**Data Mining Assignment**

**Rajat Saxena**

**2011C6PS649H**

**Arpit Sharma**

**2011A7PS078H**

**1. List of files**

* apriori.py : Association rule mining python file
* CleanData.py : Cleans the data (from file WebExtract.csv into test.csv)
* output.txt : Output of an interesting run
* Dataset.csv : Integrated Data set file
* Cuisine.txt : File having Cuisine code with associated cuisine name
* WebExtract.txt : File containing the raw data picked from the site
* Report.txt : Complete description of project

**2. Objective**

The food industry is by one of the largest and most vital industries in the world. Everyone needs food for survival, and most of us thoroughly enjoy to eat. Thus, it makes sense that the industry would take advantage of the data out there. There are a lot of websites or app like OpenTable, UrbanSpoon, LocalEats etc which suggests user where to have food on the basis of different parameters like sanitary grade, location and cuisine types.

By using data mining for these sites, we can easiy suggest apps clients where to look for best restaurants serving cuisine types of their choice in USA with best sanitary grades.

As a result of the rules found we can help application, websites rate restaurants and help customers find restaurants of their choice. This will reduce the effort, time and experience of users to find restaurants. Also the restaurants can benefit from this by deciding on which location to go or what cuisine types to serve thus leading to increase in their business.

**3. Dataset Description**

* We have chosen the new york city restaurant insepction dataset. It contains data listing the restaurants in new york city, their cuisine, the borough in which they are in and their sanitary grade.
* The dataset is stores in comma separated format.
* Link for the dataset is:

*http://www.nyc.gov/html/doh/downloads/zip/bigapps/dohmh\_restaurant-inspe ctions \_002.zip.*

**3.1 Reason for this Dataset**

* We had an intuition that certain types of restaurants are concentrated in certain boroughs in new york city.
* We wanted to see if connections existed between certain types of cuisines and sanitary grades.
* We also wanted to see the borough with maximum number of A grade restaurants.
* As there are lots of sites coming up which helps customer find restaurants of there interests so this dataset can be used to tell them good restaurants on the basis of cuisine types and sanitary grades.

**3.2 Preprocessing/ Data Cleaning**

* In the above zip file, the data is contained in WebExtract.txt. We have converted the file into an csv file as it makes easy to mine data.
* The data set contained the following columns: CAMIS, DBA, BORO, BUILDING, STREET, ZIPCODE, PHONE, CUISINECODE, INSPDATE, ACTION, VIOLCODE, SCORE, CURRENTGRADE, GRADEDATE, RECORDDATE
* We have retained only the columns BORO, CUISINECODE and CURRENTGRADE as we wanted to find interesting relations between these 3 fields. We wanted to check whether the sanitary grades are affected by the types of cuisine. We also wanted to find out the relation between borough and the types of cuisines plus the sanitary grade for those cuisines. For this we need only 3 fields which are described previously only.
* We have removed the records which are empty for any of the 3 fields as these fields won't contribute much so for better results we removed them.
* The column BORO contains borough code, which is decoded in the file *RI\_Webextract\_BigApps\_Latest.xls.* We have replaced the borough code by the name of the associated borough.
* The coloumn CUISINECODE contains cuisine code which is decoded in Cuisine.txt. We have replaced the cuisine code by the name of the corresponding cuisine. The above two tasks are done through the file CleanData.py. The file uses Cuisine.txt to read the cuisine code. This helps in better mainpulation of rules generated from the algorithm later.

