Solutions for Homework 1

2.3.1 0.5'

(c) Map: for each integer i in the file, emit key-value pair (i, 1)

Reduce: turn the value list into 1.

Note the result is obtained from the keys of the output.

6.1.1 0.5'

(a) frequent items are: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20

6.1.5 1'

(a) {5, 7} → 2

The baskets containing both item 5 and item 7 are basket 35 and basket 70, in which only basket 70 also contains item 2. Hence, the confidence of the association rule {5,7}→2 is 1/2.

(b) {2, 3, 4} → 5

The baskets whose numbers are the multiples of 12 contain item set {2,3,4} as a subset, there are 8 such baskets, while only those whose numbers are the multiples of 60 contain item set {2,3,4,5} as a subset, there are 1 such basket. Hence, the confidence of the association rule {2,3,4}→5 is 1/8.

6.2.1 0.5'

For any pair {i, j} in the triangular matrix, the corresponding index k is $(i-1)\left(20-\frac{i}{2}\right)+j-i$

Solve the equation $100=(i-1)\left(20-\frac{i}{2}\right)+j-i, with\ 1 \le i < j \le 20$, we can get the pair {7,8}.

6.2.6 1'

(a)

Candidate itemsets of size 1 (C_1):

Items with index 1~100

Truly frequent itemsets of size 1 (L_1):

Items with index 1~20

Candidate itemsets of size 2 (C_2):

$\forall 1 \le i < j \le 20\ \{i,j\}$

Truly frequent itemsets of size 2 (L_2):

The same as 6.1.1 (b)

Candidate itemsets of size 3 (C_3):

$\forall \mathrm{c}, |\mathrm{c}| = 3\ and\ \exists c_i, c_j \in L_2, \mathrm{c} = c_i \cup c_j$

Truly frequent itemsets of size 3 (L_3):

{1, 2, 3}, {1, 2, 4}, {1, 2, 5}, {1, 2, 6}, {1, 2, 7}, {1, 2, 8}, {1, 2, 9}, {1, 2, 10}, {1, 2, 12}, {1, 2, 14}, {1, 2, 16}, {1, 2, 18}, {1, 2, 20},

{1, 3, 4}, {1, 3, 5}, {1, 3, 6}, {1, 3, 9}, {1, 3, 12}, {1, 3, 15}, {1, 3, 18},

{1, 4, 5}, {1, 4, 6}, {1, 4, 8}, {1, 4, 10}, {1, 4, 12}, {1, 4, 16}, {1, 4, 20},

{1, 5, 10}, {1, 5, 15}, {1, 5, 20},
{1, 6, 9}, {1, 6, 12}, {1, 6, 18},
{1, 7, 14}, {1, 8, 16}, {1, 9, 18}, {1, 10, 20},
{2, 3, 4}, {2, 3, 6}, {2, 3, 9}, {2, 3, 12}, {2, 3, 18},
{2, 4, 5}, {2, 4, 6}, {2, 4, 8}, {2, 4, 10}, {2, 4, 12}, {2, 4, 16}, {2, 4, 20},
{2, 5, 10}, {2, 5, 20},
{2, 6, 9}, {2, 6, 12}, {2, 6, 18},
{2, 7, 14}, {2, 8, 16}, {2, 9, 18}, {2, 10, 20},
{3, 4, 12}, {3, 5, 15},
{4, 5, 10}, {4, 5, 20}, {4, 6, 12},
{5, 10, 20},
{6, 9, 18}

Candidate itemsets of size 4 (C_4):

$\forall \mathrm{c}, |\mathrm{c}| = 4 \text{ and } \exists c_i, c_j \in L_3, \mathrm{c} = c_i \cup c_j$

Truly frequent itemsets of size 4 (L_4):

