DATA MINING STUDY OF RESTAURANTS IN US

BY,
KRITHIKA RAGHAVAN
AMOL KUMTAKAR
GUNJAN BATRA

CONTENTS

INTRODUCTION

The objective of this project is to perform a study on the Restaurant Industry in US.

Using Tableau, we have analyzed the restaurant data set of 5 cities of US based on the parameters of Cost of restaurant, Cuisines, User review rating, Locality, Health Inspection Grade and Text in the Reviews. Using WEKA, we use different classification algorithms to build a model for Text Data of User reviews to be classified as Positive/ Negative

Further the project, discusses text mining results from different Data Mining algorithms. We have started the project by collecting data from Zomato (popular site for checking reviews about each restaurant) on hotels available in New York.

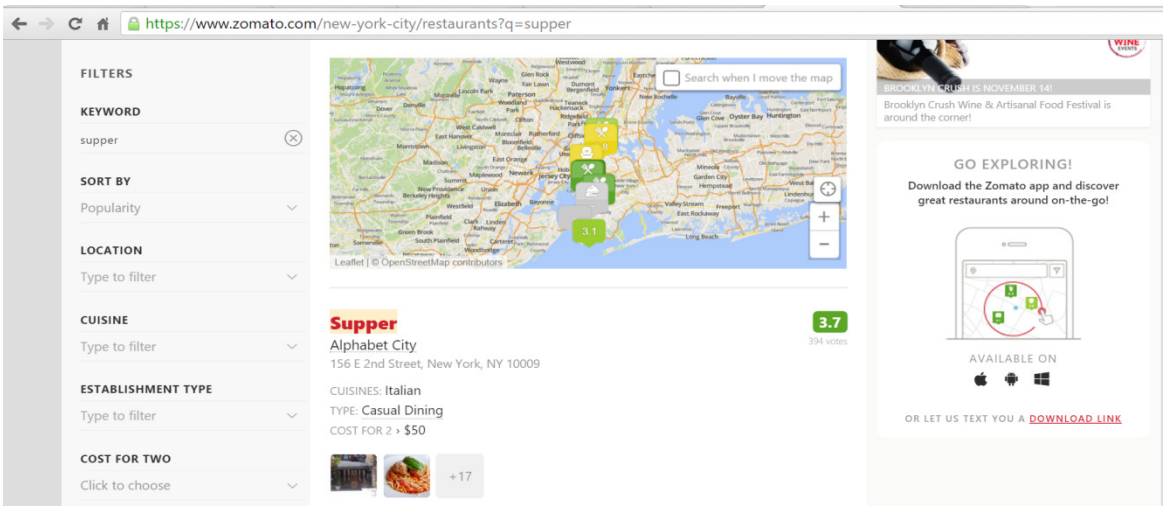
The following information were collected for 5 cities of US – New York, San Francisco, Chicago, Detroit, Houston

Restaurant_ID, Restaurant_Name, Restaurant_URL, Restaurant_Address, Restaurant_Locality, Restaurant_City, Restaurant_Zipcode, Restaurant_Cuisines, Restaurant_Avg_cost_for_two, Restaurant_UserRating, Restaurant_UserRatingText, Restaurant_UserRatingVotes

Further for New York City, Health Inspection Rating and the restaurant reviews from Zomato were collected manually from http://www.nyc.gov/html/doh/html/environmental/food-service-inspection.shtml and www.zomato.com respectively.

Each hotel has both positive as well as negative reviews. In this project we apply multiple algorithms on this data set, analyze the resulting efficiency and the ROC graphs (Weka) and conclude which has been the best classification Algorithm for our data set.

Below is a snapshot of Zomato website for a specific hotel; **Supper**



Each hotel has three tabs namely reviews, blogs and popular. Zomato rates each hotel based on the data in all the three tables. Below snapshot shows how for hotel **Supper** we have three tabs:

The data is taken only from the reviews tab and not from any other tab.

TOOLS USED -

We have used Zomato API, Ms-Access, Weka and Tableau to implement various data mining algorithms and techniques and also for data visualization.

Zomato API ➔This a tool for developers to obtain real time data from Zomato by writing web scripts

MS-Access ➔ It is a database used to store a huge data set. It makes the job easy to create forms or load data when the data set are huge. MS-Access is SQL database and we have **FORMS** option to generate a form to accept the inputs and store them in a tabular format.

Weka ➔ Weka is a data mining tool which has in built machine learning algorithms. This tool is extremely useful when we have to work on mining of data with various algorithms. This tool lets us compare the output of various algorithms on each data set and lets us decide which algorithm worked best with the given data set.

Tableau ➔ Tableau is a data visualization tool. Even though we can do visualization with Weka, Tableau is the right tool that can show each point with perfect plotting and gives clarity in the data visualization. It produces a family of interactive data visualization products focused on business intelligence.

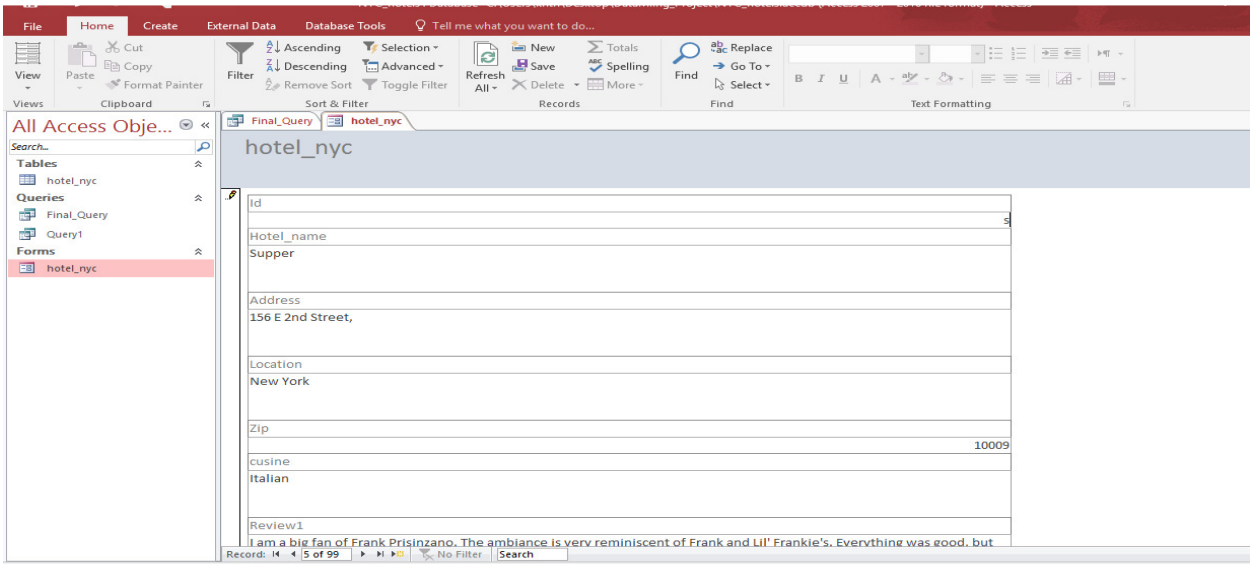
DATA COLLECTION

We collected the details of 495 restaurants in 5 cities of US from the Zomato API by writing scripts in the Zomato Developer API.

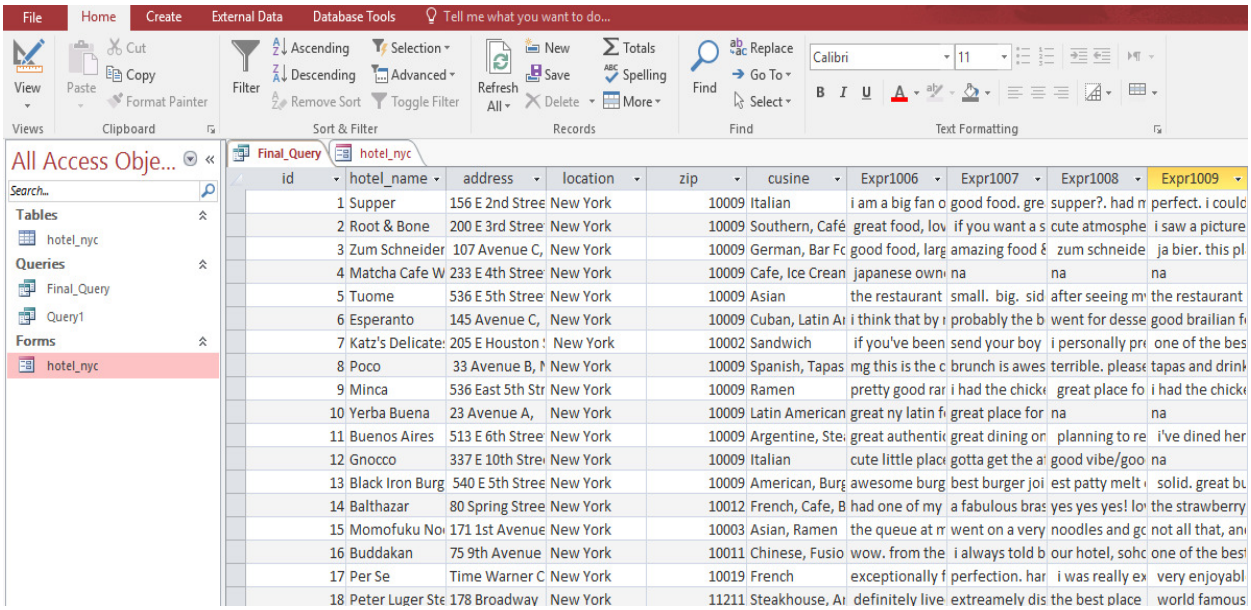
Restaurant_ID	Restaurant_Name	Restaurant_URL	Restaurant_Address	Restaurant_Locality	Restaurant_City	Restaurant_Zipcode	Restaurant_Cuisines	Restaurant_Avg_cost_for_two	Restaurant_UserRating	Restaurant_UserRatin	Restaurant_UserRatingVotes	Health Inspection
16781904	Momofuku	https://ww	171 1st Aver	East Village	New York Ci	10003	Asian, Rame	60	4.1	Excellent	1530	A
16767139	Gramercy Ta	https://ww	42 E 20th Str	Union Squa	New York Ci	10003	American	160	3.7	Very Good	1754	A
16760100	Balthazar	https://ww	80 Spring St	Soho	New York Ci	10012	French, Cafe	140	3.7	Very Good	4005	A
16775039	Peter Luger	https://ww	178 Broadw	Williamsbu	New York Ci	11211	Steakhouse,	150	3.8	Very Good	2091	A
16783153	Shake Shack	https://ww	366 Columb	Upper West	New York Ci	10024	American, B	30	4.2	Excellent	1524	A
16783998	The Halal Gu	https://ww	6th Avenue	Theater Dis	New York Ci	10019	Middle Easte	25	4.4	Excellent	493	A
16761344	Buddakan	https://ww	75 9th Aven	Meatpackin	New York Ci	10011	Chinese, Fus	150	3.9	Very Good	1483	A
16761402	Burger Joint	https://ww	Le Parker M	Theater Dis	New York Ci	10019	American, B	25	4.1	Excellent	1381	A
16785398	Shake Shack	https://ww	691 8th Ave	Hell's Kitch	New York Ci	10036	American, B	30	4.4	Excellent	856	A

Further, for New York City collected the Health Inspection Rating for restaurants, from http://www.nyc.gov/html/doh/html/environmental/food-service-inspection.shtml

Also, we have collected 1545 reviews (Weka – instances) across 90 hotels in New York City area. We created a Form in MS-Access where we loaded all the reviews to build the data set. Below is the snapshot of the form and data set.



DATA MINING PROJECT



File Home Create External Data Database Tools Tell me what you want to do...

Views: View | Clipboard: Paste, Cut, Copy, Format Painter | Sort & Filter: Filter, Ascending, Descending, Remove Sort, Selection, Advanced, Toggle Filter | Records: Refresh All, New, Save, Delete, Totals, Spelling, More | Find: Find, Replace, Go To, Select | Text Formatting: Calibri, 11

All Access Obje...

Search...

Tables
- hotel_nyc

Queries
- Final_Query
- Query1

Forms
- hotel_nyc

Final_Query | hotel_nyc

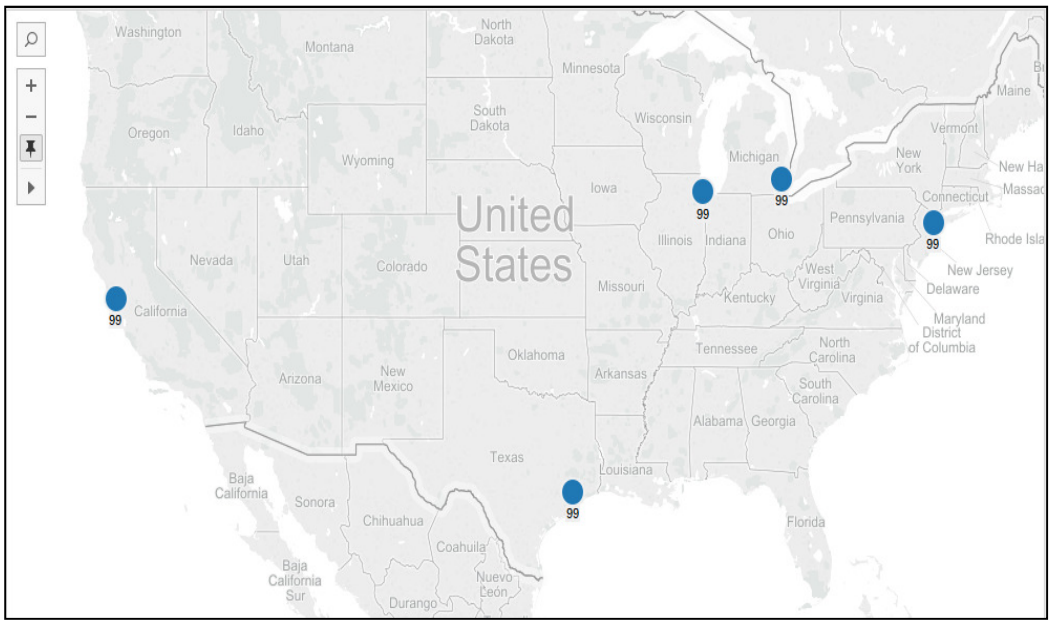
id	hotel_name	address	location	zip	cusine	Expr1006	Expr1007	Expr1008	Expr1009
1	Supper	156 E 2nd Stree	New York	10009	Italian	i am a big fan o	good food. gre	supper?. had n	perfect. i could
2	Root & Bone	200 E 3rd Stree	New York	10009	Southern, Café	great food, lov	if you want a s	cute atmosphe	i saw a picture
3	Zum Schneider	107 Avenue C,	New York	10009	German, Bar Fc	good food, larg	amazing food &	zum schneide	ja bier. this pl
4	Matcha Cafe W	233 E 4th Stree	New York	10009	Cafe, Ice Crean	japanese own	na	na	na
5	Tuome	536 E 5th Stree	New York	10009	Asian	the restaurant	small. big. sid	after seeing m	the restaurant
6	Esperanto	145 Avenue C,	New York	10009	Cuban, Latin Ar	i think that by	probably the b	went for desse	good brailian f
7	Katz's Delicate	205 E Houston	New York	10002	Sandwich	if you've been	send your boy	i personally pr	one of the bes
8	Poco	33 Avenue B, N	New York	10009	Spanish, Tapas	mg this is the c	brunch is awes	terrible. please	tapas and drink
9	Minca	536 East 5th Str	New York	10009	Ramen	pretty good rar	i had the chicke	great place fo	i had the chicke
10	Yerba Buena	23 Avenue A,	New York	10009	Latin American	great ny latin f	great place for	na	na
11	Buenos Aires	513 E 6th Stree	New York	10009	Argentine, Ste	great authentic	great dining or	planning to re	i've dined her
12	Gnocco	337 E 10th Stre	New York	10009	Italian	cute little place	gotta get the a	good vibe/goo	na
13	Black Iron Burg	540 E 5th Stree	New York	10009	American, Burg	awesome burg	best burger joi	est patty melt	solid. great bu
14	Balthazar	80 Spring Stree	New York	10012	French, Cafe, B	had one of my	a fabulous bras	yes yes yes! lo	the strawberry
15	Momofuku No	171 1st Avenue	New York	10003	Asian, Ramen	the queue at m	went on a very	noodles and gc	not all that, an
16	Buddakan	75 9th Avenue	New York	10011	Chinese, Fusio	wow. from the	i always told b	our hotel, soho	one of the best
17	Per Se	Time Warner C	New York	10019	French	exceptionally f	perfection. har	i was really ex	very enjoyabl
18	Peter Luger Ste	178 Broadway	New York	11211	Steakhouse, Ar	definitely live	extreamely dis	the best place	world famous

