Predicting Party Of a Congress Senator

- Kaushik Ghosh, Techno India Saltlake ,161300110046 of 2016-2017
- Abhishek Kumar, Techno India Saltlake ,161300110004 of 2016-2017
- Ashish Kumar, Techno India Saltlake ,161300110023 of 2016-2017
- Priyam Mukherjee,Government College Of Engineering And ceramic Technology,161130110057 of 2016-2017

Roy Chowdhury

Globsyn Knowledge Foundation Kolkata

Document sign date :Mar 20, 2019

Table of Contents

Y Roy Chowdhury

Globsyn Knowledge Foundation ★ Kolkata

Document sign date :Mar 20, 2019

Acknowledgement

I take this opportunity to express my profound gratitude and deep regards to my faculty ,Titash Ghosh for his exemplary guidance, monitoring and constant encouragement throughout the course of this project. The blessing, help and guidance given by him time to time shall carry me a long way in the journey of life on which I am about to embark. I am obliged to my project team members for the valuable information provided by them in their respective fields. I am grateful for their cooperation during the period of my assignment.

Kaushik Ghosh
Abhishek Kumar
Ashish Kumar
Priyam Mukherjee

Y Roy Chowdhury

Globsyn Knowledge Foundation
Kolkata

Document sign date :Mar 20, 2019

Project Objective

This project aims to predict the party of US Senator on the basis of his/her votes given on various occasions in the Senate.

The project uses Big Data comcept using hadoop framework, to implement and apply machine learning to the prpoject.

Y Roy Chowdhury

Globsyn Knowledge Foundation
Kolkata

Document sign date :Mar 20, 2019

Big data

Big Data includes a mix of structured, semi-structured and unstructured real time data originating from variety of sources.

It is the biggest game-changing opportunity for marketing and sales since the Internet went mainstream almost 20 years ago. The data big bang has unleashed torrents of terabytes about everything from customer behaviors to weather patterns to demographic consumer shifts in emerging markets. The trend is growing and in 2018 these numbers became only bigger. The amount of data generated each second will grow 700% by 2020, according to GDC prognosis.

Y Roy Chowdhury

Globsyn Knowledge Foundation ★ Kolkata

Document sign date :Mar 20, 2019

The big data flows can be described with 3 v's: variety, velocity, and volume. Here is how these relate to the banks:

- Variety stands for the plenitude of data types processed, and the banks do have to deal with huge numbers of various types of data. From transaction details and history to credit scores and risk assessment reports — the banks have troves of such data.
- Velocity means the speed at which new data is added to the database. Hitting the threshold of 100 transactions per minute is easy for a respectable bank.
- Volume means the amount of space this data will take to store. Huge financial institutions like the New York Stock Exchange (NYSE) generate terabytes of data daily.

The financial and banking data will be one of the cornerstones of this Big Data flood, and being able to process it means being competitive among the banks and financial institutions.

Y Roy Chowdhury

Globsyn Knowledge Foundation
Kolkata

Document sign date :Mar 20, 2019

The benefits of leveraging big data with marketing campaigns and promotion offers include:

- Increased offer conversion and response rates
- Improved up-sell and cross-sell offers of higher margin products for deeper product penetration
- More marketing occasions to deliver relevant offers
- Higher asset and portfolio values and increased customer share and advocacy
- Identify high value customers for specialized offers and predict response rates

The companies who are successful in turning data into above-market growth will excel at three things:

1. Using analytics to identify valuable business opportunities from the data to drive decisions and improve marketing return on investment (MROI)

2. Turning those insights into well-designed products and offers that delight customers

3. Delivering those products and offers effectively to the marketplace.

Y Roy Chowdhury

Globsyn Knowledge Foundation ★ Kolkata

Document sign date :Mar 20, 2019

This goldmine of data represents a pivot-point moment for marketing and sales leaders. Companies that inject big data and analytics into their operation show

productivity rates and profitability that are 5 percent to 6 percent height than those of their peers. That’s an advantage no company can afford to gnome.

The main purpose of this project is to analyse a large dataset of customers and their history of subscribing to a term deposit of a bank and predict if a target customer with certain given input characteristics or “features” will subscribe to a term deposit or not.

Y Roy Chowdhury

Globsyn Knowledge Foundation ★ Kolkata

Document sign date :Mar 20, 2019

Project Scope

Adopting the Big Data analytics and imbuing it into the existing banking sector workflows is one of the key elements of surviving and prevailing in the rapidly evolving business environment of the digital millennium.

The main Big Data Technologies used for this project are:

1. Apache Hadoop
2. Pig

Other tehnologies like Hive, SparkML, can be further used in the future to analyse the data better.

Y Roy Chowdhury

Globsyn Knowledge Foundation ✶ Kolkata

Document sign date :Mar 20, 2019

1. Apache Hadoop

Apache Hadoop is a java based free software framework that can effectively store large amount of data in a cluster. This framework runs in parallel on a cluster and has an ability to allow us to process data across all nodes. Hadoop Distributed File System (HDFS) is the storage system of Hadoop which splits big data and distribute across many nodes in a cluster. This also replicates data in a cluster thus providing high availability.

2. Pig

Pig is a high level scripting language that is used with Apache Hadoop. Pig enables data workers to write complex data transformations without knowing Java.

Pig’s simple SQL-like scripting language is called Pig Latin, and appeals to developers already familiar with scripting languages and SQL.

Pig is complete, so you can do all required data manipulations in Apache Hadoop with Pig. Through the User Defined Functions(UDF) facility in Pig, Pig can invoke code in many languages like JRuby, Jython and Java. You can also embed Pig scripts in other languages. The result is that you can use Pig as a component to build larger and more complex applications that tackle real business problems.

Pig works with data from many sources, including structured and unstructured data, and store the results into the Hadoop Data File System.

Pig scripts are translated into a series of MapReduce jobs that are run on the Apache Hadoop cluster.

Y Roy Chowdhury

Document sign date :Mar 20, 2019

Requirement Specifications

Hardware Requirements:

1) Intel Core 2 Duo/Quad/hex/Octa or higher end 64 bit processor PC or Laptop (Minimum operating frequency of 2.5GHz)

2) Hard Disk capacity of 1- 4TB.

3) 64-512 GB RAM

4) 10 Gigabit Ethernet or Bonded Gigabit Ethernet

Y Roy Chowdhury

Globsyn Knowledge Foundation Kolkata

Document sign date :Mar 20, 2019

Software Requirements

1. Hadoop and MapReduce

Hadoop is an open source software framework for storing and processing big data across large clusters of commodity hardware. MapReduce is a programming paradigm that allows for massive scalability across hundreds or thousands of servers in a Hadoop cluster.

Popular Hadoop offerings include Edureka, Cloudera, Hortonworks and MapR, among others.

2. Database/File System

Hadoop Distributed File System (HDFS) manages the retrieval and storing of data and metadata required for computation. Other popular file system and database approaches include HBase or Cassandra

Y Roy Chowdhury

Globsyn Knowledge Foundation Kolkata

3. Pig

`Pig` is a high level scripting language that is used with Apache Hadoop. Pig enables data workers to write complex data transformations without knowing Java. Pig's simple SQL-like scripting language is called Pig Latin, and appeals to developers already familiar with scripting languages and SQL.

Pig is complete, so you can do all required data manipulations in Apache Hadoop with Pig. Through the User Defined Functions(UDF) facility in Pig, Pig can invoke code in many languages like JRuby, Jython and Java. You can also embed Pig scripts in other languages. The result is that you can use Pig as a component to build larger and more complex applications that tackle real business problems.

Y Roy Chowdhury

Globsyn Knowledge Foundation ★ Kolkata

Document sign date :Mar 20, 2019

Pig’s simple SQL-like scripting language is called Pig Latin, and appeals to developers already familiar with scripting languages and SQL.

Pig is complete, so you can do all required data manipulations in Apache Hadoop with Pig. Through the User Defined Functions(UDF) facility in Pig, Pig can invoke code in many languages like JRuby, Jython and Java. You can also embed Pig scripts in other languages. The result is that you can use Pig as a component to build larger and more complex applications that tackle real business problems.

Y Roy Chowdhury

Database Design

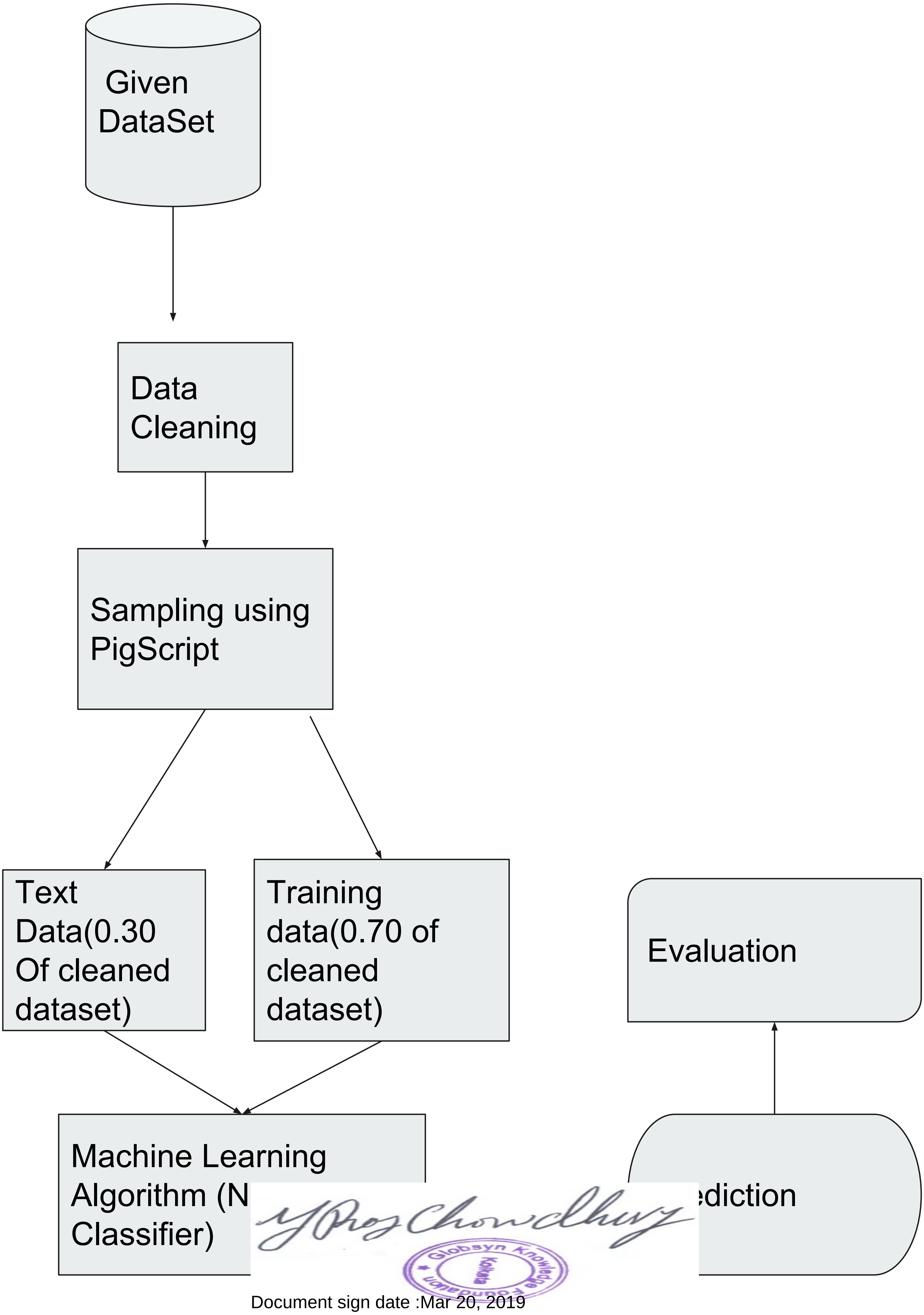
A small dataset was provided to us.