{1, 2, 3, 4}, {1, 2, 3, 6}, {1, 2, 3, 9}, {1, 2, 3, 12}, {1, 2, 3, 18},
{1, 2, 4, 5}, {1, 2, 4, 6}, {1, 2, 4, 8}, {1, 2, 4, 10}, {1, 2, 4, 12}, {1, 2, 4, 16}, {1, 2, 4, 20},
{1, 2, 5, 10}, {1, 2, 5, 20}, {1, 2, 6, 9}, {1, 2, 6, 12}, {1, 2, 6, 18},
{1, 2, 7, 14}, {1, 2, 8, 16}, {1, 2, 9, 18}, {1, 2, 10, 20},
{1, 3, 4, 12}, {1, 3, 5, 15},
{1, 4, 5, 10}, {1, 4, 5, 20}, {1, 4, 6, 12},
{1, 5, 10, 20}, {1, 6, 9, 18}
{2, 3, 4, 12}, {2, 4, 5, 10}, {2, 4, 5, 20}, {2, 4, 6, 12},
{2, 5, 10, 20}, {2, 6, 9, 18}
{4, 5, 10, 20}

Candidate itemsets of size 5 (C_5):

$\forall \mathrm{c}, |\mathrm{c}| = 5 \text{ and } \exists c_i, c_j \in L_4, \mathrm{c} = c_i \cup c_j$

Truly frequent itemsets of size 5 (L_5):

{1, 2, 3, 4, 12}, {1, 2, 4, 5, 10}, {1, 2, 4, 5, 20}, {1, 2, 4, 6, 12}, {1, 2, 5, 10, 20}, {1, 2, 6, 9, 18},
{1, 4, 5, 10, 20}, {2, 4, 5, 10, 20}

Candidate itemsets of size 6 (C_6):

$\forall \mathrm{c}, |\mathrm{c}| = 6 \text{ and } \exists c_i, c_j \in L_5, \mathrm{c} = c_i \cup c_j$

Truly frequent itemsets of size 6 (L_6):

{1, 2, 4, 5, 10, 20}

Since there’s only one item set, there will be no candidate itemsets of size 7, the algorithm stops here.

6.3.1 1’

(a)

item	1	2	3	4	5	6
support	4	6	8	8	6	4

pair	{1, 2}	{1, 3}	{1, 4}	{1, 5}	{1, 6}	{2, 3}	{2, 4}	{2, 5}
support	2	3	2	1	0	3	4	2

pair	{2, 6}	{3, 4}	{3, 5}	{3, 6}	{4, 5}	{4, 6}	{5, 6}	
support	1	4	4	2	3	3	2	

(b)

pair	{1, 2}	{1, 3}	{1, 4}	{1, 5}	{1, 6}	{2, 3}	{2, 4}	{2, 5}
bucket	2	3	4	5	6	6	8	10
pair	{2, 6}	{3, 4}	{3, 5}	{3, 6}	{4, 5}	{4, 6}	{5, 6}	
bucket	1	1	4	7	9	2	8	

(c)

bucket	0	1	2	3	4	5	6	7	8	9	10
support	0	5	5	3	6	1	3	2	6	3	2

The frequent buckets are those with support above 4, thus 1, 2, 4, 8.

(d)

As only pairs in frequent buckets will be counted on the second pass of PCY, they are {1, 2}, {1, 4}, {2, 4}, {2, 6}, {3, 4}, {3, 5}, {4, 6}, {5, 6}

11.1.3 0.5’

$(\mathrm{a}-\lambda)(\mathrm{d}-\lambda)(\mathrm{f}-\lambda)+2\mathrm{bce}-(\mathrm{a}-\lambda)\mathrm{e}^2-(d-\lambda)c^2-(f-\lambda)b^2=0;$

11.2.1 0.5’

(a) $\mathrm{M}^T M=\begin{bmatrix}1 & 2 & 3 & 4\\ 1 & 4 & 9 & 16\end{bmatrix}\begin{bmatrix}1 & 1\\ 2 & 4\\ 3 & 9\\ 4 & 16\end{bmatrix}=\begin{bmatrix}30 & 100\\ 100 & 354\end{bmatrix}$

$M\mathrm{M}^T=\begin{bmatrix}1 & 1\\ 2 & 4\\ 3 & 9\\ 4 & 16\end{bmatrix}\begin{bmatrix}1 & 2 & 3 & 4\\ 1 & 4 & 9 & 16\end{bmatrix}=\begin{bmatrix}2 & 6 & 12 & 20\\ 6 & 20 & 42 & 72\\ 12 & 42 & 90 & 156\\ 20 & 72 & 156 & 272\end{bmatrix}$

