CSE601: Data Mining

Homework2

Group: 31

Anuj Rastogi – 5013 4324

Nalin Kumar – 5017 0479

Pranshu Pancholi – 5016 9864

Implementation Details:

1) Implementation of Apriori Algorithm –

Reading Dataset -

We used Java programming language to implement the Apriori Algorithm. First task in the implementation of Apriori Algorithm was to read the dataset from gene_expression.txt file. Each row in the dataset was input one at a time. Within each row the first entry i.e. sample number was skipped and gene data (UP/Down) was input by iterating through a row 100 times. Gene data was appended with the Gene number for e.g. 1UP, 4Down etc. and stored in a HashMap "setCount" which contained the single length itemset and its count in the entire dataset. Last column in the dataset i.e. disease name was also added to the HashMap with its count in the dataset.

Generating Frequent Itemsets –

In this task we generated frequent itemsets of length 1, 2, 3 and so on, such that each item has a support >= each of the threshold support of 30%, 40%, 50%, 60%, 70% given in the homework description. First task in the generation of frequent itemsets was to generate frequent itemsets of length-1. After reading the dataset, we have a fully populated HashMap of single length items with their count. We removed the items from this HashMap whose support was less than the threshold support given. We calculated the support of the itemset by simple formula –

$$\text{Support of itemset} = \frac{\text{No. of occurrence of itemset in the samples}}{\text{Total number of samples}}$$

By removing the infrequent itemsets from the HashMap we were able to achieve pruning of supersets of the pruned itemset. After this process we had the HashMap of Frequent Itemset of length-1.

To create itemsets of length-2 or more we followed a recursive approach. We first create an itemset of length-k (k being the length of itemset starting from 2) by grouping the items from the so as to form an itemset of length-k. Then, we calculated the support of the length-k itemset and if the support is greater than the threshold. Doing this we pruned the supersets of pruned itemset. Then we added the itemset to the HashMap "highLevelItemsets" with its support count value. Once we are finished generating all possible length-k frequent itemsets, we recursively call the function to create frequent itemsets of length-k+1. We do this until we generate 0, length-k frequent itemsets.

2) Implementation of Association Rule Generation –

To generate the association rules of the frequent itemsets with support >= 50% and confidence >= 60%, we considered the frequent itemsets of length-2 and length-3 generated for support 50% in the previous step. We ignored the length-1 frequent itemsets because single length itemsets cannot be used to generate an association rule.

To generate association rules of size 2 and 3 we used the recursive technique. First we created all the non-empty subsets of the each frequent itemset excluding itself. For e.g. for length-2 frequent itemset {A, B}, subsets generated were {A}, {B}. For length-3 frequent itemset {A, B, C}, subsets generated were {A}, {B}, {C}, {A, B}, {B, C}, {C, A}. Combination of these subsets were used to form an association rule. For e.g. for length-2 frequent itemsets {A}, {B} subsets can be used to form 2 association rules A->B and B->A, similarly length-3 frequent itemsets can be used to form 6 association rules of size 3.

After creating an association rule we calculated the confidence of each rule. All the rules whose confidence is >= the threshold confidence were stored others rejected. We used the simple formula to calculate the confidence of each rule –

Confidence of a rule = <u>No. of occurrence of rule items together in the samples</u>
No. of occurrence of body items together in the samples

To store the generated association rule whose confidence >= 60%, we created an inner class named "AssocationRule". This class has 2 properties "body" and "head" to store the body and head of the generated association rule. To store a rule we created an object of the class initializing it with the head and body of the generated association rule and stored the object in an ArrayList. For the given dataset we were able to generate 138 distinct association rules. Thus, once we have generated all the association rules we have an ArrayList of rules which can be iterated to find rules which satisfy the conditions given in the template query.