Our dataset consists of following fields

- State of the US Congress Senator
- Name of the US Congress Senator
- Party to which they belong
- The Voting fields consists of 4’columns
    - Vote 1
    - Vote 2
    - Vote 3
    - Vote 4

★ There are 434 entries in the dataset.

Application WorkFlow



Given DataSet

Data Cleaning

Sampling using PigScript

Text Data(0.30 Of cleaned dataset)

Training data(0.70 of cleaned dataset)

Machine Learning Algorithm (N

Classifier)

ediction

Evaluation

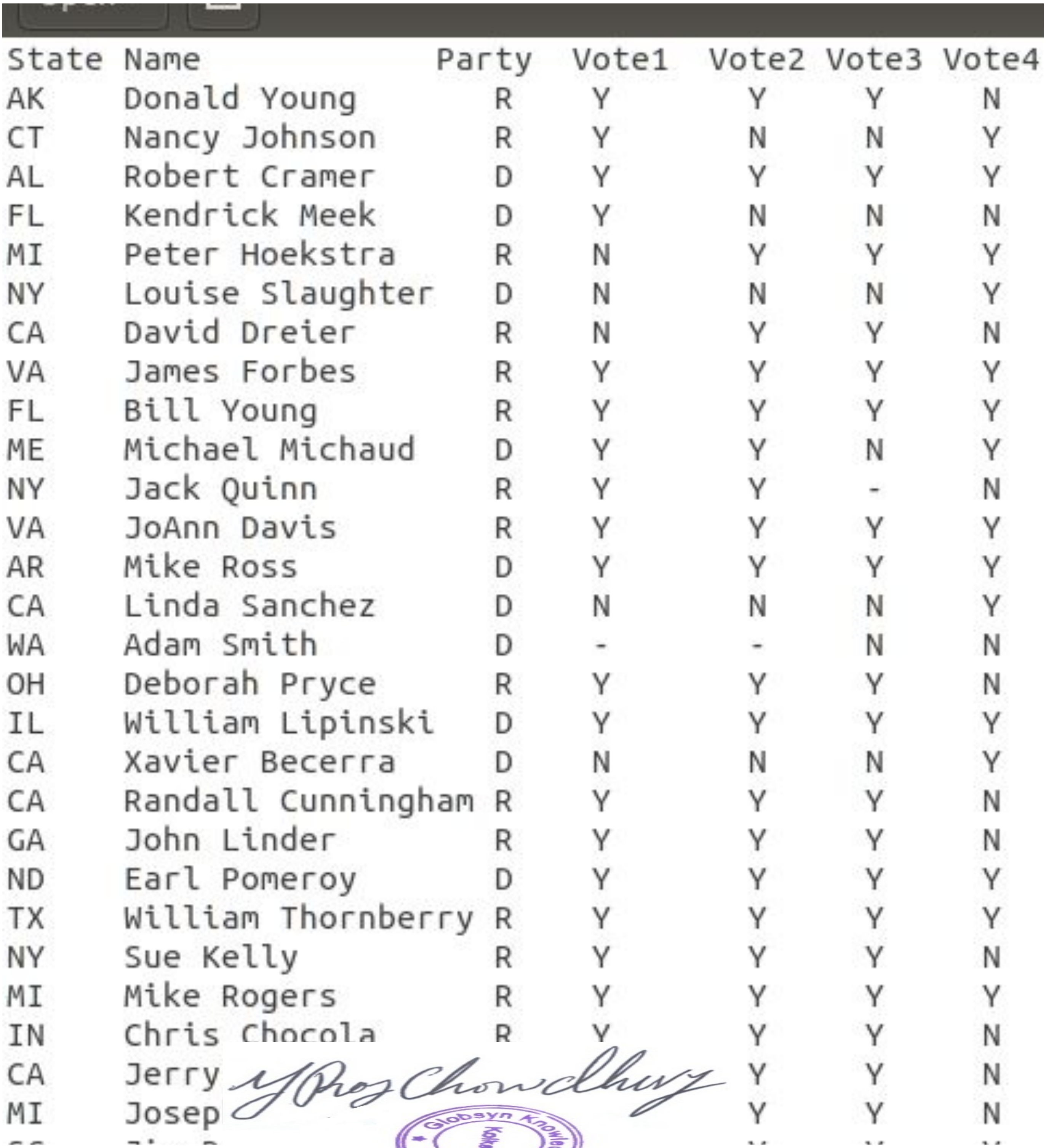
Y Roy Chowdhury

Globsyn Knowledge Foundation ★ Kolkata

Document sign date :Mar 20, 2019

Screenshots

Uncleaned Raw Dataset VOTE.CSV



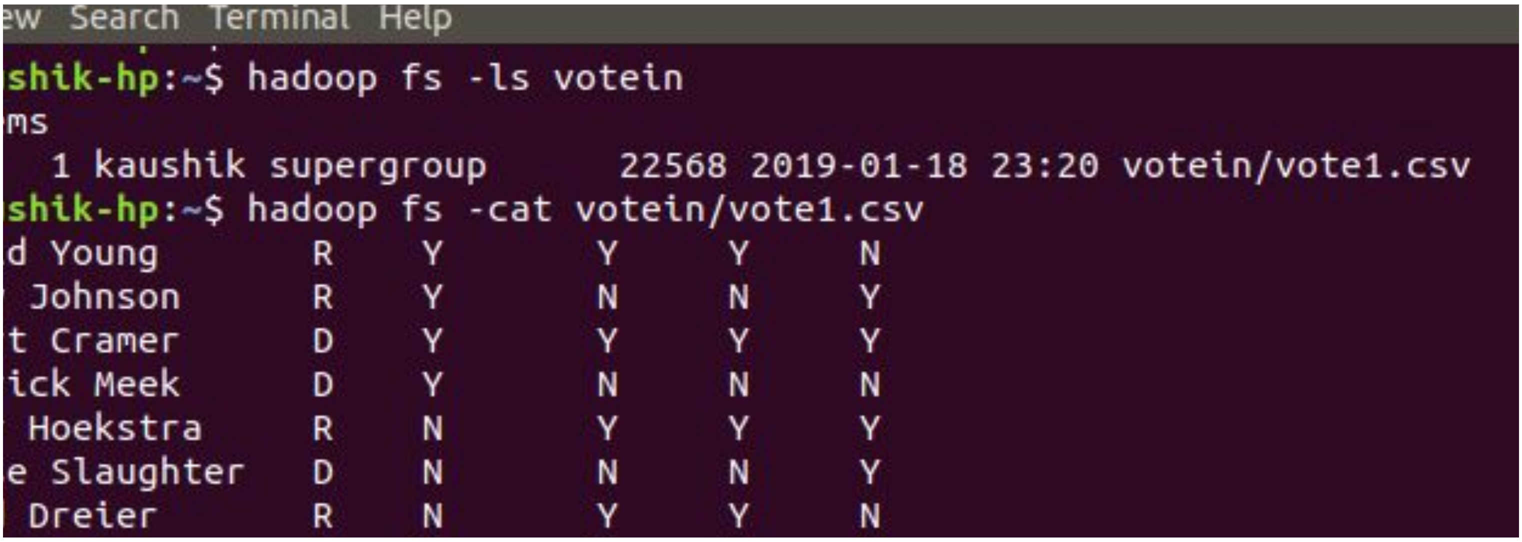
State	Name	Party	Vote1	Vote2	Vote3	Vote4
AK	Donald Young	R	Y	Y	Y	N
CT	Nancy Johnson	R	Y	N	N	Y
AL	Robert Cramer	D	Y	Y	Y	Y
FL	Kendrick Meek	D	Y	N	N	N
MI	Peter Hoekstra	R	N	Y	Y	Y
NY	Louise Slaughter	D	N	N	N	Y
CA	David Dreier	R	N	Y	Y	N
VA	James Forbes	R	Y	Y	Y	Y
FL	Bill Young	R	Y	Y	Y	Y
ME	Michael Michaud	D	Y	Y	N	Y
NY	Jack Quinn	R	Y	Y	-	N
VA	JoAnn Davis	R	Y	Y	Y	Y
AR	Mike Ross	D	Y	Y	Y	Y
CA	Linda Sanchez	D	N	N	N	Y
WA	Adam Smith	D	-	-	N	N
OH	Deborah Pryce	R	Y	Y	Y	N
IL	William Lipinski	D	Y	Y	Y	Y
CA	Xavier Becerra	D	N	N	N	Y
CA	Randall Cunningham	R	Y	Y	Y	N
GA	John Linder	R	Y	Y	Y	N
ND	Earl Pomeroy	D	Y	Y	Y	Y
TX	William Thornberry	R	Y	Y	Y	Y
NY	Sue Kelly	R	Y	Y	Y	N
MI	Mike Rogers	R	Y	Y	Y	Y
IN	Chris Chocola	R	Y	Y	Y	N
CA	Jerry [illegible]	[illegible]	[illegible]	Y	Y	N
MI	Josep[illegible]	[illegible]	[illegible]	Y	Y	N

YRoy Chowdhury

Globsyn Knowledge Foundation Kolkata

Document sign date :Mar 20, 2019

The commands *$hadoop fs -put vote1.csv* votein puts the file into the hadoop file system

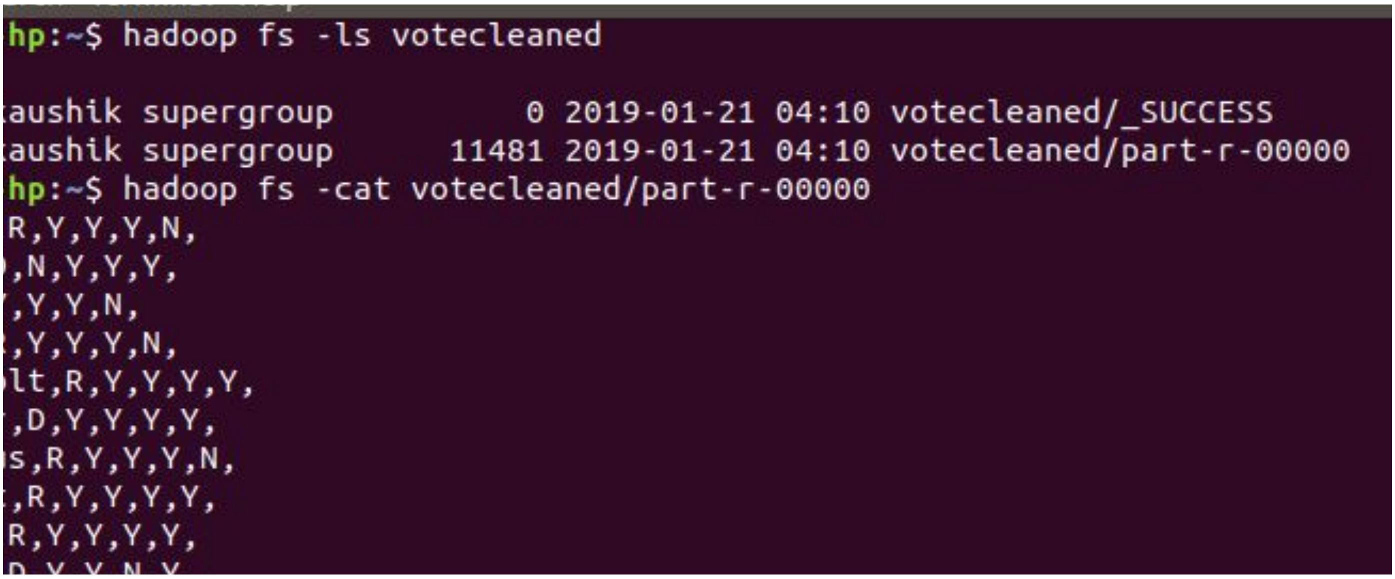


```
shik-hp:~$ hadoop fs -ls votein
ms
  1 kaushik supergroup      22568 2019-01-18 23:20 votein/vote1.csv
shik-hp:~$ hadoop fs -cat votein/vote1.csv
d Young       R    Y        Y      Y      N
 Johnson      R    Y        N      N      Y
t Cramer      D    Y        Y      Y      Y
ick Meek      D    Y        N      N      N
 Hoekstra     R    N        Y      Y      Y
e Slaughter   D    N        N      N      Y
 Dreier       R    N        Y      Y      N
```

- The title bar was removed from the dataset, to have only the predictions and no anomalies.