TABLEAU

Using Tableau, we analyze the restaurant data set of 5 cities of US based on following parameters

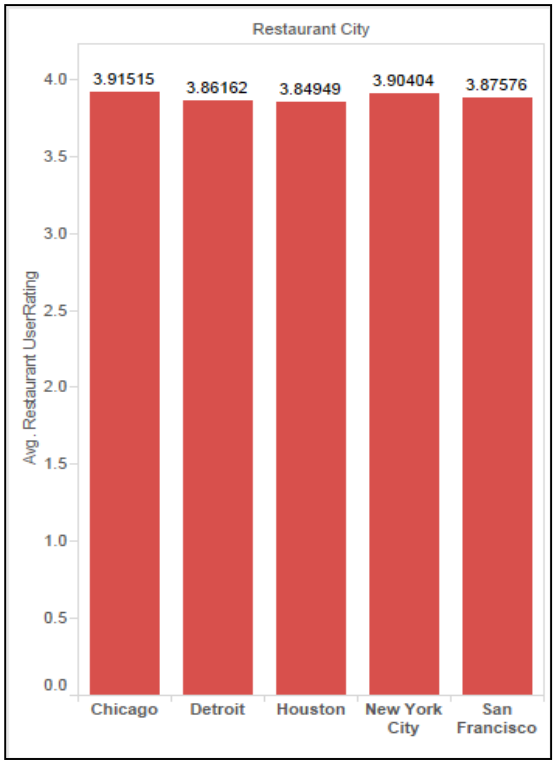
- Cost of restaurant
- Cuisines
- User review rating
- Locality
- Health Inspection Grade
- Text in the Reviews

Description of Data

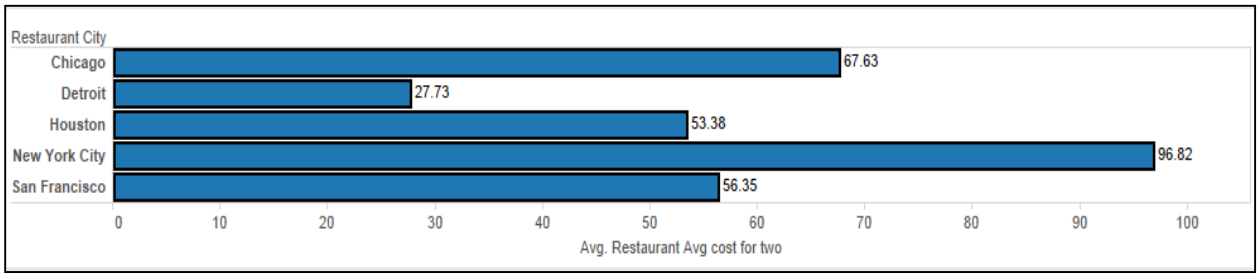


We took 99 Restaurant records for each city – New York, Houston, Detroit, San Francisco, and Chicago. 20 data records at each time were obtained for each city at a time in the form of json files which was converted to csv we combined the data set to obtain a collective data of 496 restaurants for exploration in Tableau and Weka. After exploration in Weka and Tableau the attributes were reduces from 31 to 13

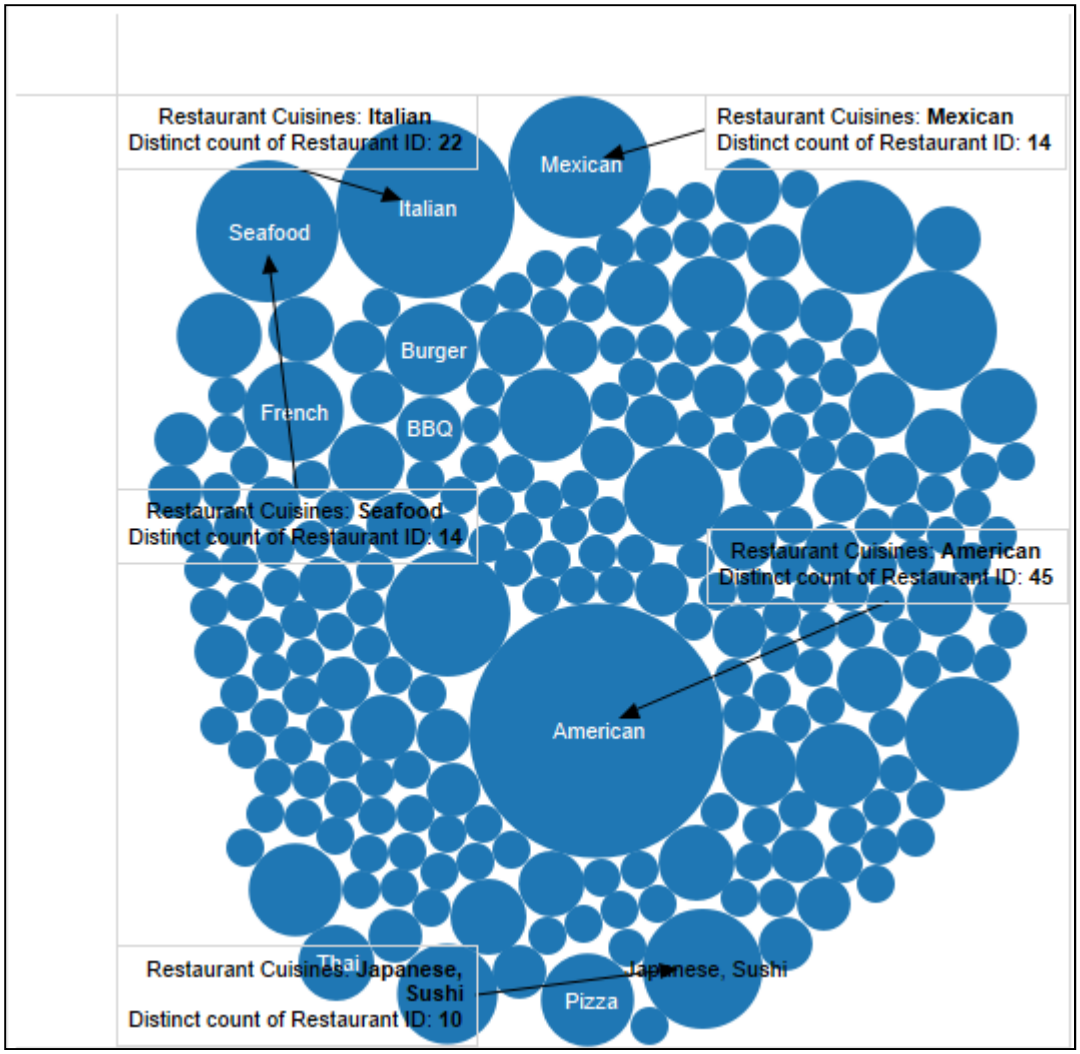
Data Analysis



The graph between average user rating and the 5 cities shows that Chicago restaurants have the best user ratings



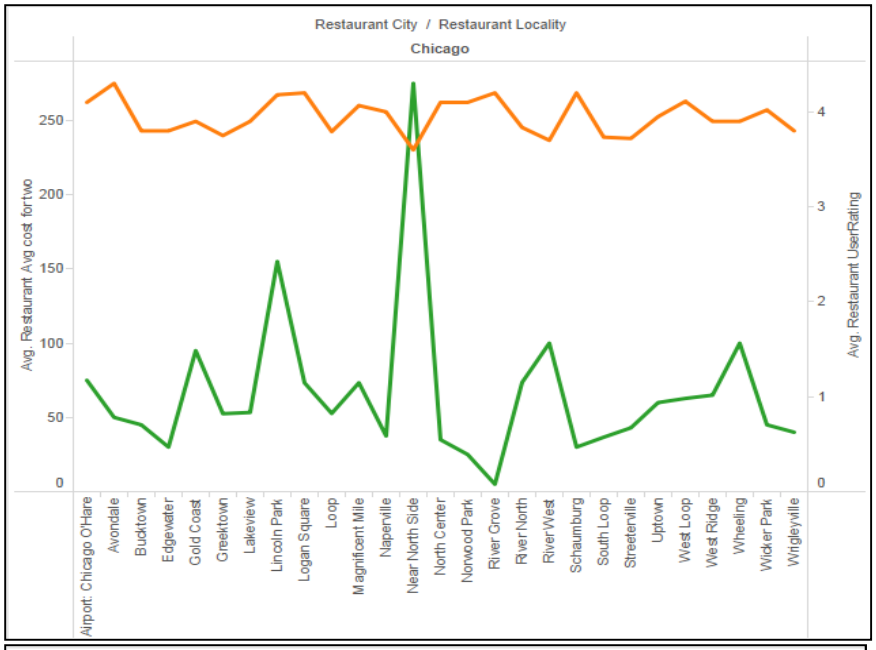
This graph between the Restaurant cities and Average of Average cost for two people shows that New York City has most expensive restaurants

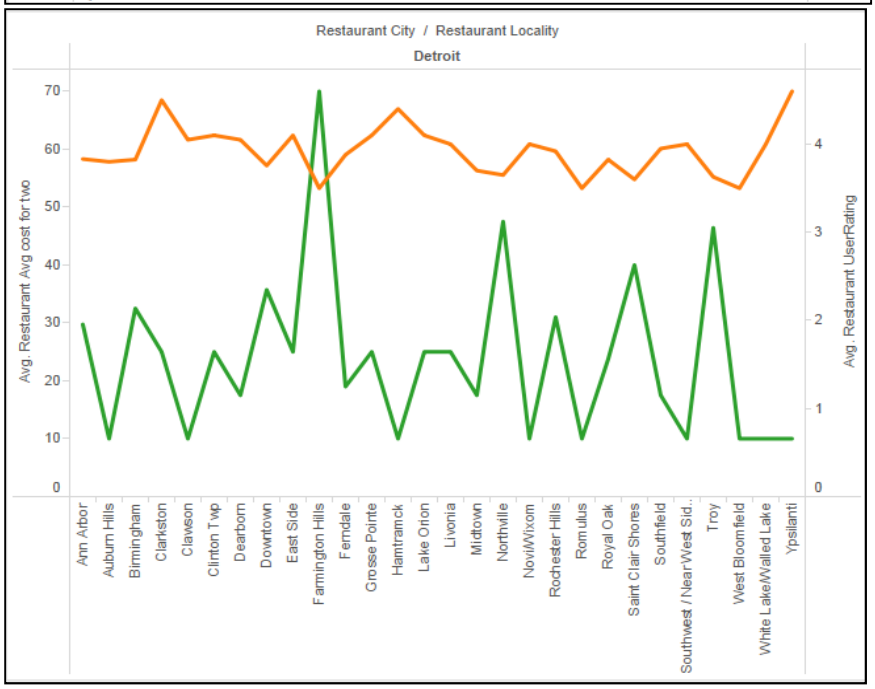


We took the count of restaurant for every cuisine to get the most common cuisine. The most popular cuisine in our data set is American which is in 45 restaurants, followed by Italian, Mexican, Seafood and Japanese

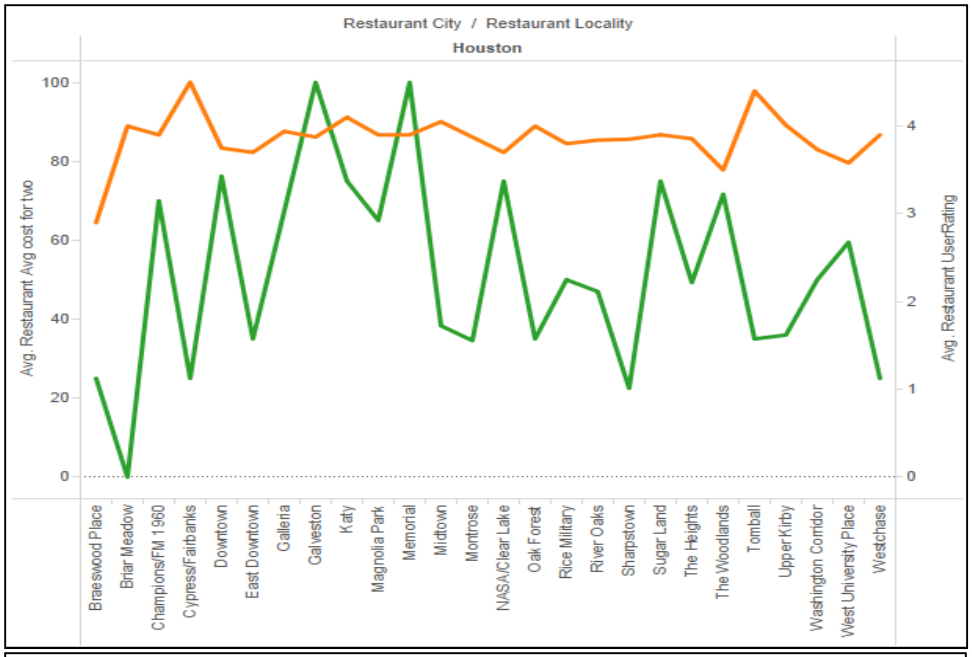
We did a study **of relationship between Average Cost for two and Average User Rating of a Restaurant** for the five different cities

