CSE 601 – FALL 2016
DATA MINING AND BIO INFORMATICS

HOMEWORK – 2
APRIORI ALGORITHM
IMPLEMENTATION AND RESULTS REPORT

Submitted By:
Arti Gupta
(50170010)
artigupt@buffalo.edu
Sudharchith Sonty (50169912)
sudharch@buffalo.edu

IMPLEMENTATION OF THE ALGORITHM

The Implementation of the algorithm is as follows.

PHASE 1: READING THE FILE

We read the gene_expression.txt file into Java. Using a Buffered Reader in combination with a File Reader, we created a Map with the key as String which in this case is the Sample number and the value as a String builder with the rest of the genes flowing into it.

The Map is named as gene_expression and contains the Sample ID as the key and the Specified number of genes as the value.

PHASE 2: GENERATING DATA SETS

We have written a small method to calculate the minimum support count based on the number of lines of the input file.

Once we read the Map, we iterate over the value set of the Map and split the stringbuilder object by a comma and then individually count the number of “UP” and “DOWN” values for each Gene across all the samples and store them in two separate array lists, one for up and one for down.

Then, we compare each of those values with the minimum support derived earlier and only those values greater than equal to the minimum support value are stored in a separate ArrayList called FrequentDataSet.

Then, we enumerate the FrequentDataSet values to produce all possible combinations without overlap which are stored in a list of Integer Arrays.

This is done while looping over the length value which ranges from 2 to the length of the FrequentDataSet derived in the previous step.

We use a method called generate(int r) to generate the possible combinations of sets from a given set of length 1 elements.

Once we, get this data. We store it in a Map with the Key as the length and the value as the corresponding list of Integer Arrays which hold the combinations.

RESULTS FOR THE REQUIREMENTS:

1) Support = 30%
    - Support is set to be 30%
    - Number of Length:1 Frequent Item Sets :194
2) Support = 40%
    - Support is set to be 30%
    - Number of Length:1 Frequent Item Sets :167
3) Support = 50%
    - Support is set to be 30%
    - Number of Length:1 Frequent Item Sets :109
4) Support = 60%
    - Support is set to be 30%
    - Number of Length:1 Frequent Item Sets :34
5) Support = 70%
    - Support is set to be 30%
    - Number of Length:1 Frequent Item Sets :7