C)

(d) Repeat Part (c), but use the cosine distance.

Cosine Similarity = ∑AiBi /(√Ai2√ Bi2)

Cosine Distance = cos-1(Similarity)/π

Cosine Similarity between A and B = 0.5774

Cosine Distance = 0.4226

Cosine Similarity between A and C = 0.5

Cosine Distance = 0.5

Cosine Similarity between B and C = 0.2887

Cosine Distance = 0.7113

(e) Normalize the matrix by subtracting from each nonblank entry the average value for its user.

Normalizing the utility matrix

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | a | b | c | d | e | f | g | h |
| A | 0.66 | 1.66 |  | 1.66 | 1 |  | -0.34 | -1.34 |
| B |  | 0.66 | 1.66 | 0.66 | -1.34 | -0.34 | -1.34 |  |
| C | -1 |  | -2 | 0 |  | 1 | 2 | 0 |

(f) Using the normalized matrix from Part (e), compute the cosine distance between each pair of users

Cosine similarity between A and B = 0.584

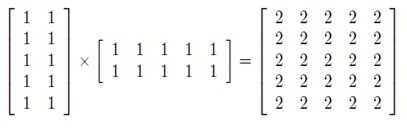
Cosine similarity between A and C = -0.1154

Cosine similarity between B and C = -0.73955

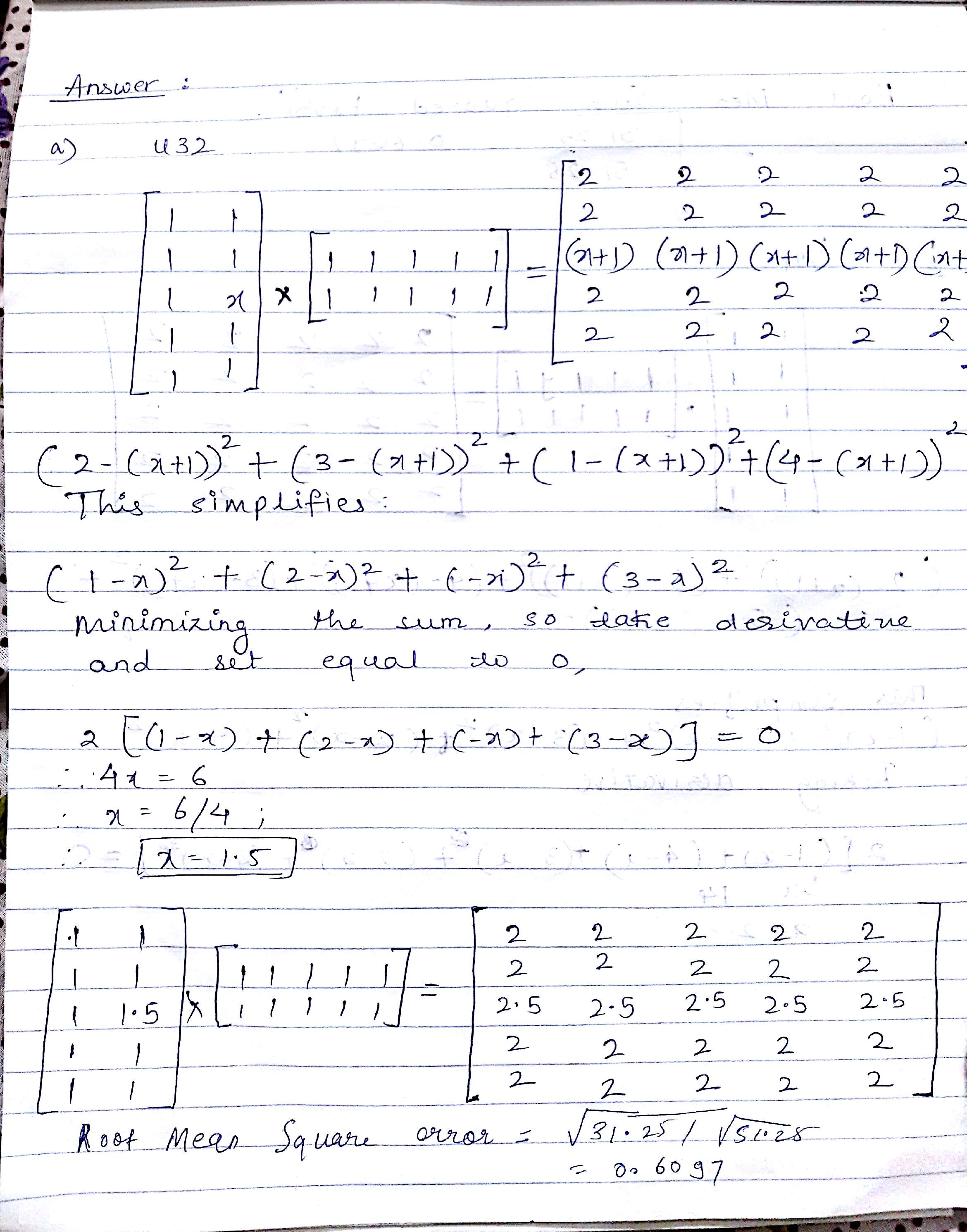
**Exercise 9.4.1** : Starting with the decomposition of Fig. 9.10, we may choose