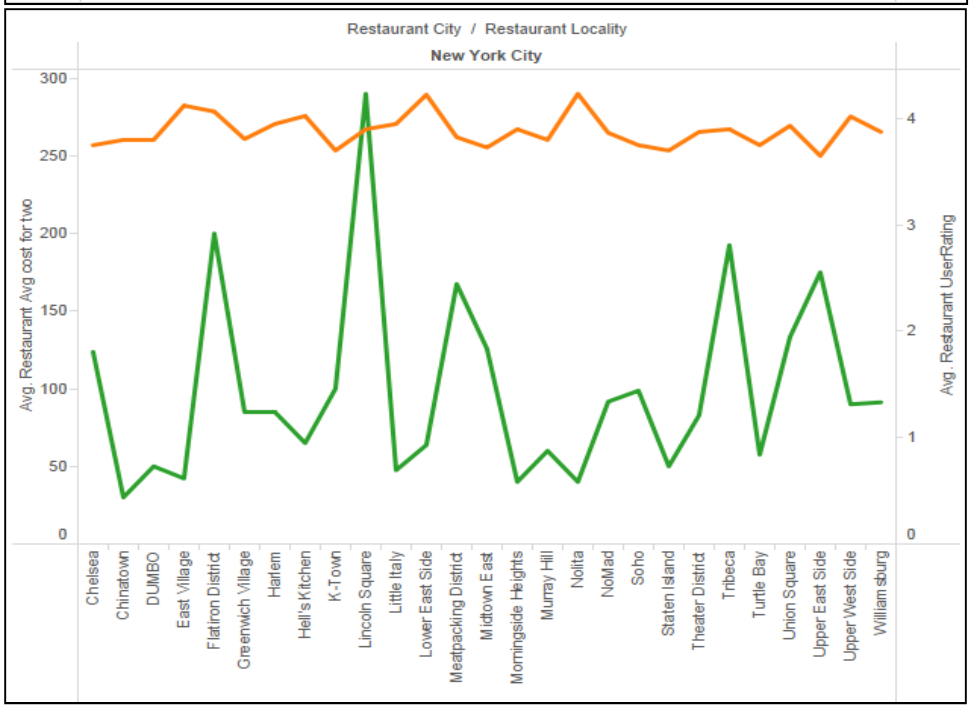
DATA MINING PROJECT



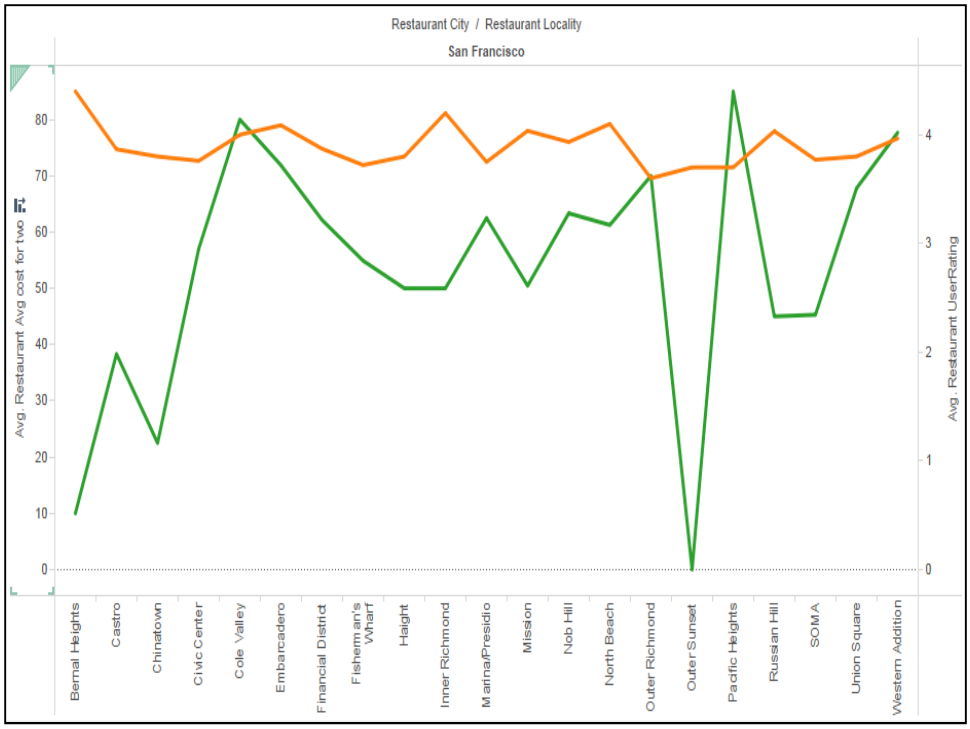


DATA MINING PROJECT





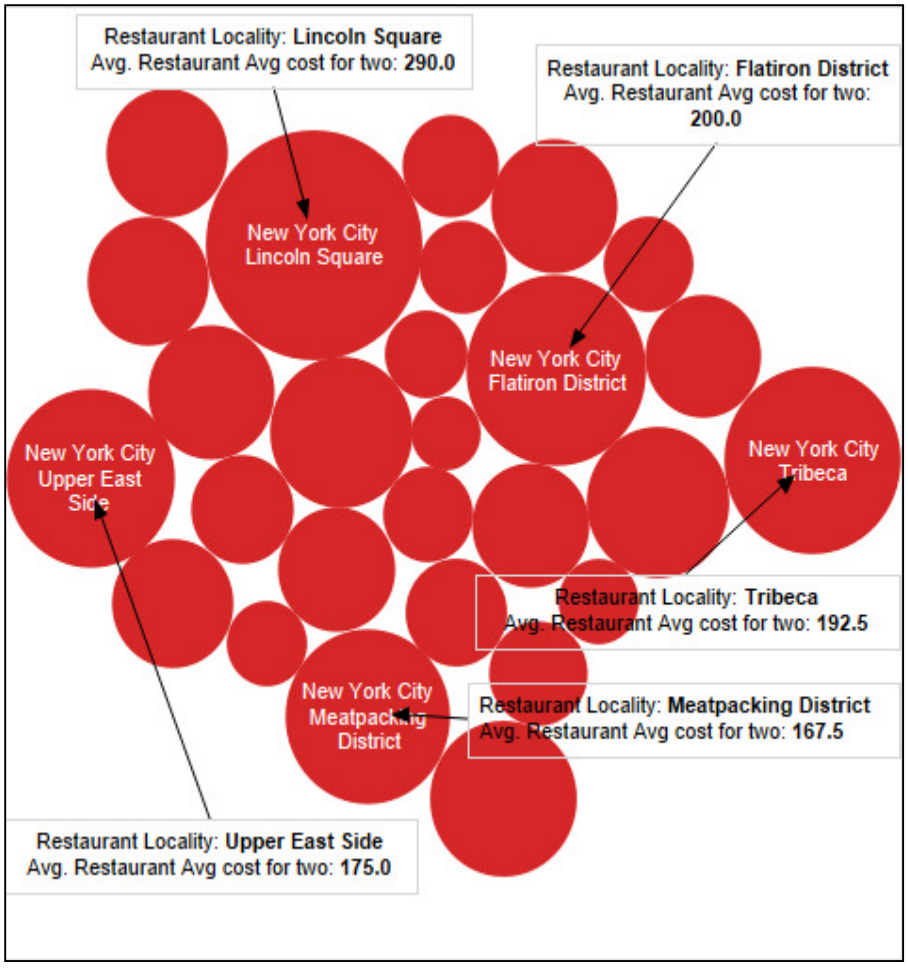
DATA MINING PROJECT



We couldn't see much correlation between the two variables. We also calculated Pearson Correlation Coefficient between the entire data set of Average Cost for two and User review rating which came out as -0.09466647716057779 which mean that they are probably not correlated.

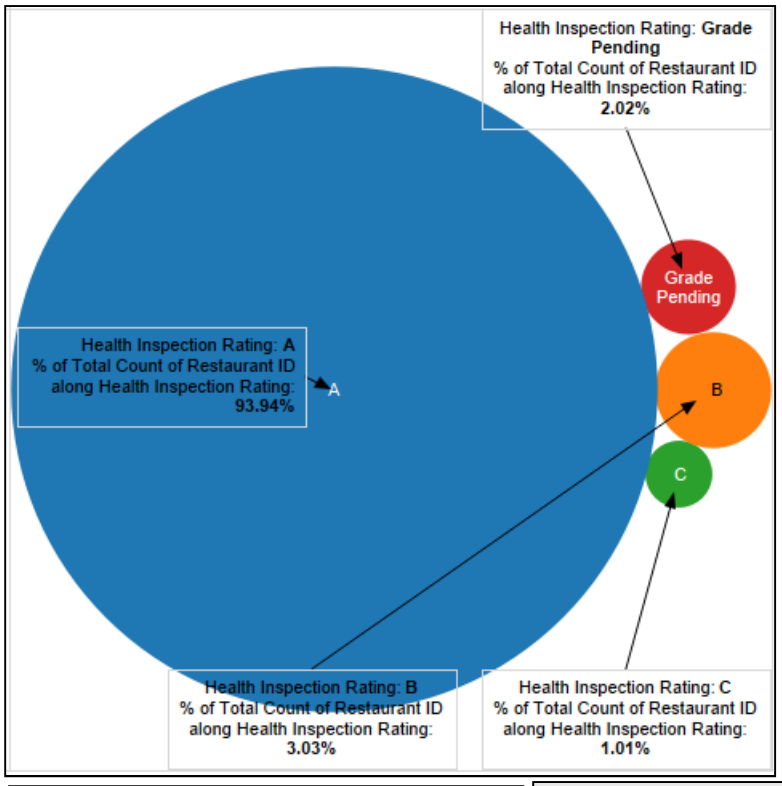
New York City

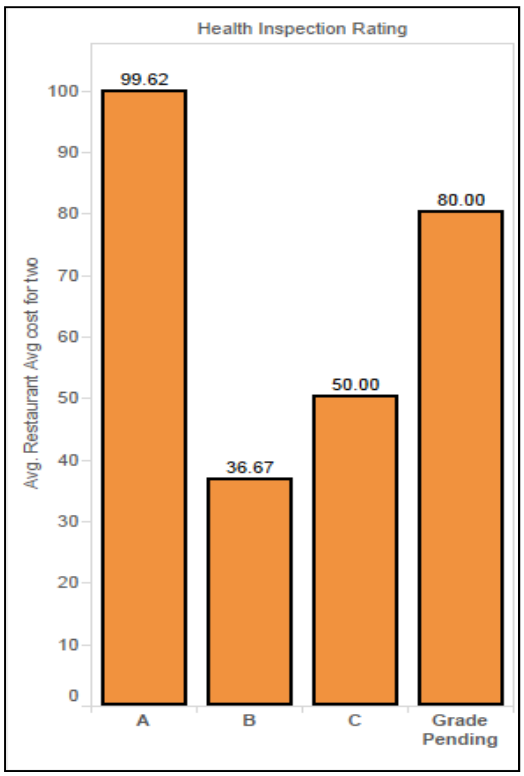
Locality wise – Average cost of two people

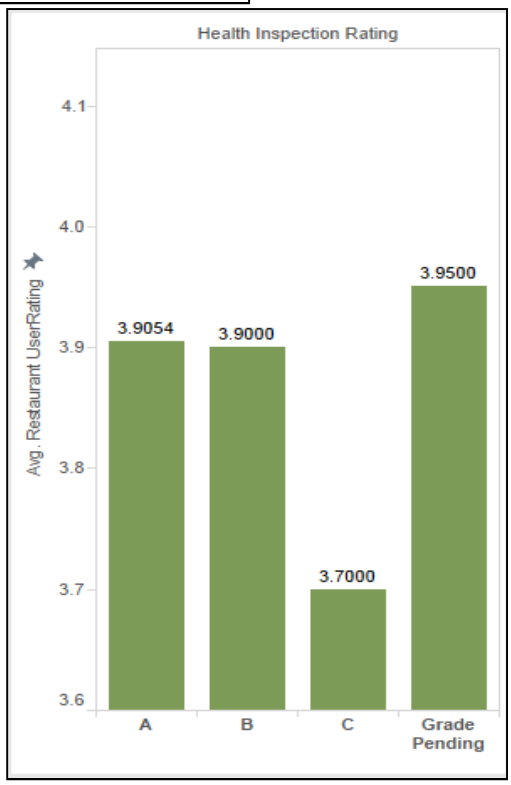


We found the average cost per locality in New York City and found that the most expensive restaurants in New York City are in Lincoln Square, followed by Flatiron District, Upper East Side, Tribeca and Meat Packing District

Results of Health Inspection Grades





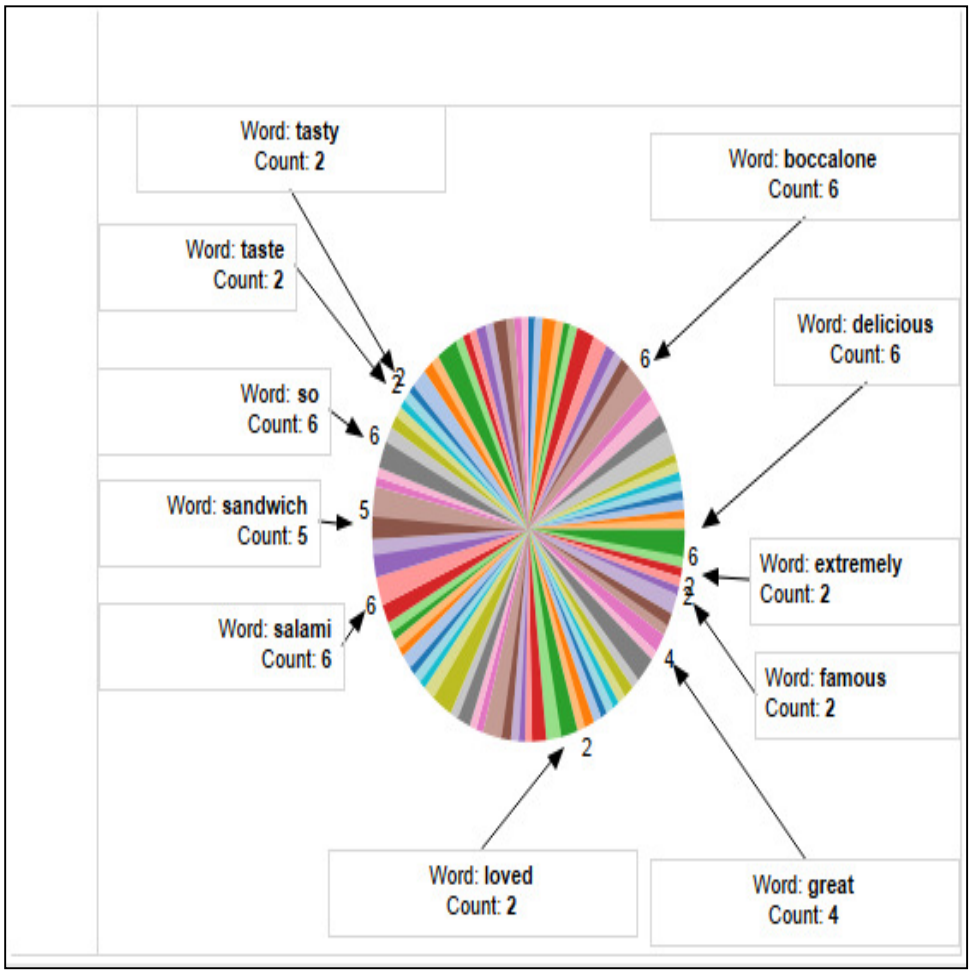


DATA MINING PROJECT

We took the percentage of Restaurants in each grade A, B, C for New York City and also found the average User rating of restaurants in each grade. We also found the Average Health Inspection rating of restaurants in each grade

