Patient Care analysis - HealthCare

Problem statement

Health Care Industry desires to classify the Patients using their pathology data for their CARE (Self-management, Doctor-Advise, Further Diagnostic and Chronic Medication) management improvement that facilitates to build a multi-classification model to build **CARE Management Model (CMM)** with right classification of patient.

Tools Used:

Spark Cluster
Spark MLLIB (Machine Learning Library)
R & Python Language

Apache Spark

Apache Spark is a lightning-fast cluster computing technology, designed for fast computation. It is based on Hadoop MapReduce and it extends the MapReduce model to efficiently use it for more types of computations, which includes interactive queries and stream processing. The main feature of Spark is its in-memory cluster computing that increases the processing speed of an application. Spark is designed to cover a wide range of workloads such as batch applications, iterative algorithms, interactive queries and streaming. Apart from supporting all these workload in a respective system, it reduces the management burden of maintaining separate tools.

Spark SQL

Spark SQL is a component on top of Spark Core that introduces a new data abstraction called Schema RDD, which provides support for structured and semi-structured data.

Spark Streaming

Spark Streaming leverages Spark Core's fast scheduling capability to perform streaming analytics. It ingests data in mini-batches and performs RDD (Resilient Distributed Datasets) transformations on those mini-batches of data.

MLlib (Machine Learning Library)

MLlib is a distributed machine learning framework above Spark because of the distributed memory based Spark architecture. It is, according to benchmarks, done by the MLlib developers against the Alternating Least Squares (ALS) implementations. Spark MLlib is nine times as fast as the Hadoop

disk-based version of Apache Mahout(before Mahout gained a Spark interface).

GraphX

GraphX is a distributed graph-processing framework on top of Spark. It provides an API forexpressing graph computation that can model the user-defined graphs by using Pregel abstraction

Python

Python is a dynamic, interpreted (bytecode-compiled) language. There are no type declarations of variables, parameters, functions, or methods in source code. This makes the code short and flexible, and you lose the compile-time type checking of the source code. Universally, Python has gained a reputation because of it's easy to learn. The syntax of Python programming language is designed to be easily readable. Python has significant popularity in scientific computing. The people working in this field are scientists first, and programmers second. Nowadays working on bulk amount of data, popularly known as big data. The more data you have to process, the more important it becomes to manage the memory you use. Here Python will work very efficiently.

Scala

If you're the kind of data scientist who deals with large datasets, Scala will be invaluable. It's practically the de facto language for the current Big Data tools like Apache Spark, Finagle, Scalding, etc. Many of the high performance data science frameworks that are built on top of Hadoop usually are written and use Scala or Java. The reason Scala is used in these environments is because of its amazing concurrency support, which is key in parallelizing a lot of the processing needed for large data sets. It also runs on the JVM, which makes it almost a no-brainer when paired with Hadoop. Like Java, Scala is object-oriented, and uses a curly-brace syntax reminiscent of the C programming language. Unlike Java, Scala has many features of functional programming languages like Scheme, Standard ML and Haskell, including currying, type inference, immutability, lazy evaluation, and pattern matching.

Problem Understanding

This study is commissioned with the following objectives:

The care management model system gives computer the ability to learn without being explicitly programmed. (Machine Learning) To classify patients based on different criteria. The Care management models will facilities for

1) Personalized Medicine
2) Predictive Analytics and Preventive Measures
3) Self-Motivated Care
4) Disease Modelling and Mapping

Sample Source:

Naive Bayes Classification Algorithm For Data Accuracy

NB.py

```
# Import required packages
from __future__ import print_function, division
from pyspark.mllib.classification import NaiveBayes, NaiveBayesModel
from pyspark.mllib.util import MLUtils
from pyspark import SparkContext
# Load and parse the data file.
sc = SparkContext("local","classification")
data = MLUtils.loadLibSVMFile(sc, '/home/mcalab/mini/HealthSVM.txt')
# Split data approximately into training (50%) and test (50%)
training, test = data.randomSplit([0.5, 0.5])
# Train a naive Bayes model.
model = NaiveBayes.train(training, 1.0)
predictions = model.predict(test.map(lambda x: x.features))
predictions.foreach(print)
predictions.coalesce(1).saveAsTextFile("/home/mcalab/mini/output1")
# Make prediction and test accuracy.
predictionAndLabel = test.map(lambda p: (model.predict(p.features), p.label))
predictionAndLabel.foreach(print)
cp = predictionAndLabel.filter(lambda (x, v): x == v).count()
tc = test.count()
accuracy = (cp/tc)*100
print('\n')
print('*****************Output********************')
```