Results of Part 1:

1) Support set to be **30%**

   Number of length-1 Frequent Itemsets: 196
   Number of length-2 Frequent Itemsets: 5340
   Number of length-3 Frequent Itemsets: 5287
   Number of length-4 Frequent Itemsets: 1518
   Number of length-5 Frequent Itemsets: 438
   Number of length-6 Frequent Itemsets: 88
   Number of length-7 Frequent Itemsets: 11
   Number of length-8 Frequent Itemsets: 1
   Total: 12879

2) Support set to be **40%**

   Number of length-1 Frequent Itemsets: 167
   Number of length-2 Frequent Itemsets: 753
   Number of length-3 Frequent Itemsets: 149
   Number of length-4 Frequent Itemsets: 7

Number of length-5 Frequent Itemsets: 1
Total: 1077

3) Support set to be **50%**

Number of length-1 Frequent Itemsets: 109
Number of length-2 Frequent Itemsets: 63
Number of length-3 Frequent Itemsets: 2
Total: 174

4) Support set to be **60%**
Number of length-1 Frequent Itemsets: 34
Number of length-2 Frequent Itemsets: 2
Total: 36

5) Support set to be **70%**
Number of length-1 Frequent Itemsets: 7
Total: 7

Results of Part 2:

Template 1 Results –

S. No.	Query	Result
1.	RULE HAS ANY OF G6_UP	10
2.	RULE HAS 1 OF G1_UP	14
3.	RULE HAS 1 OF (G1_UP, G10_DOWN)	28
4.	BODY HAS ANY OF G6_UP	5
5.	BODY HAS NONE OF G72_UP	124
6.	BODY HAS 1 OF (G1_UP, G10_DOWN)	15
7.	HEAD HAS ANY OF G6_UP	5
8.	HEAD HAS NONE OF (G1_UP, G6_UP)	126
9.	HEAD HAS 1 OF (G6_UP, G8_UP)	6
10.	RULE HAS 1 OF (G1_UP, G6_UP, G72_UP)	48
11.	RULE HAS ANY OF (G1_UP, G6_UP, G72_UP)	50

Template 2 Results –

S. No.	Query	Result
1.	SIZE OF RULE >= 3	12
2.	SIZE OF BODY >= 2	6
3.	SIZE OF HEAD >= 2	6

Template 3 Results –

S. No.	Query	Result
1.	BODY HAS ANY OF G1_UP AND HEAD HAS 1 OF G59_UP	1
2.	BODY HAS ANY OF G1_UP OR HEAD HAS 1 OF G6_UP	12
3.	BODY HAS 1 OF G1_UP OR HEAD HAS 2 OF G6_UP	7
4.	HEAD HAS 1 OF G1_UP AND BODY HAS 0 OF DISEASE	7
5.	HEAD HAS 1 OF DISEASE OR RULE HAS 1 OF (G72_UP, G96_DOWN)	30
6.	BODY HAS 1 of (G59_UP, G96_DOWN) AND SIZE OF RULE >=3	7

Rules satisfying the above queries:

Template 1 queries:

1. RULE HAS ANY OF G6_UP

    ```
    [6UP]->[59UP]
    [59UP]->[6UP]
    [6UP]->[32Down]
    [32Down]->[6UP]
    [6UP]->[13Down]
    [13Down]->[6UP]
    [6UP]->[38Down]
    [38Down]->[6UP]
    [28Down]->[6UP]
    [6UP]->[28Down]
    ```

2. RULE HAS 1 OF G1_UP

    ```
    [70Down]->[1UP]
    [1UP]->[70Down]
    [38Down]->[1UP]
    [1UP]->[38Down]
    [10Down]->[1UP]
    [1UP]->[10Down]
    [54UP]->[1UP]
    [1UP]->[54UP]
    [67UP]->[1UP]
    [1UP]->[67UP]
    [59UP]->[1UP]
    [1UP]->[59UP]
    [1UP]->[72UP]
    [72UP]->[1UP]
    ```

3. RULE HAS 1 OF (G1_UP, G10_DOWN)

    ```
    [28Down]->[10Down]
    [10Down]->[28Down]
    [10Down]->[88Down]
    [88Down]->[10Down]
    ```

```
[70Down]->[1UP]
[1UP]->[70Down]
[38Down]->[1UP]
[1UP]->[38Down]
[10Down]->[1UP]
[1UP]->[10Down]
[10Down]->[38Down]
[38Down]->[10Down]
[54UP]->[1UP]
[1UP]->[54UP]
[10Down]->[94UP]
[94UP]->[10Down]
[67UP]->[1UP]
[1UP]->[67UP]
[10Down]->[47UP]
[47UP]->[10Down]
[59UP]->[1UP]
[1UP]->[59UP]
[1UP]->[72UP]
[72UP]->[1UP]
[10Down]->[59UP]
[59UP]->[10Down]
[10Down]->[70Down]
[70Down]->[10Down]
```