The Output file part-r-00000 in http://localhost:50070/explorer.html#/user/kaushik/votecleaned

Output from the mapper program, to clean and add commas to the data set,(It had tabs which added extra commas to when the pig script was run)



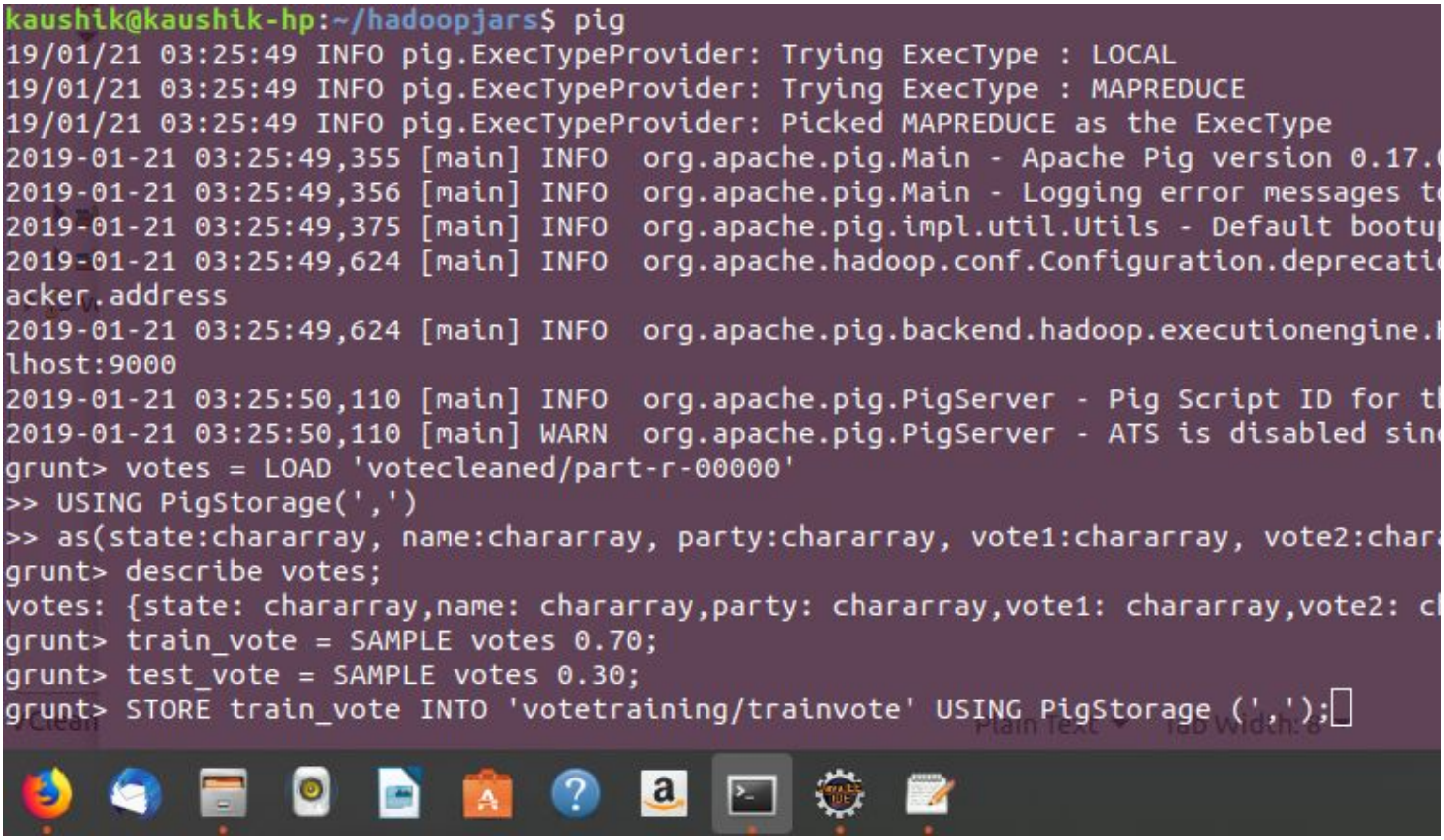
```
hp:~$ hadoop fs -ls votecleaned

aushik supergroup          0 2019-01-21 04:10 votecleaned/_SUCCESS
aushik supergroup      11481 2019-01-21 04:10 votecleaned/part-r-00000
hp:~$ hadoop fs -cat votecleaned/part-r-00000
R,Y,Y,Y,N,
,N,Y,Y,Y,
,Y,Y,N,
,Y,Y,Y,N,
lt,R,Y,Y,Y,Y,
,D,Y,Y,Y,Y,
s,R,Y,Y,Y,N,
,R,Y,Y,Y,Y,
R,Y,Y,Y,Y,
D,Y,Y,N,Y,
```

Y Roy Chowdhury

Globsyn Knowledge Foundation ★ Kolkata

Document sign date :Mar 20, 2019

Splitting the Training and Testing Data Set

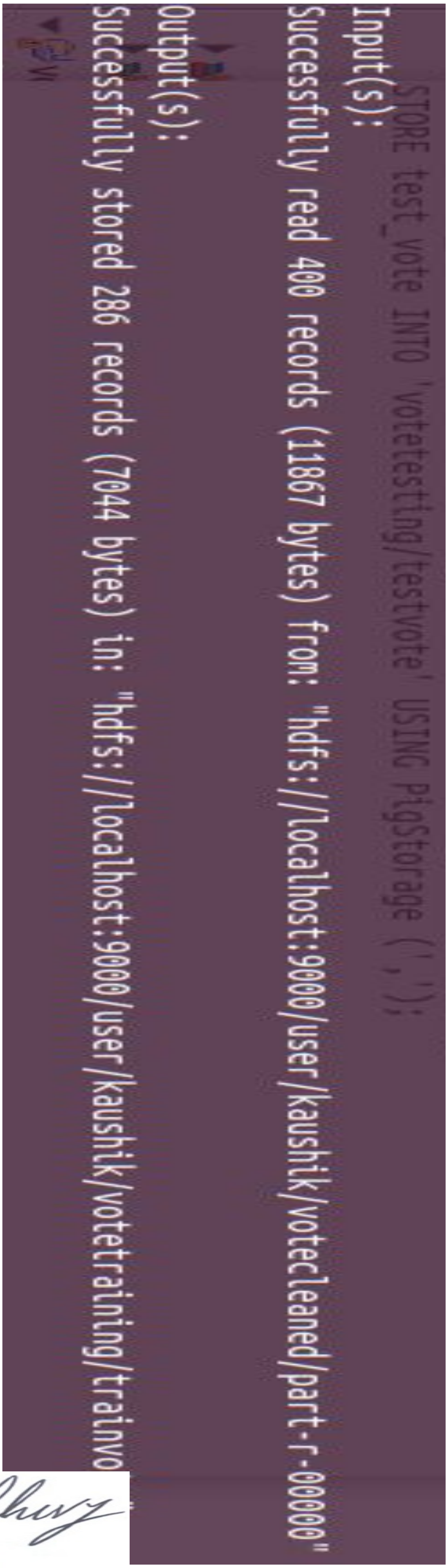


```
kaushik@kaushik-hp:~/hadoopjars$ pig
19/01/21 03:25:49 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
19/01/21 03:25:49 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
19/01/21 03:25:49 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2019-01-21 03:25:49,355 [main] INFO  org.apache.pig.Main - Apache Pig version 0.17.
2019-01-21 03:25:49,356 [main] INFO  org.apache.pig.Main - Logging error messages t
2019-01-21 03:25:49,375 [main] INFO  org.apache.pig.impl.util.Utils - Default bootu
2019-01-21 03:25:49,624 [main] INFO  org.apache.hadoop.conf.Configuration.deprecati
acker.address
2019-01-21 03:25:49,624 [main] INFO  org.apache.pig.backend.hadoop.executionengine.
lhost:9000
2019-01-21 03:25:50,110 [main] INFO  org.apache.pig.PigServer - Pig Script ID for t
2019-01-21 03:25:50,110 [main] WARN  org.apache.pig.PigServer - ATS is disabled sin
grunt> votes = LOAD 'votecleaned/part-r-00000'
>> USING PigStorage(',')
>> as(state:chararray, name:chararray, party:chararray, vote1:chararray, vote2:chara
grunt> describe votes;
votes: {state: chararray,name: chararray,party: chararray,vote1: chararray,vote2: c
grunt> train_vote = SAMPLE votes 0.70;
grunt> test_vote = SAMPLE votes 0.30;
grunt> STORE train_vote INTO 'votetraining/trainvote' USING PigStorage (',');
```

Globsyn Knowledge Foundation Kolkata

Document sign date :Mar 20, 2019

Similarly , the test vote was stored in user/kaushik/testvote

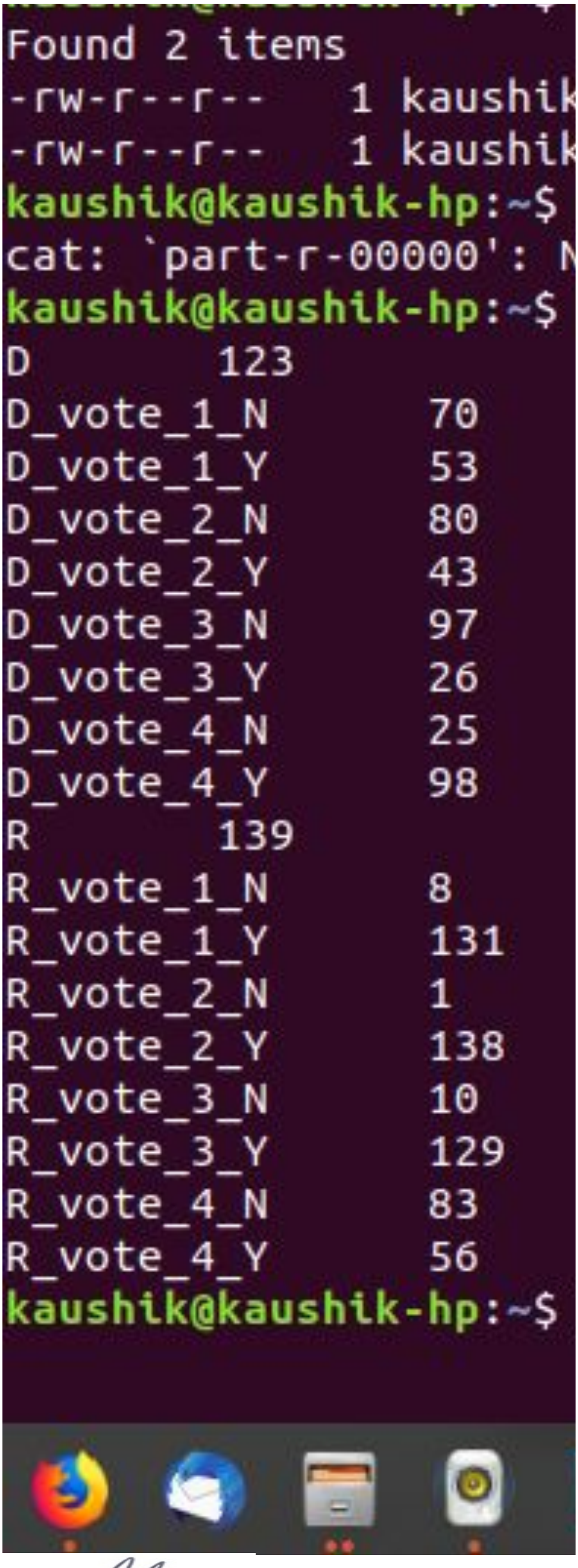


```
STORE test_vote INTO 'votetesting/testvote' USING PigStorage (',');
Input(s):
Successfully read 400 records (11867 bytes) from: "hdfs://localhost:9000/user/kaushik/votecleaned/part-r-00000"
Output(s):
Successfully stored 286 records (7044 bytes) in: "hdfs://localhost:9000/user/kaushik/votetraining/trainvo
```

Globsyn Knowledge Foundation Kolkata

Document sign date :Mar 20, 2019

Output of the mapreduce program to find the counts from the refined training data.