**4. Rule Mining Process**

* We implemented Apriori algorithm exactly as given in the paper "Agrawal, R., & Srikant, R. Fast algorithms for mining association rules. In Proc. 20th Int. Conf. Very Large Data Bases, VLDB (Vol. 1215, pp. 487-499)"
* First we calculated candidate itemsets of size 1 and then depending upon if they satisified minimum support criteria, we added them in the set of large itemsets (of size 1).
* The functions getSupport() and getConfidence() calculate the Support and Confidence of itemsets/rules respectively according to the standard formulae. Support of an itemset is the fraction of times it appears in the dataset. Confidence of an association rule LHS=>RHS is equal to Support(LHS union RHS)/Support(LHS).
* Large itemsets (L[k]) is calculated from smaller itemsets L[k-1] as given in Section 2.1.1 of the Research Paper. A join operation is done , where L[k-1] is joined with itself such that in the resulting L[k],k-2 elements should be common from the two L[k-1] itemsets.
* Once this is done, pruning is the next important step where we remove all those large itemsets (in L[k])newly formed, whose any subset is not present in L[k-1]. This removes noise.
* Now we have our large itemset. To form the association rules from the large itemset, we take all these sets and generate permutations from them (such that atleast one item is present in the RHS of any rule). We keep only those permutations in the rules whose confidence exceeds the given threshold value.
* Thus we have our association rules which we output to **output.txt.**

**5. Interesting Run**

The following run with Support: 0.005 and Confidence: 0.55 will give interesting results:

*python apriori.py Dataset.csv 0.005 0.55*

**6. Resulting Rules**

Complete listing of the rules can be found in the Output.txt files which contains 30 high-confidence association rules along with the frequent itemsets having support greater than the minimum suppport. The Association rules found out after executing helps us in finding out the best cuisine available in different parts of city. The rules also helps us in rating the restaurant on the basis of sanitary grade.

The following 6 rules were selected by us from the total 30 rules as they seemed more appealing than others:

1. [**Ice Cream Gelato Yogurt Ices] => [A] (Conf: 80.5%, Supp: 0.6%)**

“Ice Cream Gelato Yogurt Ices” as a cuisine type has the highest percentage of A grade restaurants.

2. **[French] => [MANHATTAN] (Conf: 79.8%, Supp: 0.9%)**

Most french restaurants in New York City are in Manhattan.

3. **[Korean] => [QUEENS] (Conf: 62.7%, Supp: 0.7%)**

Most Korean restaurants in New York City are in Queens. The high confidence of this rule is quite surprising.

4. **[Japanese] => [MANHATTAN] (Conf: 60.3%, Supp: 1.9%)**

Most Japanese restaurants in New York City are in Manhattan.

5. **[STATENISLAND] => [A] (Conf: 62.5%, Supp: 2.2%)**

We thought Manhattan would have highest percentage of A grade restaurants, but it turned out to be Staten Island.

6. **[Jewish/Kosher] => [BROOKLYN] (Conf: 55.1%, Supp: 0.7%)**

Most Jewish restaurants in New York City are in Brooklyn. This was definitely something new learnt!

There is one more deduction which we found out from above code that if you want to try out different tyoes of cuisines at one time, you should go to Manhattan as it has maximum number of different cuisine serving restaurants.

So from the total 30 rules found out by executing the code, we will show the above found out rules to the customer as these show the cuisine tyes with highest percentage of A grade restaurants and the location of restaurants serving different cuisine types. Also the customer can easily find out the restaurants in America with best sanitary grades.

**7. Recommendations**

My recommendations to the application or website company will be following:

* Suggest users to go to Manhattan for dining.
* They can rate different restaurants on the basis of cuisine types they are serving.
* They can also tell them what place to go to on the basis of sanitary grades.
* There are certain boroughs in USA where the users can go to enjoy a variety of cuisine and top sanitary grades like Manhatten, Staten Island.

My recommendation to the users will be following:

* If the user wants to dine in the restaurants with best sanitary grade, he should go to Staten Island or go to restaurant which serves Ice Cream Gelato Yogurt Ices.
* If the user wants to try out different cuisine types at a single time i will ask them to go to Manhattan.
* If the user wants to have french or japanese he should go to Manhattan.
* Also if someone wishes to eat Jewish or Kosher we will recommend them to go to Brooklyn.

**8. Recreate Results**

* **Dependencies**: You will need Python 2.7 + version installed on your system to run the submitted code.
* Unzip the folder DataMiningAssign.tar.gz
* Open a new terminal and go to the unzipped folder directory.
* Write down the following command in the command terminal:

*python apriori.py Dataset.csv <support> <confidence>*