The average user review rating for A grade restaurants is much higher than C grade restaurants

The average cost for two in A grade restaurants is much higher than in B and C grade restaurants



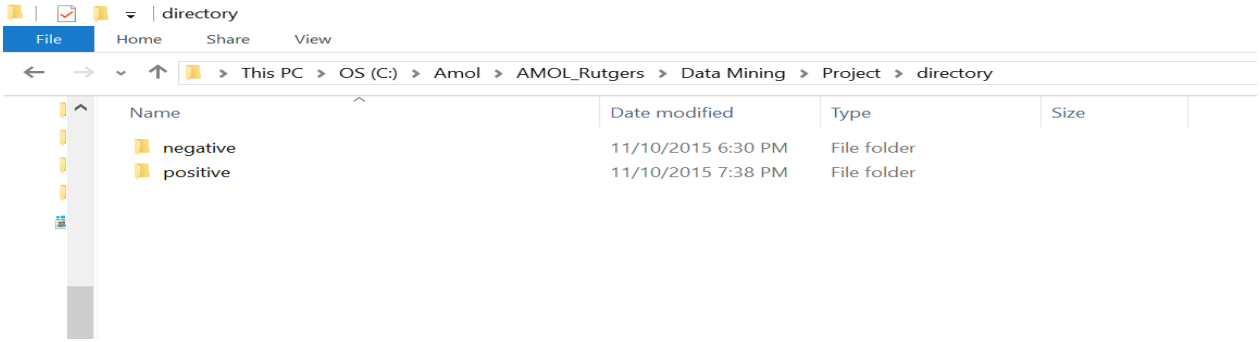
We did a study on Buccalone the restaurant with highest rating and found that the words adding to positive review of Buccalone: delicious, tasty, taste, famous, great, loved…

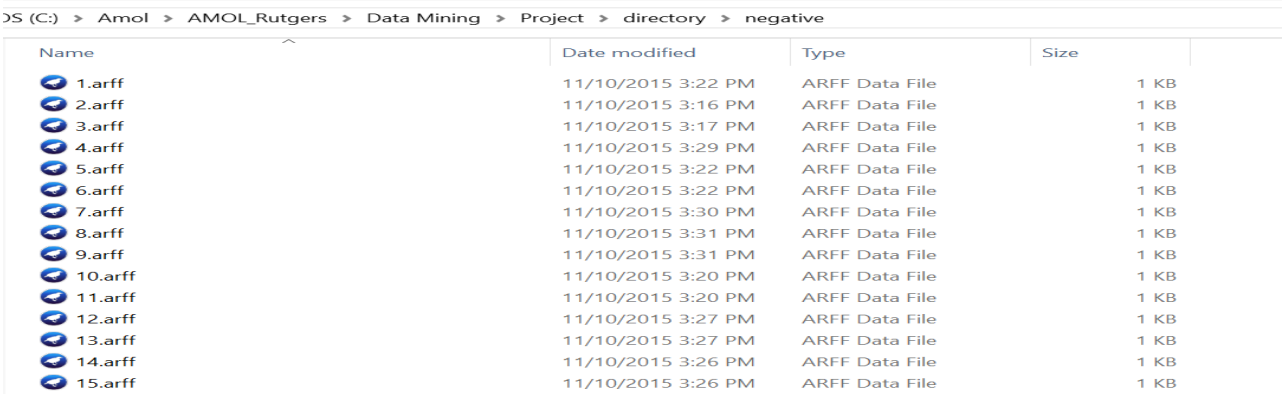
WEKA

SCRIPTING

The following snapshots will show the positive and negative arff files for a specific hotel. Below we have shown the directory of both positive and negative reviews with few instances of arff files for both positive reviews and negative reviews.

All reviews are classified under two folders under one directory. One positive folder which contains **67 files of positive reviews** for each hotel and the second is **negative folder which contains 70 files** of negative reviews for each hotel. The 70 files in each folder are arff files which were written manually for both positive and negative reviews.





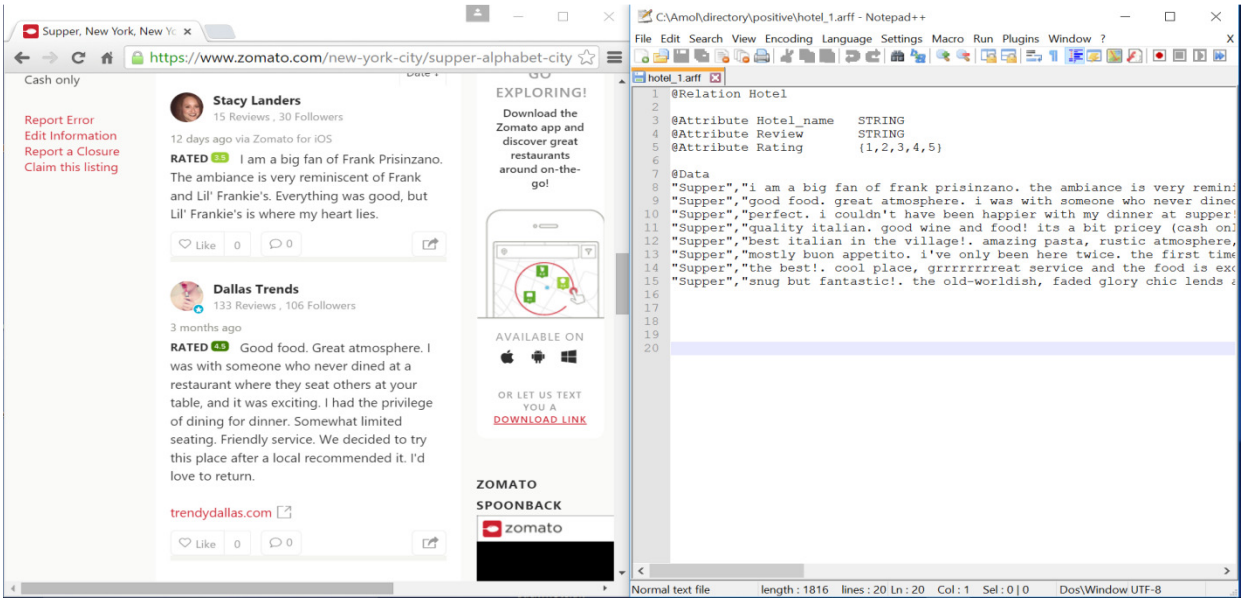
DATA MINING PROJECT

OS (C:) > Amol > AMOL_Rutgers > Data Mining > Project > directory > positive

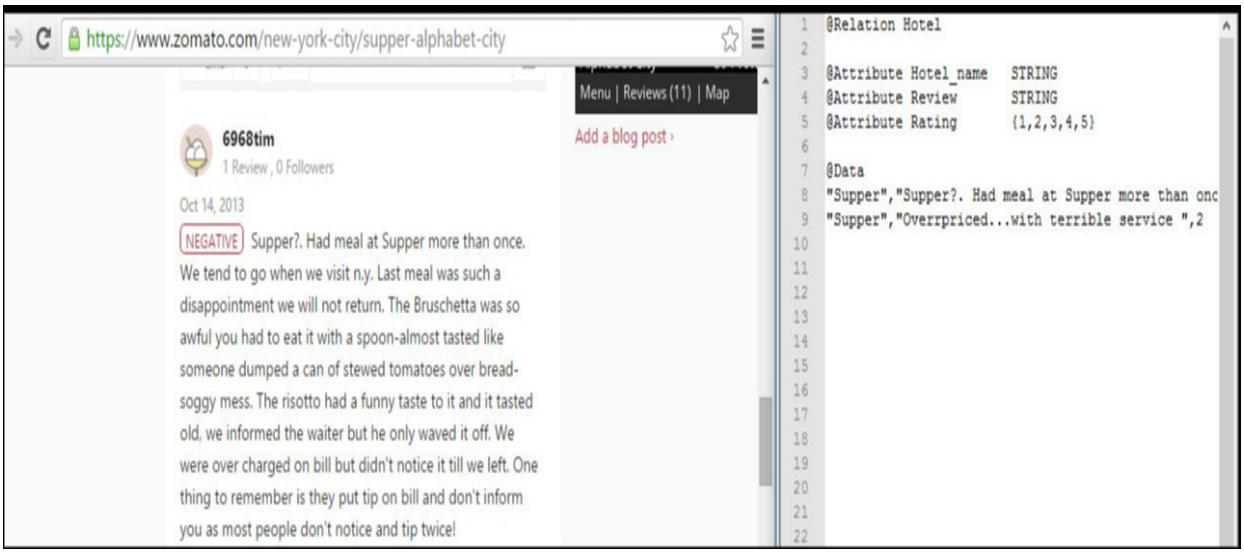
Name	Date modified	Type	Size
hotel_1.arff	11/10/2015 4:56 PM	ARFF Data File	2 KB
hotel_2.arff	11/10/2015 3:38 PM	ARFF Data File	1 KB
hotel_3.arff	11/10/2015 3:42 PM	ARFF Data File	1 KB
hotel_4.arff	11/10/2015 3:44 PM	ARFF Data File	1 KB
hotel_5.arff	11/10/2015 3:46 PM	ARFF Data File	1 KB
hotel_6.arff	11/10/2015 3:59 PM	ARFF Data File	1 KB
hotel_7.arff	11/10/2015 4:04 PM	ARFF Data File	1 KB
hotel_8.arff	11/10/2015 4:18 PM	ARFF Data File	1 KB
hotel_9.arff	11/10/2015 4:54 PM	ARFF Data File	1 KB
hotel_10.arff	11/10/2015 4:51 PM	ARFF Data File	1 KB
hotel_11.arff	11/10/2015 4:54 PM	ARFF Data File	1 KB
hotel_12.arff	11/10/2015 5:00 PM	ARFF Data File	1 KB
hotel_13.arff	11/10/2015 5:04 PM	ARFF Data File	1 KB
hotel_14.arff	11/10/2015 5:06 PM	ARFF Data File	1 KB
hotel_15.arff	11/10/2015 5:09 PM	ARFF Data File	1 KB

We have shown it accordingly with the Zomato site for better reference.

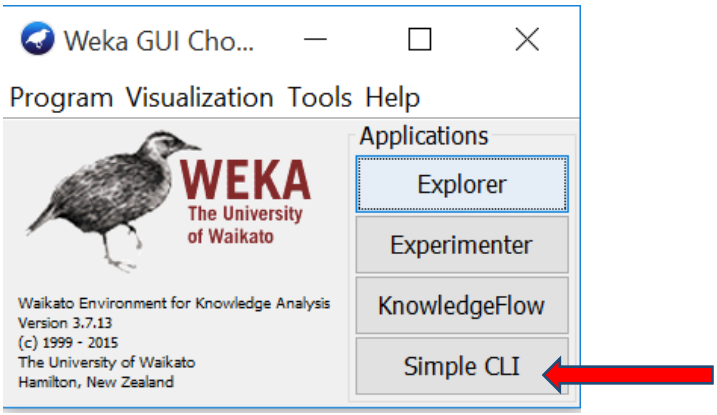
Positive review arff →



Negative Review arff →



After the arff files are created, using Weka CLI (command line interface) script was written to convert all the arff files into single data. Below is the snapshot of the CLI command used to convert the arff files into single data:



SimpleCLI

```
Welcome to the WEKA SimpleCLI

Enter commands in the textfield at the bottom of
the window. Use the up and down arrows to move
through previous commands.
Command completion for classnames and files is
initiated with <Tab>. In order to distinguish
between files and classnames, file names must
be either absolute or start with '.\' or '~/'
(the latter is a shortcut for the home directory).
<Alt+BackSpace> is used for deleting the text
in the commandline in chunks.

> help

Command must be one of:
          java <classname> <args> [ > file]
          break
          kill
          capabilities <classname> <args>
          cls
          history
          exit
          help <command>
```

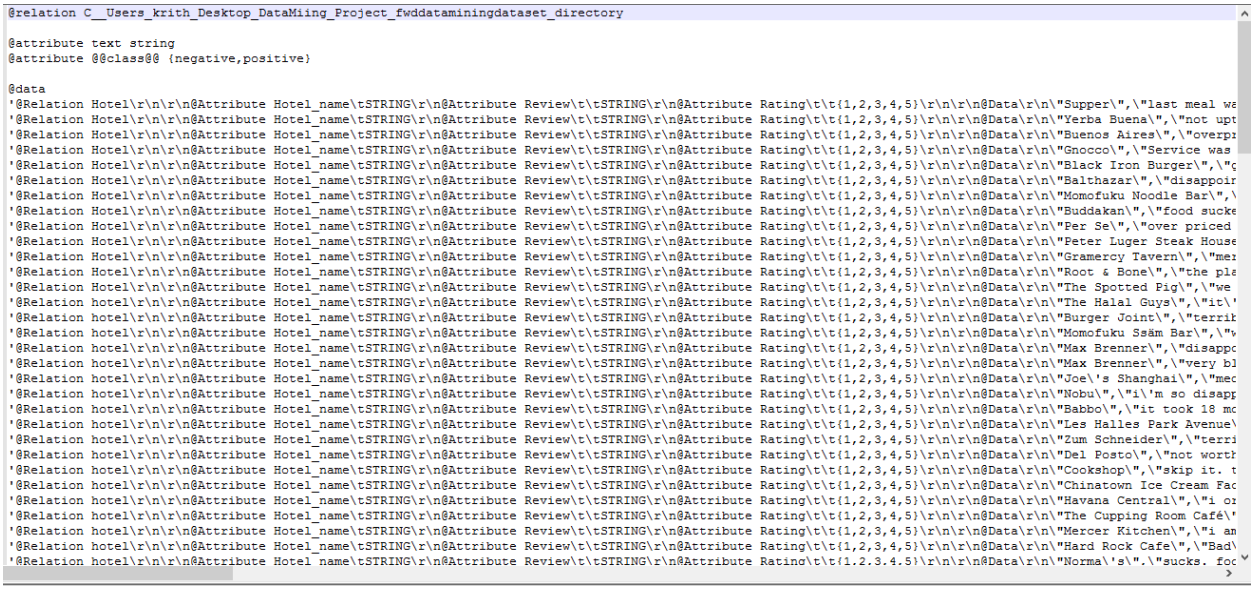
```
java weka.core.converters.TextDirectoryLoader -dir C:\Amol\directory > C:C:\Amol\directory\hotel_nyc.arff
```

The command tells from where the data of positive reviews and negative reviews are to be referred from. We have also given the path where the single arff file indicating the path where the complete data should be stored.

