**COMS E6111-Advanced Database Systems, Spring 2014**

**Project 3**

Abhinav Bajaj (ab3900), Rafica Abdul Rahim (ra2688)

# List of Files for submission -

1. apriori.py - File with the implementation code of A-priori algorithm.
2. INTEGRATED-DATASET.csv – generated data set file for our project.
3. README.pdf - This file.
4. Makefile - File containing the steps to compile and run the code on CLIC machines.
5. example-run.txt- File containing output of the interesting sample run .
6. cleaningData.py – File used for extracting the INTEGRATED-DATASET.csv from raw data.

# Description of Data-Set

## Data-Set Selection

We used “Restaurant Inspection Results” data from NYC Open Data data set(s).

The data set is available [here](https://data.cityofnewyork.us/Health/Restaurant-Inspection-Results/4vkw-7nck?)

The data-set is downloaded as a zip file.

Below is the description of the files included in the data-set.

* WebExtract.txt – It’s the main file that contains the data in csv format. It has many columns for data of inspection results of the restaurants. We found the columns - BORO, ZIPCODE, CUISINE, VIOLCODE & CURRENTGRADE interesting and considered them for our data-set.
* Action.txt – This file contains the mapping of action code and descriptions in csv format. Did not use this file.
* Cuisine.txt – This file contained the mapping of Cuisine Code and description in csv format. We used this file and replaced the Cuisine Code, in WebExtract.txt file mentioned above, with the description given in this file.
* RI\_Webextract\_BigApps\_Latest.xls – This file contains the metadata about the columns in other data files. Basically we used the mapping of BORO code to the borough name.
* Violation.txt - This file contained the mapping of Violation Code and description in csv format. We used this file and replaced the Violation Code, in WebExtract.txt file mentioned above, with the description given in this file.

## INTEGRATED-DATASET Extraction

Initially we created the INTEGRATED-DATASET.csv by using MS Excel. We used the VLOOK-UP function to join the Cuisine Code, Violation Code and BORO Code fields in the WebExtract.txt, with the corresponding data in Cuisine.txt and Violation.txt files.

Then we wanted to remove the transactions with missing fields (reason explained below). We faced performance issues in MS Excel on removing such entries because the size of the large number of transactions in the file. MS Excel could not remove all the data in once. So we had to manually filter and select thousands of such rows and deleted them.

This part of cleaning step proved very inconvenient and hence we decided to write a python script to generate this final INTEGRATED\_DATASET.csv file. The python script is part of our submission.

**NOTE:** The steps we followed in MS Excel are given in APPPENDIX A.

Run the cleaningData.py file with below command –

python cleaningData.py <Path to WebExtract.txt> <Path to Cuisine.txt> <Path to Violation.txt> <INTEGRATED-DATASET output file path>'

For example –

python cleaningData.py "/home/WebExtract.txt" "/home/Cuisine.txt" "/home/Violation.txt" "/home/INTEGRATED-DATASET.csv"

### Why removed the transactions with missing fields?

The intuition here is that if there is some information missing in a transaction then we cannot either count or ignore it in support of the particular item-set. Also this will result into incorrect confidence scores. For example, consider the transaction below –

"QUEENS","11374","Chinese","Hot food item not held at or above 140º F.",<**MISSING**>

So here the last column of CURRENTGRADE is missing. So let assume that we get an interesting association rule out of this data as below –

[QUEENS, Chinese] => [B] (Conf: 66.0%, Supp: 1.5%)

If the above transaction is considered then the support for L.H.S will increase but since we don’t know the rating in this transaction, we cannot decide whether to increase the support of R.H.S or not. So, the support and confidence score will be incorrect if we consider this transaction. Hence, we decide to remove such transactions.

## “INTERESTING” about Data-Set

We considered many data-sets before finalizing this data-set. Few of them are the “School Demographics and Accountability”, “NYS Math Test Results”, “ELA Test Results” and more. The other most interesting data we could think of was combining the “Demographic Statistics By Zip Code” and “Natural Gas Consumption by ZIP Code”.