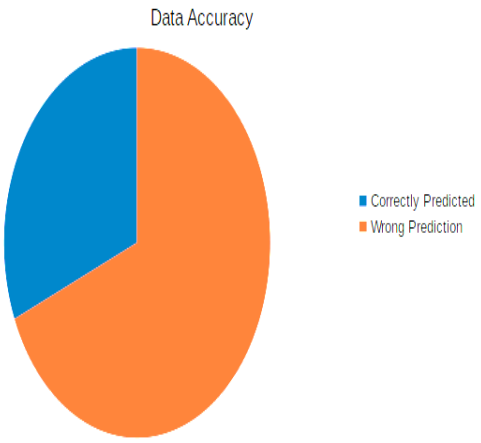
SCREEN SHOTS

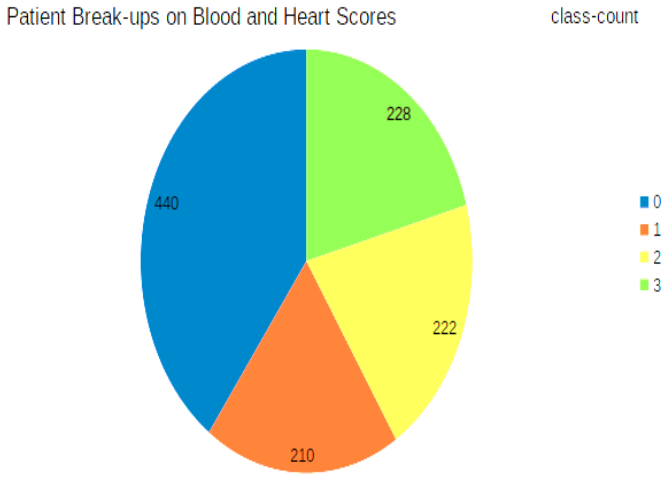
Naive Bayes Classification Algorithm For Data Accuracy

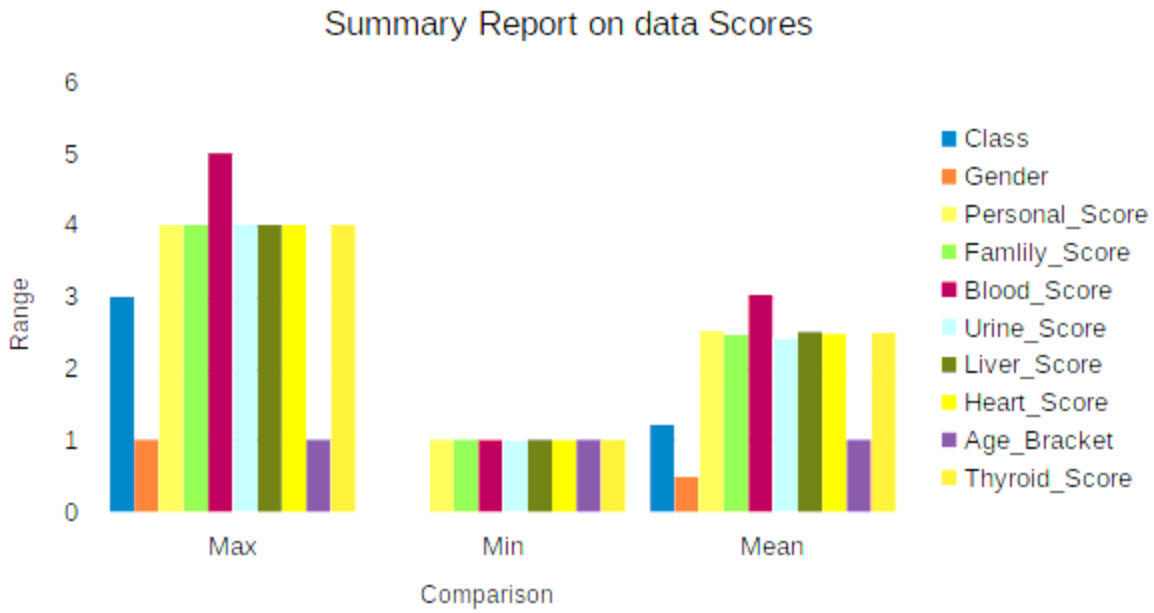
Running the problem on Spark

Output Confusion Matrix

```
mcalab@mcalab: ~/mini
16/10/22 15:31:51 INFO TaskSchedulerImpl: Removed TaskSet 22.0, whose tasks have all c
ompleted, from pool
16/10/22 15:31:51 INFO DAGScheduler: ResultStage 22 (count at /home/mcalab/mini/DT.py:
33) finished in 0.023 s
16/10/22 15:31:51 INFO DAGScheduler: Job 17 finished: count at /home/mcalab/mini/DT.py
:33, took 0.031801 s


############## OutPut ##############

                  True | False
         --------------------------
          True   | 92     | 0     |
         --------------------------
          False  | 0      | 183   |

############### OutPut #################

Label & prediction Error Count  :  0
Test Count                      :  464.0
Test Error                      :  0.0 %

*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*_*

16/10/22 15:31:51 INFO SparkContext: Invoking stop() from shutdown hook
16/10/22 15:31:51 INFO SparkUI: Stopped Spark web UI at http://127.0.0.1:4040
16/10/22 15:31:51 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint
stopped!
```







CONCLUSION

Big data is the reality and is going to stay there for a long time. While have a care management model by any hospital they can predict the future condition of any patient without any wait. Time is a major problem for health care. Within the time anything may happens. So life is saved by care management model. So the model created by this project is so important in the everyday life. But the project model wants to more accurate. The only way to improve the accuracy of this model is collect more data and applies it for modelling. So in future hospital have to concentrate on big data preparation and management in order to create a good model for the system. The project suggest to gain success hospital need to change service manner and increase service on patients who are professionally unbalanced and try to efficiently utilise care management model for patients, serious and quick action needed when pathology values get changing abnormally.

Customer Segmentation - Banking

Problem:

BANK wants to sell the new product in insurance/loans so analyse and report customer's money spending trends. This segmentation/information can be used to provide solution to meet the requirement of the customer which might be very unique and product can be offered.

Tools Used:

Hadoop Cluster
Sqoop in Hadoop
Mysql database
PIG scripts in Hadoop
Hive Data warehouse in Hadoop
R Language

Sqoop in Hadoop Framework

The evaluated data from MYSQL is being imported using Sqoop into Hadoop environment for further tuning and analysis. The sqoop import command is used to bring the table from MYSQL to HDFS.

