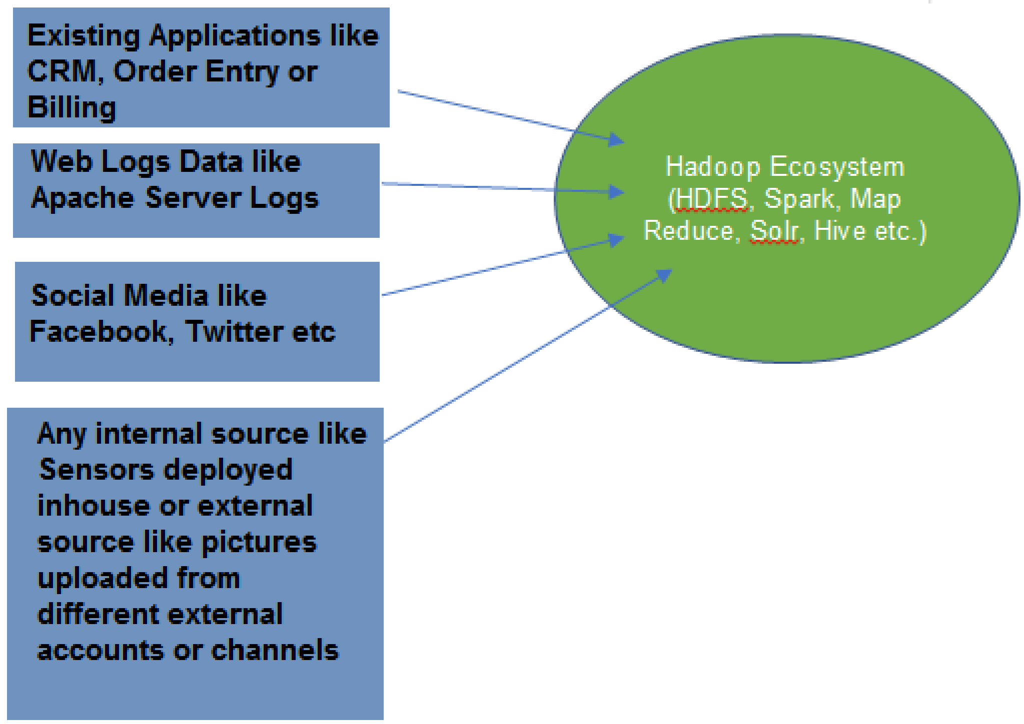
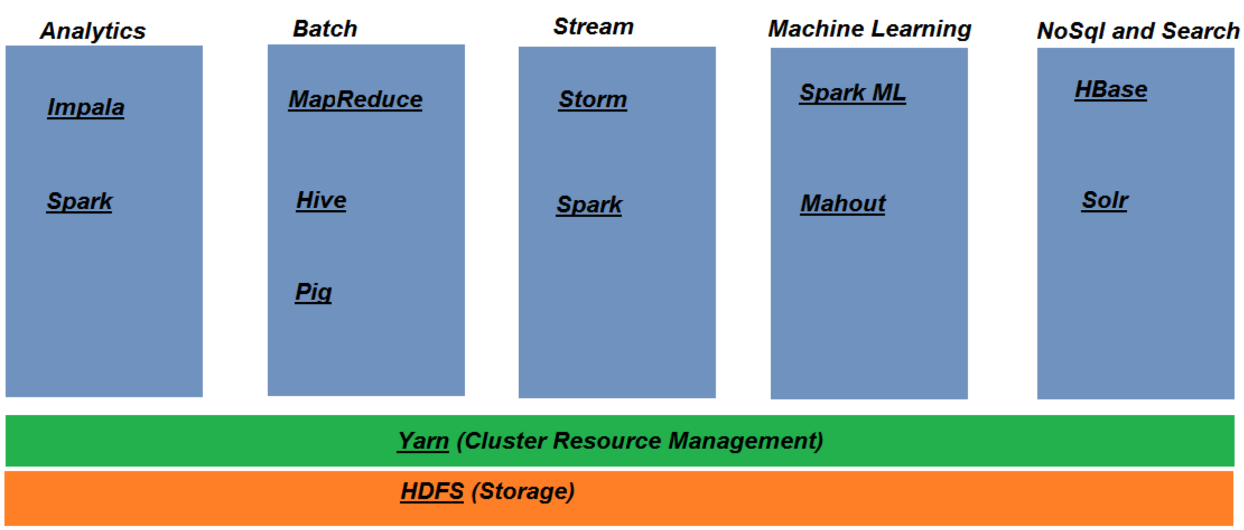
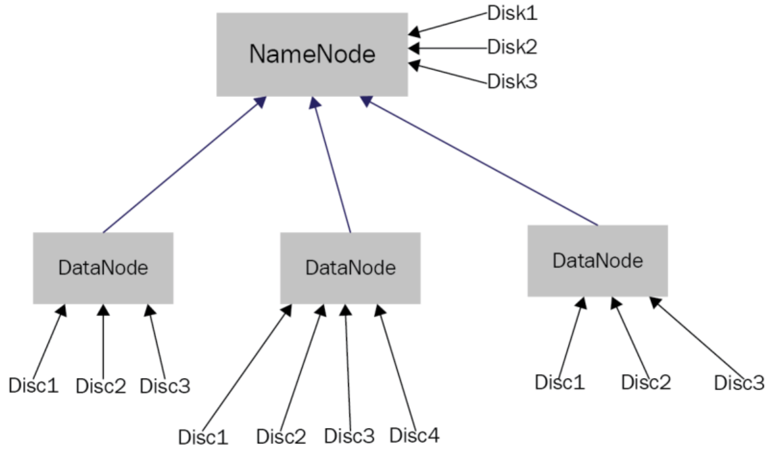
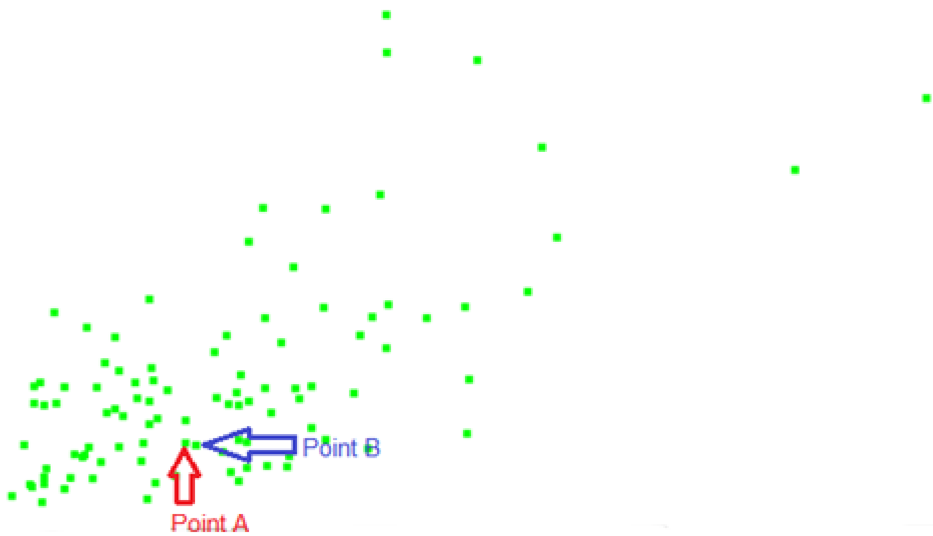
Chapter 1: Big Data Analytics with Java






NameNode
Disk1
Disk2
Disk3
DataNode
DataNode
DataNode
Disc1
Disc2
Disc3
Disc1
Disc2
Disc3
Disc4
Disc1
Disc2
Disc3

Chapter 2: First Steps in Data Analysis




```
+------------+----------------+----------------+--------------+
|make_country|    make_display|         make_id|make_is_common|
+------------+----------------+----------------+--------------+
|       Italy|          Abarth|          abarth|             0|
|          UK|              AC|              ac|             0|
|         USA|           Acura|           acura|             1|
|       Italy|       AlfaRomeo|      alfa-romeo|             1|
|          UK|          Allard|          allard|             0|
+------------+----------------+----------------+--------------+
only showing top 5 rows
```

The results printed are

```
//Result of Selected columns :
+------------+----------------+
|make_country|    make_display|
+------------+----------------+
|       Italy|          Abarth|
|          UK|              AC|
|         USA|           Acura|
|       Italy|       AlfaRomeo|
|          UK|          Allard|
+------------+----------------+
only showing top 5 rows
```

```
+------------+------------+-----------+--------------+
|make_country|make_display|    make_id|make_is_common|
+------------+------------+-----------+--------------+
|       Italy|      Abarth|     abarth|             0|
|       Italy|   AlfaRomeo| alfa-romeo|             1|
|       Italy| Autobianchi|autobianchi|             0|
|       Italy|   Bizzarrini| bizzarrini|             0|
|       Italy|     Bugatti|    bugatti|             1|
|       Italy|   De Tomaso|   de-tomaso|             0|
|       Italy|     Ferrari|    ferrari|             1|
|       Italy|        Fiat|       fiat|             1|
|…
+------------+------------+-----------+--------------+
```