SimpleCLI

```
Welcome to the WEKA SimpleCLI

Enter commands in the textfield at the bottom of
the window. Use the up and down arrows to move
through previous commands.
Command completion for classnames and files is
initiated with <Tab>. In order to distinguish
between files and classnames, file names must
be either absolute or start with '.\' or '~/'
(the latter is a shortcut for the home directory).
<Alt+BackSpace> is used for deleting the text
in the commandline in chunks.

> help

Command must be one of:
          java <classname> <args> [ > file]
          break
          kill
          capabilities <classname> <args>
          cls
          history
          exit
          help <command>

> java weka.core.converters.TextDirectoryLoader -dir C:\Amol\directory > C:\Amol\directory\hotel_nyc.arff

Finished redirecting output to 'C:\Amol\directory\hotel_nyc.arff'.
```

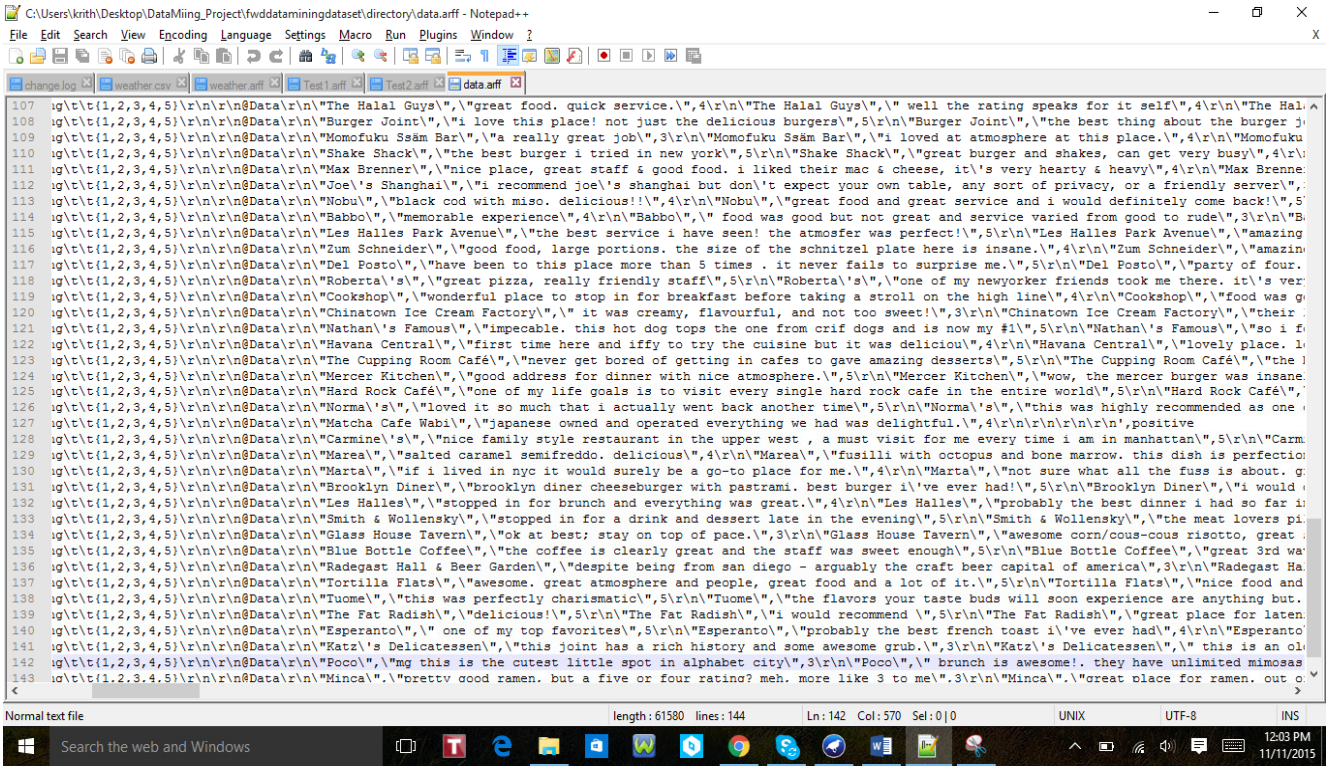
Once we merge all the positive and negative arff files to build one Master arff files, we see the data organised as below



The Master arff file contains the metadata of the data in the file. The Meta data says the datatype of the attributes we have in the file along with the classes present in the file. The file lists the two classes namely positive and negative from the data set.

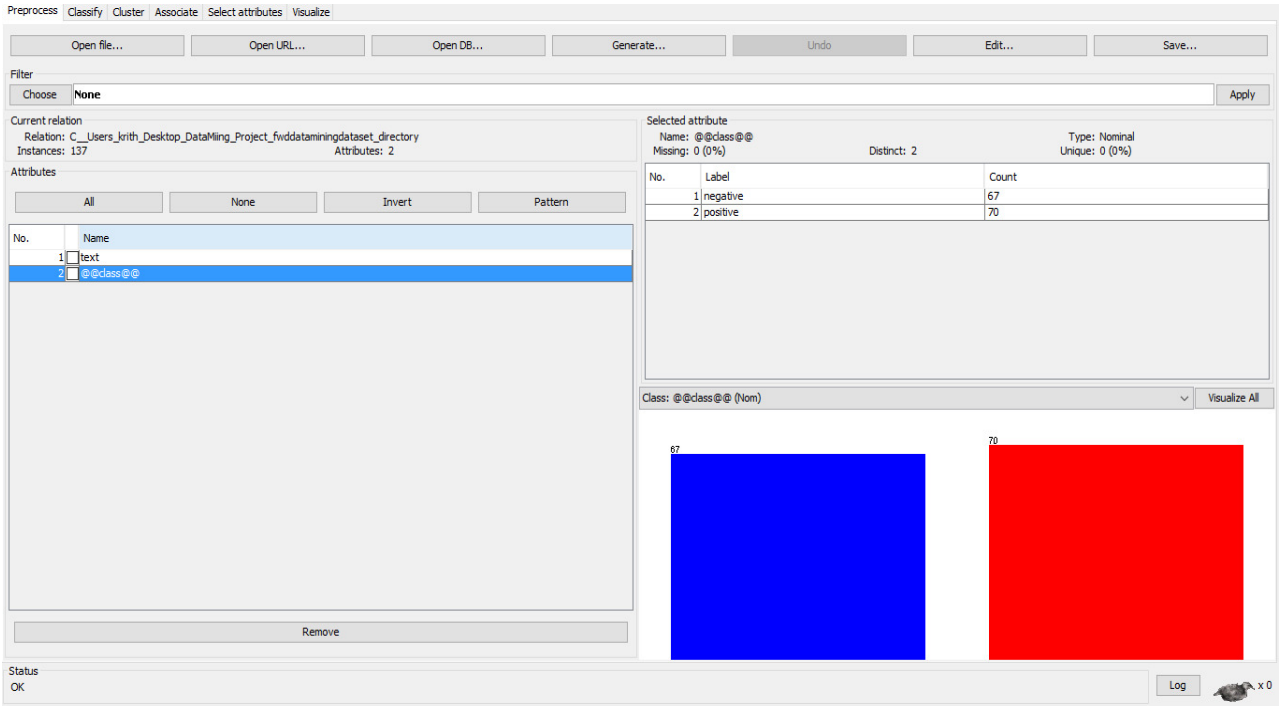
It can be noted that the master data details are contained inside the **data** (@data) attribute. The same data can be seen for each review of a hotel as shown in the below snapshot:

DATA MINING PROJECT



PRE-PROCESSING

The first step in Weka is Pre-processing. The data once loaded generates a histogram based on the class (positive and negative) as below.

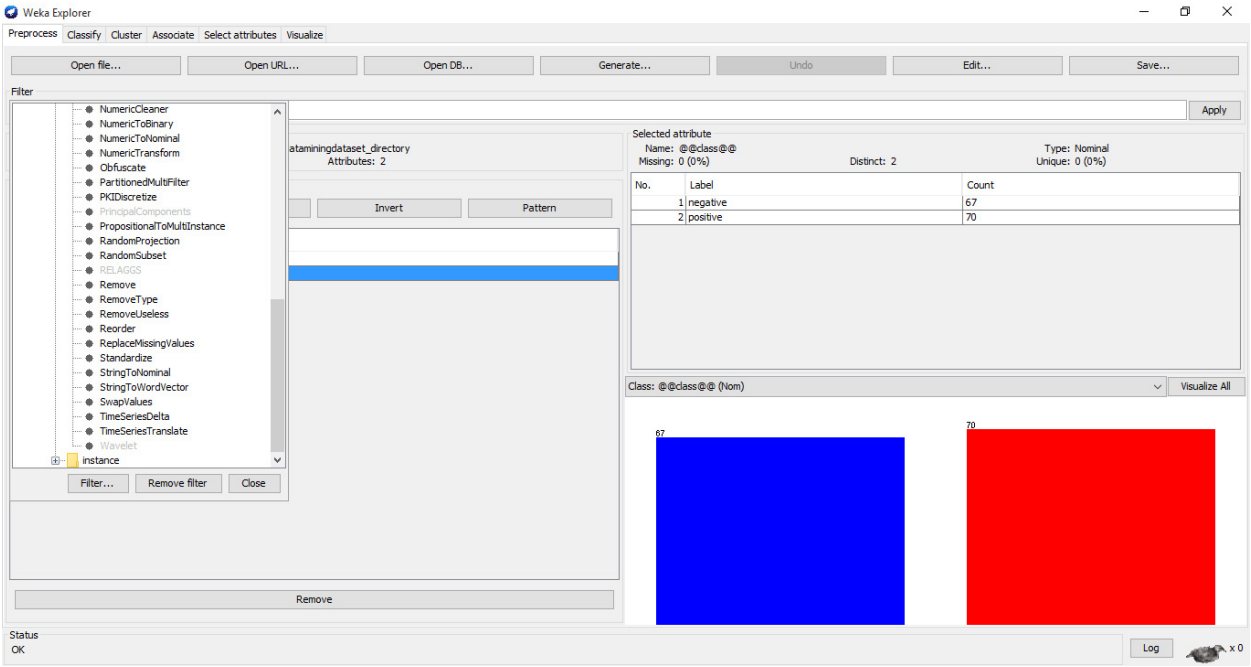


The count on each graph shows the number of instances we have in each class. The reviews are split into each word and the words are compared and classified to check if each word falls either in positive or negative class. We can choose the filers on which data pre-processing can be done. Weka gives us three options to filter the data on. Firstly, data can be filtered on all available filters or multiple filters can be selected. If not, Weka lets us filter data on supervised or unsupervised filters.

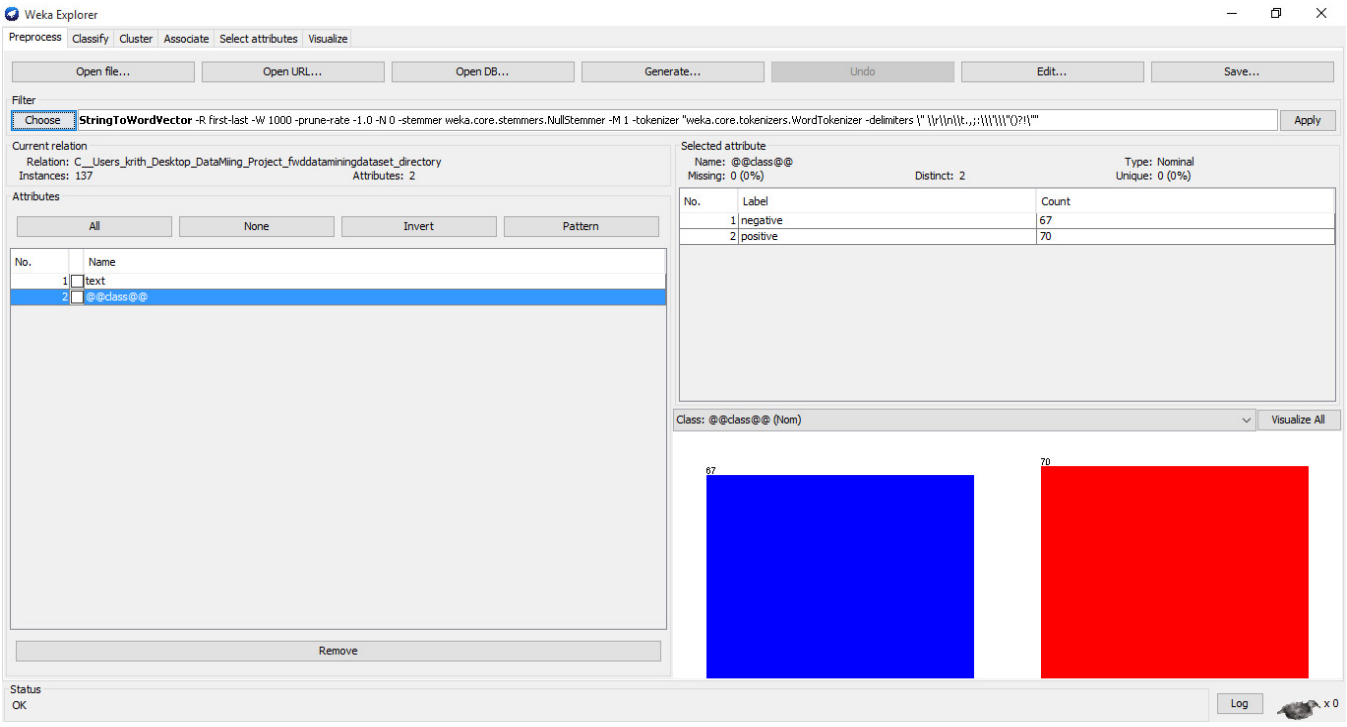
Each supervised and unsupervised filters have different in built functions to filter the data on. For instance, supervised filter has functions called **Attribute classification** based on which we classify the attribute for each class.

Here, we have used unsupervised **StringToWordVector** filter to split each word in the reviews to form a vector of individual words. **StringToWordVector** is used to Converts String attributes into a set of attributes representing word occurrence information from the text contained in the strings.