Pig

The imported data from sqoop is given an input to the Pig grunt shell. The log data is loaded in to pig shell using load command. The loaded log data may contain some of the text data in an unstructured format. These unstructured data are cleansed using pig filter commands. This fine-tuned data are stored as comma separated csv(Comma Separated) file for further analysis using pig storage.

Hive

The fine tuned data which is in csv format is loaded into a database inside hive. A table with suitable data type is created inside the database using HQL create query, after analysing the data. The data in the csv file is loaded in to the table created for taking adhoc reports using select query. The result values are stored in to HDFS as csv file for further analysis in R.

R LANGUAGE

The resultant values stored as csv file from Hive based on the query is taken as an input to R using read command. The loaded data is used to create

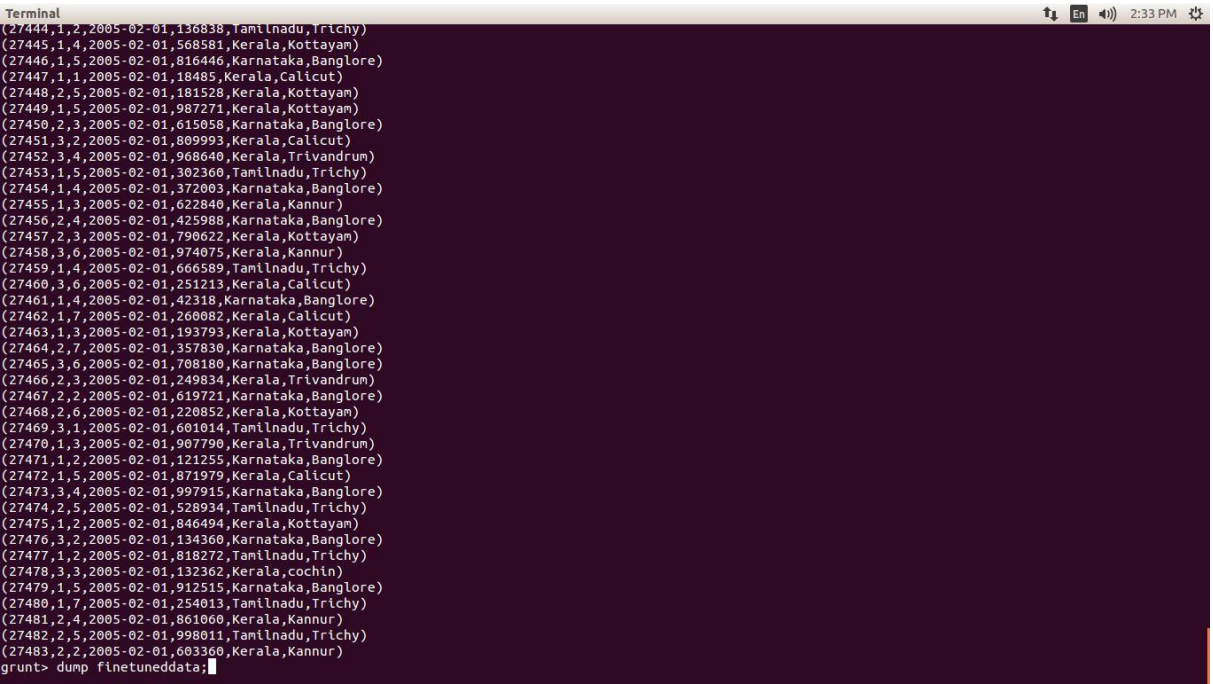
summary of the data for analysis, and for creation of the Chart or Graphs for Visualization and easy understanding about the analysis.