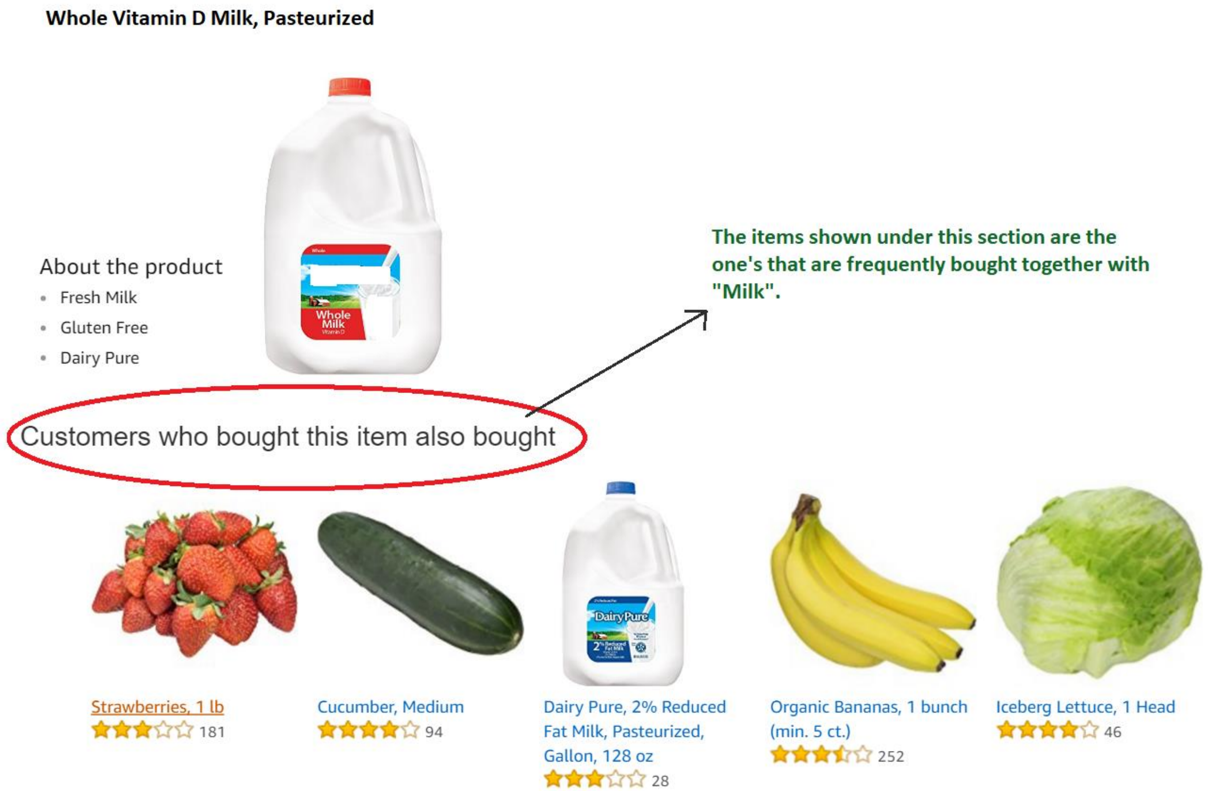
```
Number of cars from Italy in this dataset --> 17
```

```
Japan car dataset ------> 15
```

```
Printing Distinct Countries below
Denmark
Serbia
India
UK
…

Total Distinct Countries : 23
```

```
UK , 39
USA , 29
Italy , 17
Germany , 16
Japan , 15
France , 8
South Korea , 5
Netherlands , 3
China , 3
Sweden , 3
Russia , 3
India , 2
Czech Republic , 2
…
```

```
+------------+---+
|make_country|cnt|
+------------+---+
|          UK| 39|
+------------+---+
```

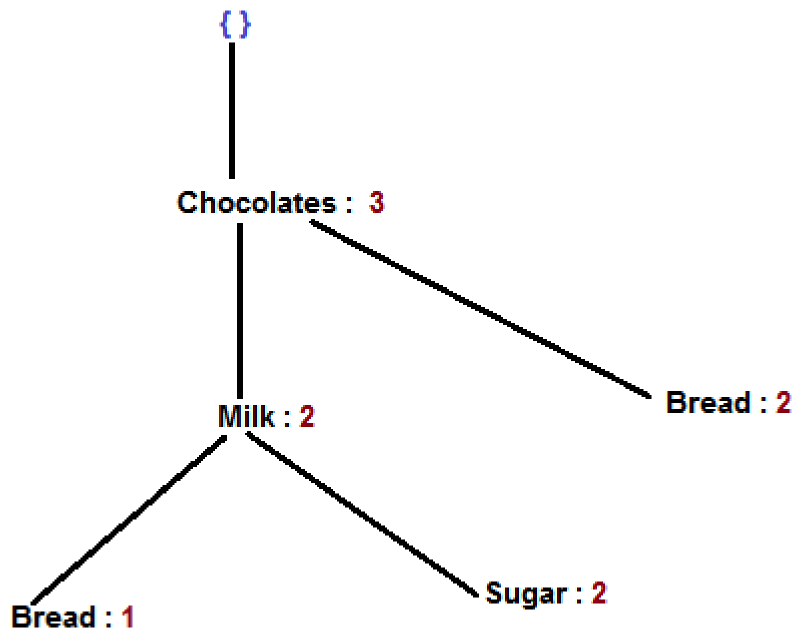


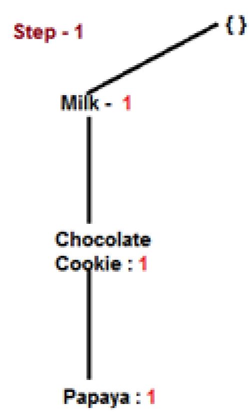
Support for an item ➡ $$\frac{\text{Number of transactions where this item or items were sold}}{\text{Total Number of Transactions}}$$

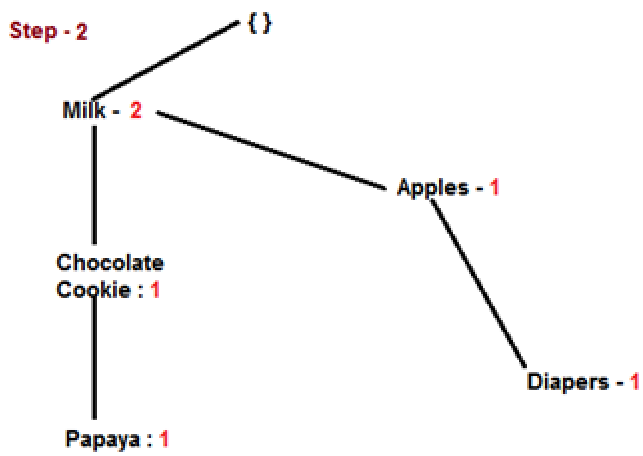
example :

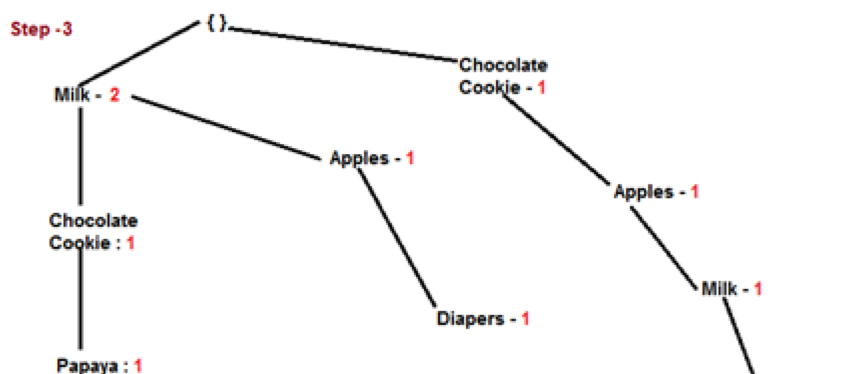
Support for Milk ➡ $$\frac{\text{Number of Transactions containing Milk}}{\text{Total Number of Transactions}} = 3 / 4 = 0.75$$

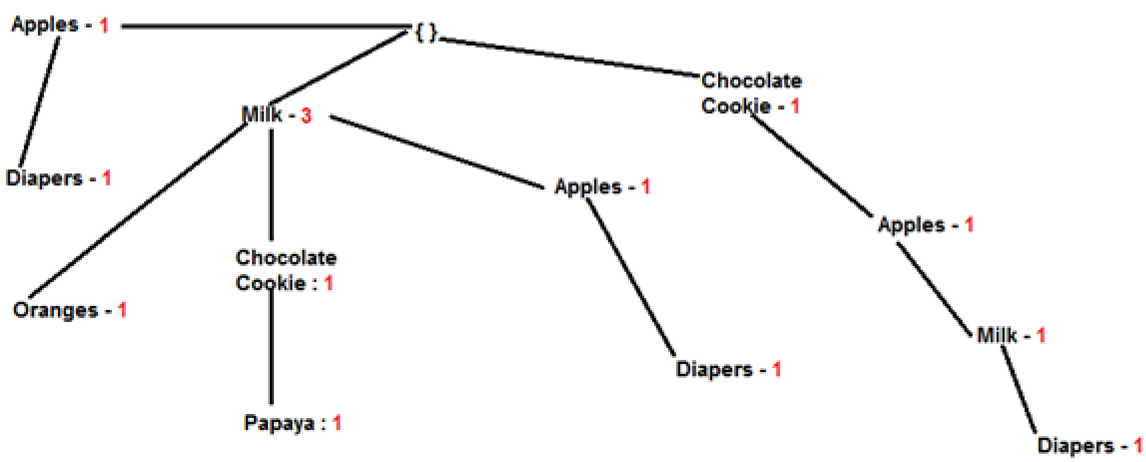
Confidence for {A} => {B} is ➡ $$\frac{\text{Support of (\{A\} U \{B\})}}{\text{Support of \{A\}}}$$