4. BODY HAS ANY OF G6_UP

```
[6UP]->[59UP]
[6UP]->[32Down]
[6UP]->[13Down]
[6UP]->[38Down]
[6UP]->[28Down]
```

5. BODY HAS NONE OF G72_UP

```
[28Down]->[10Down]
[10Down]->[28Down]
[10Down]->[88Down]
[88Down]->[10Down]
[2Down]->[38Down]
[38Down]->[2Down]
[59UP]->[13Down]
[13Down]->[59UP]
[28Down]->[88Down]
[88Down]->[28Down]
[70Down]->[1UP]
[1UP]->[70Down]
[38Down]->[52Down]
[52Down]->[38Down]
[6UP]->[59UP]
[59UP]->[6UP]
[38Down]->[72UP]
[91UP]->[38Down]
[38Down]->[91UP]
```

```
[38Down]->[1UP]
[1UP]->[38Down]
[10Down]->[1UP]
[1UP]->[10Down]
[97Down]->[72UP]
[28Down]->[2Down]
[2Down]->[28Down]
[67UP]->[38Down]
[38Down]->[67UP]
[54UP]->[24Down]
[24Down]->[54UP]
[10Down]->[38Down]
[38Down]->[10Down]
[6UP]->[32Down]
[32Down]->[6UP]
[70Down]->[38Down]
[38Down]->[70Down]
[54UP]->[1UP]
[1UP]->[54UP]
[87UP]->[59UP]
[59UP]->[87UP]
[10Down]->[94UP]
[94UP]->[10Down]
[41Down]->[38Down]
[38Down]->[41Down]
[67UP]->[1UP]
[1UP]->[67UP]
[59UP]->[38Down]
[38Down]->[59UP]
[6UP]->[13Down]
[13Down]->[6UP]
[13Down]->[72UP]
[28Down]->[52Down]
[52Down]->[28Down]
[87UP]->[28Down]
[28Down]->[87UP]
[97Down]->[82Down]
[82Down]->[97Down]
[87UP]->[88Down]
[88Down]->[87UP]
[10Down]->[47UP]
[47UP]->[10Down]
[6UP]->[38Down]
[38Down]->[6UP]
[28Down]->[47UP]
[47UP]->[28Down]
[65Down]->[38Down]
[38Down]->[65Down]
[59UP]->[72UP]
[38Down]->[94UP]
[94UP]->[38Down]
[47UP]->[38Down]
[38Down]->[47UP]
[41Down]->[28Down]
[28Down]->[41Down]
```

```
[8UP]->[88Down]
[88Down]->[8UP]
[28Down]->[6UP]
[6UP]->[28Down]
[59UP]->[1UP]
[1UP]->[59UP]
[88Down]->[24Down]
[24Down]->[88Down]
[32Down]->[72UP]
[96Down]->[82Down]
[82Down]->[96Down]
[41Down]->[88Down]
[88Down]->[41Down]
[1UP]->[72UP]
[28Down]->[32Down]
[32Down]->[28Down]
[10Down]->[59UP]
[59UP]->[10Down]
[13Down]->[82Down]
[82Down]->[13Down]
[88Down]->[59UP]
[59UP]->[88Down]
[59UP]->[32Down]
[32Down]->[59UP]
[96Down]->[72UP]
[54UP]->[88Down]
[88Down]->[54UP]
[28Down]->[59UP]
[59UP]->[28Down]
[96Down]->[59UP]
[59UP]->[96Down]
[28Down]->[38Down]
[38Down]->[28Down]
[88Down]->[38Down]
[38Down]->[88Down]
[32Down]->[38Down]
[38Down]->[32Down]
[59UP]->[82Down]
[82Down]->[59UP]
[28Down]->[13Down]
[13Down]->[28Down]
[82Down]->[72UP]
[10Down]->[70Down]
[70Down]->[10Down]
[59UP]->[82Down, 72UP]
[59UP, 82Down]->[72UP]
[82Down]->[59UP, 72UP]
[96Down]->[59UP, 72UP]
[96Down, 59UP]->[72UP]
[59UP]->[96Down, 72UP]
```