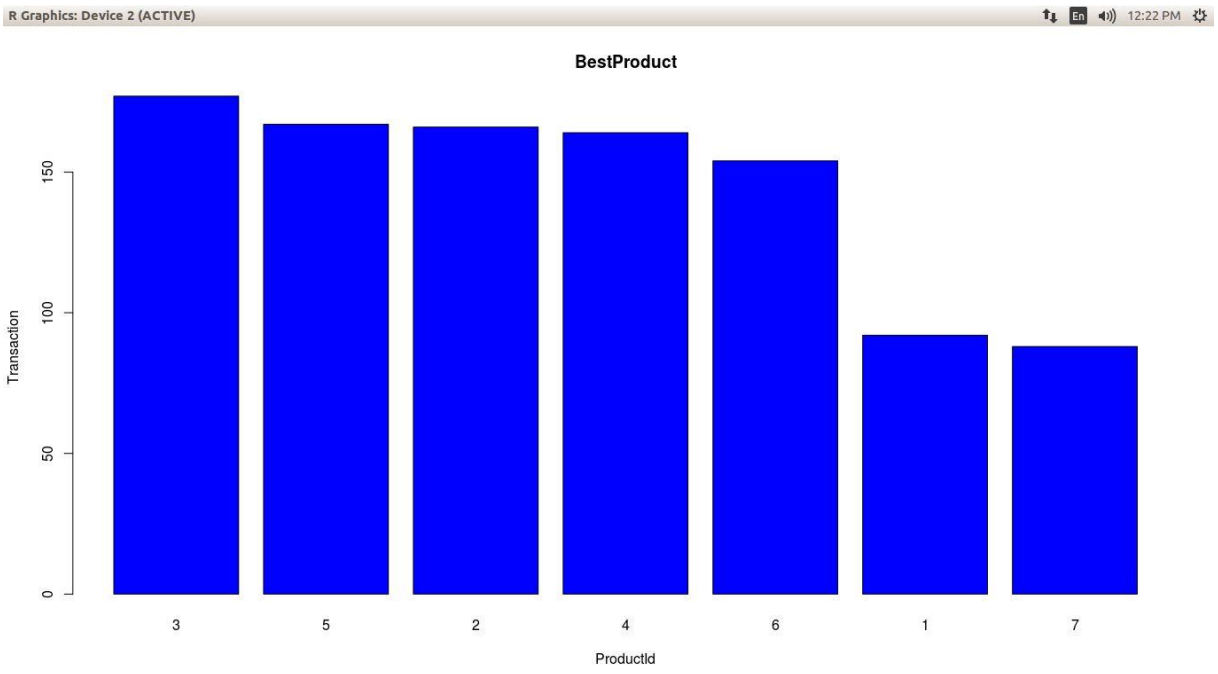
MySql

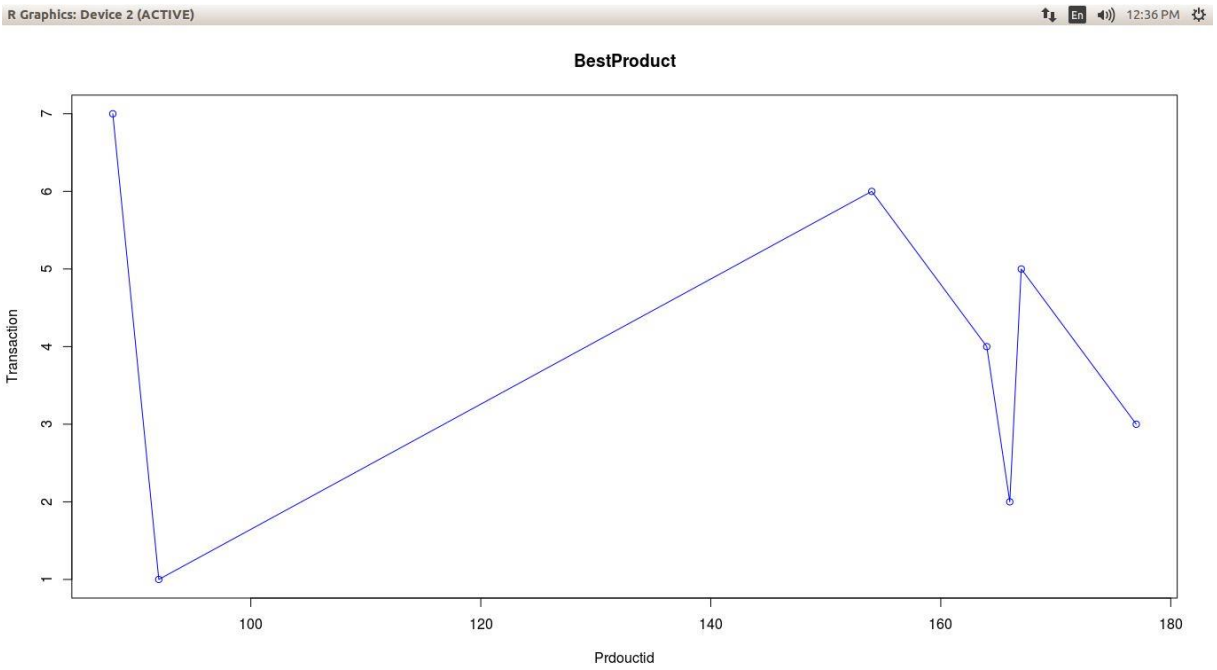
Un-structured tower data is being captured from tower data management system. The log file containing the tower utilization details of different towers is copied into mySql. A single linear table is created in mySql and the data is copied for quick process and data evalution using my-sql capabilities.

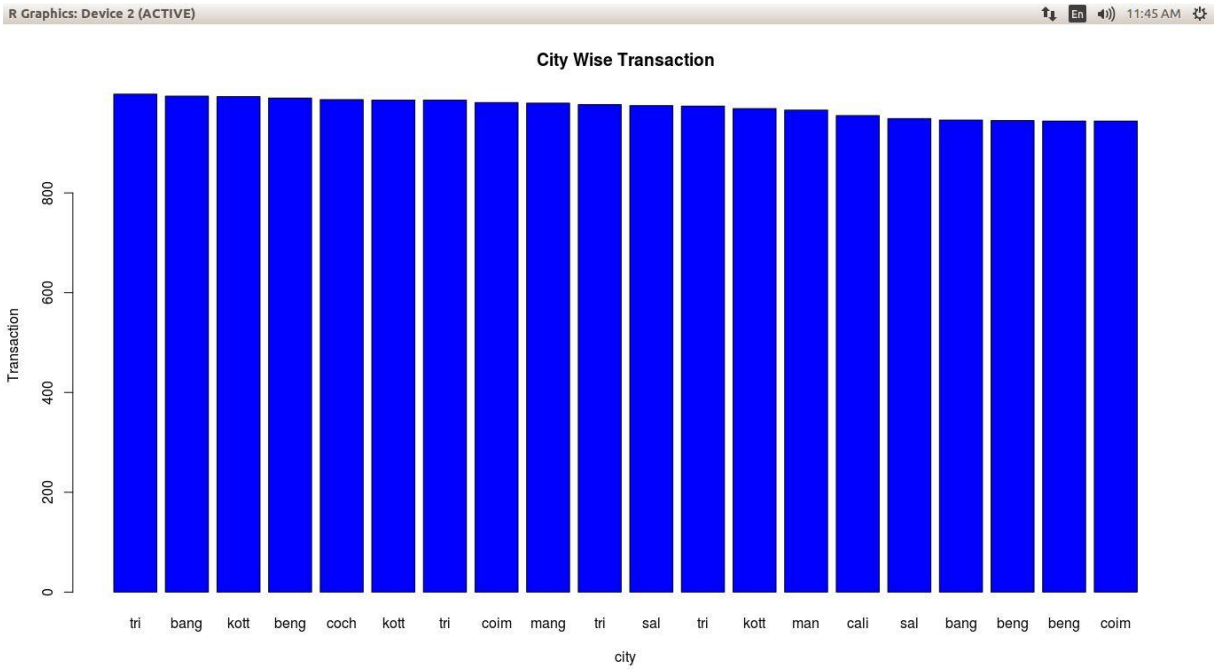
DATA PROFILING

Product Item	No of Records	Transaction Amount
Savings Account	183	92609400
Current Account	321	166077747
Term Deposits	350	177386416
Loans	338	164462435
Credit Cards	331	167896232
Gold	312	154077082
Insurance	165	88702632
Total	**2000**	**1011211944**