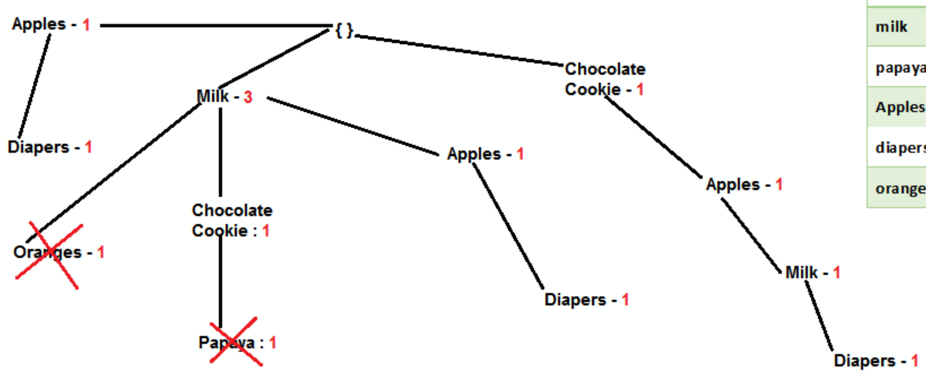

{ }
Chocolates : 3
Milk : 2
Bread : 2
Bread : 1
Sugar : 2


Step - 1
{ }
Milk - 1
Chocolate Cookie : 1
Papaya : 1


Step - 2
{ }
Milk - 2
Apples - 1
Chocolate
Cookie : 1
Diapers - 1
Papaya : 1

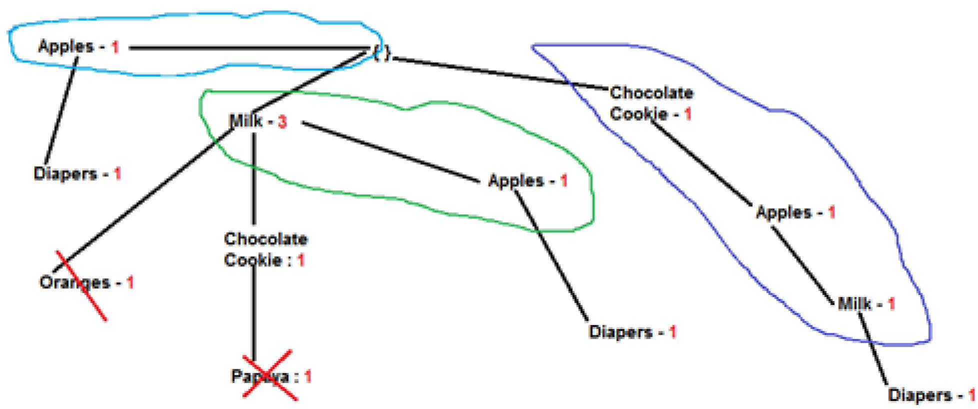

Step -3
{ }
Chocolate
Cookie - 1
Milk - 2
Apples - 1
Apples - 1
Chocolate
Cookie : 1
Milk - 1
Diapers - 1
Papaya : 1
Diapers - 1


Apples - 1
{ }
Chocolate
Cookie - 1
Milk - 3
Diapers - 1
Apples - 1
Apples - 1
Chocolate
Cookie : 1
Oranges - 1
Milk - 1
Diapers - 1
Papaya : 1
Diapers - 1



chocolate cookie	2
milk	4
papaya	1
Apples	3
diapers	3
oranges	1

Less Than Support So Remove.

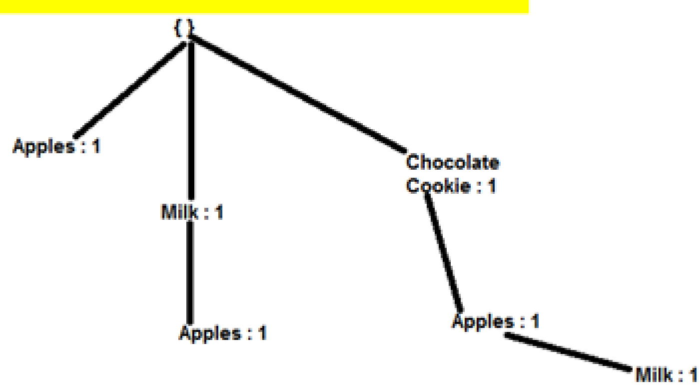


Conditional Patterns for 'Diapers'

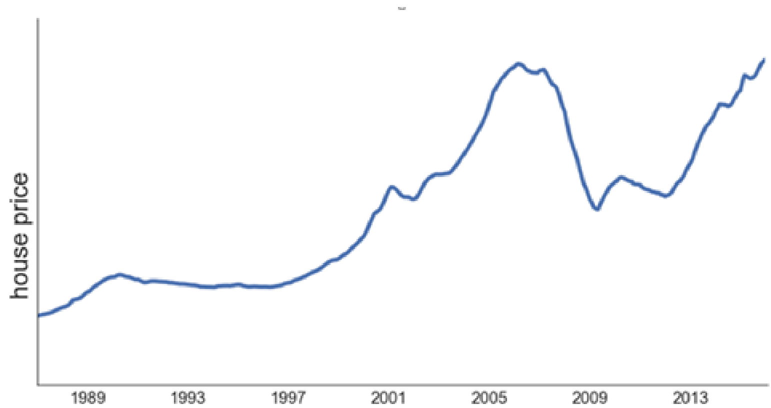
1. {Apples : 1}
2. {Milk, Apples : 1}
3. {Chocolate Cookie, Apples, Milk : 1}

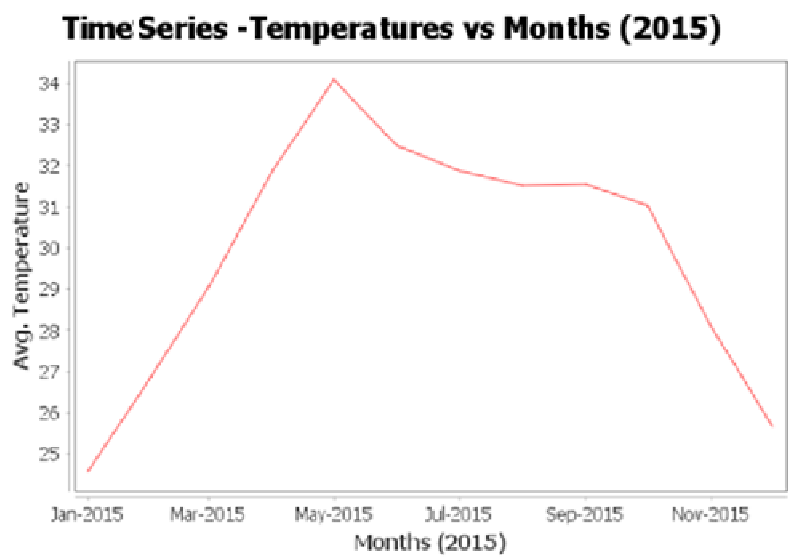
Conditional FP-Tree for 'Diapers' using transactions

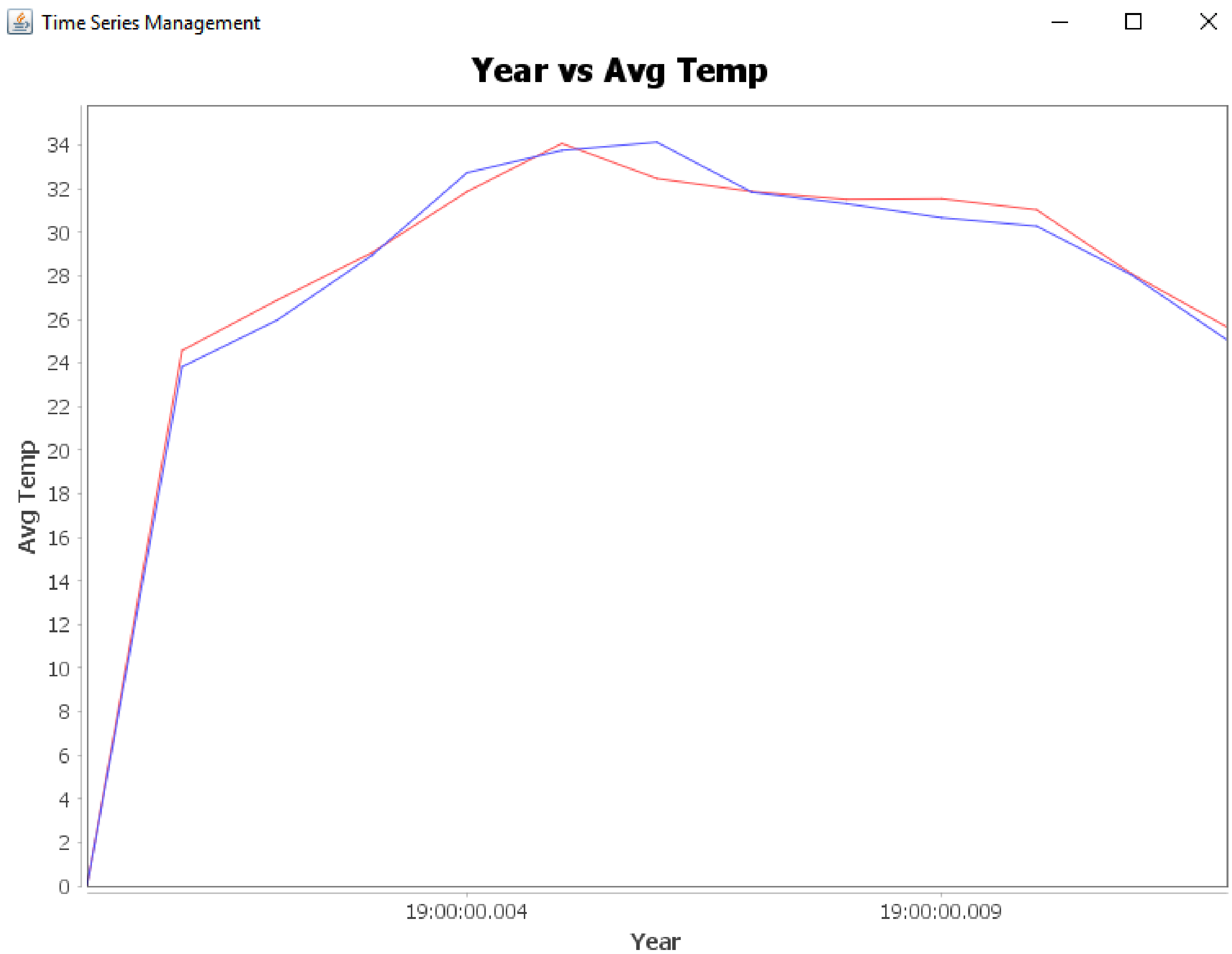
{Apples : 1} , {Milk, Apples : 1}, {Chocolate Cookie, Apples, Milk : 1}