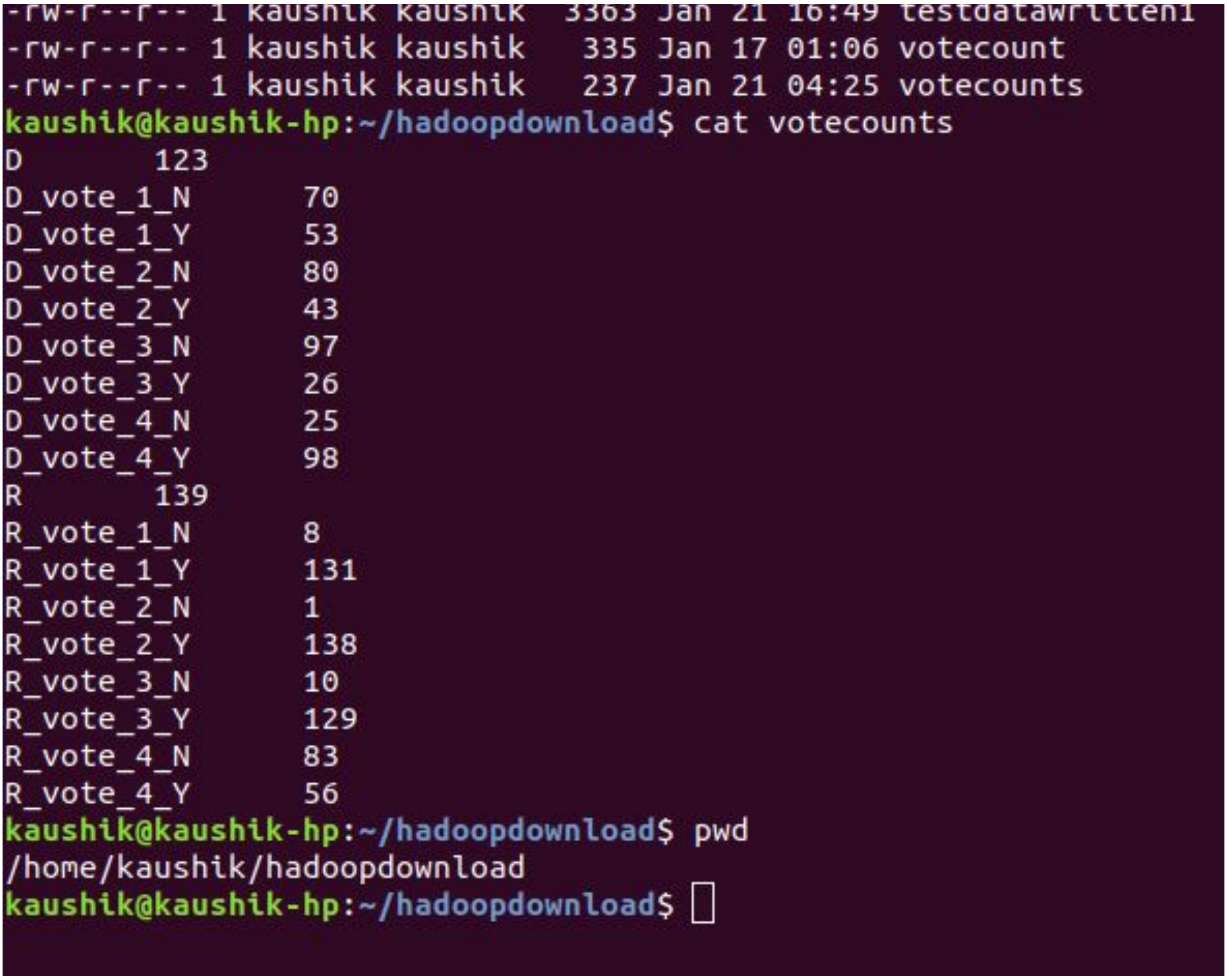


```
Found 2 items
-rw-r--r--   1 kaushik
-rw-r--r--   1 kaushik
kaushik@kaushik-hp:~$
cat: `part-r-00000': N
kaushik@kaushik-hp:~$
D       123
D_vote_1_N      70
D_vote_1_Y      53
D_vote_2_N      80
D_vote_2_Y      43
D_vote_3_N      97
D_vote_3_Y      26
D_vote_4_N      25
D_vote_4_Y      98
R       139
R_vote_1_N      8
R_vote_1_Y      131
R_vote_2_N      1
R_vote_2_Y      138
R_vote_3_N      10
R_vote_3_Y      129
R_vote_4_N      83
R_vote_4_Y      56
kaushik@kaushik-hp:~$
```

Y Roy Chowdhury

Globsyn Knowledge Foundation Kolkata

Document sign date :Mar 20, 2019



```
-rw-r--r-- 1 kaushik kaushik  3363 Jan 21 16:49 testdatawritten1
-rw-r--r-- 1 kaushik kaushik   335 Jan 17 01:06 votecount
-rw-r--r-- 1 kaushik kaushik   237 Jan 21 04:25 votecounts
kaushik@kaushik-hp:~/hadoopdownload$ cat votecounts
D       123
D_vote_1_N      70
D_vote_1_Y      53
D_vote_2_N      80
D_vote_2_Y      43
D_vote_3_N      97
D_vote_3_Y      26
D_vote_4_N      25
D_vote_4_Y      98
R       139
R_vote_1_N      8
R_vote_1_Y      131
R_vote_2_N      1
R_vote_2_Y      138
R_vote_3_N      10
R_vote_3_Y      129
R_vote_4_N      83
R_vote_4_Y      56
kaushik@kaushik-hp:~/hadoopdownload$ pwd
/home/kaushik/hadoopdownload
kaushik@kaushik-hp:~/hadoopdownload$ 
```

Downloading and renaming the file into the local database for the prediction program to read and process named ‘votecounts’

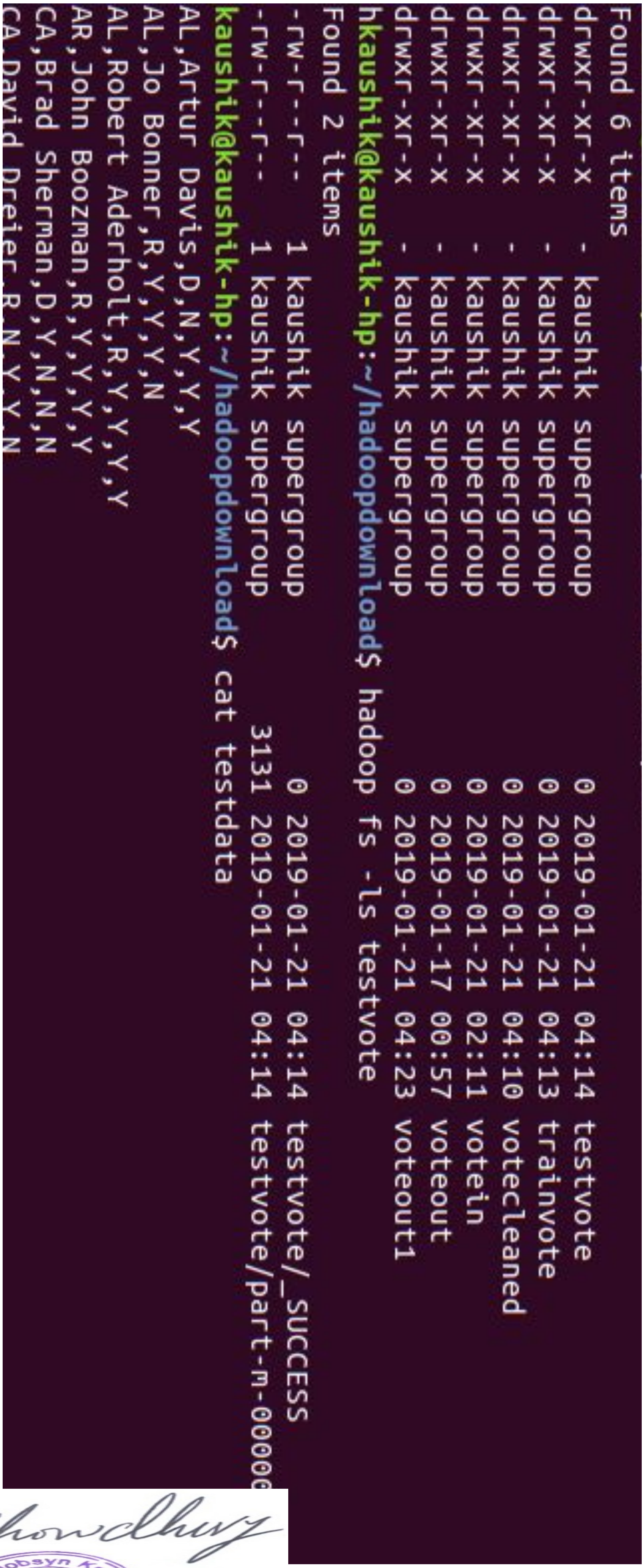
```
$hadoop fs -get
voteout1/part-r-00000
$mv p                          s
```