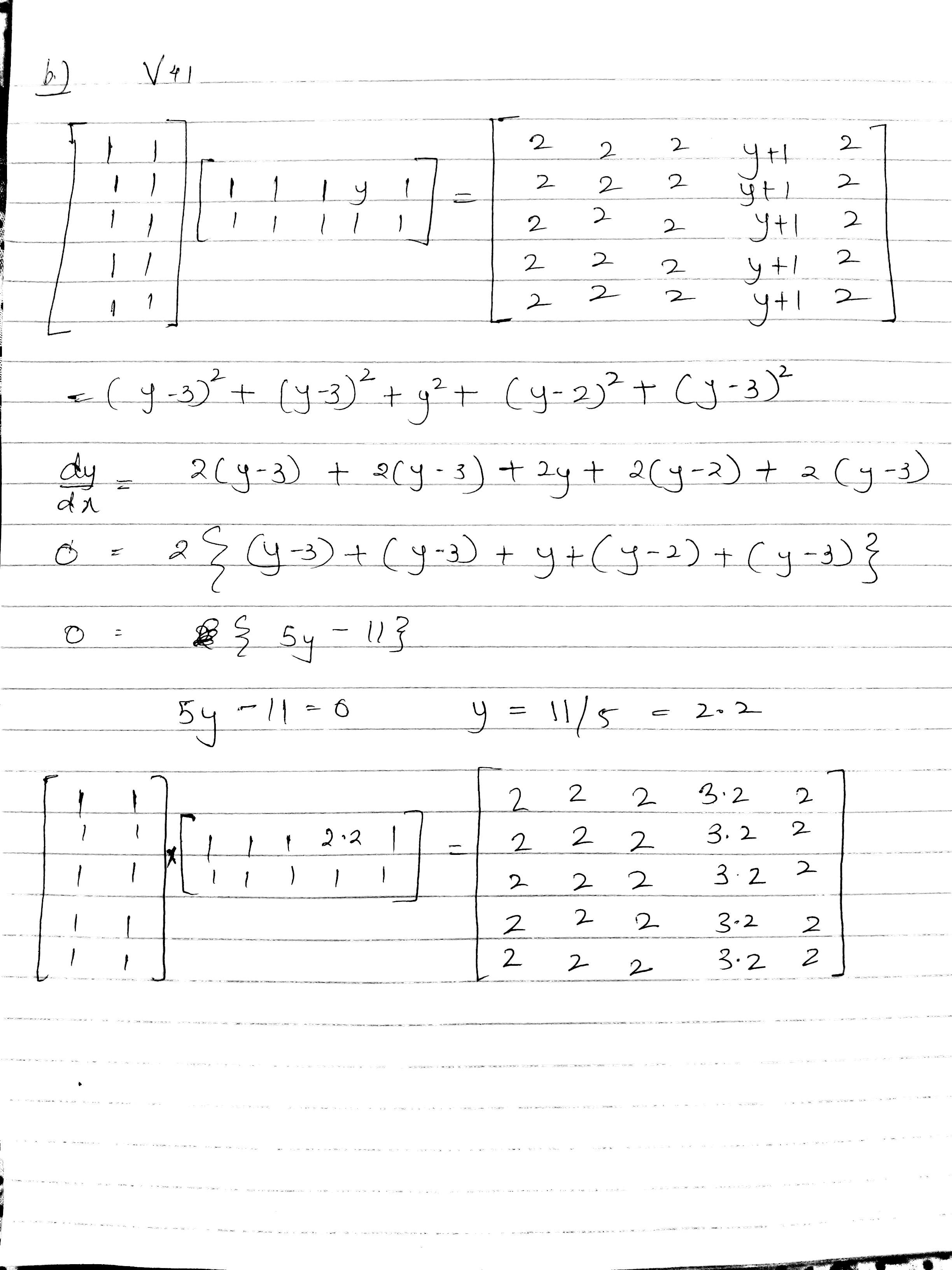
any of the 20 entries in U or V to optimize first. Perform this first optimization

step assuming we choose: (a) u32 (b) v41.