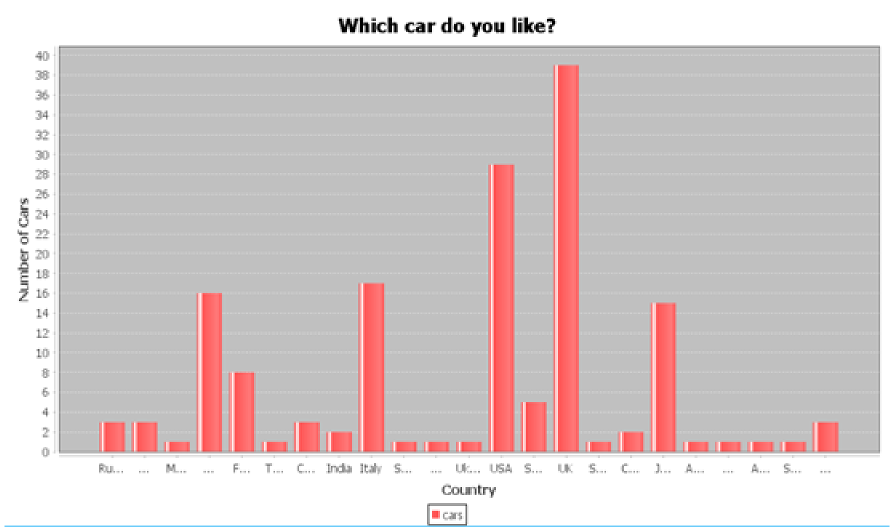
Chapter 3: Data Visualization



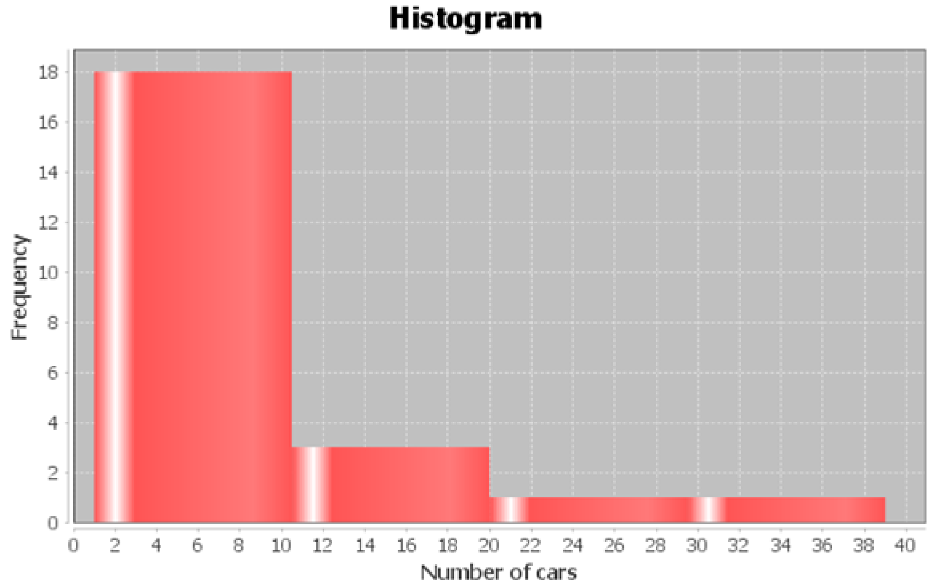


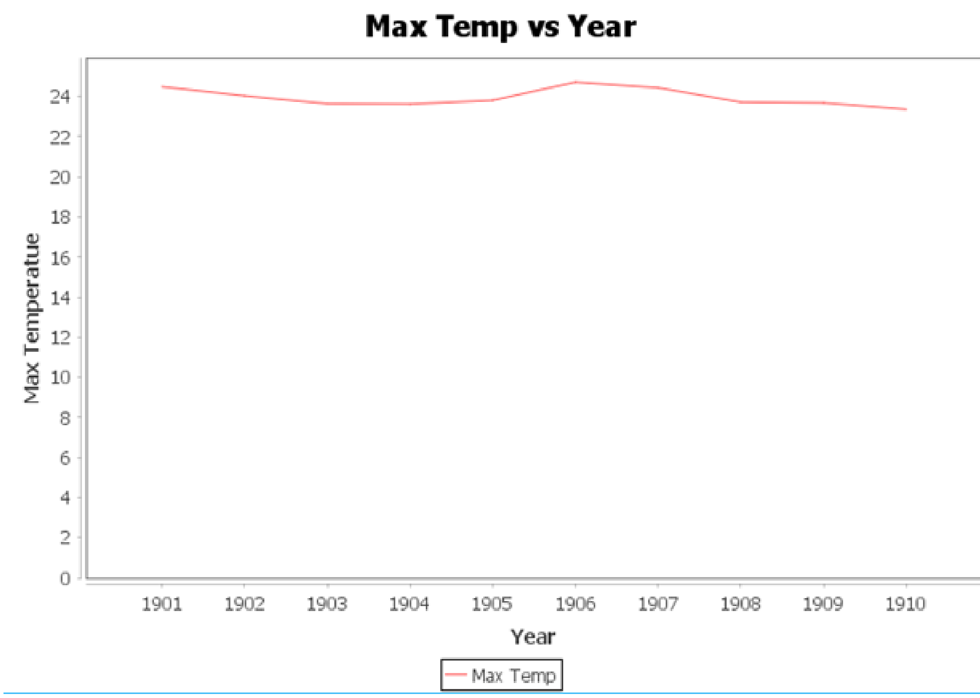


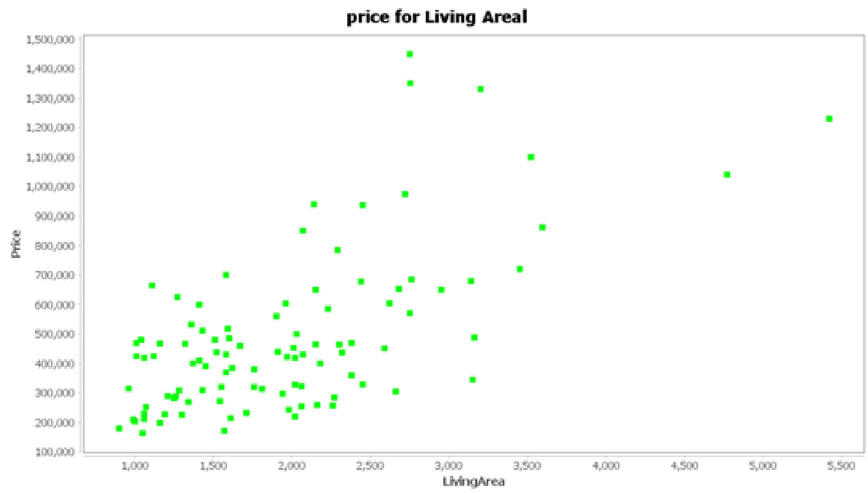
Time Series Management
Year vs Avg Temp
Avg Temp
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
19:00:00.004
19:00:00.009
Year