(b) Eigenpairs of $\mathrm{M}^T M$ are:

$$\lambda_1=382.3786,\lambda_2=1.6214,x_1=\begin{bmatrix}0.273\\ 0.962\end{bmatrix},x_2=\begin{bmatrix}-0.962\\ 0.273\end{bmatrix}$$

(c) Eigenvalues of $M\mathrm{M}^T$ are:

$$\lambda_1=382.3786,\lambda_2=1.6214,\lambda_3=\lambda_4=0$$

(d)Eigenvectors of $M\mathrm{M}^T$ are:

$$\mathrm{x}_1=\begin{bmatrix}0.0632\\ 0.2247\\ 0.4847\\ 0.8430\end{bmatrix},\mathrm{x}_2=\begin{bmatrix}0.5411\\ 0.6534\\ 0.3369\\ -0.4084\end{bmatrix},\mathrm{x}_3=\begin{bmatrix}-0.7613\\ 0.1570\\ 0.5519\\ -0.3021\end{bmatrix},\mathrm{x}_4=\begin{bmatrix}-0.3517\\ 0.7056\\ -0.5891\\ 0.1769\end{bmatrix}$$

11.3.2 0.5’

We can map [0, 3, 0, 0, 4] into “concept space” by multiplying it by V, getting the representation of Leslie in concept space which is [1.74, 2.84]. Multiplying [1.74, 2.84] by V^{T}, we get [1.0092, 1.0092, 1.0092, 2.0164, 2.0164] which can be used to represent how well Leslie would like the other movies.

11.4.2 1’

(a)

The columns for both Matrix and Alien are $[1\ 3\ 4\ 5\ 0\ 0\ 0]^T$, the row for Jim is $[3\ 3\ 3\ 0\ 0\ ]$, and the row for John is $[4\ 4\ 4\ 0\ 0\ ]$.

Scale the two columns by $\sqrt{rq_1} = \sqrt{rq_2} = \sqrt{2 \times 51/243} = 0.6478835$ and get matrix C:

$$\mathrm{C} = \begin{bmatrix} 1.54 & 1.54 \\ 4.63 & 4.63 \\ 6.17 & 6.17 \\ 7.72 & 7.72 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

Scale the row for Jim by $\sqrt{rp_2} = \sqrt{2 \times 27/243} = 0.471$ and the row for John by $\sqrt{rp_3} = \sqrt{2 \times 48/243} = 0.6285$

$$\mathrm{R} = \begin{bmatrix} 6.37 & 6.37 & 6.37 & 0 & 0 \\ 6.37 & 6.37 & 6.37 & 0 & 0 \end{bmatrix}$$

The W is $\begin{bmatrix} 3 & 3 \\ 4 & 4 \end{bmatrix}$.

Applying SVD on W, we get

$$\mathrm{W} = \mathrm{X\Sigma Y}^T = \begin{bmatrix} -0.6 & -0.8 \\ -0.8 & -0.6 \end{bmatrix} \begin{bmatrix} \sqrt{50} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} -0.707 & -0.707 \\ -0.707 & 0.707 \end{bmatrix}$$

$$\Sigma^+ = \begin{bmatrix} 1/\sqrt{50} & 0 \\ 0 & 0 \end{bmatrix}$$

$$\mathrm{U} = \mathrm{Y}(\Sigma^+)^2\mathrm{X}^\mathrm{T} = \begin{bmatrix} -0.707 & -0.707 \\ -0.707 & 0.707 \end{bmatrix} \begin{bmatrix} 0.02 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} -0.6 & -0.8 \\ -0.8 & -0.6 \end{bmatrix} = \begin{bmatrix} 0.0085 & 0.0085 \\ -0.0113 & 0.0113 \end{bmatrix}$$