Y Roy Chowdhury

Globsyn Knowledge Foundation
Kolkata

Document sign date :Mar 20, 2019

Test data downloaded into local system



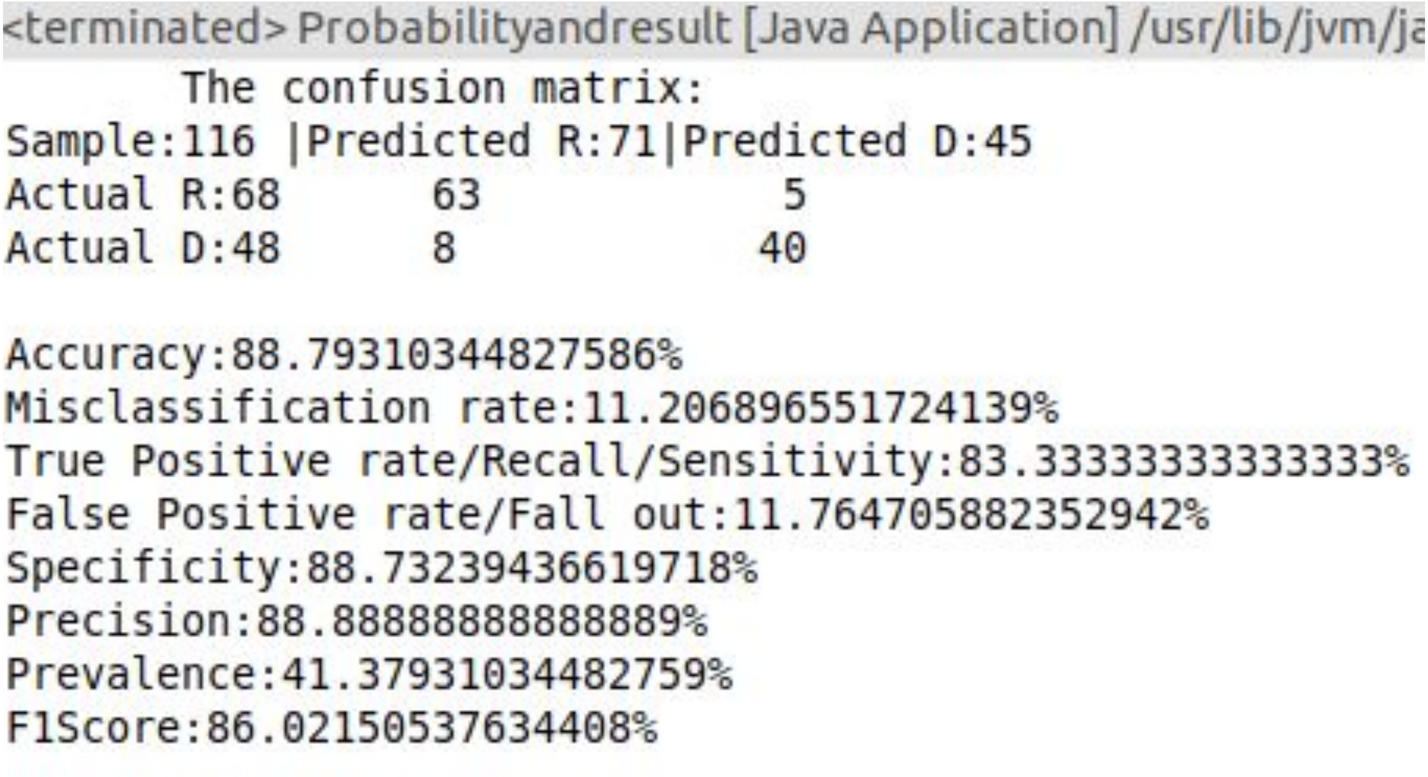
```
Found 6 items
drwxr-xr-x   - kaushik supergroup          0 2019-01-21 04:14 testvote
drwxr-xr-x   - kaushik supergroup          0 2019-01-21 04:13 trainvote
drwxr-xr-x   - kaushik supergroup          0 2019-01-21 04:10 votecleaned
drwxr-xr-x   - kaushik supergroup          0 2019-01-21 02:11 votein
drwxr-xr-x   - kaushik supergroup          0 2019-01-17 00:57 voteout
drwxr-xr-x   - kaushik supergroup          0 2019-01-21 04:23 voteout1
hkaushik@kaushik-hp:~/hadoopdownload$ hadoop fs -ls testvote
Found 2 items
-rw-r--r--   1 kaushik supergroup          0 2019-01-21 04:14 testvote/_SUCCESS
-rw-r--r--   1 kaushik supergroup       3131 2019-01-21 04:14 testvote/part-m-0000
kaushik@kaushik-hp:~/hadoopdownload$ cat testdata
AL,Artur Davis,D,N,Y,Y,Y
AL,Jo Bonner,R,Y,Y,Y,N
AL,Robert Aderholt,R,Y,Y,Y,Y
AR,John Boozman,R,Y,Y,Y,Y
CA,Brad Sherman,D,Y,N,N,N
CA,David Dreier,R,N,Y,N
```

Y Roy Chowdhury

Globsyn Knowledge Foundation Kolkata

Document sign date :Mar 20, 2019

Final output with confusion matrix



```
<terminated> Probabilityandresult [Java Application] /usr/lib/jvm/ja
       The confusion matrix:
Sample:116 |Predicted R:71|Predicted D:45
Actual R:68      63             5
Actual D:48      8             40

Accuracy:88.79310344827586%
Misclassification rate:11.206896551724139%
True Positive rate/Recall/Sensitivity:83.33333333333333%
False Positive rate/Fall out:11.764705882352942%
Specificity:88.73239436619718%
Precision:88.88888888888889%
Prevalence:41.37931034482759%
F1Score:86.02150537634408%
```

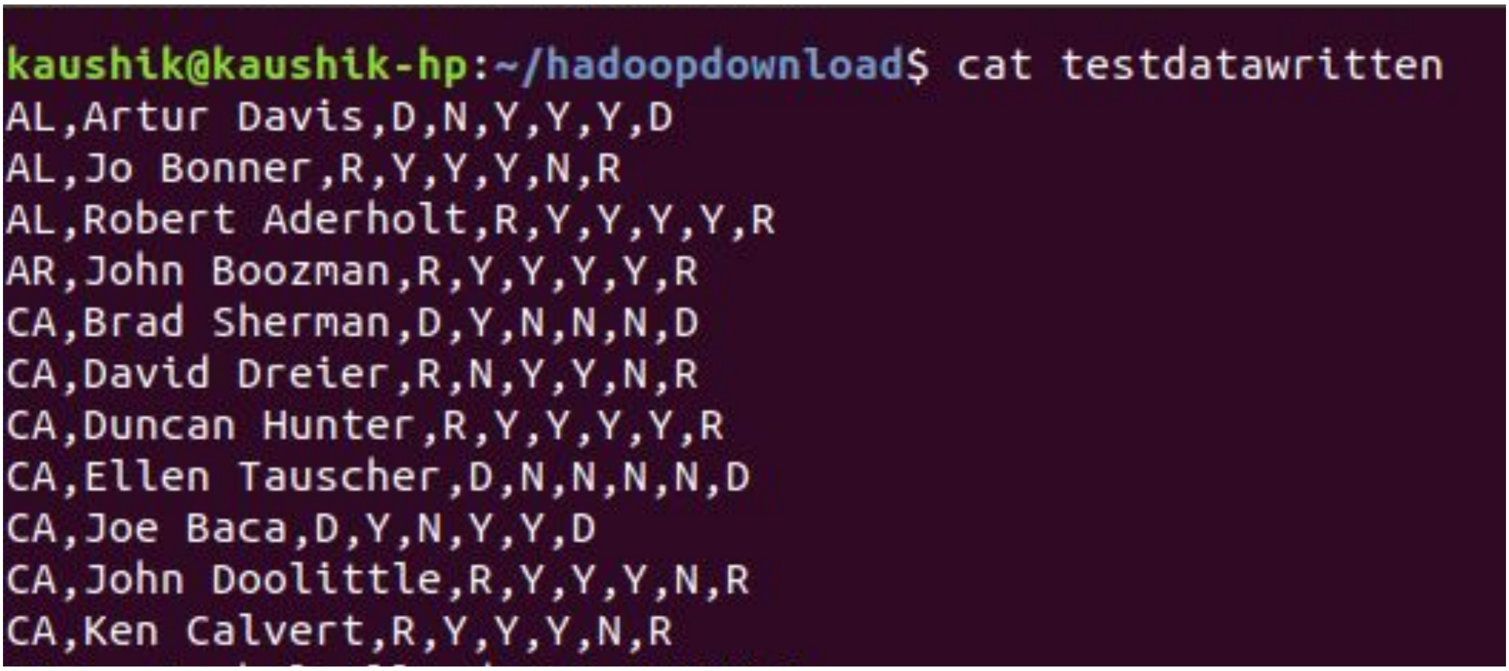
```
       The confusion matrix:
Sample:116 |Predicted R:71|Predicted D:45
Actual R:68       63        5
Actual D:48       8         40

Accuracy:88.79310344827586%
Misclassification rate:11.206896551724139%
True Positive rate/Recall/Sensitivity:83.33333333333333%
False Positive rate/Fall out:11.764705882352942%
Specificity:88.73239436619718%
Precision:88.88888888888889%
Prevalence:41.37931034482759%
F1Score:86.0
```

Roy Chowdhury

Globsyn Knowledge Foundation Kolkata

Document sign date :Mar 20, 2019

Final Output with predicted values in a file printed on local disk “testdatawritten”



```
kaushik@kaushik-hp:~/hadoopdownload$ cat testdatawritten
AL,Artur Davis,D,N,Y,Y,Y,D
AL,Jo Bonner,R,Y,Y,Y,N,R
AL,Robert Aderholt,R,Y,Y,Y,Y,R
AR,John Boozman,R,Y,Y,Y,Y,R
CA,Brad Sherman,D,Y,N,N,N,D
CA,David Dreier,R,N,Y,Y,N,R
CA,Duncan Hunter,R,Y,Y,Y,Y,R
CA,Ellen Tauscher,D,N,N,N,N,D
CA,Joe Baca,D,Y,N,Y,Y,D
CA,John Doolittle,R,Y,Y,Y,N,R
CA,Ken Calvert,R,Y,Y,Y,N,R
```

Predicted data is in the last column

Roy Chowdhury

Globsyn Knowledge Foundation Kolkata

Document sign date :Mar 20, 2019

Future Scope of Improvements

1. In this project the processing and predictions are done offline. In future it could be done online using Apache Spark ML.

2. With further increase in the fields or measures of getting a party, such as facebook activities, twitter activities the prediction can be improved.

3. The increase in sample size will help in better choice of algorithm.

YRoy Chowdhury

Document sign date :Mar 20, 2019

Code

PART 1 - STRUCTURING AND CLEANING

```
/*
Driver program to clean the dataset and add commas for it
to be easily processed by pig scripts.
*/


package vc;

import java.io.IOException;
import java.net.URI;

import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.*;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;

public class VCDriver {
```

Chowdhury

Globsyn Knowledge Foundation Kolkata

Document sign date :Mar 20, 2019

```
    public static void main(String[] args) throws IOException {

        Configuration conf = new Configuration();
        //it contains configuration data taken from xml files
        //hdfs-site.xml,yarn-site.xml
        Job job = Job.getInstance(conf);



        job.setJarByClass(VCDriver.class);
        job.setMapperClass(VCMapper.class); //Reducer not required
        job.setOutputKeyClass(Text.class);
        job.setOutputValueClass(Text.class);

        FileInputFormat.addInputPath(job,new Path("votein"));
        //above line specifies where to get input data(in hdfs);
        //an object of path class represents a file or folder in hdfs
        //why add? (not set?)
        FileOutputFormat.setOutputPath(job,new Path("votecleaned"));
        //"wcout" must not exist before. if exists, we will get error
        try {
                job.waitForCompletion(true);
            } catch (ClassNotFoundException e) {
                // TODO Auto-generated catch block
                e.printStackTrace();
            } catch (InterruptedException e) {
                // TODO Auto-generated catch block
                e.printStackTrace();
            }

    }
}
```

Y Roy Chowdhury

Globsyn Knowledge Foundation
Kolkata

Document sign date :Mar 20, 2019

```
/*
The Mapper program to clean the data
and assign commas in place of spaces
The output of the mapper program is
(<record with commas>,<empty string>)
 */



package vc;


import java.io.IOException;
import java.util.ArrayList;
import java.util.StringTokenizer;

import org.apache.hadoop.io.*;
import
org.apache.hadoop.mapreduce.Mapper;
```

Y Roy Chowdhury

Globsyn Knowledge Foundation Kolkata

```
public class VCMapper extends
Mapper<LongWritable,Text,Text,Text>
{

   @Override
   protected void map(LongWritable
key, Text value,Context context)
         throws IOException,
InterruptedException {
     String record =
value.toString().trim();
     StringTokenizer tokenizer = new
StringTokenizer(record);
     String s = "";
     ArrayList<String> elements = new
ArrayList<String>();
     String x = "";
     int flag = 0;
```

Y Roy Chowdhury

Globsyn Knowledge Foundation Kolkata

```
while(tokenizer.hasMoreTokens()) {
//discarding records that have no
value

         x = tokenizer.nextToken();
         int r = x.compareTo("-");
         if(r!=0)
         {
        elements.add(x);
         }
         else
         {
             flag = 1; //discarding
parameter
             elements.clear();
             break;
         }
     }
```