But we decided to work on this Data-set because of below of the following “INTERESTING” reasons:-

* The data relates to everyone as all of us like to eat out and New York is famous of the variety of cuisines available in the city. Probably we can find food of every ethnic group available in the city.
* The data gives you the quality rating given to restaurants. This rating is very important information and gives statistics of overall food quality of restaurants categorized by cuisine or boroughs or zip code.
* We expect to get interesting statistics like which borough has most Indian restaurants or how many American restaurants are rated ‘A’ or Common rules violated by Korean restaurants or most of the Italian restaurant in 10027 zip code are rated ‘A’.

# Running the program -

The project is in Python.

Run Command Options -

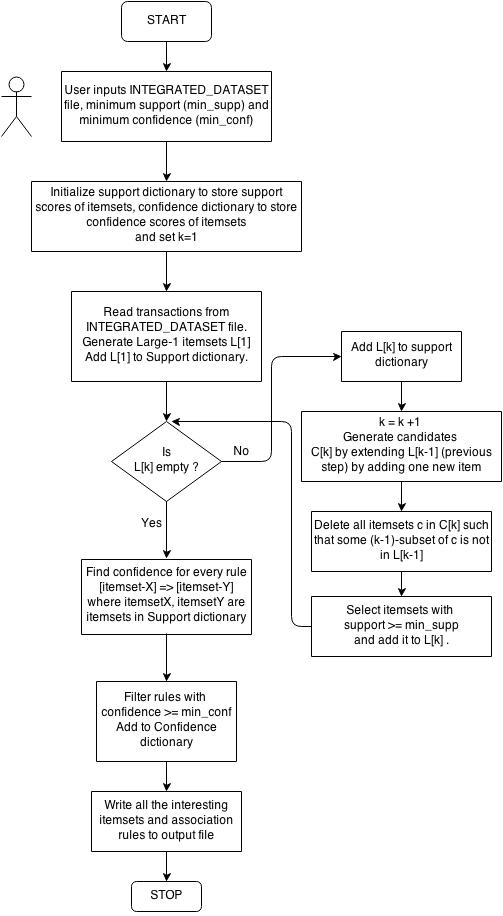
python apriori.py <path to INTEGRATED-DATASET.csv> <min\_sup> <min\_conf>

For example -

python apriori.py INTEGRATED-DATASET.csv 0.3 0.5

# Design Description

We implemented the a-priori algorithm described in Section 2.1 of the [Agrawal and Srikant paper in VLDB 1994](http://www.cs.columbia.edu/~gravano/Qual/Papers/agrawal94.pdf) to compute these frequent item-sets. The major work flow or design of the program is presented in the flow diagram below –



# Interesting Sample Run A. Command line specification

python apriori.py INTEGRATED-DATASET.csv 0.0068 0.61

## B. Explanation of “interestingness” from sample run

Below are few examples that give interesting observations from the association rule mining on this data.

1. [ CUISINE|Donuts ] => [ CURRENTGRADE|A ] (Conf: 83.6442281698 %, Supp: 1.33491532736 %)

*Observation* - 83% of the “Donuts” cuisine restaurants have ‘A’ rating. The confidence is very high here.

1. [ BORO|MANHATTAN ,CUISINE|Café/Coffee/Tea ] => [ CURRENTGRADE|A ] (Conf: 81.9288613034 %, Supp: 1.48074641355 %)

*Observation* – 82% of the “Café/Coffee/Tea” cuisine restaurants in MANHATTAN have ‘A’ rating.

1. [ CUISINE|French ] => [ BORO|MANHATTAN ] (Conf: 81.2785388128 %, Supp: 1.07518067091 %)

*Observation* – 81% of the French restaurants are in Manhattan.