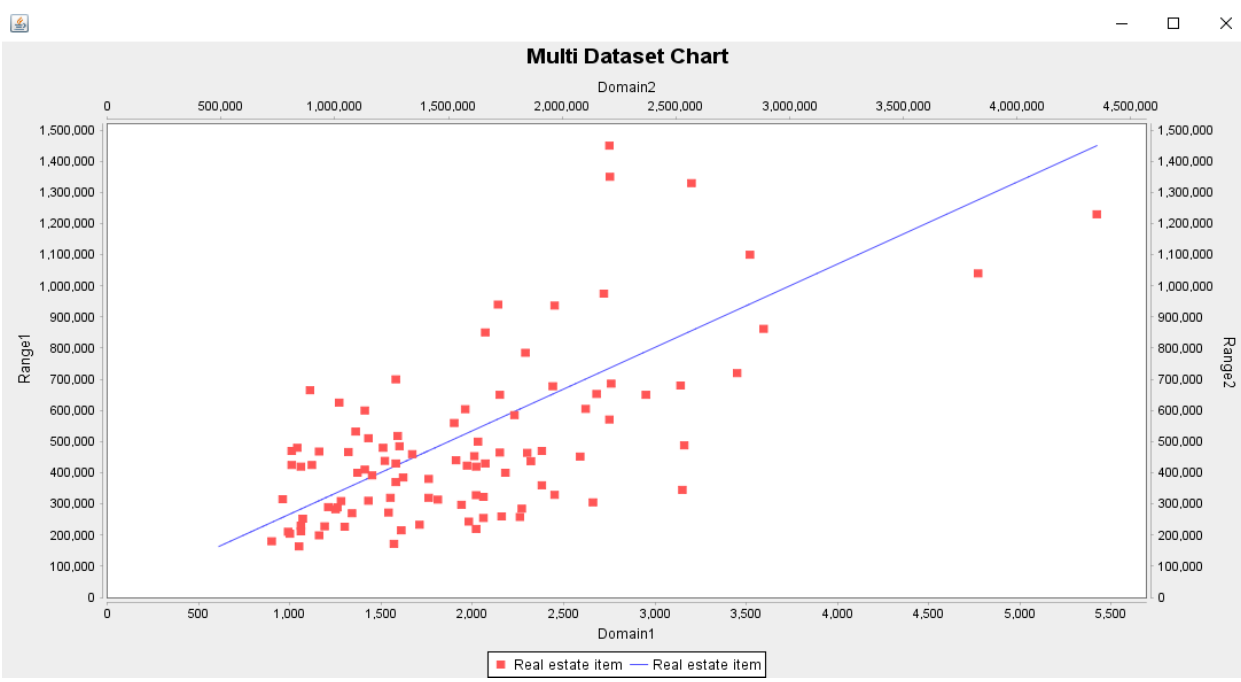


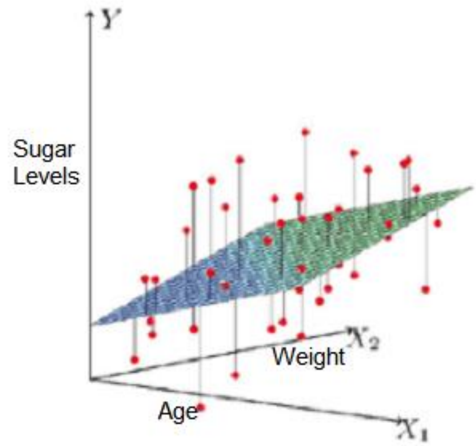
Which car do you like?
Number of Cars
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
India
Italy
USA
UK
Country
cars

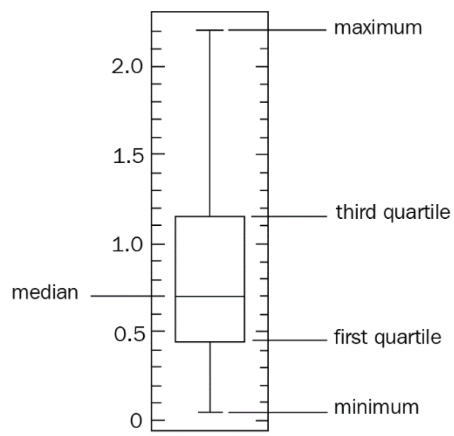

Histogram
Frequency
18
16
14
12
10
8
6
4
2
0
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40
Number of cars

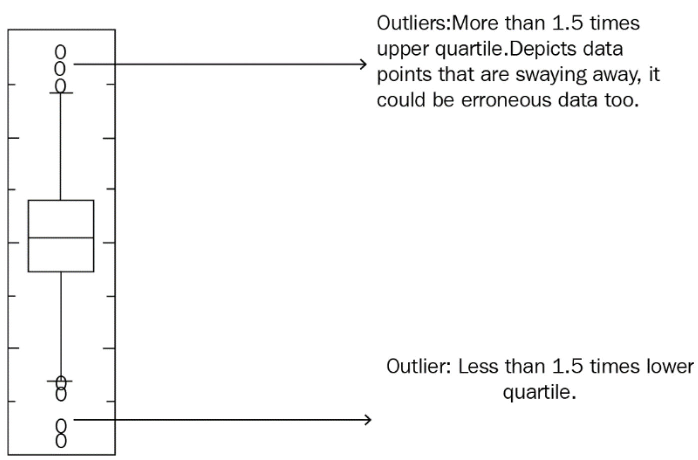

Max Temp vs Year
Max Temperatue
24
22
20
18
16
14
12
10
8
6
4
2
0
1901 1902 1903 1904 1905 1906 1907 1908 1909 1910
Year
Max Temp

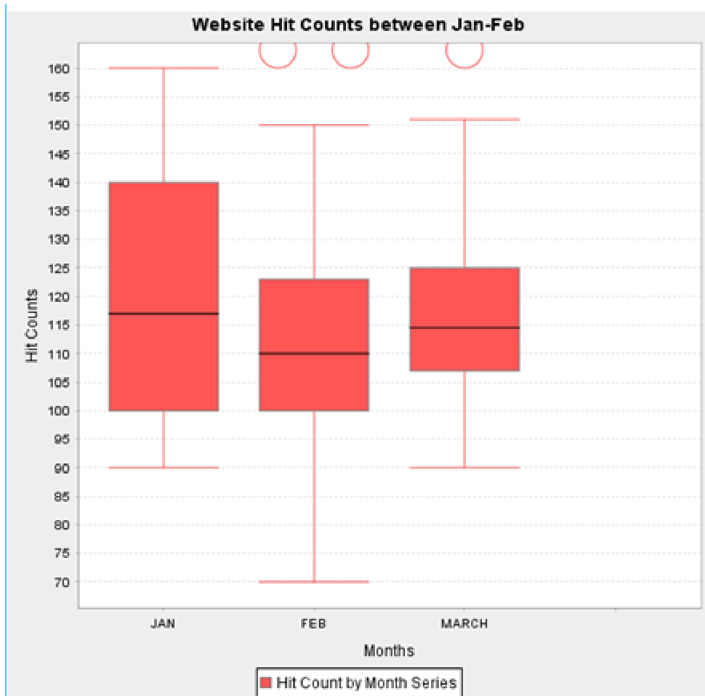

price for Living AreaI
Price
LivingArea
1,500,000
1,400,000
1,300,000
1,200,000
1,100,000
1,000,000
900,000
800,000
700,000
600,000
500,000
400,000
300,000
200,000
100,000
1,000
1,500
2,000
2,500
3,000
3,500
4,000
4,500
5,000
5,500

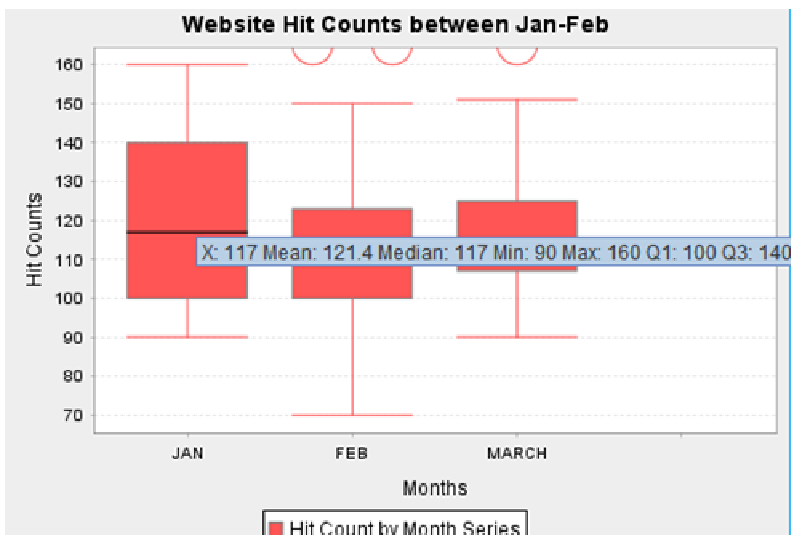

Multi Dataset Chart
Domain2
0
500,000
1,000,000
1,500,000
2,000,000
2,500,000
3,000,000
3,500,000
4,000,000
4,500,000
Range1
Range2
1,500,000
1,400,000
1,300,000
1,200,000
1,100,000
1,000,000
900,000
800,000
700,000
600,000
500,000
400,000
300,000
200,000
100,000
0
0
500
1,000
1,500
2,000
2,500
3,000
3,500
4,000
4,500
5,000
5,500
Domain1
Real estate item
Real estate item


