NATIONAL INSTITUTE OF TECHNOLOGY,

TIRUCHIRAPPALLI

CSOE 17

BIG DATA ANALYTICS

ASSIGNMENT 2

Yagnesh L Pazhaniyappan

114117098

Task : To implement the Apriori algorithm and use it to mine category sets that are frequent in the input data

Input: A dataset(“categories.txt”) that consists of the category lists of 77,185 places in the US. Each line corresponds to the category list of one place, where the list consists of a number of category instances (e.g., hotels, restaurants, etc.) that are separated by semicolons.

Output: Two files named Pattern.txt in the folder /results, one having length 1 frequent categories and the other having all frequent category sets.

Implementation:

We create a class called Apriori which has the following methods to implement the apriori algorithm.

read_transactions_from_file() : Read transactions from the input file.

get_one_itemset() : Gets unique items from the list of transactions.

self_cross() : Takes union of a set with itself to form bigger sets.

get_min_supp_itemsets(): Returns those itemsets whose support is > minSupport

subsets(): Returns subsets of a set.

write_part1() and write_part2() : To write the results into a text file.

The class constructor takes in the value of minimum support (0.01 in our case) and assigns to the object. We then run the methods of the class in order to generate the result and write them onto separate files as per the requirements

We first check the values with a test data set basket.txt and when we are satisfied with the result, we run it on the given data set.

Result: Two files named Pattern.txt in the folder /results, one having length 1 frequent categories and the other having all frequent category sets are generated by the application.

Conclusion : Apriori algorithm is implemented and the result is obtained in form of a text file.