Answer –





**2.1 Problem 1**

Load the Movielens 100k dataset (ml-100k.zip) into Python using Pandas dataframes. Build a user profile on unscaled data for both users 200 and 15, and calculate the cosine similarity and distance between the user's preferences and the item/movie 95. Which user would a recommender system suggest this movie to?

**Solution:**

This problem has been solved by using the methodology of recommendation system building based on prediction of the ratings. Once the user profile was built and we got the movies they rated we were able to find the feature/genre they based on the other side we trained the system with cross validation on 75% of the data and tested on 25% for prediction. However, to get that prediction we used the mean average user ratings and found the dot similarity to the item/movie, we found that the similarity in ratings to movie 95 for both user 200 and 15.

User 200 similarity : 0.79654824927062673

User 15 similarity : 0.64315408653336836

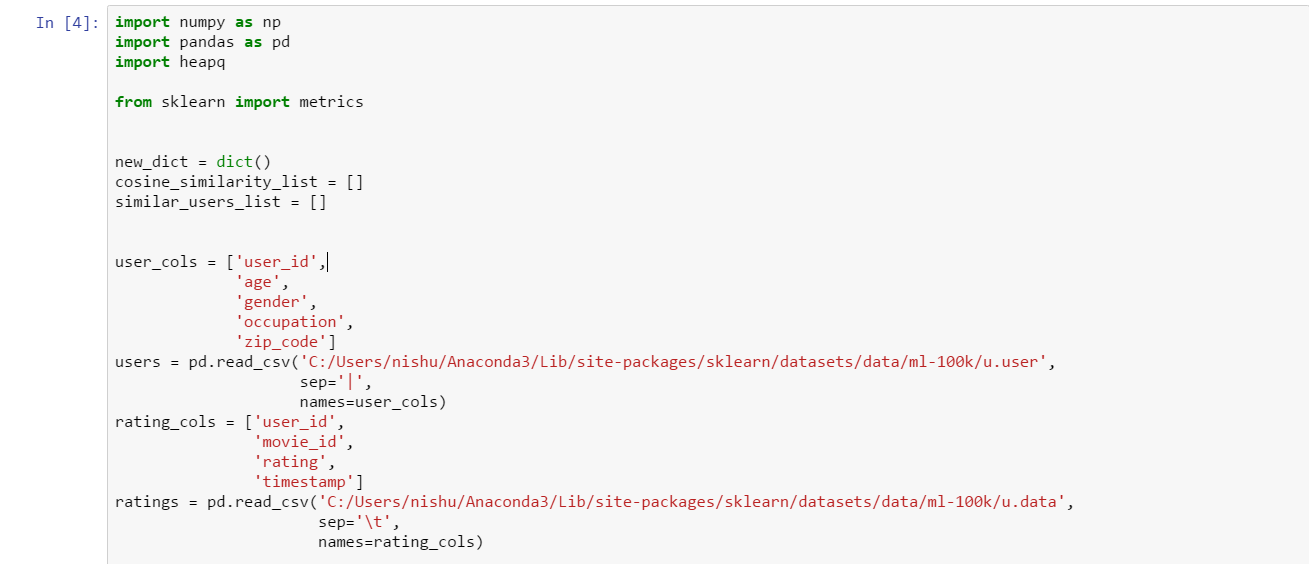
Here user 200 has higher similarity which led the recommendation system choose user 200.

**2.2 Problem 2**

Load the Movielens 100k dataset (ml-100k.zip) into Python using Pandas dataframes. Convert the ratings data into a utility matrix representation, and the 10 most similar users for user 1 based on cosine similarity of the user ratings data. Based on the average of of the ratings for item 508 from the similar users, what is the expected rating for this item for user 1?