6. BODY HAS 1 OF (G1_UP, G10_DOWN)

```
[10Down]->[28Down]
[10Down]->[88Down]
```

```
[1UP]->[70Down]
[1UP]->[38Down]
[10Down]->[1UP]
[1UP]->[10Down]
[10Down]->[38Down]
[1UP]->[54UP]
[10Down]->[94UP]
[1UP]->[67UP]
[10Down]->[47UP]
[1UP]->[59UP]
[1UP]->[72UP]
[10Down]->[59UP]
[10Down]->[70Down]
```

7. HEAD HAS ANY OF G6_UP

```
[59UP]->[6UP]
[32Down]->[6UP]
[13Down]->[6UP]
[38Down]->[6UP]
[28Down]->[6UP]
```

8. HEAD HAS NONE OF (G1_UP, G6_UP)

```
[28Down]->[10Down]
[10Down]->[28Down]
[10Down]->[88Down]
[88Down]->[10Down]
[2Down]->[38Down]
[38Down]->[2Down]
[59UP]->[13Down]
[13Down]->[59UP]
[28Down]->[88Down]
[88Down]->[28Down]
[1UP]->[70Down]
[38Down]->[52Down]
[52Down]->[38Down]
[6UP]->[59UP]
[38Down]->[72UP]
[72UP]->[38Down]
[91UP]->[38Down]
[38Down]->[91UP]
[1UP]->[38Down]
[1UP]->[10Down]
[97Down]->[72UP]
[72UP]->[97Down]
[28Down]->[2Down]
[2Down]->[28Down]
[67UP]->[38Down]
[38Down]->[67UP]
[54UP]->[24Down]
[24Down]->[54UP]
[10Down]->[38Down]
```

```
[38Down]->[10Down]
[6UP]->[32Down]
[70Down]->[38Down]
[38Down]->[70Down]
[1UP]->[54UP]
[87UP]->[59UP]
[59UP]->[87UP]
[10Down]->[94UP]
[94UP]->[10Down]
[41Down]->[38Down]
[38Down]->[41Down]
[1UP]->[67UP]
[59UP]->[38Down]
[38Down]->[59UP]
[6UP]->[13Down]
[13Down]->[72UP]
[72UP]->[13Down]
[28Down]->[52Down]
[52Down]->[28Down]
[87UP]->[28Down]
[28Down]->[87UP]
[97Down]->[82Down]
[82Down]->[97Down]
[87UP]->[88Down]
[88Down]->[87UP]
[10Down]->[47UP]
[47UP]->[10Down]
[6UP]->[38Down]
[28Down]->[47UP]
[47UP]->[28Down]
[65Down]->[38Down]
[38Down]->[65Down]
[59UP]->[72UP]
[72UP]->[59UP]
[38Down]->[94UP]
[94UP]->[38Down]
[47UP]->[38Down]
[38Down]->[47UP]
[41Down]->[28Down]
[28Down]->[41Down]
[8UP]->[88Down]
[88Down]->[8UP]
[6UP]->[28Down]
[1UP]->[59UP]
[88Down]->[24Down]
[24Down]->[88Down]
[32Down]->[72UP]
[72UP]->[32Down]
[96Down]->[82Down]
[82Down]->[96Down]
[41Down]->[88Down]
[88Down]->[41Down]
[1UP]->[72UP]
[28Down]->[32Down]
[32Down]->[28Down]
```

```
[10Down]->[59UP]
[59UP]->[10Down]
[13Down]->[82Down]
[82Down]->[13Down]
[88Down]->[59UP]
[59UP]->[88Down]
[59UP]->[32Down]
[32Down]->[59UP]
[96Down]->[72UP]
[72UP]->[96Down]
[54UP]->[88Down]
[88Down]->[54UP]
[28Down]->[59UP]
[59UP]->[28Down]
[96Down]->[59UP]
[59UP]->[96Down]
[28Down]->[38Down]
[38Down]->[28Down]
[88Down]->[38Down]
[38Down]->[88Down]
[32Down]->[38Down]
[38Down]->[32Down]
[59UP]->[82Down]
[82Down]->[59UP]
[28Down]->[13Down]
[13Down]->[28Down]
[82Down]->[72UP]
[72UP]->[82Down]
[10Down]->[70Down]
[70Down]->[10Down]
[59UP]->[82Down, 72UP]
[59UP, 82Down]->[72UP]
[82Down]->[59UP, 72UP]
[82Down, 72UP]->[59UP]
[72UP]->[59UP, 82Down]
[59UP, 72UP]->[82Down]
[96Down]->[59UP, 72UP]
[96Down, 59UP]->[72UP]
[59UP]->[96Down, 72UP]
[59UP, 72UP]->[96Down]
[72UP]->[96Down, 59UP]
[96Down, 72UP]->[59UP]
```