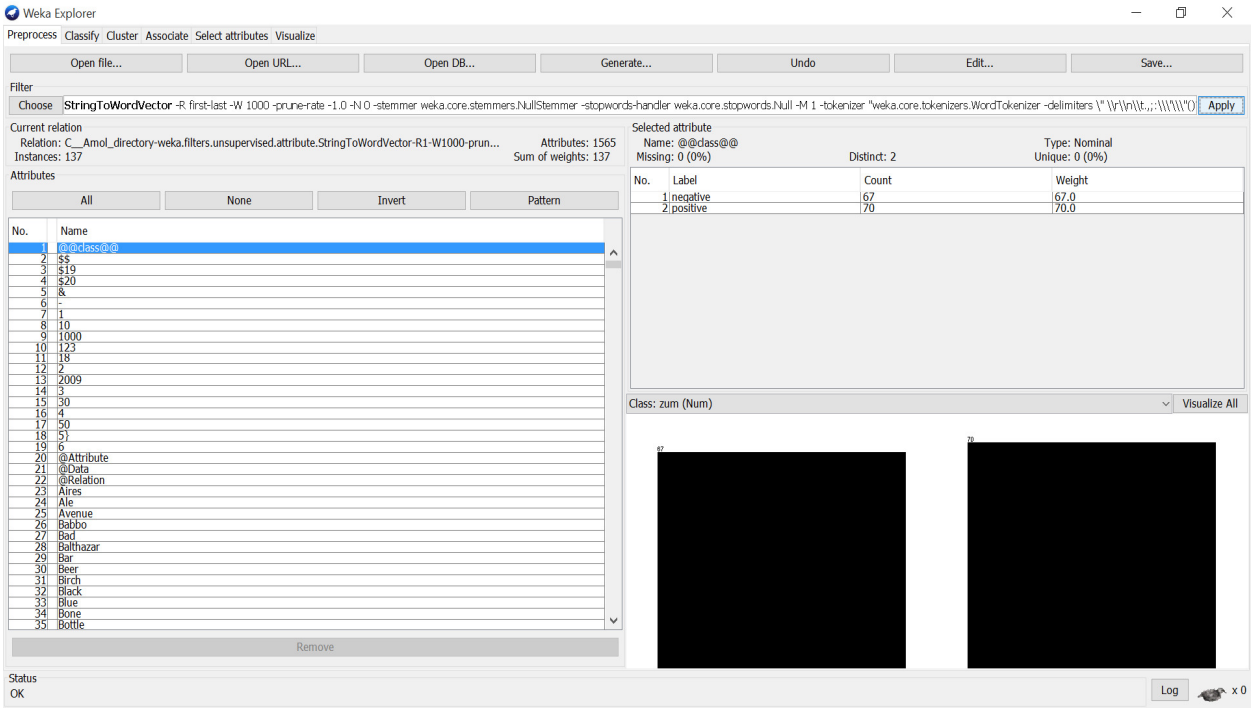
Following are the steps to select and apply the unsupervised filter **StringToWordVector.**



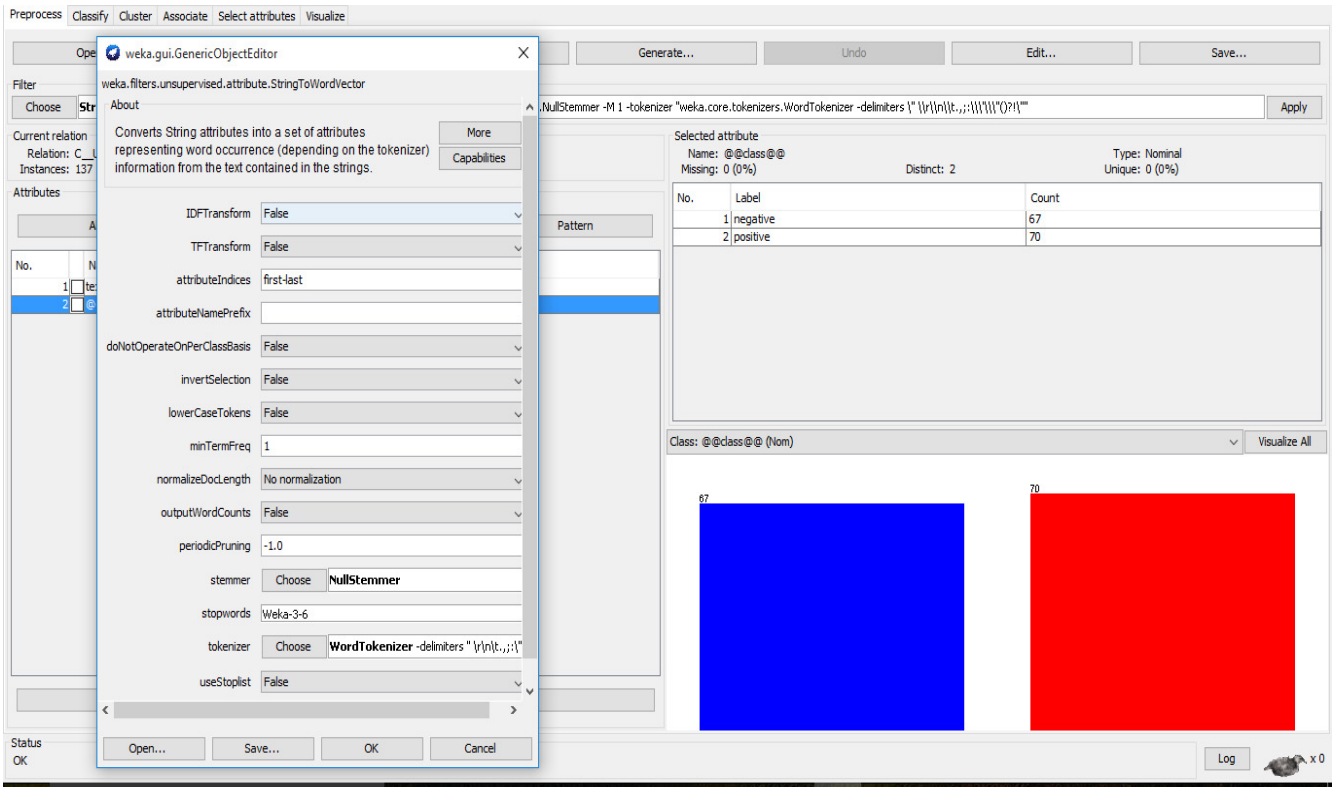
The filter once selected can be applied by clicking the attribute for which we want to apply the filter for. In our case, we require the filter to be applied on the class that has positive and negative instances. Once we apply the filter we can see the number of instances and attributes displayed as below



DATA MINING PROJECT

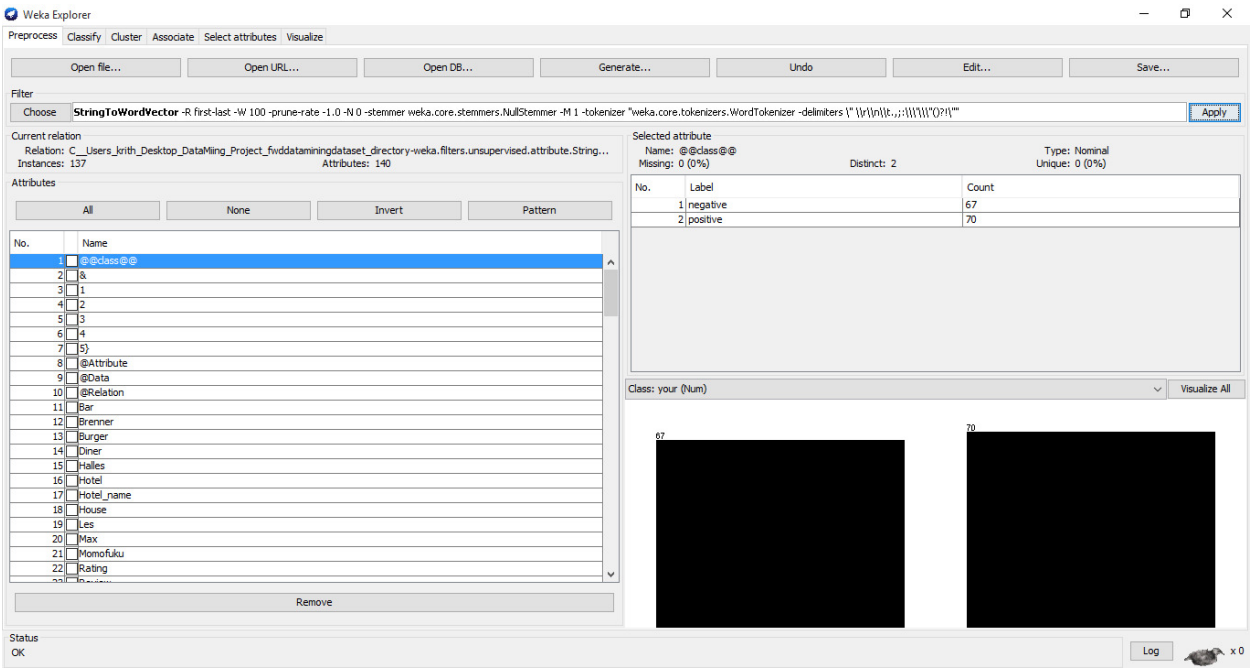


We can modify the filter options by clicking the field that shows the selected filter name to open the edit menu of the filter.



The first few options provided in edit option are briefed below:

1. IDFTransform ➔ Inverse document frequency gives the count of the words in the documents.
2. TFT ➔ Term frequency transform gives the count of each term in all the documents.
3. Minimum term frequency ➔ when set, gives those terms which has atleast the number of times set for the MTF.
4. Stemmer ➔ Used to combine words together like **Works, Working** and so on. It gives us option to choose between four different stemmers.
5. StopwordsHandler ➔ We can provide the word list to filter the document with those words. By default Weka has English language stop words. We can also select different language stop words.
6. WordToKeep ➔ If set lets us to retain only those number of words



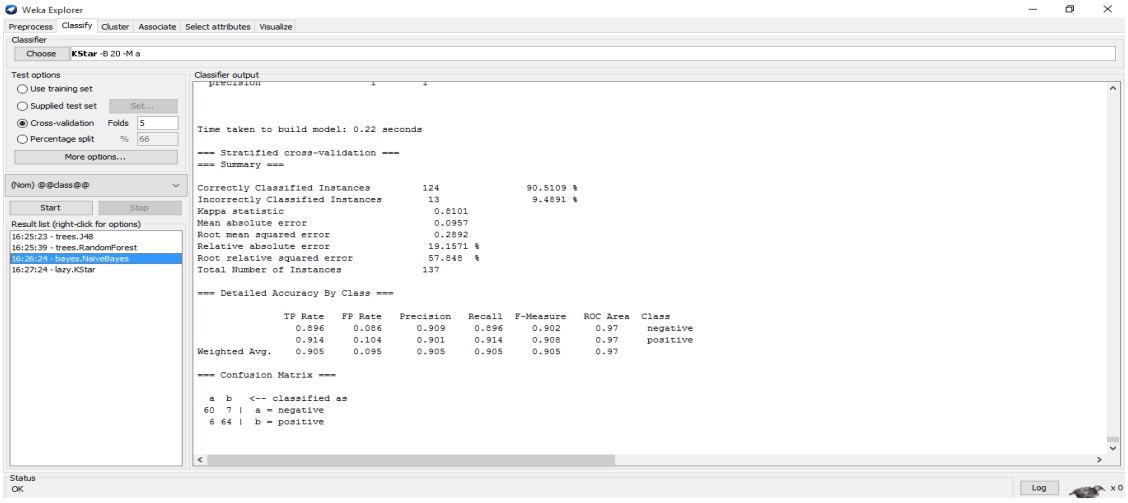
After the filter is applied, we can see the number of attributes and number of instances depending on the filer options you set. We can see the change in the count of attributes before and after applying the filers.

CLASSIFICATION

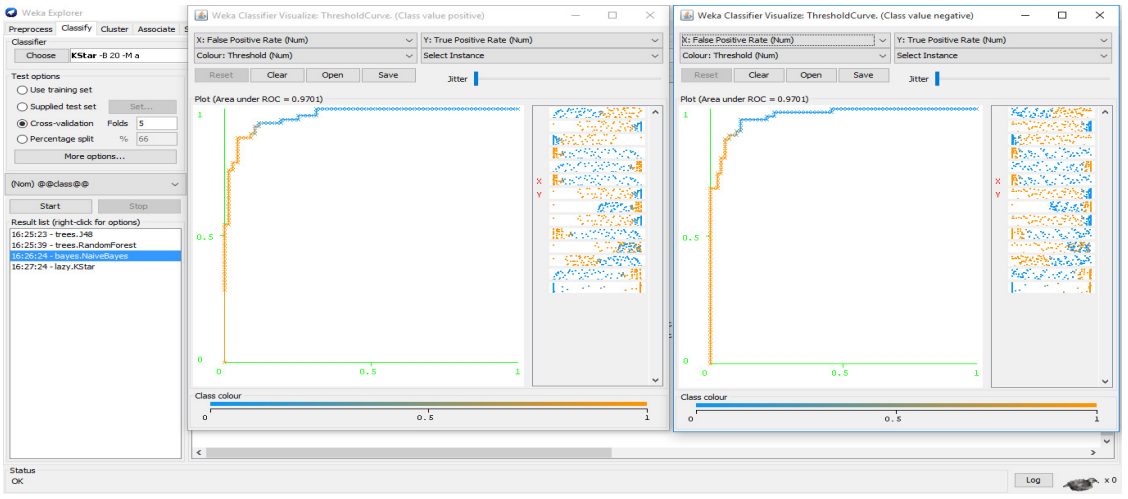
After pre-processing the data, data needs to be classified using the data mining algorithms to check if the words that are split fall in the respective class. We have used the below four algorithms to classify the texts. We have plotted their respective ROC curves.

ROC (Receiver Operating Characteristics) curve is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. This is plotted against “False Positive Rate” (X-axis) and “True Positive Rate” (Y-axis). True Positive Rate is also called as the **Recall**.

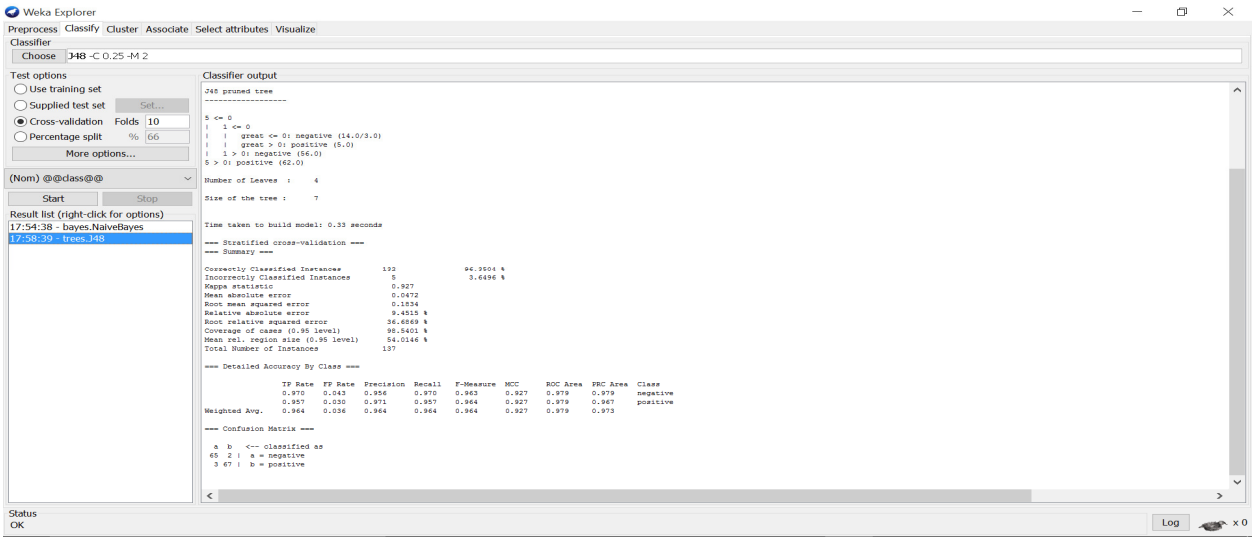
1. Naïve Bayes ➔ In machine learning, **naive Bayes classifiers** are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features. Naïve bayes are highly scalable requiring a number of parameters linear in the number of variables. This comes under bayes inside the classifiers.
   The below snapshot shows the naïve bayes algorithm applied on the input data set. We can see the confusion matrix at the end that shows how many data were correctly classified as positive and negative. Shows the accuracy and the error generated with the algorithm on the data set.