Y Roy Chowdhury

Document sign date :Mar 20, 2019

```
int n = elements.size();
     String name="";
     if(n==9)//taking into consideration the
variable names' lengths
     {
          name = name + elements.get(1)
+" " +elements.get(2) +" "+
elements.get(3);
          elements.set(1,name);
          elements.set(2, elements.get(4));
          elements.set(3, elements.get(5));
          elements.set(4, elements.get(6));
          elements.set(5, elements.get(7));
          elements.set(6, elements.get(8));
          elements.set(7,"");
          elements.set(8,"");

     }
```

Y Roy Chowdhury

Globsyn Knowledge Foundation Kolkata

```
if(n==8)
      {
            name = name + elements.get(1) + " " +
elements.get(2);
            elements.set(1,name);
            elements.set(2, elements.get(3));
            elements.set(3, elements.get(4));
            elements.set(4, elements.get(5));
            elements.set(5, elements.get(6));
            elements.set(6, elements.get(7));
            elements.set(7, "");
      }
      int d;
      for(int i=0;i<n;i++)//for loop to add the commas
      {
            x = elements.get(i);
            d = x.compareTo("");
            if(d != 0)
            s = s + x +",";
      }

      Text keyout = new Text(s);
      Text valueout = new Text("");
      if(flag==0)//condition to write into the map
output
      context.write(keyout,valueout);

   }}
```

Y Roy Chowdhury

Document sign date :Mar 20, 2019

PIG SCRIPT TO DIVIDE THE DATA SET INTO TRAINING AND TESTING

```
votes1 = LOAD 'vote/vote1.csv'
USING PigStorage('\t')
as(state:chararray, name:chararray,
party:chararray, vote1:chararray, vote2:chararray,
vote3:chararray, vote4:chararray);
train_vote = SAMPLE votes 0.70;
test_vote = SAMPLE votes 0.30;
STORE train_vote INTO 'vote/trainvote' USING
PigStorage ('\t');
STORE test_vote INTO 'vote/testvote1' USING
PigStorage (',');
```

Y Roy Chowdhury

Document sign date :Mar 20, 2019

Part 2. MAPREDUCE PROGRAM - COUNTING

```
/*
The Driver program to count the number of
instances of
'R'
'D'
'R' and 'vote1 = Y'
'R' and 'vote1 = N'
'R' and 'vote2 = Y'
'R' and 'vote2 = N'
'R' and 'vote3 = Y'
'R' and 'vote3 = N'
'R' and 'vote4 = Y'
'D' and 'vote4 = N'
'D' and 'vote1 = Y'
'D' and 'vote1 = N'
'D' and 'vote2 = Y'
'D' and 'vote2 = N'
'D' and 'vote3 = Y'
'D' and 'vote3 = N'
'D' and 'vote4 = Y'
'D' and 'v
*/
;
```

```
package vot;

import java.io.IOException;
import java.net.URI;

import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.*;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat
```

Y Roy Chowdhury

```
public class VDriver  {

    public static void main(String[] args)
throws IOException {
      Configuration conf = new
Configuration();
      //it contains configuration data taken
from xml files
      //hdfs-site.xml,yarn-site.xml
      Job job = Job.getInstance(conf);



      job.setJarByClass(VDriver.class);
      job.setMapperClass(VMapper.class);
      job.setReducerClass(VReducer.class);
      job.setOutputKeyClass(Text.class);

job.setOutputValueClass(IntWritable.class
);
```

Y Roy Chowdhury

Globsyn Knowledge Foundation ★ Kolkata

```
FileInputFormat.addInputPath(job,new
Path("trainvote"));
    //above line specifies where to get input
data(in hdfs);
    //an object of path class represents a file
or folder in hdfs
    FileOutputFormat.setOutputPath(job,new
Path("voteout1"));
    //"voteout1" must not exist before. if
exists, we will get error
    try {
            job.waitForCompletion(true);
        } catch (ClassNotFoundException e) {
            // TODO Auto-generated catch
block
            e.printStackTrace();
        } catch (InterruptedException e) {
            // TODO Auto-generated catch
block
            e.printStackTrace();
        }


   }

}
```

Y Roy Chowdhury

Globsyn Knowledge Foundation Kolkata

Document sign date :Mar 20, 2019

```
/*Mapper program to calculate
instances of
'R'
'D'
'R' and 'vote1 = Y'
'R' and 'vote1 = N'
'R' and 'vote2 = Y'
'R' and 'vote2 = N'
'R' and 'vote3 = Y'
'R' and 'vote3 = N'
'R' and 'vote4 = Y'
'D' and 'vote4 = N'
'D' and 'vote1 = Y'
'D' and 'vote1 = N'
'D' and 'vote2 = Y'
'D' and 'vote2 = N'
'D' and 'vote3 = Y'
'D' and 'vote3 = N'
'D' and 'vote4 = Y'
'D' and 'vote4 = N'
by sending the key value pair as
(<party>_vote_<number>_<Y/N>,1)
 */
```

Y Roy Chowdhury

Globsyn Knowledge Foundation Kolkata

Document sign date :Mar 20, 2019

```
package vot;

import java.io.IOException;
import java.util.ArrayList;
import java.util.StringTokenizer;

import org.apache.hadoop.io.*;
import
org.apache.hadoop.mapreduce.Mapp
er;

public class VMapper extends
Mapper<LongWritable,Text,Text,IntWr
itable> {
```

Y Roy Chowdhury

Globsyn Knowledge Foundation Kolkata

```
    @Override
    protected void
map(LongWritable key, Text
value,Context context)
          throws IOException,
InterruptedException {
      String record =
value.toString().trim();//value = one
record in the file
      String words[] =
record.split(","); //the array that
stores the value splitted from
```

Y Roy Chowdhury

```
 int c = 8;
     String word = "";
     String y="";

     for(int i=words.length-1;i>=0;i--) //for loop to get the party content, iteration from back
     {

          word= words[i];

          if(c==4)
          {
          Text keyout = new Text(word);
          IntWritable valueout = new IntWritable(1);
          context.write(keyout,valueout);
          break;
          }

          c--;
     }
```

Y Roy Chowdhury

Globsyn Knowledge Foundation Kolkata

Document sign date :Mar 20, 2019

```
c = 8;
     String x = "";
     for(int i=words.length-1;c>=5;i--)
//string generation of the combinations
     {
          String s =  words[i];
          if(s.equals("Y")) //writing the
key-value pair as :(string generated,1)
          {
               x = x + (c-4);
               Text keyout = new
Text(word+"_vote_"+x+"_Y");
               IntWritable valueout = new
IntWritable(1);

context.write(keyout,valueout);
               x = "";
          }
```

Y Roy Chowdhury

Globsyn Knowledge Foundation
Kolkata

Document sign date :Mar 20, 2019

```
else if(s.equals("N"))//writing the
key-value pair as :(string generated,1)
          {
               x = x + (c-4);
               Text keyout = new
Text(word+"_vote_"+x+"_N");
               IntWritable valueout = new
IntWritable(1);

context.write(keyout,valueout);
               x = "";
          }
```

Chowdhury

Globsyn Knowledge Foundation ★ Kolkata

```
 else//handling the other values if
any (already handled in cleaning)
        {
            x = x + (c-4);
            Text keyout = new
Text(word+"_vote_"+x+"NoVote");
            IntWritable valueout =
new IntWritable(1);

context.write(keyout,valueout);
            x = "";
        }

        c--;
    }

    }
}
```

Chowdhury

Globsyn Knowledge Foundation
Kolkata

Document sign date :Mar 20, 2019

```
/*
 Reducer program to count the number
of 1's per key
 accepting the key value pair from
mapper as
 (<party>_vote_<number>_<Y/N>,1)
 returning
(<party>_vote_<number>_<Y/N>,K)
where K = number of instances
 */




package vot;

import java.io.IOException;

import org.apache.hadoop.io.*;
import
org.apache.hadoop.mapreduce.Reduce
r;
```

Y Roy Chowdhury

Globsyn Knowledge Foundation Kolkata

```
public class VReducer extends
Reducer<Text,IntWritable,Text,IntWritable>
{

   @Override
   protected void reduce(Text key,
Iterable<IntWritable> values,Context
context)
         throws IOException,
InterruptedException {
      // TODO Auto-generated method stub
      int sum = 0;
      for(IntWritable v : values) //values of
the format '1'
      {
         sum = sum + v.get();
      }
      context.write(key, new
IntWritable(sum)); //key of the format
(<party>_vote_<number>_<Y/N>)
   }

  }
```

Y Roy Chowdhury

Globsyn Knowledge Foundation
Kolkata

Java program to calculate the probabilities and predictions

```
/*
The java program to calculate the
probabilities
of the instances for example
of a record being 'R' given its
combination of votes

 This program also predicts the value
by taking the testing set
 as input and prints the accuracy
 */



package voprob;
import java.io.*;
import java.util.*;
public class Probabilityandresult
{
    static Hashtable<String,Double>
h1;
```

Y Roy Chowdhury

```
public static void main(String[] args)throws Exception
{
File file = new File("/home/kaushik/hadoopdownload/votecounts");
Hashtable<String,Integer> h = new Hashtable<String,Integer>();
BufferedReader br = new BufferedReader(new FileReader(file));

String st; String vr; String[] ar = new String[2];
while ((st = br.readLine()) != null)
{
    ar = st.split("\t");
    h.put(ar[0],Integer.parseInt(ar[1]));
}

//System.out.println("values are"+h);
int ttl = h.get("R")+h.get("D");
```

Y Roy Chowdhury

Globsyn Knowledge Foundation Kolkata

```
//System.out.println(ttl);
double p_v1_y_R =
(double)h.get("R_vote_1_Y")/h.get("R");
double p_v1_n_R =
(double)h.get("R_vote_1_N")/h.get("R");
double p_v2_y_R =
(double)h.get("R_vote_2_Y")/h.get("R");
double p_v2_n_R =
(double)h.get("R_vote_2_N")/h.get("R");
double p_v3_y_R =
(double)h.get("R_vote_3_Y")/h.get("R");
double p_v3_n_R =
(double)h.get("R_vote_3_N")/h.get("R");
double p_v4_y_R =
(double)h.get("R_vote_4_Y")/h.get("R");
double p_v4_n_R =
(double)h.get("R_vote_4_N")/h.get("R");
double p_v1_y_D =
(double)h.get("D_vote_1_Y")/h.get("D");
double p_v1_n_D =
(double)h.get("D_vote_1_N")/h.get("D");
double p_v2_y_D =
(double)h.get("D_vote_2_Y")/h.get("D");
double p_v2_n_D =
(double)h.get("D_vote_2_N")/h.get("D");
double p_v3_y_D =
(double)h.get("D_vote_3_Y")/h.get("D");
double p_v3_n_D =
(double)h.get("D_vote_3_N")/h.get("D");
double p_v4_y_D =
(double)h.get("D_vote_4_Y")/h.get("D");
double p_v4_n_D =
(double)h.get("D_vote_4_N")/h.get("D");
```

```
double p_R = (double)h.get("R")/ttl;
double p_D = (double)h.get("D")/ttl;
h1 = new Hashtable<String,Double>();
double p_R_nnnn =
p_v1_n_R*p_v2_n_R*p_v3_n_R*p_v4_n_R*p_R
;
double p_R_nnny =
p_v1_n_R*p_v2_n_R*p_v3_n_R*p_v4_y_R*p_R;
double p_R_nnyn =
p_v1_n_R*p_v2_n_R*p_v3_y_R*p_v4_n_R*p_R;
double p_R_nnyy =
p_v1_n_R*p_v2_n_R*p_v3_y_R*p_v4_y_R*p_R;
double p_R_nynn =
p_v1_n_R*p_v2_y_R*p_v3_n_R*p_v4_n_R*p_R;
double p_R_nyny =
p_v1_n_R*p_v2_y_R*p_v3_n_R*p_v4_y_R*p_R;
double p_R_nyyn =
p_v1_n_R*p_v2_y_R*p_v3_y_R*p_v4_n_R*p_R;
double p_R_nyyy =
p_v1_n_R*p_v2_y_R*p_v3_y_R*p_v4_y_R*p_R;
double p_R_ynnn =
p_v1_y_R*p_v2_n_R*p_v3_n_R*p_v4_n_R*p_R;
```