1. [ VIOLATION|Food contact surface not properly maintained. ] => [ CURRENTGRADE|A ] (Conf: 75.9691912709 %, Supp: 1.27666918348 %)

*Observation* – 75% of the ‘A’ restaurants were found to have the violation “Food contact surface not properly maintained.”

1. [ BORO|STATEN ISLAND ] => [ CURRENTGRADE|A ] (Conf: 66.0031225605 %, Supp: 2.18876065149 %)

*Observation* – 66% of the restaurants in STATEN ISLAND are ‘A’ rated. Other borough are not in the results.

1. [ CUISINE|Korean ] => [ BORO|QUEENS ] (Conf: 64.9947753396 %, Supp: 0.805091144429 %)

*Observation* – 65% of the “Korean” restaurants are in Queens. This reflects that Korean population is quite high in Queens.

1. [ BORO|QUEENS ,CUISINE|Pizza ] => [ CURRENTGRADE|A ] (Conf: 62.8318584071 %, Supp: 0.765828928918 %)

*Observation* – 63% of “Pizza” restaurants in Queens are rated ‘A’.

1. [ BORO|MANHATTAN ,ZIPCODE|10019 ] => [ CURRENTGRADE|A ] (Conf: 61.7855378027 %, Supp: 1.61837989429 %)

*Observation* – 62% of restaurants in zip code 10019 in Manhattan are rated ‘A’.

# APPENDIX A

Steps to create the INTEGRATED\_DATASET.csv file in MS Excel are below –

Steps to create the combined file from Cuisine.txt, RI\_Webextract\_BigApps\_Latest.xls, Violation.txt and WebExtract.txt files.

1. Change the file extension of WebExtract.txt from txt to csv. Open the file in MS Excel.
2. Add a new sheet to this excel. Open the cuisine.txt in an editor and copy the data from this file to the new sheet (Let’s call it CUISINE sheet) created.
3. Go to first sheet and create a new column next to CUISINE. Select this column and add a VLOOKUP function which will take CUISINE column as look-up value, CUISINE sheet (CUISINECODE and CODEDESC) as table array, CODEDESC as Col\_index\_num and fourth parameter “False”.
4. After step 3, Cuisine description values will come in the new column created.
5. Copy and paste (values) this column to the CUISINE column.
6. Delete the column we created in step 3 and the sheet created in Step 2.
7. Repeat the steps 2 to 6 for Violation description.
8. For getting the borough information copy the mapping from RI\_Webextract\_BigApps\_Latest.xls to a new sheet with column for BOROCODE and NAME. Then follow the steps from 3 to 6.

For details on VLOOKUP function follow this link - <http://office.microsoft.com/en-us/excel-help/vlookup-HP005209335.aspx>.

Steps to delete the transactions with missing fields from the combined file are below.

1. Select the heading line (line no 1). Select filter from drop down option of "Sort & Filter" button on Home tab. The column heading will turn into drop-downs.
2. Select the drop down of Column "BORO". Uncheck "Select All" and then check "#N/A". Now only the data having "#N/A" in "BORO" column will be displayed.
3. Select all the rows and then right-click.
4. Select "Delete row" option. This will delete all the filtered rows.
5. Click on the drop down of Column "BORO" again. Check "Select All" and all the data will be displayed again.
6. Select the drop down of Column "ZIPCODE". Uncheck "Select All" and then check "(Blanks)". Now only the data having no information in "ZIPCODE" column will be displayed.
7. Select all the rows and then right-click.
8. Select "Delete row" option. This will delete all the filtered rows.
9. Click on the drop down of Column "ZIPCODE" again. Check "Select All" and all the data will be displayed again.
10. Repeat the steps 6-9 for the remaining columns.
11. After these steps we will total of 231775 rows (excluding the heading) in the file.

NOTE – We faced the issue in step 10 for column “CURRENTGRADE”. The number of such transactions was huge and MS Excel will not respond during this step. So, we had to delete all those rows in batch of size manageable by MS Excel.