9. HEAD HAS 1 OF (G6_UP, G8_UP)

```
[59UP]->[6UP]
[32Down]->[6UP]
[13Down]->[6UP]
[38Down]->[6UP]
[88Down]->[8UP]
[28Down]->[6UP]
```

10. RULE HAS 1 OF (G1_UP, G6_UP, G72_UP)

```
[70Down]->[1UP]
```

```
[1UP]->[70Down]
[6UP]->[59UP]
[59UP]->[6UP]
[38Down]->[72UP]
[72UP]->[38Down]
[38Down]->[1UP]
[1UP]->[38Down]
[10Down]->[1UP]
[1UP]->[10Down]
[97Down]->[72UP]
[72UP]->[97Down]
[6UP]->[32Down]
[32Down]->[6UP]
[54UP]->[1UP]
[1UP]->[54UP]
[67UP]->[1UP]
[1UP]->[67UP]
[6UP]->[13Down]
[13Down]->[6UP]
[13Down]->[72UP]
[72UP]->[13Down]
[6UP]->[38Down]
[38Down]->[6UP]
[59UP]->[72UP]
[72UP]->[59UP]
[28Down]->[6UP]
[6UP]->[28Down]
[59UP]->[1UP]
[1UP]->[59UP]
[32Down]->[72UP]
[72UP]->[32Down]
[96Down]->[72UP]
[72UP]->[96Down]
[82Down]->[72UP]
[72UP]->[82Down]
[59UP]->[82Down, 72UP]
[59UP, 82Down]->[72UP]
[82Down]->[59UP, 72UP]
[82Down, 72UP]->[59UP]
[72UP]->[59UP, 82Down]
[59UP, 72UP]->[82Down]
[96Down]->[59UP, 72UP]
[96Down, 59UP]->[72UP]
[59UP]->[96Down, 72UP]
[59UP, 72UP]->[96Down]
[72UP]->[96Down, 59UP]
[96Down, 72UP]->[59UP]
```

11. RULE HAS ANY OF (G1_UP, G6_UP, G72_UP)

```
[70Down]->[1UP]
[1UP]->[70Down]
[6UP]->[59UP]
[59UP]->[6UP]
[38Down]->[72UP]
```

```
[72UP]->[38Down]
[38Down]->[1UP]
[1UP]->[38Down]
[10Down]->[1UP]
[1UP]->[10Down]
[97Down]->[72UP]
[72UP]->[97Down]
[6UP]->[32Down]
[32Down]->[6UP]
[54UP]->[1UP]
[1UP]->[54UP]
[67UP]->[1UP]
[1UP]->[67UP]
[6UP]->[13Down]
[13Down]->[6UP]
[13Down]->[72UP]
[72UP]->[13Down]
[6UP]->[38Down]
[38Down]->[6UP]
[59UP]->[72UP]
[72UP]->[59UP]
[28Down]->[6UP]
[6UP]->[28Down]
[59UP]->[1UP]
[1UP]->[59UP]
[32Down]->[72UP]
[72UP]->[32Down]
[1UP]->[72UP]
[72UP]->[1UP]
[96Down]->[72UP]
[72UP]->[96Down]
[82Down]->[72UP]
[72UP]->[82Down]
[59UP]->[82Down, 72UP]
[59UP, 82Down]->[72UP]
[82Down]->[59UP, 72UP]
[82Down, 72UP]->[59UP]
[72UP]->[59UP, 82Down]
[59UP, 72UP]->[82Down]
[96Down]->[59UP, 72UP]
[96Down, 59UP]->[72UP]
[59UP]->[96Down, 72UP]
[59UP, 72UP]->[96Down]
[72UP]->[96Down, 59UP]
[96Down, 72UP]->[59UP]
```