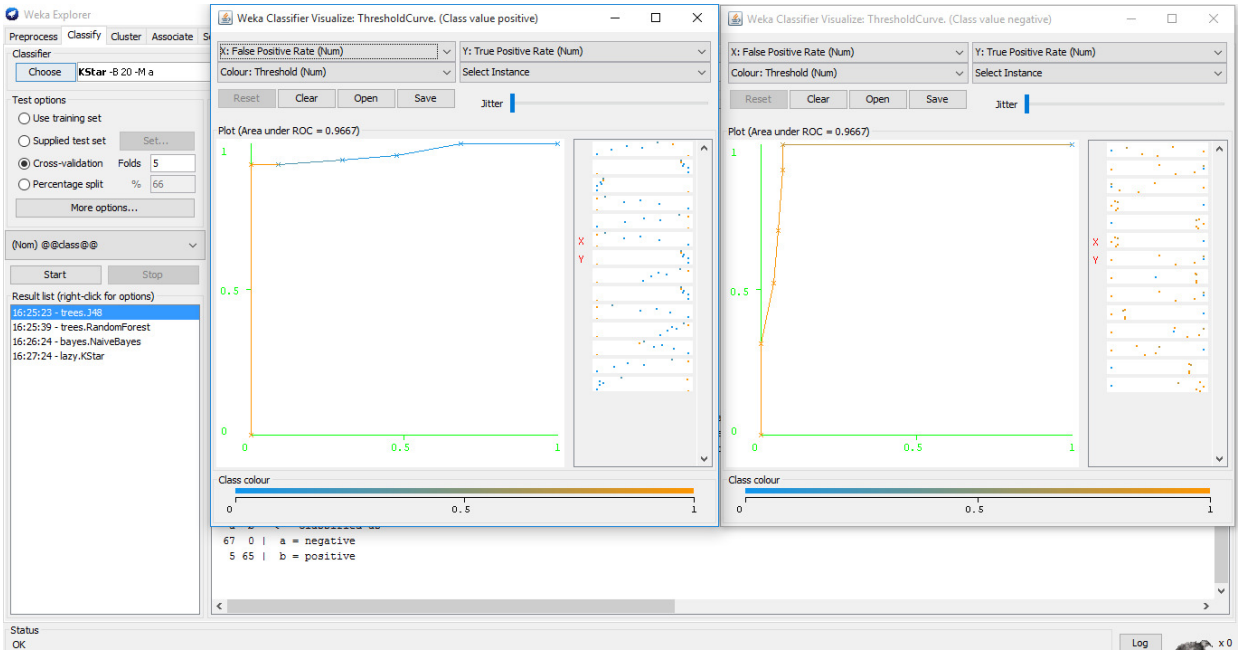
ROC Curve -



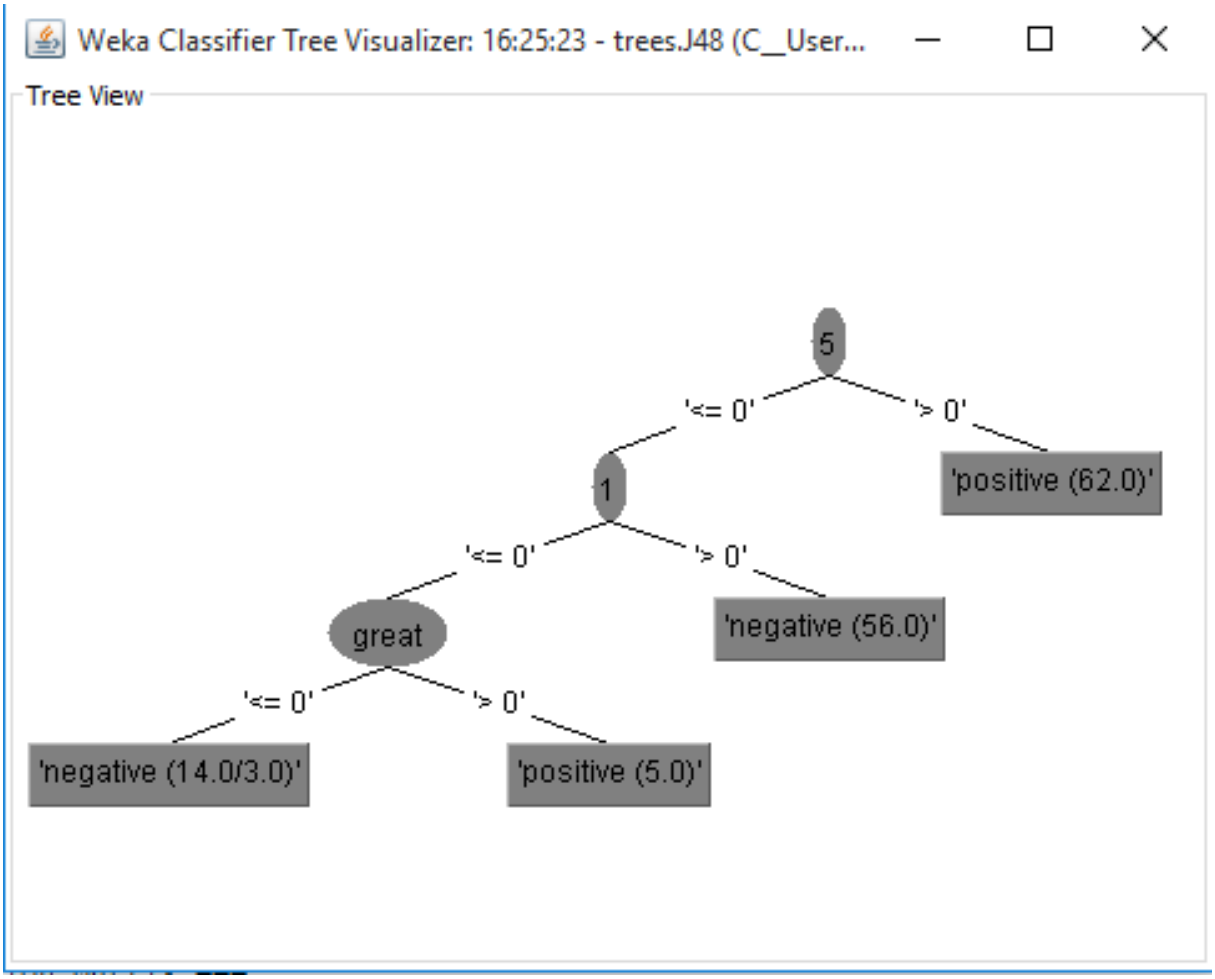
2. J48 ➔ This is inside the trees within classifier. The J48 is an algorithm which is an open source java implementation inside C4.5 algorithm. The C4.5 algorithm creates a decision tree based on a set of labeled data. For instance here we have two classes positive and negative.
   The below snapshot shows the amount of data correctly classified. Total of 132 instances were correctly classified and 5 were incorrectly classified. We can also note the difference in the confusion matrix where there are 2 misclassified data in negative and 3 misclassified data in positive.



ROC Curve:

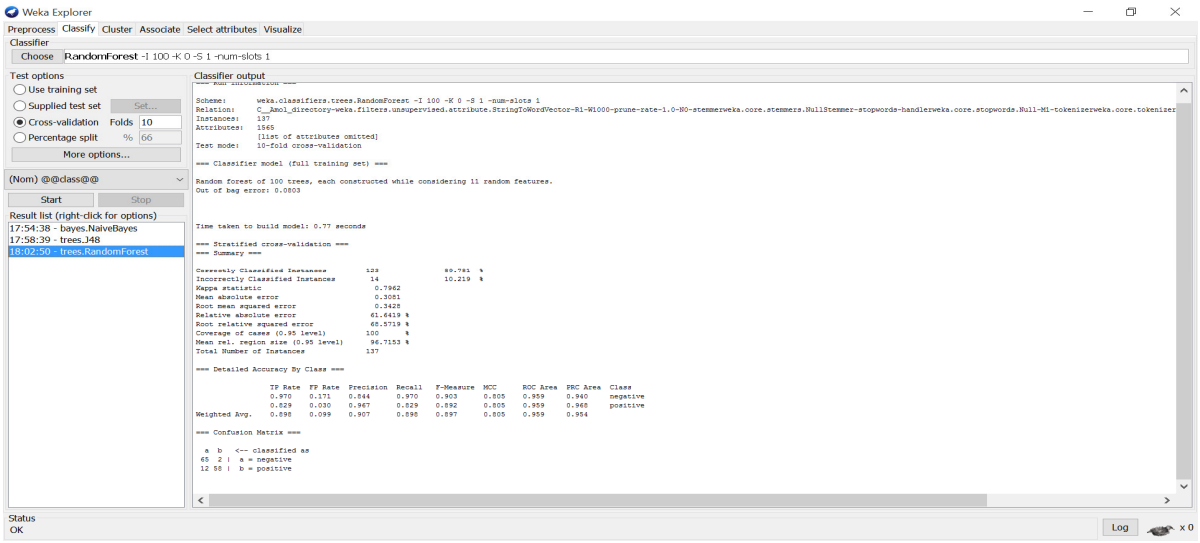


The J48 builds a decision tree based on the classes and the instances.

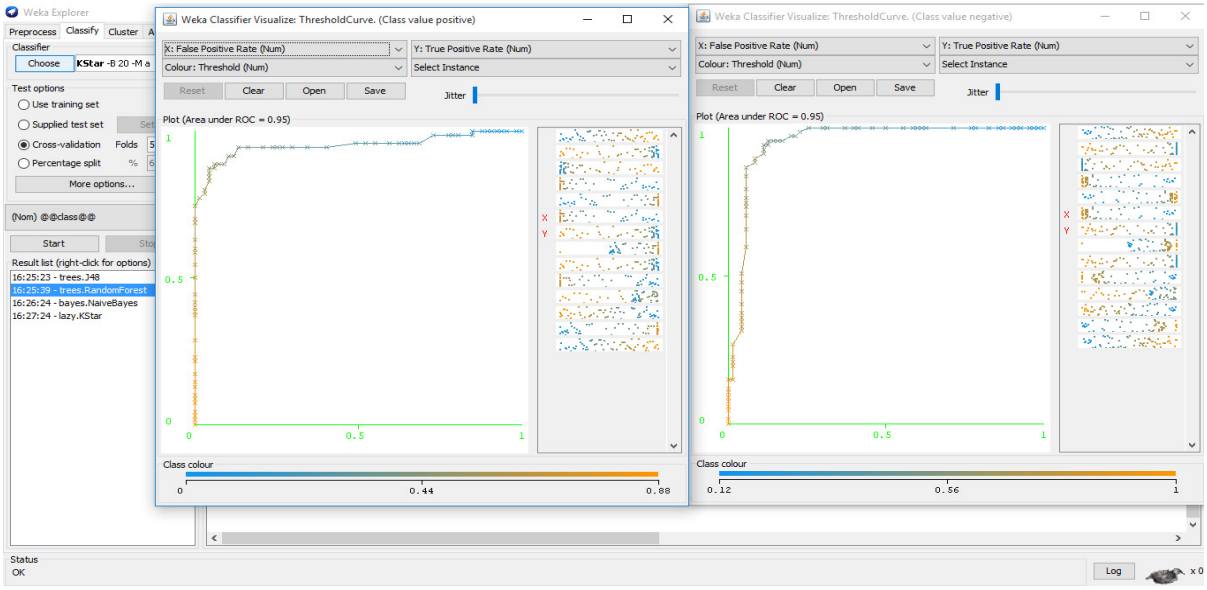


3. Random Forest ➔ This is again available inside the trees within classifier. This is an ensemble model where we create a decision tree by modifying the input features. In this approach a subset of input features is chosen to form each training set. The subset can be chosen randomly or based on the recommendation of domain experts.

   The below snapshot shows the amount of incorrectly classified to correctly classified. We can see that Random forests have incorrectly classified 14 instances out of the total 137 instances where 2 are from negative class and 12 are from positive class. The confusion matrix can be found below with the difference.

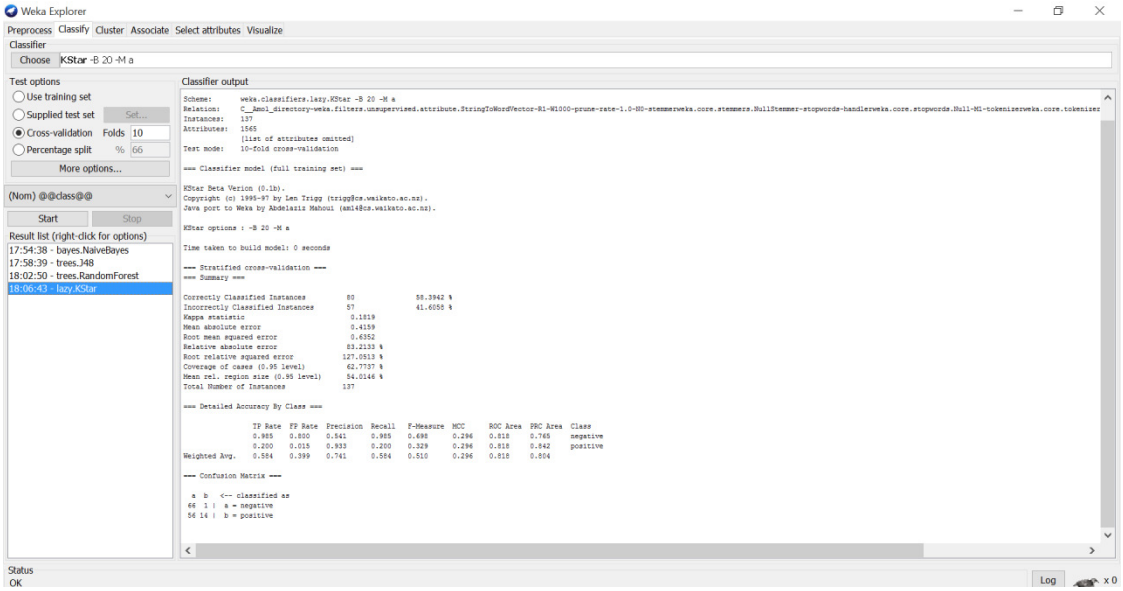


ROC Curve:

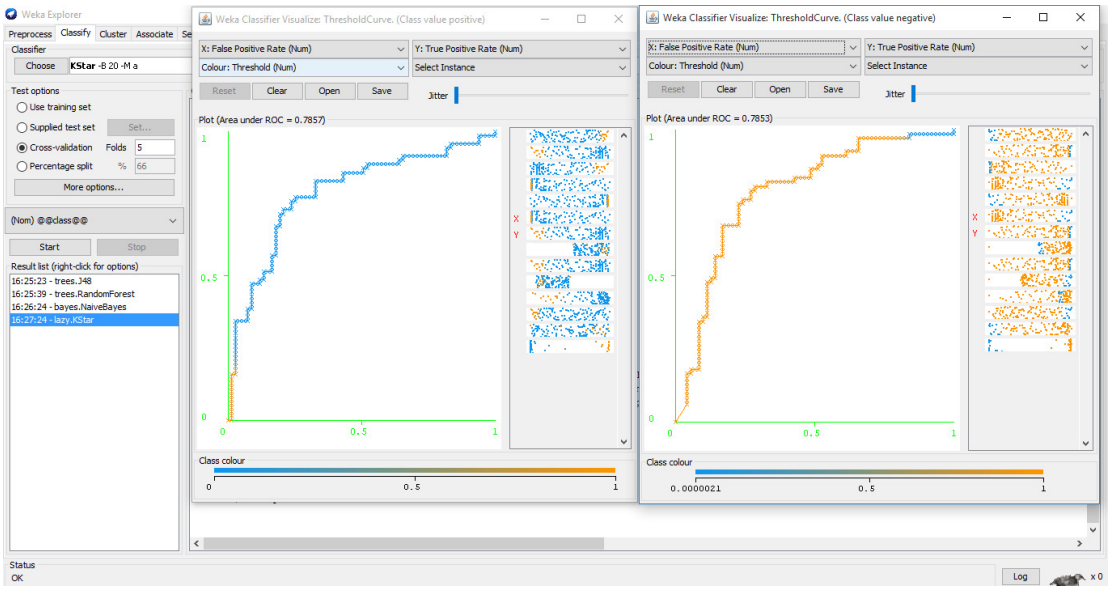


4. KStar ➔ Kstar is a lazy classifier. **Lazy learning** is a learning method in which generalization beyond the training data is delayed until a query is made to the system, as opposed to in eager learning, where the system tries to generalize the training data before receiving queries.