Y
Sugar
Levels
Weight
X_2
Age
X_1

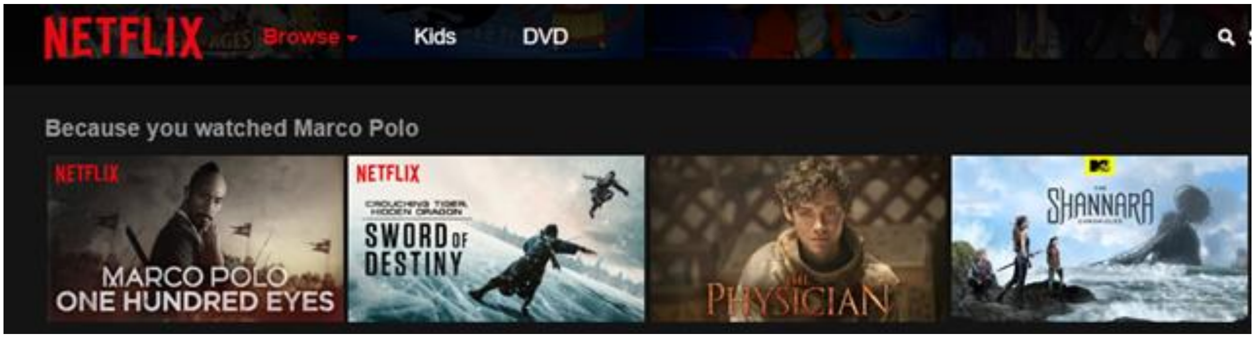

maximum
2.0
1.5
third quartile
1.0
median
0.5
first quartile
0
minimum

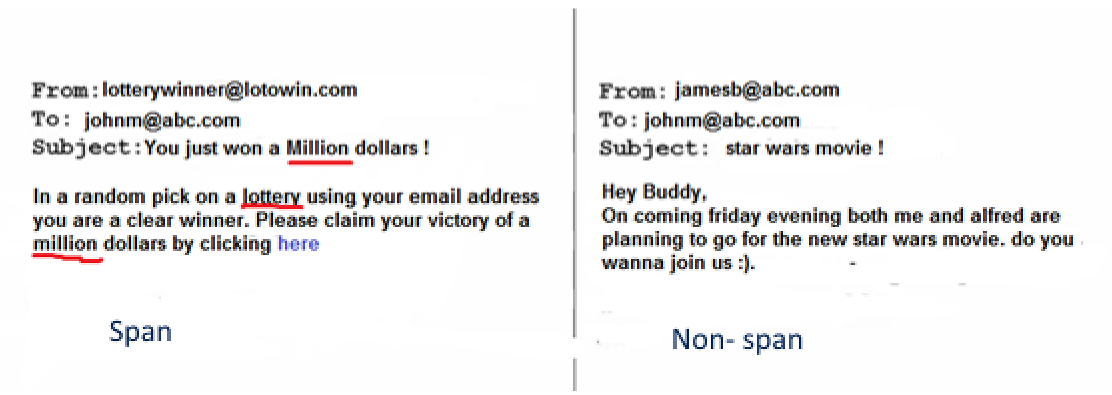

Outliers:More than 1.5 times
upper quartile.Depicts data
points that are swaying away, it
could be erroneous data too.
Outlier: Less than 1.5 times lower
quartile.


Website Hit Counts between Jan-Feb
Hit Counts
160
155
150
145
140
135
130
125
120
115
110
105
100
95
90
85
80
75
70
JAN
FEB
MARCH
Months
Hit Count by Month Series

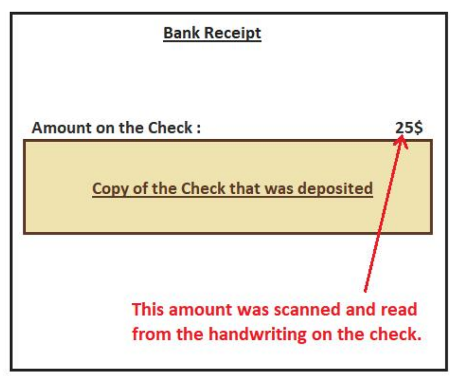

Website Hit Counts between Jan-Feb
Hit Counts
160
150
140
130
120
110
100
90
80
70
X: 117 Mean: 121.4 Median: 117 Min: 90 Max: 160 Q1: 100 Q3: 140
JAN
FEB
MARCH
Months
Hit Count by Month Series