Template 2 queries:

1. SIZE OF RULE >= 3

```
[59UP]->[82Down, 72UP]
[59UP, 82Down]->[72UP]
[82Down]->[59UP, 72UP]
[82Down, 72UP]->[59UP]
[72UP]->[59UP, 82Down]
```

```
[59UP, 72UP]->[82Down]
[96Down]->[59UP, 72UP]
[96Down, 59UP]->[72UP]
[59UP]->[96Down, 72UP]
[59UP, 72UP]->[96Down]
[72UP]->[96Down, 59UP]
[96Down, 72UP]->[59UP]
```

2. SIZE OF BODY >= 2

```
[59UP, 82Down]->[72UP]
[82Down, 72UP]->[59UP]
[59UP, 72UP]->[82Down]
[96Down, 59UP]->[72UP]
[59UP, 72UP]->[96Down]
[96Down, 72UP]->[59UP]
```

3. SIZE OF HEAD >= 2

```
[59UP]->[82Down, 72UP]
[82Down]->[59UP, 72UP]
[72UP]->[59UP, 82Down]
[96Down]->[59UP, 72UP]
[59UP]->[96Down, 72UP]
[72UP]->[96Down, 59UP]
```

Template 3 queries:

1. BODY HAS ANY OF G1_UP AND HEAD HAS 1 OF G59_UP

```
[1UP]->[59UP]
```

2. BODY HAS ANY OF G1_UP OR HEAD HAS 1 OF G6_UP

```
[1UP]->[70Down]
[59UP]->[6UP]
[1UP]->[38Down]
[1UP]->[10Down]
[32Down]->[6UP]
[1UP]->[54UP]
[1UP]->[67UP]
[13Down]->[6UP]
[38Down]->[6UP]
[28Down]->[6UP]
[1UP]->[59UP]
[1UP]->[72UP]
```

3. BODY HAS 1 OF G1_UP OR HEAD HAS 2 OF G6_UP

```
[1UP]->[70Down]
[1UP]->[38Down]
[1UP]->[10Down]
[1UP]->[54UP]
[1UP]->[67UP]
[1UP]->[59UP]
[1UP]->[72UP]
```

4. HEAD HAS 1 OF G1_UP AND BODY HAS 0 OF DISEASE

```
[70Down]->[1UP]
[38Down]->[1UP]
[10Down]->[1UP]
[54UP]->[1UP]
[67UP]->[1UP]
[59UP]->[1UP]
[72UP]->[1UP]
```

5. HEAD HAS 1 OF DISEASE OR RULE HAS 1 OF (G72_UP, G96_DOWN)

```
[38Down]->[72UP]
[72UP]->[38Down]
[97Down]->[72UP]
[72UP]->[97Down]
[13Down]->[72UP]
[72UP]->[13Down]
[59UP]->[72UP]
[72UP]->[59UP]
[32Down]->[72UP]
[72UP]->[32Down]
[96Down]->[82Down]
[82Down]->[96Down]
[1UP]->[72UP]
[72UP]->[1UP]
[96Down]->[72UP]
[72UP]->[96Down]
[96Down]->[59UP]
[59UP]->[96Down]
[82Down]->[72UP]
[72UP]->[82Down]
[59UP]->[82Down, 72UP]
[59UP, 82Down]->[72UP]
[82Down]->[59UP, 72UP]
[82Down, 72UP]->[59UP]
[72UP]->[59UP, 82Down]
[59UP, 72UP]->[82Down]
[96Down]->[59UP, 72UP]
[96Down, 59UP]->[72UP]
[59UP, 72UP]->[96Down]
[72UP]->[96Down, 59UP]
```

6. BODY HAS 1 of (G59_UP, G96_DOWN) AND SIZE OF RULE >=3

```
[59UP]->[82Down, 72UP]
[59UP, 82Down]->[72UP]
[59UP, 72UP]->[82Down]
[96Down]->[59UP, 72UP]
[59UP]->[96Down, 72UP]
[59UP, 72UP]->[96Down]
[96Down, 72UP]->[59UP]
```