R Graphics: Device 2 (ACTIVE)
12:22 PM
BestProduct
Transaction
0
50
100
150
3
5
2
4
6
1
7
ProductId


R Graphics: Device 2 (ACTIVE)
12:36 PM
BestProduct
Transaction
1
2
3
4
5
6
7
100
120
140
160
180
Prdouctid


R Graphics: Device 2 (ACTIVE)
11:45 AM
City Wise Transaction
Transaction
0
200
400
600
800
tri
bang
kott
beng
coch
kott
tri
coim
mang
tri
sal
tri
kott
man
cali
sal
bang
beng
beng
coim
city

Sentiment Analysis – Text mining

Problem statement

Text mining is gaining importance due to problem of discovering useful information from the data deluge that the organizations are facing today. This project intends to present a broad overview of text mining and its components and techniques and their use in various business applications. This project gives a description about text mining and the reasons for its increased importance over the years. This is followed with presenting a generic process framework for text mining and describes its different components and sub-components, business applications, and brief description of text mining tools available in the market. Text mining involves the pre-processing of document collections (text categorization, feature/term extraction, etc.), the storage of the intermediate representations, the techniques to analyze these intermediate representations (such as distribution, analysis, clustering, trend analysis, and association rules), and visualization of the results.

DATA PREPARATION

Data Set Description (Select, Clean, Construct, Integrate and Format Data)

This data's are collected from twitter and doing some cleansing and analysis. After it going to be text mining by using python or R to do some process and finally we get exact output from the result.

Text Mining Framework Components

The different stages in the text mining framework are described below:

Textual Data Sources

The textual data is available in numerous internal and external data source like electronic text, call centre logs, social media, corporate documents, research papers, application forms, service notes, emails, etc.

Pre-processing

Pre-processing tasks include methods to collect data from the disparate data sources. This is the preliminary step of identifying the textual information for mining and analysis. Pre-processing tasks apply various feature extraction methods against the data. Pre-processing tasks include different types of techniques to transform the raw, unstructured, original format data into structured, intermediate data format. Knowledge discovery operations are conducted against the structured intermediate data.

For the preparation of unstructured data into a structured data format, different techniques are needed than those of traditional data mining systems where the knowledge discovery is done against the structured data sources. Various preprocessing techniques exist and can be used in combination to create structured data representation from raw textual data. Therefore different combinations of techniques can be used based on the type of the raw textual data.

Text Cleansing

Text cleansing is the process of cleansing noisy text from the textual sources. Noisy textual data can be found in SMS, email, online chat, news articles, blogs and web pages. Such text may have spelling errors, abbreviations, non-standard terminology, missing punctuation, misleading case information, as well as false starts, repetitions, and special characters.

Noise can be defined as any kind of difference in the surface form of an electronic text from the original, intended or actual text. The text used in the in short message service (SMS) and on-line forums like twitter, chat and discussion boards and social networking sites is often distorted mainly because the recipients can very well understand the shorter form of the longer words and also reduces the time and effort of the sender. Most of the text is created and stored so that humans can understand it, and it is not always easy for a computer to process that text.

Removing stop words

Stop words are words which are filtered before or after processing of textual data. There is not one definite list of stop words which all tools use, if even used. Some tools specifically avoid removing them to support phrase search. The most common stop words found in the text are “the”, “is”, “at”, “which” and “on”. These kinds of stop words can sometimes cause problems when looking for the phrases that include them. Some search engines remove some of the most common words from the query on order to improve performance.

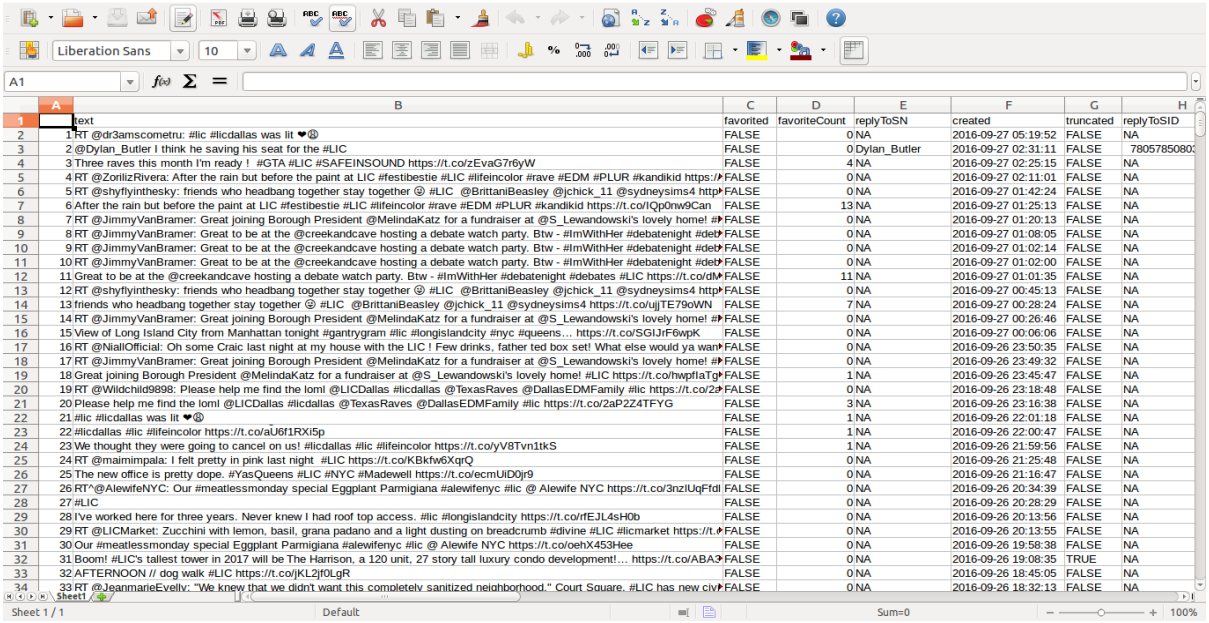
Tools Used:

Python Language

Python

Python is a dynamic, interpreted (bytecode-compiled) language. There are no type declarations of variables, parameters, functions, or methods in source code. This makes the code short and flexible, and you lose the compile-time type checking of the source code. Universally, Python has gained a reputation because of it’s easy to learn. The syntax of Python programming language is designed to be easily readable. Python has significant popularity in

scientific computing. The people working in this field are scientists first, and programmers second. Now a days working on bulk amount of data, popularly known as big data. The more data you have to process, the more important it becomes to manage the memory you use. Here Python will work very efficiently.



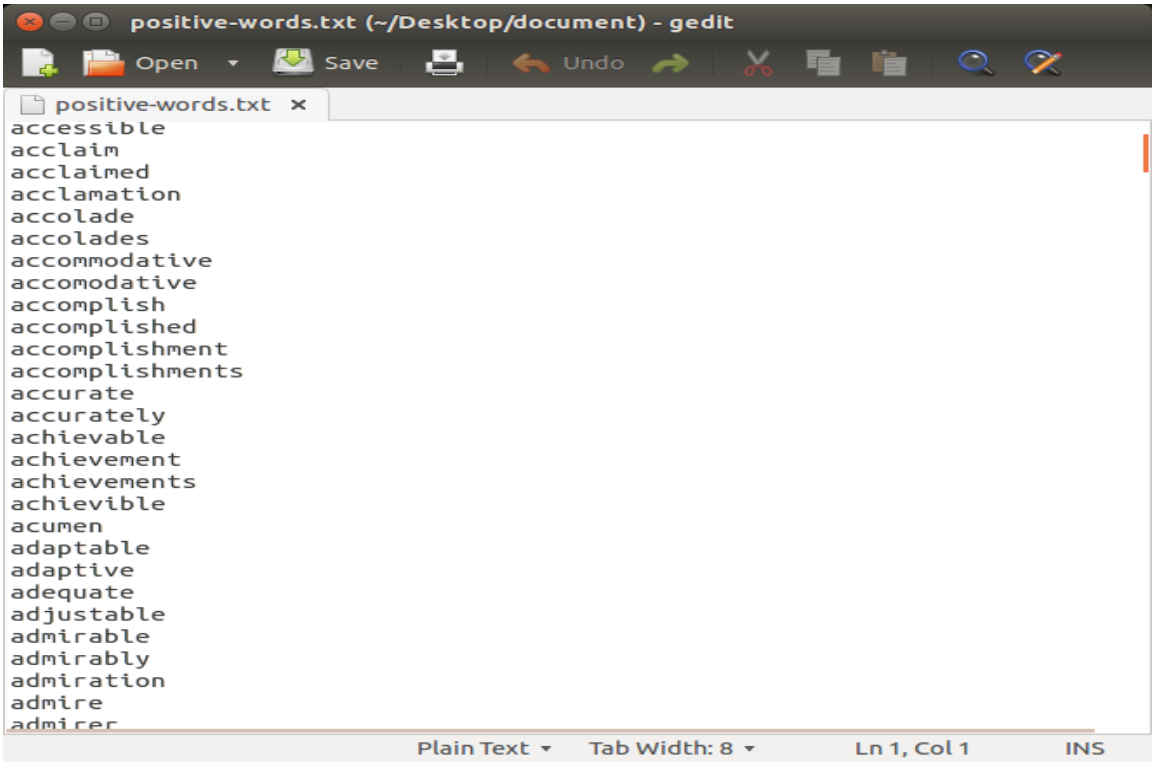
	A	B	C	D	E	F	G	H
1		text	favorited	favoriteCount	replyToSN	created	truncated	replyToSID
2	1	RT @dr3amscometru: #lic #licdallas was lit ❤Ⓢ	FALSE	0	NA	2016-09-27 05:19:52	FALSE	NA
3	2	@Dylan_Butler I think he saving his seat for the #LIC	FALSE	0	Dylan_Butler	2016-09-27 02:31:11	FALSE	780578508
4	3	Three raves this month I'm ready ! #GTA #LIC #SAFEINSOUND https://t.co/zEvaG7r6yW	FALSE	4	NA	2016-09-27 02:25:15	FALSE	NA
5	4	RT @ZorilizRivera: After the rain but before the paint at LIC #festibestie #LIC #lifeincolor #rave #EDM #PLUR #kandikid https://	FALSE	0	NA	2016-09-27 02:11:01	FALSE	NA
6	5	RT @shyflyinthesky: friends who headbang together stay together Ⓖ #LIC @BrittaniBeasley @jchick_11 @sydneysims4 http	FALSE	0	NA	2016-09-27 01:42:24	FALSE	NA
7	6	After the rain but before the paint at LIC #festibestie #LIC #lifeincolor #rave #EDM #PLUR #kandikid https://t.co/IQp0nw9Can	FALSE	13	NA	2016-09-27 01:25:13	FALSE	NA
8	7	RT @JimmyVanBramer: Great joining Borough President @MelindaKatz for a fundraiser at @S_Lewandowski's lovely home! #I	FALSE	0	NA	2016-09-27 01:20:13	FALSE	NA
9	8	RT @JimmyVanBramer: Great to be at the @creekandcave hosting a debate watch party. Btw - #ImWithHer #debatenight #deb	FALSE	0	NA	2016-09-27 01:08:05	FALSE	NA
10	9	RT @JimmyVanBramer: Great to be at the @creekandcave hosting a debate watch party. Btw - #ImWithHer #debatenight #deb	FALSE	0	NA	2016-09-27 01:02:14	FALSE	NA
11	10	RT @JimmyVanBramer: Great to be at the @creekandcave hosting a debate watch party. Btw - #ImWithHer #debatenight #deb	FALSE	0	NA	2016-09-27 01:02:00	FALSE	NA
12	11	Great to be at the @creekandcave hosting a debate watch party. Btw - #ImWithHer #debatenight #debates #LIC https://t.co/dM	FALSE	11	NA	2016-09-27 01:01:35	FALSE	NA
13	12	RT @shyflyinthesky: friends who headbang together stay together Ⓖ #LIC @BrittaniBeasley @jchick_11 @sydneysims4 http	FALSE	0	NA	2016-09-27 00:45:13	FALSE	NA
14	13	friends who headbang together stay together Ⓖ #LIC @BrittaniBeasley @jchick_11 @sydneysims4 https://t.co/ujjTE79oWN	FALSE	7	NA	2016-09-27 00:28:24	FALSE	NA