Y Roy Chowdhury

Document sign date :Mar 20, 2019

```
double p_R_ynny =
p_v1_y_R*p_v2_n_R*p_v3_n_R*p_v4_y_R*p_R;
double p_R_ynyn =
p_v1_y_R*p_v2_n_R*p_v3_y_R*p_v4_n_R*p_R;
double p_R_ynyy =
p_v1_y_R*p_v2_n_R*p_v3_y_R*p_v4_y_R*p_R;
double p_R_yynn =
p_v1_y_R*p_v2_y_R*p_v3_n_R*p_v4_n_R*p_R;
double p_R_yyny =
p_v1_y_R*p_v2_y_R*p_v3_n_R*p_v4_y_R*p_R;
double p_R_yyyn =
p_v1_y_R*p_v2_y_R*p_v3_y_R*p_v4_n_R*p_R;
double p_R_yyyy =
p_v1_y_R*p_v2_y_R*p_v3_y_R*p_v4_y_R*p_R;
//democratic
double p_D_nnnn =
p_v1_n_D*p_v2_n_D*p_v3_n_D*p_v4_n_D*p_D;
double p_D_nnny =
p_v1_n_D*p_v2_n_D*p_v3_n_D*p_v4_y_D*p_D;
double p_D_nnyn =
p_v1_n_D*p_v2_n_D*p_v3_y_D*p_v4_n_D*p_D;
double p_D_nnyy =
p_v1_n_D*p_v2_n_D*p_v3_y_D*p_v4_y_D*p_D;
double p_D_nynn =
p_v1_n_D*p_v2_y_D*p_v3_n_D*p_v4_n_D*p_D;
```

Y Roy Chowdhury

Globsyn Knowledge Foundation Kolkata

Document sign date :Mar 20, 2019

```
double p_D_nyny =
p_v1_n_D*p_v2_y_D*p_v3_n_D*p_v4_y_D*p_D;
double p_D_nyyn =
p_v1_n_D*p_v2_y_D*p_v3_y_D*p_v4_n_D*p_D;
double p_D_nyyy =
p_v1_n_D*p_v2_y_D*p_v3_y_D*p_v4_y_D*p_D;
double p_D_ynnn =
p_v1_y_D*p_v2_n_D*p_v3_n_D*p_v4_n_D*p_D;
double p_D_ynny =
p_v1_y_D*p_v2_n_D*p_v3_n_D*p_v4_y_D*p_D;
double p_D_ynyn =
p_v1_y_D*p_v2_n_D*p_v3_y_D*p_v4_n_D*p_D;
double p_D_ynyy =
p_v1_y_D*p_v2_n_D*p_v3_y_D*p_v4_y_D*p_D;
double p_D_yynn =
p_v1_y_D*p_v2_y_D*p_v3_n_D*p_v4_n_D*p_D;
double p_D_yyny =
p_v1_y_D*p_v2_y_D*p_v3_n_D*p_v4_y_D*p_D;
double p_D_yyyn =
p_v1_y_D*p_v2_y_D*p_v3_y_D*p_v4_n_D*p_D;
double p_D_yyyy =
p_v1_y_D*p_v2_y_D*p_v3_y_D*p_v4_y_D*p_D;
```

Y Roy Chowdhury

Globsyn Knowledge Foundation * Kolkata

```
//assigning to hashtable
h1.put("p_D_nnnn", p_D_nnnn);
h1.put("p_D_nnny", p_D_nnny);
h1.put("p_D_nnyn", p_D_nnyn);
h1.put("p_D_nnyy", p_D_nnyy);
h1.put("p_D_nynn", p_D_nynn);
h1.put("p_D_nyny", p_D_nyny);
h1.put("p_D_nyyn", p_D_nyyn);
h1.put("p_D_nyyy", p_D_nyyy);
h1.put("p_D_ynnn", p_D_ynnn);
h1.put("p_D_ynny", p_D_ynny);
h1.put("p_D_ynyn", p_D_ynyn);
h1.put("p_D_ynyy", p_D_ynyy);
h1.put("p_D_yynn", p_D_yynn);
h1.put("p_D_yyny", p_D_yyny);
h1.put("p_D_yyyn", p_D_yyyn);
h1.put("p_D_yyyy", p_D_yyyy);
```

```
h1.put("p_R_nnnn", p_R_nnnn);
h1.put("p_R_nnny", p_R_nnny);
h1.put("p_R_nnyn", p_R_nnyn);
h1.put("p_R_nnyy", p_R_nnyy);
h1.put("p_R_nynn", p_R_nynn);
h1.put("p_R_nyny", p_R_nyny);
h1.put("p_R_nyyn", p_R_nyyn);
h1.put("p_R_nyyy", p_R_nyyy);
h1.put("p_R_ynnn", p_R_ynnn);
h1.put("p_R_ynny", p_R_ynny);
h1.put("p_R_ynyn", p_R_ynyn);
h1.put("p_R_ynyy", p_R_ynyy);
h1.put("p_R_yynn", p_R_yynn);
h1.put("p_R_yyny", p_R_yyny);
h1.put("p_R_yyyn", p_R_yyyn);
h1.put("p_R_yyyy", p_R_yyyy);
```

```
//System.out.println(h1);

//prediction
File file1 = new File("/home/kaushik/hadoopdownload/testdata");
int are=0,ad=0;int pd=0,pr=0;
BufferedReader br1 = new BufferedReader(new FileReader(file1));
BufferedWriter writer = new BufferedWriter(new FileWriter("/home/kaushik/hadoopdownload/testdatawritten"))
;
```

Roy Chowdhury

```
String arry[] = new String[7];
String ss="";String x="",y="";int
c=0;int d=0;int
rr=0,dd=0,rd=0,dr=0;
String sl;
while ((sl = br1.readLine()) !=
null)
{
    arry = sl.split(",");
    //System.out.println(arry[6]);
    ss =
(arry[3]+arry[4]+arry[5]+arry[6]).t
oLowerCase();
    x = "p_D_"+ss;
    y = "p_R_"+ss;
```

Y Roy Chowdhury

Globsyn Knowledge Foundation
* Kolkata

```
//System.out.println(x+y);
    double m = h1.get(x);
    double n = h1.get(y);
    if(m>n)
    {
      if(arry[2].equals("D"))
      {
           writer.write(sl+",D");
           writer.append("\n");
           dd++;
           d++;
           ad++;
           pd++;
      }
      else
      {
           are++;
           writer.write(sl+",D");
           writer.append("\n");
           rd++;
           pd++;
      }
    }
```

Y Roy Chowdhury

Document sign date :Mar 20, 2019

```
  else
    {
      if(arry[2].equals("R"))
      {
           writer.write(sl+",R");
           writer.append("\n");
           rr++;
           d++;
           are++;
           pr++;
      }
      else
      {
           writer.write(sl+",R");
           writer.append("\n");
           dr++;
           pr++;
           ad++;
      }
    }
    c++;ss = "";x = "p_D_";y="p_R_";
}
```

```
writer.close();
//Printing the confusion matrix
System.out.println("      The confusion
matrix:");
System.out.print("Sample:" +c+ "
|Predicted R:"/*Predicted
No*/+pr+"|Predicted D:"/*Predicted
Yes*/+pd+"\n");
System.out.print("Actual
R:"+are/*Actual No*/+"      "+rr/*True
Negative*/+"          "+rd/*False
Negative*/+"\n");
System.out.print("Actual
D:"+ad/*Actual Yes*/+"      "+dr/*False
Positive*/+"             "+dd/*True
Positive*/+"\n");
double accuracy = (double)((rr +
dd)*100)/*(TN+TP)*100/total*//c;
```

Globsyn Knowledge Foundation Kolkata

```
double
misclassificationrate/*(FN+FP)*100
/total*/= (double)((rd+dr)*100)/c;
double
truePositiverate/*(TP)*100/Actual
Yes*/= (double)(dd*100)/ad;
double
falsePositiverate/*(FP)*100/Actual
No*/= (double)(dr*100)/are;
double specificity/*(TN)*100/true
negative + false positive*/=
(double)((rr)*100)/(rr+dr);
double
precision/*(TP)*100/Predicted
Yes*/= (double)((dd)*100)/pd;
double prevalence/*(Actual
Yes)*100/Total*/=
(double)((ad)*100)/c;
```

Y Roy Chowdhury

Globsyn Knowledge Foundation Kolkata

Document sign date :Mar 20, 2019

```
double fscore =
(double)2*(precision*truePositiverate)/(pre
cision+truePositiverate);
System.out.println();
System.out.println("Accuracy:"+accuracy+
"%");
System.out.println("Misclassification
rate:"+misclassificationrate+"%");
System.out.println("True Positive
rate/Recall/Sensitivity:"+truePositiverate+"
%");
System.out.println("False Positive
rate/Fall out:"+falsePositiverate+"%");
System.out.println("Specificity:"+specificity
+"%");
System.out.println("Precision:"+precision+
"%");
System.out.println("Prevalence:"+prevalen
ce+"%");
System.out.println("F1Score:"+fscore+"%"
);
}
}
```

Y Roy Chowdhury

OUTPUT

```
The confusion matrix:
Sample:116 |Predicted R:71|Predicted D:45
Actual R:68     63              5
Actual D:48     8          40

Accuracy:88.79310344827586%
Misclassification rate:11.206896551724139%
True Positive
rate/Recall/Sensitivity:83.33333333333333%
False Positive rate/Fall
out:11.764705882352942%
Specificity:88.73239436619718%
Precision:88.88888888888889%
Prevalence:41.37931034482759%
F1Score:86.02150537634408%
```

Y Roy Chowdhury

Globsyn Knowledge Foundation Kolkata

Document sign date :Mar 20, 2019

Features of the Project

➢ Prediction Analysis

- Feature Fields vote1,vote2,vote3,vote4 used to help classify the record into ‘R’ or ‘D’.
- Vote1, vote2 , vote3, vote 4 are independent of each other
- Classification based algorithm needed.
- **<u>Naive Bayes Algorithm</u>** suits the above needs.
- Bayes formula :

P(A/B)=(P(B/A)*P(A))/P(B)

Roy Chowdhury

➢ Prediction of ‘party’ using Naive Bayes Algorithm

- Formulae used sample:
- We take a single instance from our dataset,where R=Republican and D=Democratic
- Here , we take the voting pattern arbitrarily as vote1 = Y,vote2 = Y, vote3 = Y, vote4 = N.