Chapter 4: Basics of Machine Learning

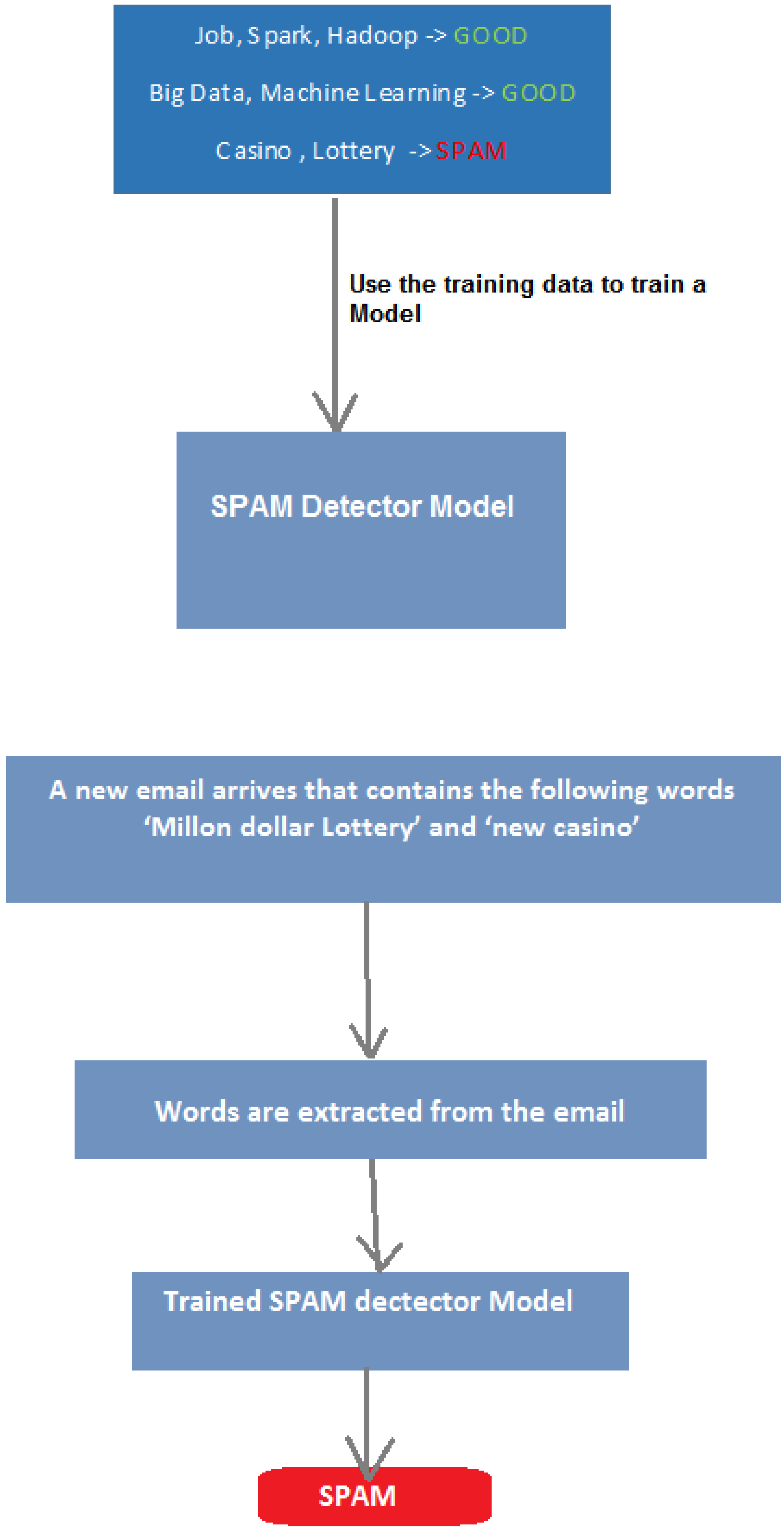




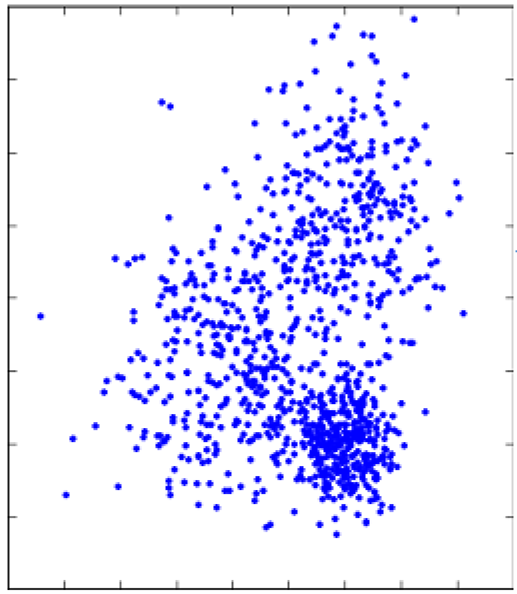
Check deposit through the ATM machine

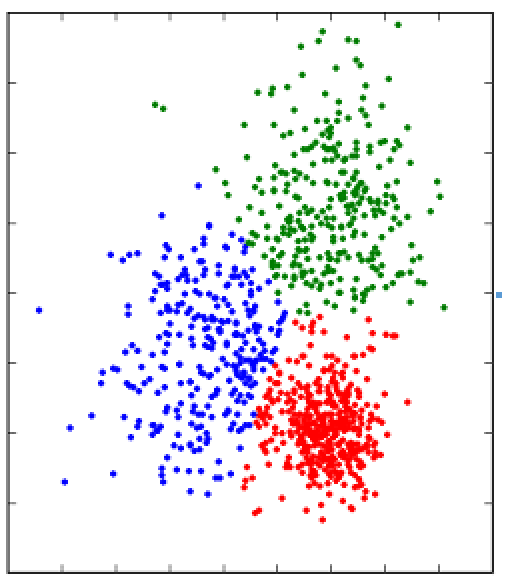


A bank receipt sample



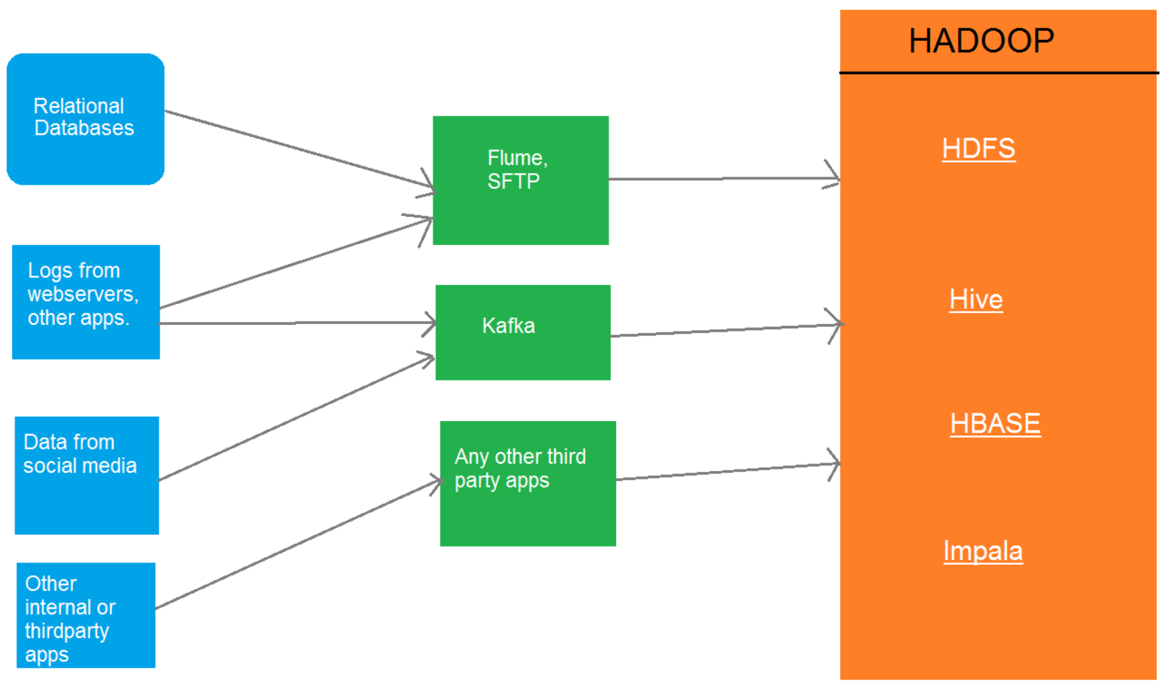
Job, Spark, Hadoop -> GOOD
Big Data, Machine Learning -> GOOD
Casino , Lottery -> SPAM
Use the training data to train a Model
SPAM Detector Model
A new email arrives that contains the following words 'Millon dollar Lottery' and 'new casino'
Words are extracted from the email
Trained SPAM dectector Model
SPAM



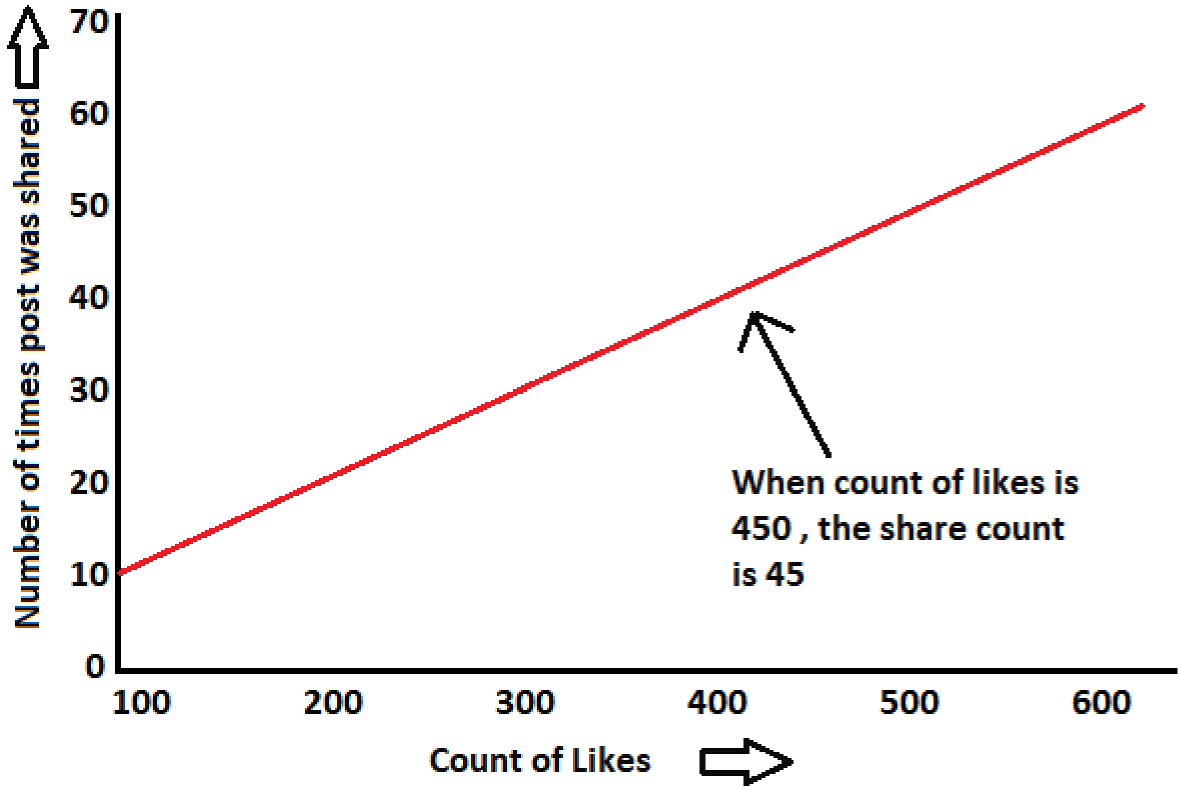


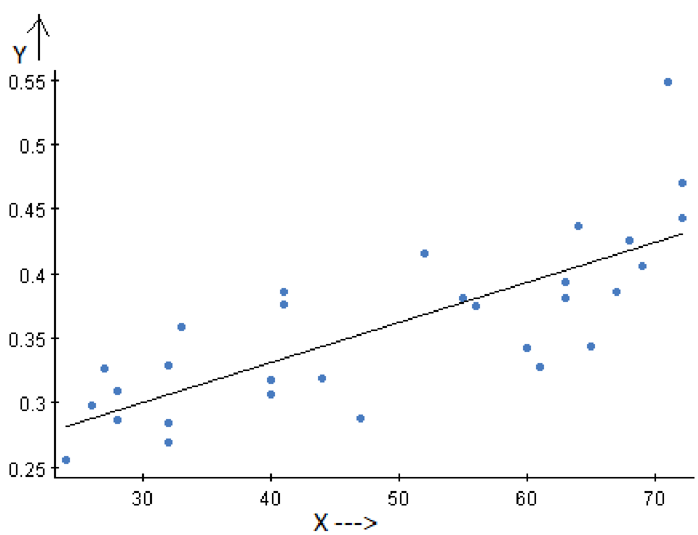
Correlation is: 0.8500286768773001

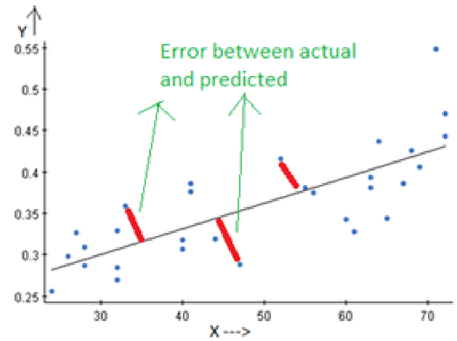
id	features	clicked	selectedFeatures
7	[0.0,0.0,18.0,1.0]	1.0	[18.0]
8	[0.0,1.0,12.0,0.0]	0.0	[12.0]
9	[1.0,0.0,15.0,0.1]	0.0	[15.0]

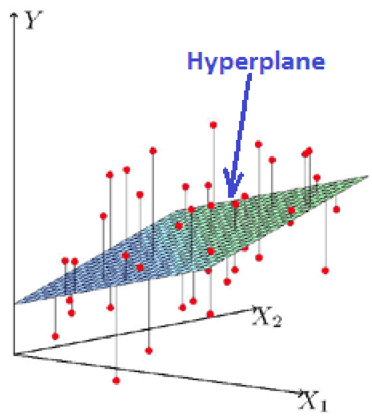


Chapter 5: Regression on Big Data





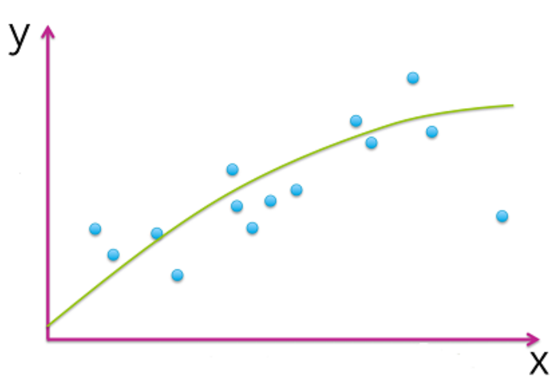




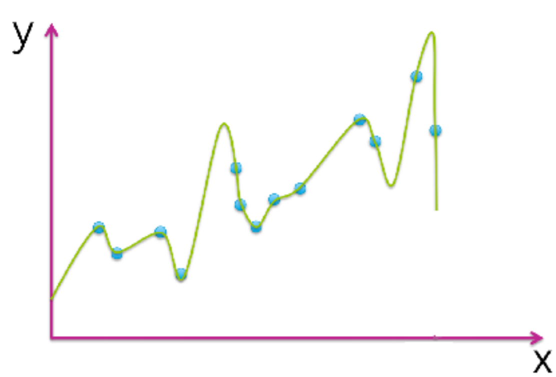
X1 : Number of facebook likes on Post

X2 : Number of comments on Post

Y : Number of Shares of Post (dependent variable)