**Solution:**

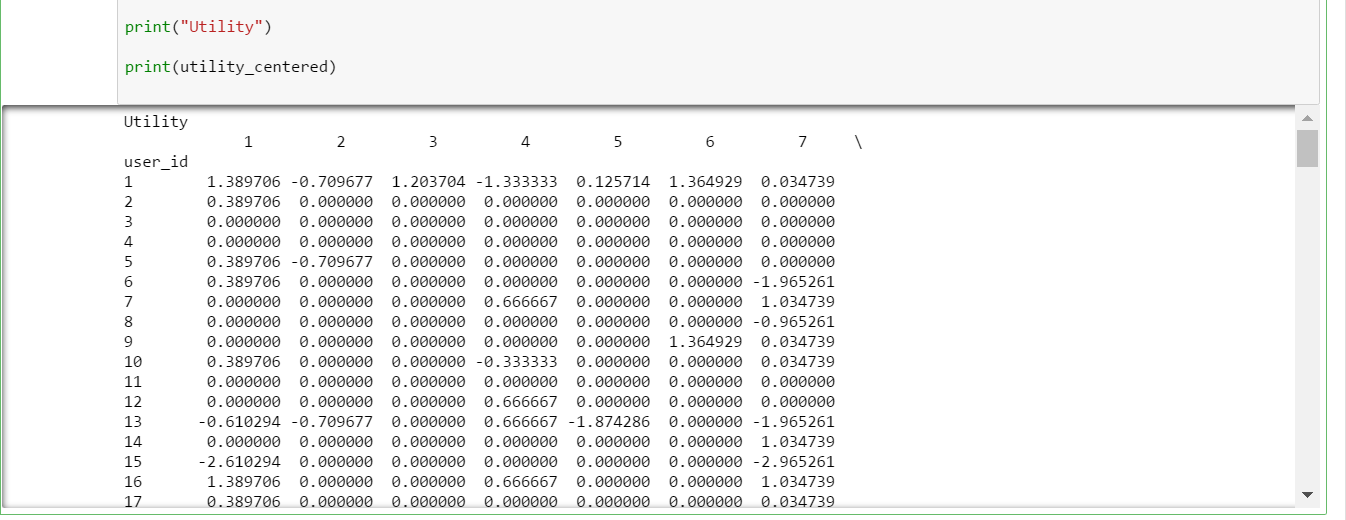
Import data

****

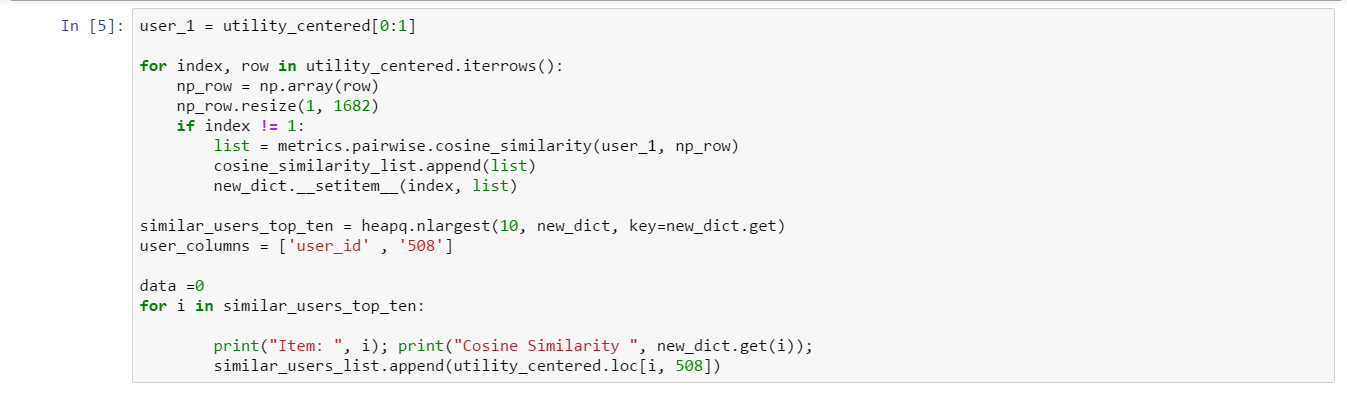
Create the utility matrix

****

Utility matrix

****

Calculate the cosine similarity and find out the 10 most similar users

****

Similar users and their expected ratings. 10 most similar users based on cosine similarity are as follows.

****

Hence, the expected rating of user 1 for item 508 will be 0.268965517241.