  ○ **Probability of ‘R’ for one arbitrary record**

$$\frac{P(Party = R)}{P(vote1 =' Y', vote2 =' Y', vote3 =' Y', vote4 =' N')} =$$

$$P(vote1 =' Y'|R).P(vote2 =' Y'|R).P(vote3 =' Y'|R).P(vote4 =' N'|R).P(R)$$

Y Roy Chowdhury

Globsyn Knowledge Foundation Kolkata

Document sign date :Mar 20, 2019

*where,

$$P(vote1 =' Y'|R) = \frac{count(Party = R and vote1 =' Y'}{No of appear ances of R}$$

$$P(vote2 =' Y'|R) = \frac{count(Party = R and vote2 =' Y'}{No of appear ances of R}$$

$$P(vote3 =' Y'|R) = \frac{count(Party = R and vote3 =' Y'}{No of appear ances of R}$$

$$P(vote4 =' N'|R) = \frac{count(Party = R and vote4 =' N'}{No of appear ances of R}$$

Y Roy Chowdhury

Globsyn Knowledge Foundation ★ Kolkata

Document sign date :Mar 20, 2019

- Probability of ‘D’ for one arbitrary record

$$\frac{P(Party = D)}{P(vote1 =' Y', vote2 =' Y', vote3 =' Y', vote4 =' N')}$$

*Where,

$$P(vote1 =' Y' | D) = \frac{count(Party = D \, and \, vote1 =' Y'}{No \, of \, appearances \, of \, D}$$

$$P(vote2 =' Y' | D) = \frac{count(Party = D \, and \, vote2 =' Y'}{No \, of \, appearances \, of \, D}$$

Y Roy Chowdhury

Document sign date :Mar 20, 2019

$$P(vote3 =' Y'|D) = \frac{count(Party = D \, and \, vote3 =' Y')}{No \, of \, appearances \, of \, D}$$

$$P(vote4 =' N'|D) = \frac{count(Party = D \, and \, vote4 =' N')}{No \, of \, appearances \, of \, D}$$

There will be a total of 32 formulas(including all possible combinations of both Republican and Democratic)

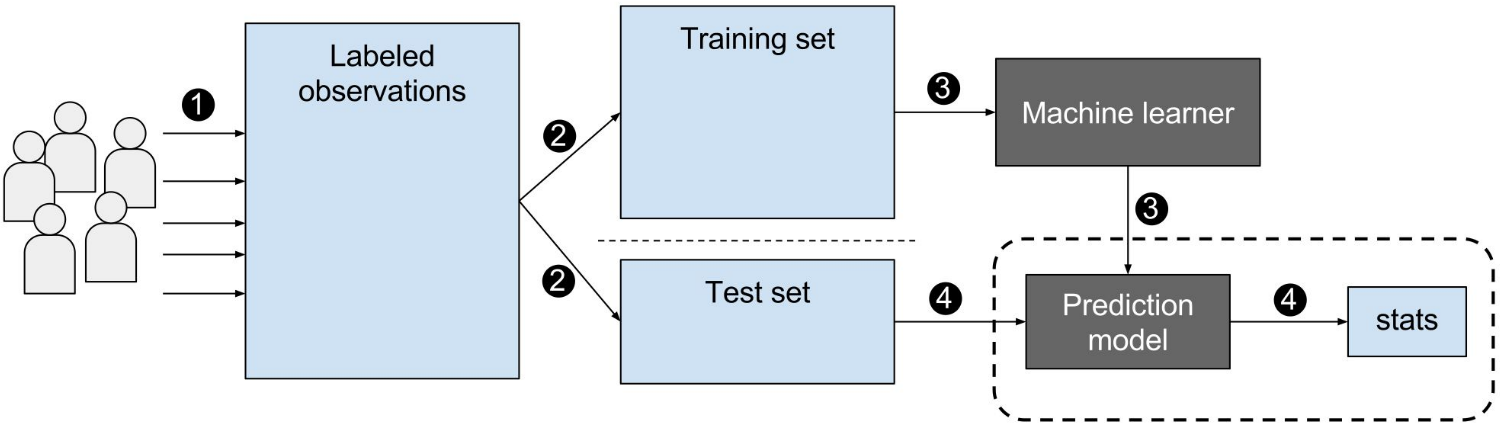
➢ **Training and Testing Model**

- 70% of total d Pig script used to divide the data set into training and testing data
- ata as Training Set
- rest 30% of total data as Testing set

Y Roy Chowdhury

Globsyn Knowledge Foundation Kolkata

Training and Testing Model

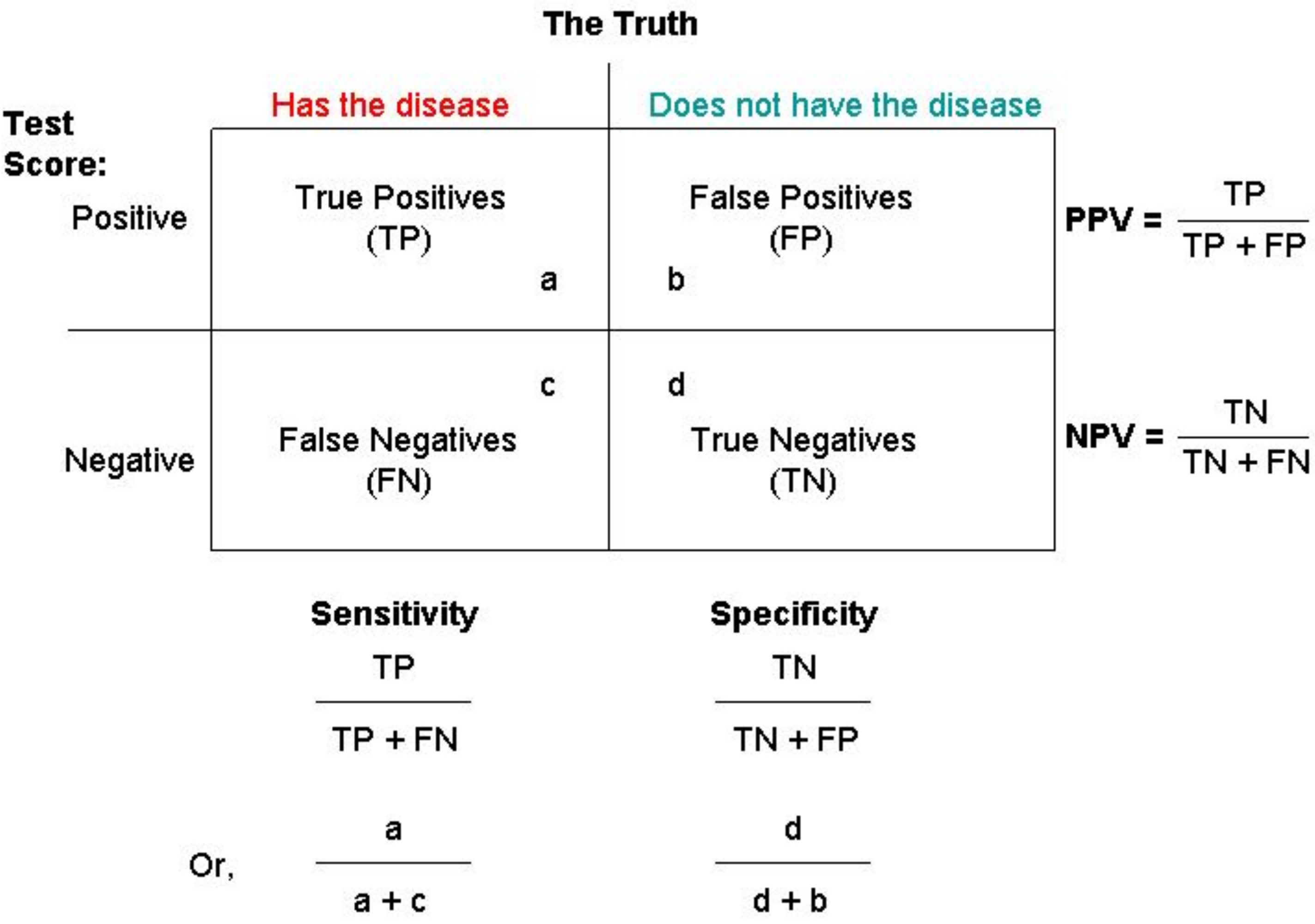


Understanding the confusion matrix

A confusion matrix (Kohavi and Provost, 1998) contains information about actual and predicted classifications done by a classification system. Performance of such systems is commonly evaluated using the data in the matrix. The following table shows the confusion matrix for a two class classifier. The entries in the confusion matrix have the following meaning in the context of our study:

- *a* is the number of **correct** predictions that an instance is **negative**,
- *b* is the number of **incorrect** predictions that an instance is **positive**,
- *c* is the number of **incorrect** of predictions that an instance **negative**, and
- *d* is the number of **correct** predictions that an instance is **positive**.

Y Roy Chowdhury



The Truth

Test Score:	Has the disease	Does not have the disease	
Positive	True Positives (TP) a	False Positives (FP) b	**PPV =** $\frac{TP}{TP + FP}$
Negative	c False Negatives (FN)	d True Negatives (TN)	**NPV =** $\frac{TN}{TN + FN}$

Sensitivity

$$\frac{TP}{TP + FN}$$

Specificity

$$\frac{TN}{TN + FP}$$

Or, $\frac{a}{a + c}$ $\qquad$ $\frac{d}{d + b}$

Y Roy Chowdhury

Globsyn Knowledge Foundation ★ Kolkata

Document sign date :Mar 20, 2019

CONFUSION MATRIX

The *accuracy* (*AC*) is the proportion of the total number of predictions that were correct. It is determined using the equation:

$$AC = \frac{a+d}{a+b+c+d}$$

The *recall* or *true positive rat*e (*TP*) is the proportion of positive cases that were correctly identified, as calculated using the equation:

$$TP = \frac{d}{c+d}$$

Y Roy Chowdhury

The *false positive rate* (*FP*) is the proportion of negatives cases that were incorrectly classified as positive, as calculated using the equation:

$$FP = \frac{b}{a + b}$$

The *true negative rate* (*TN*) is defined as the proportion of negatives cases that were classified correctly, as calculated using the equation:

$$TN = \frac{a}{a + b}$$

Y Roy Chowdhury

Globsyn Knowledge Foundation ★ Kolkata

Document sign date :Mar 20, 2019

The *false negative rate* (*FN*) is the proportion of positives cases that were incorrectly classified as negative, as calculated using the equation:

$$FN = \frac{c}{c + d}$$

Finally, *precision* (*P*) is the proportion of the predicted positive cases that were correct, as calculated using the equation:

$$P = \frac{d}{b + d}$$

Y Roy Chowdhury

CONCLUSION

This project successfully predicts the party with 88% accuracy with 70% training and 30 % Testing which is a enough to label Naive Bayes as a good option as a classification algorithm

Roy Chowdhury

Globsyn Knowledge Foundation
Kolkata

Document sign date :Mar 20, 2019

Certificate

This is to certify that Mr Kaushik Ghosh of Techno India Saltlake, registration number: 161300110046 of 2016-17, has successfully completed a project titled “**Predicting Party Of a Congress Senator”** using Big Data Programming Hadoop under the guidance of Mr Titash Ghosh.

_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _

Titash Ghosh
Globsyn Finishing School

Y Roy Chowdhury

Globsyn Knowledge Foundation ★ Kolkata

Document sign date :Mar 20, 2019

Certificate

This is to certify that Mr Abhishek Kumar of Techno India Saltlake, registration number: 161300110004 of 2016-17, has successfully completed a project titled “**Predicting Party Of a Congress Senator”** using Big Data Programming Hadoop under the guidance of Mr Titash Ghosh.

_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _

Titash Ghosh
Globsyn Finishing School

Y Roy Chowdhury

Globsyn Knowledge Foundation
Kolkata

Document sign date :Mar 20, 2019

Certificate

This is to certify that Mr Ashish Kumar of Techno India Saltlake, registration number: 161300110023 of 2016-17, has successfully completed a project titled "**Predicting Party Of a Congress Senator**" using Big Data Programming Hadoop under the guidance of Mr Titash Ghosh.

_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _

Titash Ghosh
Globsyn Finishing School

Y Roy Chowdhury

Globsyn Knowledge Foundation
Kolkata

Document sign date :Mar 20, 2019

Certificate

This is to certify that Mr Priyam Mukherjee of Government College Of Engineering and Ceramic Technology, registration number: 161130110057 of 2016-17, has successfully completed a project titled "**Predicting Party Of a Congress Senator"** using Big Data Programming Hadoop under the guidance of Mr Titash Ghosh.

Titash Ghosh
Globsyn Finishing School

Y Roy Chowdhury

Globsyn Knowledge Foundation
Kolkata

Document sign date :Mar 20, 2019