   KStar uses entropic distance measure to evaluate the data set. Kstar is an instance based classifier that is the class of a test instance is based upon the class of those training instances similar to it.

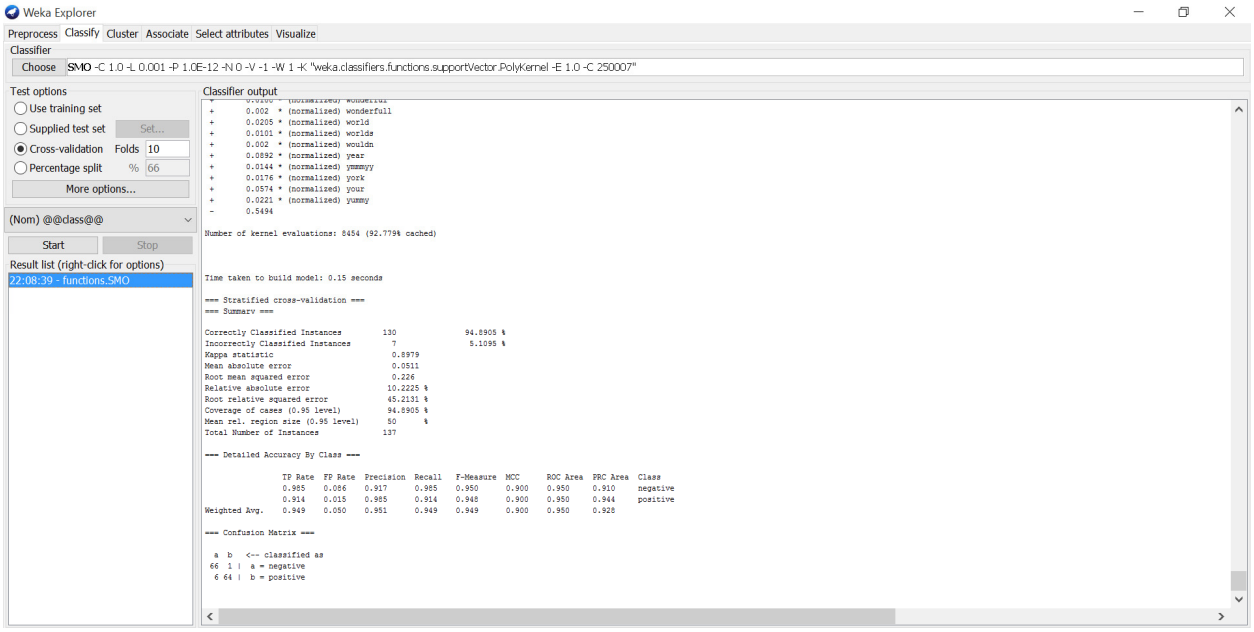


ROC Curve:

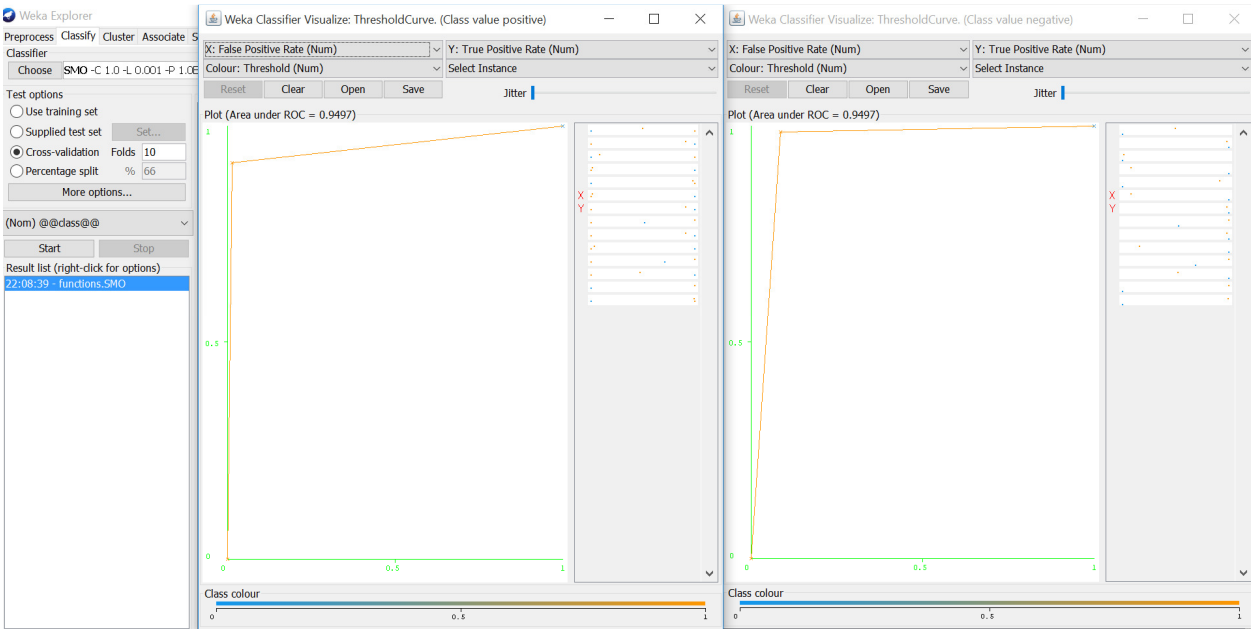


5) Support Vector Machine →

SVM also works very well with high-dimensional data and avoids the curse of dimensionality problem. Another unique aspect of this approach is that it represents the decision boundary using a subset of the training examples, known as the support vectors.



ROC Curve:



COMPARISION OF CLASSIFICATION MODELS ➔

Sr.No.	Naïve Bayes	J 48	Random Forest	K - Star	Support Vector Machine
Area under ROC	0.9701	0.966	0.95	0.78	0.94
Correctly classified Instance (%)	90.5	96.35	89.7	58.3	94.8
Time taken to build model (secs)	0.22	0.33	0.77	0	0.15

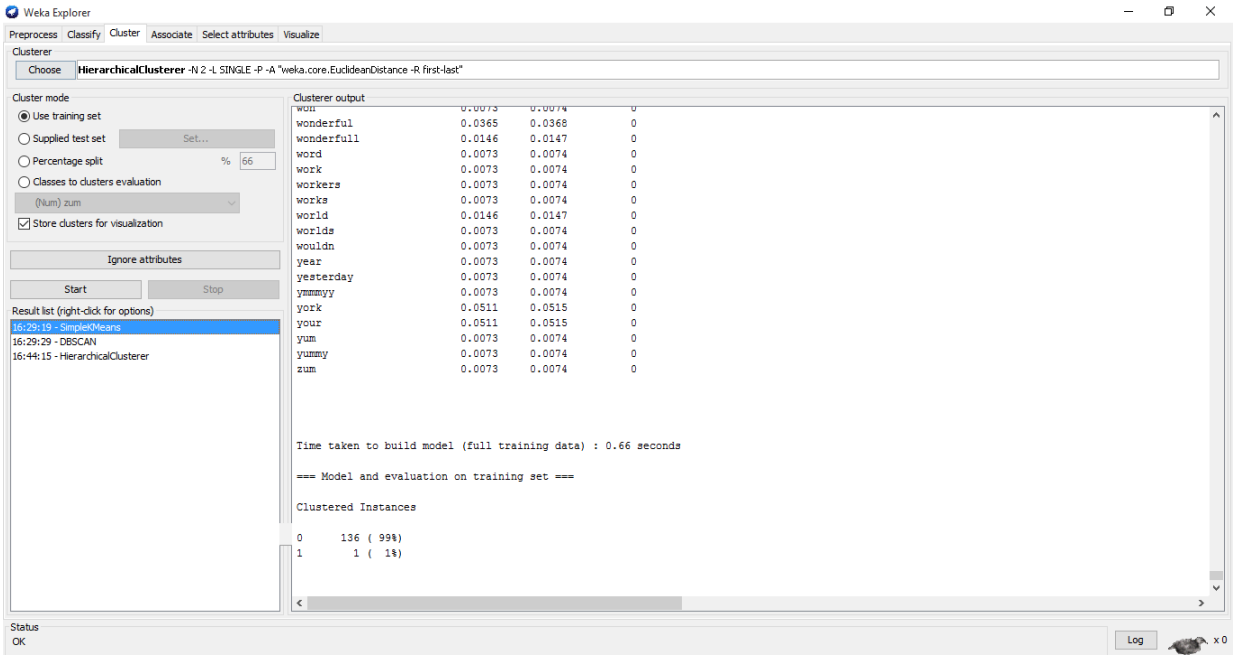
Conclusion ➔

J48 tree and **Support Vector Machine** have done the best classification for our textual data set.

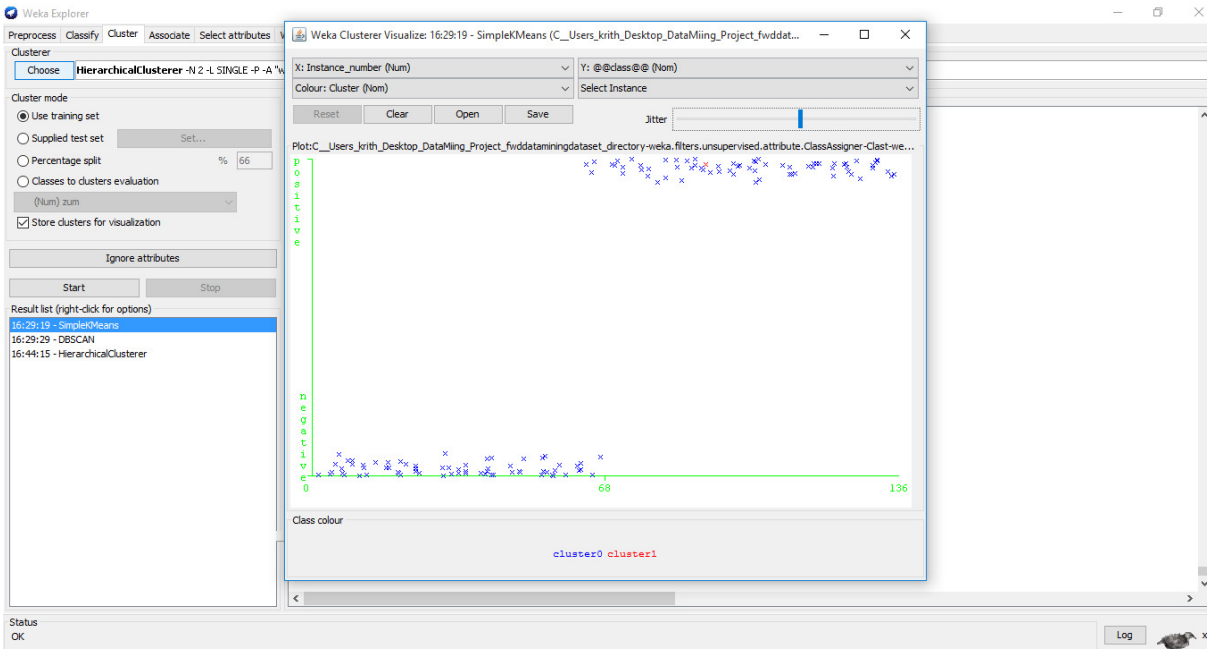
CLUSTERING

1) SimpleKMeans →

This is a prototype-based, partitional clustering technique that attempts to find a user-specified number of clusters (K), which are represented by their centroids.



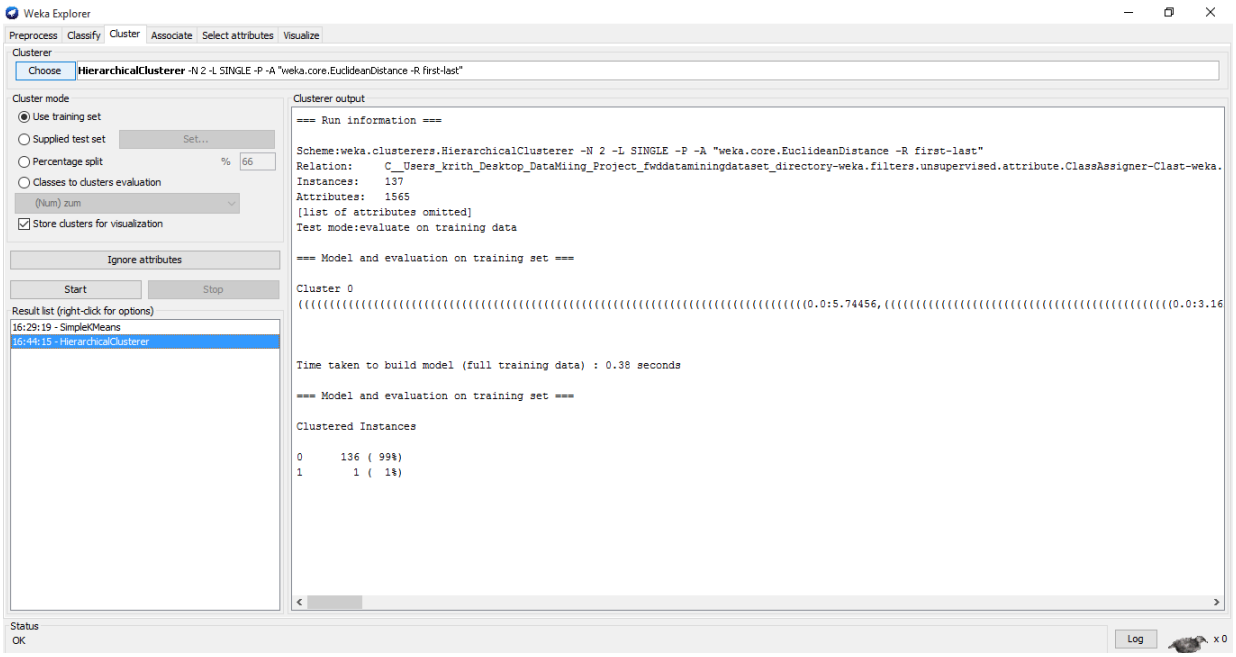
Simple Kmeans visualization of cluster -



Here we see one instance has been wrongly clustered.

2) Hierarchical Cluster →

This clustering approach refers to a collection of closely related clustering techniques that produce a hierarchical clustering by starting with each point as a singleton cluster and then repeatedly merging the two closest clusters until a single, all encompassing cluster remains.



Dendogram-