15	14	RT @JimmyVanBramer: Great joining Borough President @MelindaKatz for a fundraiser at @S_Lewandowski's lovely home! #I	FALSE	0	NA	2016-09-27 00:26:46	FALSE	NA
16	15	View of Long Island City from Manhattan tonight #gantrygram #lic #longislandcity #nyc #queens… https://t.co/SGIJrF6wpK	FALSE	0	NA	2016-09-27 00:06:06	FALSE	NA
17	16	RT @NiallOfficial: Oh some Craic last night at my house with the LIC ! Few drinks, father ted box set! What else would ya want	FALSE	0	NA	2016-09-26 23:50:35	FALSE	NA
18	17	RT @JimmyVanBramer: Great joining Borough President @MelindaKatz for a fundraiser at @S_Lewandowski's lovely home! #I	FALSE	0	NA	2016-09-26 23:49:32	FALSE	NA
19	18	Great joining Borough President @MelindaKatz for a fundraiser at @S_Lewandowski's lovely home! #LIC https://t.co/hwpf1aTg	FALSE	1	NA	2016-09-26 23:45:47	FALSE	NA
20	19	RT @Wildchild9898: Please help me find the loml @LICDallas #licdallas @TexasRaves @DallasEDMFamily #lic https://t.co/2a	FALSE	0	NA	2016-09-26 23:18:48	FALSE	NA
21	20	Please help me find the loml @LICDallas #licdallas @TexasRaves @DallasEDMFamily #lic https://t.co/2aP2Z4TFYG	FALSE	3	NA	2016-09-26 23:16:38	FALSE	NA
22	21	#lic #licdallas was lit ❤Ⓢ	FALSE	1	NA	2016-09-26 22:01:18	FALSE	NA
23	22	#licdallas #lic #lifeincolor https://t.co/aU6f1RXi5p	FALSE	1	NA	2016-09-26 22:00:47	FALSE	NA
24	23	We thought they were going to cancel on us! #licdallas #lic #lifeincolor https://t.co/yV8Tvn1tkS	FALSE	1	NA	2016-09-26 21:59:56	FALSE	NA
25	24	RT @maimiimpala: I felt pretty in pink last night #LIC https://t.co/KBkfw6XqrQ	FALSE	0	NA	2016-09-26 21:25:48	FALSE	NA
26	25	The new office is pretty dope. #YasQueens #LIC #NYC #Madewell https://t.co/ecmUiD0jr9	FALSE	0	NA	2016-09-26 21:16:47	FALSE	NA
27	26	RT*@AlewifeNYC: Our #meatlessmonday special Eggplant Parmigiana #alewifenyc #lic @ Alewife NYC https://t.co/3nzlUqFfdl	FALSE	0	NA	2016-09-26 20:34:39	FALSE	NA
28	27	#LIC	FALSE	0	NA	2016-09-26 20:28:29	FALSE	NA
29	28	I've worked here for three years. Never knew I had roof top access. #lic #longislandcity https://t.co/rfEJL4sH0b	FALSE	0	NA	2016-09-26 20:13:56	FALSE	NA
30	29	RT @LICMarket: Zucchini with lemon, basil, grana padano and a light dusting on breadcrumb #divine #LIC #licmarket https://t.c	FALSE	0	NA	2016-09-26 20:13:55	FALSE	NA
31	30	Our #meatlessmonday special Eggplant Parmigiana #alewifenyc #lic @ Alewife NYC https://t.co/oehX453Hee	FALSE	0	NA	2016-09-26 19:58:38	FALSE	NA
32	31	Boom! #LIC's tallest tower in 2017 will be The Harrison, a 120 unit, 27 story tall luxury condo development!… https://t.co/ABA3	FALSE	0	NA	2016-09-26 19:08:35	TRUE	NA
33	32	AFTERNOON // dog walk #LIC https://t.co/jKL2jf0LgR	FALSE	0	NA	2016-09-26 18:45:05	FALSE	NA
34	33	RT @JeanmarieEvelly: "We knew that we didn't want this completely sanitized neighborhood." Court Square, #LIC has new civ	FALSE	0	NA	2016-09-26 18:32:13	FALSE	NA