9.2.1(a) 0.5'

$$\cos(\mathrm{A,B}) = \frac{8.2008 + 160000\alpha^2 + 24\beta^2}{\sqrt{9.3636 + 250000\alpha^2 + 36\beta^2}\sqrt{7.1824 + 102400\alpha^2 + 16\beta^2}}$$

$$\cos(\mathrm{B,C}) = \frac{7.8256 + 204800\alpha^2 + 24\beta^2}{\sqrt{7.1824 + 102400\alpha^2 + 16\beta^2}\sqrt{8.5264 + 409600\alpha^2 + 36\beta^2}}$$

$$\cos(\mathrm{A,C}) = \frac{8.9352 + 320000\alpha^2 + 36\beta^2}{\sqrt{9.3636 + 250000\alpha^2 + 36\beta^2}\sqrt{8.5264 + 409600\alpha^2 + 36\beta^2}}$$

9.2.3 0.5'

(a) avg = (4+2+5)/3=11/3

A: 4-11/3=1/3

B: 2-11/3=-5/3

C: 5-11/3=4/3

(b)Processor Speed: 3.06*1/3-2.68*5/3+2.92*4/3=0.4467

Disk Size: 500*1/3-320*5/3+640*4/3= 486.6667

Main-Memory Size: 6*1/3-4*5/3+6*4/3=3.3333

9.3.1 1'

(a) $\mathrm{Jaccard(A,B)} = \frac{4}{8} = \frac{1}{2}$

$$\text{Jaccard}(B, C) = \frac{4}{8} = \frac{1}{2}$$

$$\text{Jaccard}(A, C) = \frac{4}{8} = \frac{1}{2}$$

(b)

$$\cos(A, B) = \frac{17}{20\sqrt{2}} = 0.601$$

$$\cos(B, C) = \frac{11}{8\sqrt{10}} = 0.435$$

$$\cos(A, C) = \frac{11}{8\sqrt{5}} = 0.615$$

9.4.1 1'

(a)Start with the U and V in Fig. 9.10

$$\begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & x \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \times \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 2 & 2 & 2 & 2 \\ 2 & 2 & 2 & 2 & 2 \\ 1+x & 1+x & 1+x & 1+x & 1+x \\ 2 & 2 & 2 & 2 & 2 \\ 2 & 2 & 2 & 2 & 2 \end{bmatrix}$$

The contribution to the sum of squares from the third row is

$$(\mathrm{x}-1)^2 + (x-2)^2 + x^2 + (x-3)^2$$

We find the minimum value of this expression by differentiating and equating to 0, as:

$$2 \times \big((\mathrm{x}-1) + (\mathrm{x}-2) + \mathrm{x} + (\mathrm{x}-3)\big) = 0$$

The solution for x is x = 1.5

Thus after the first step

$$\begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1.5 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \times \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 2 & 2 & 2 & 2 \\ 2 & 2 & 2 & 2 & 2 \\ 2.5 & 2.5 & 2.5 & 2.5 & 2.5 \\ 2 & 2 & 2 & 2 & 2 \\ 2 & 2 & 2 & 2 & 2 \end{bmatrix}$$

(b)Start with the U and V in Fig. 9.10

$$\begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \times \begin{bmatrix} 1 & 1 & 1 & y & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 2 & 2 & y+1 & 2 \\ 2 & 2 & 2 & y+1 & 2 \\ 2 & 2 & 2 & y+1 & 2 \\ 2 & 2 & 2 & y+1 & 2 \\ 2 & 2 & 2 & y+1 & 2 \end{bmatrix}$$

The contribution to the sum of squares from the forth column is

$$(\mathrm{y}-3)^2 + (y-3)^2 + y^2 + (y-2)^2 + (y-3)^2$$

We find the minimum value of this expression by differentiating and equating to 0, as:

$$2 \times \big((\mathrm{y}-3) + (\mathrm{y}-3) + \mathrm{y} + (\mathrm{y}-2) + (\mathrm{y}-3)\big) = 0$$

The solution for x is y=2.2

Thus after the first step

$$\begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \times \begin{bmatrix} 1 & 1 & 1 & 2.2 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 2 & 2 & 3.2 & 2 \\ 2 & 2 & 2 & 3.2 & 2 \\ 2 & 2 & 2 & 3.2 & 2 \\ 2 & 2 & 2 & 3.2 & 2 \\ 2 & 2 & 2 & 3.2 & 2 \end{bmatrix}$$