Sheet 1 / 1 Default Sum=0 100%

Sample Source Code:

```
from tweepy import Stream
from tweepy import OAuthHandler
from tweepy.streaming import StreamListener
#use consumer key, consumer secret, access token, access secret as follows.
ckey="28befderyHlepe25CA"
csecret="u3xE1SdgQQxzOd97BzgssrygXZ3ezjh4MbirBQO6yzJ9SVQGm"
atoken="468010305-KDYJJvdbeY6pxrX9ssdgteI26YmhBxHjGwuknEmsuGBwJ"
asecret="0R8NEyFPasssPKfcP2sdstbDoQ6OsstejbUXHTpGCxg9pAh1i"
class listener(StreamListener):
    def on_data(self, data):
        print(data)
        return(True)
    def on_error(self, status):
        print status
```

```
auth = OAuthHandler(ckey, csecret)
auth.set_access_token(atoken, asecret)
twitterStream = Stream(auth, listener())
tweet = twitterStream.filter(track=["#SBI"])
```



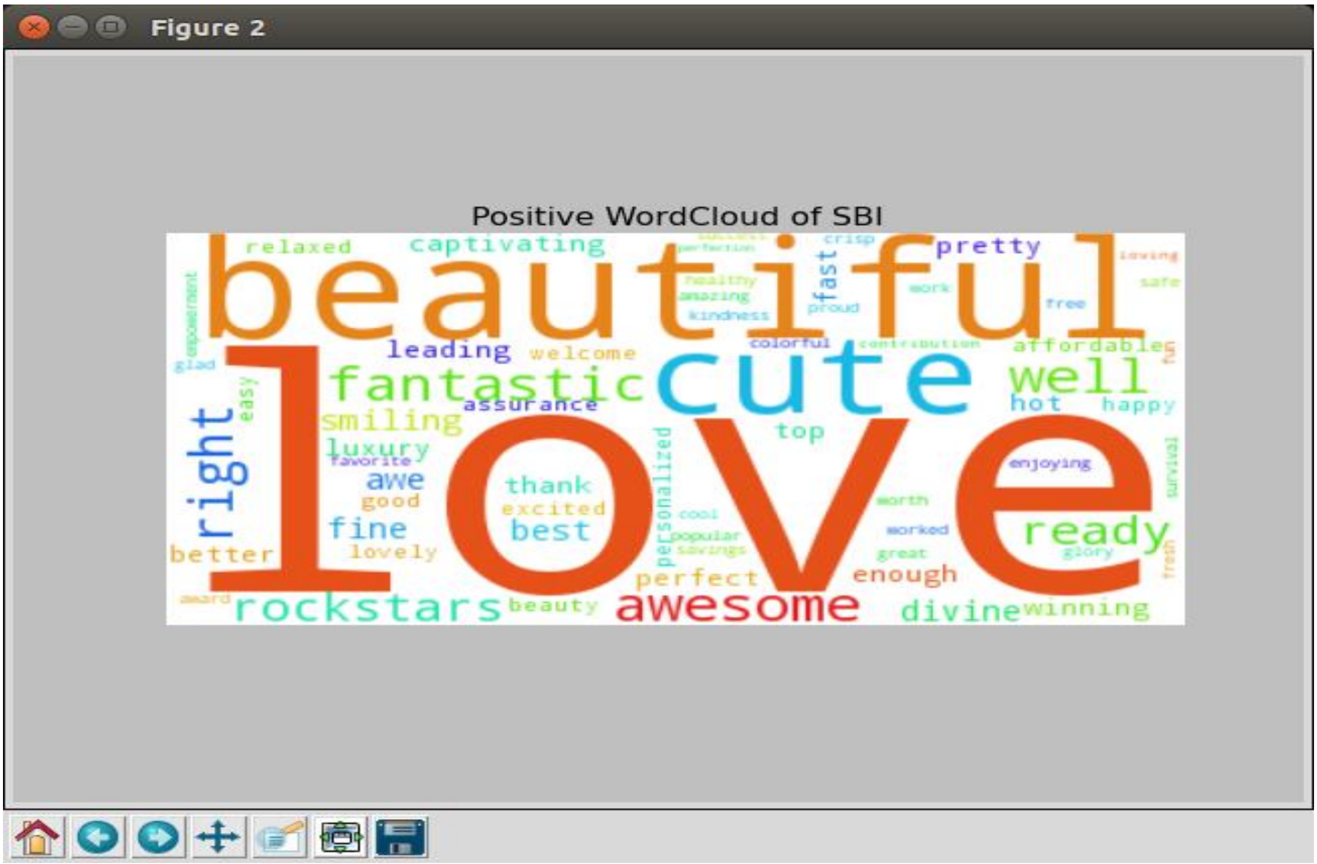


Figure 2
Positive WordCloud of SBI
beautiful
love
cute
fantastic
well
right
awesome
rockstars
ready
captivating
relaxed
pretty
leading
affordable
smiling
assurance
luxury
top
hot
happy
awe
thank
best
fine
good
excited
personalized
better
lovely
perfect
enough
divine
winning
beauty
welcome
fast
enjoying
favorite
easy