Y is depicted by Quadratic function here



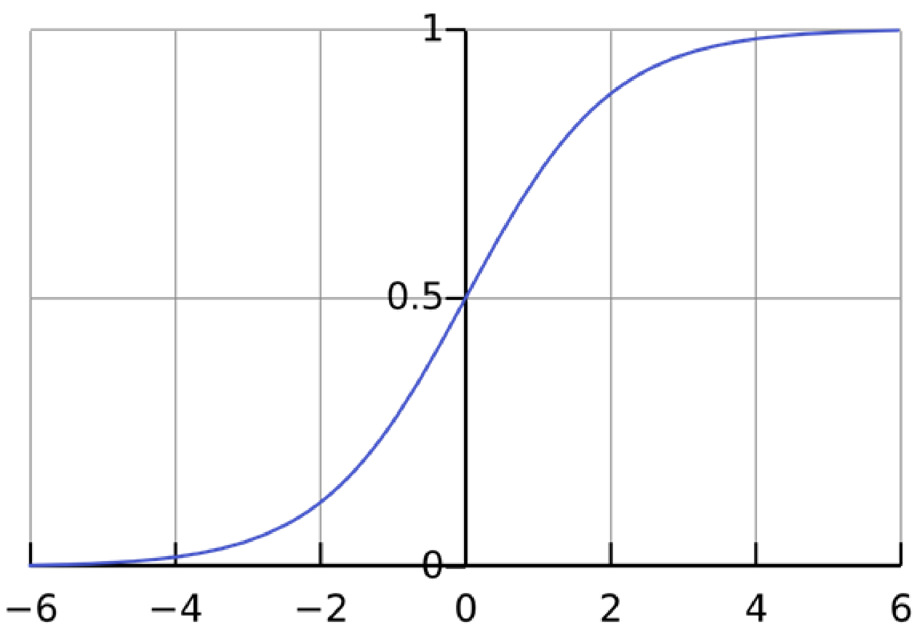
Y is depicted by Polynomial function here

```
17/03/03 01:17:32 INFO DAGScheduler: Job 1 finished: count at HousingDataExplore.j
17/03/03 01:17:32 INFO CodeGenerator: Code generated in 31.059641 ms
Number of Rows --> 21614
17/03/03 01:17:32 INFO SparkContext: Invoking stop() from shutdown hook
17/03/03 01:17:32 INFO SparkUI: Stopped Spark web UI at http://192.168.1.6:4040
```

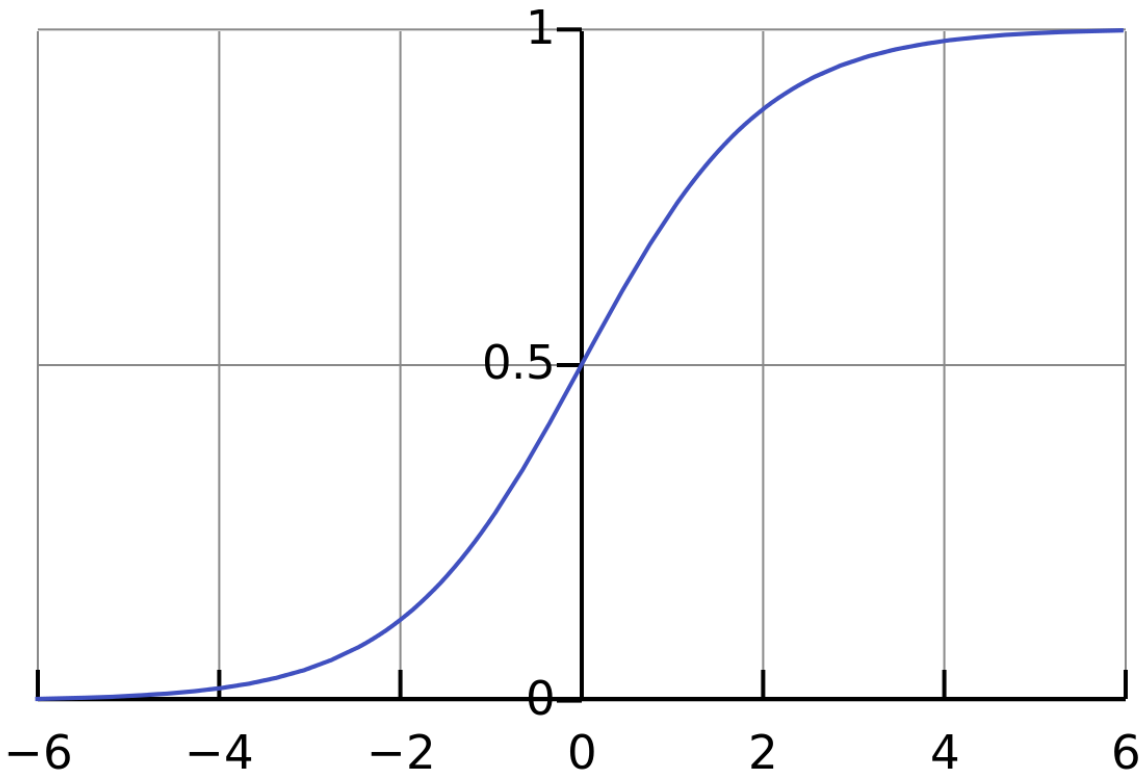
```
17/03/03 01:34:31 INFO CodeGenerator: Code generated in 36.:
+-------+-----------------+
|zipcode|         avgPrice|
+-------+-----------------+
|  98039|        2160606.6|
|  98004|1355927.0820189274|
|  98040|1194230.0212765958|
|  98112| 1095499.342007435|
"|  98102| 901258.2666666667|
```

```
+---------+--------------------+
|    label|            features|
+---------+--------------------+
| 221900.0|[3.0,1.0,1180.0,5...|
| 538000.0|[3.0,2.25,2570.0,...|
| 180000.0|[2.0,1.0,770.0,10...|
| 604000.0|[4.0,3.0,1960.0,5...|
| 510000.0|[3.0,2.0,1680.0,8...|
|1225000.0|[4.0,4.5,5420.0,1...|
```

```
Coefficients: [-64534.41523975349,8313.685443179289,314.60167605061406,-0.37592410488893885] Intercept: 90080.21844645533
RMSE: 257059.12249611464
```



$$y = \frac{1}{1+e^{-x}}$$




```
Number of Rows --> 303
```

```
+---+--------+
|sex|count(1)|
+---+--------+
|1.0|     206|
|0.0|      97|
+---+--------+
```

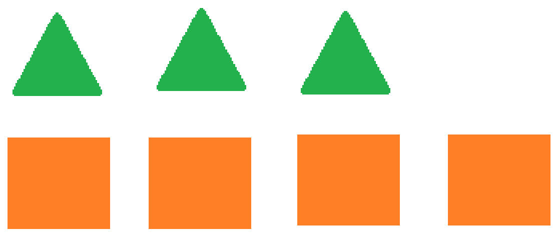
```
+---+----------------------+
|sex|avg(CAST(_c0 AS DOUBLE))|
+---+----------------------+
|1.0|     53.83495145631068|
|0.0|     55.72164948453608|
+---+----------------------+
```

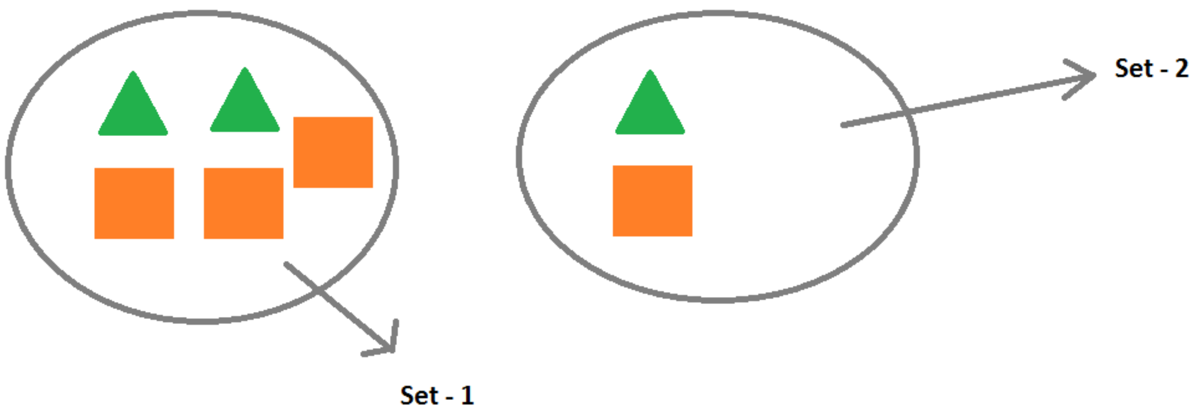
```
+---+--------+
|sex|count(1)|
+---+--------+
|1.0|     206|
|0.0|      97|
+---+--------+
```

```
+-----+--------------------+
|label|            features|
+-----+--------------------+
|  0.0|[63.0,1.0,1.0,145...|
|  1.0|[67.0,1.0,4.0,160...|
|  1.0|[67.0,1.0,4.0,120...|
|  0.0|[37.0,1.0,3.0,130...|
|  0.0|[41.0,0.0,2.0,130...|
```

```
([52.0,1.0,3.0,138.0,223.0], 0.0) , prediction=0.0
([55.0,1.0,2.0,130.0,262.0], 0.0) , prediction=0.0
([56.0,0.0,2.0,140.0,294.0], 0.0) , prediction=0.0
([56.0,1.0,2.0,120.0,236.0], 0.0) , prediction=0.0
([60.0,0.0,3.0,120.0,178.0], 0.0) , prediction=0.0
([67.0,0.0,3.0,115.0,564.0], 0.0) , prediction=1.0
([67.0,0.0,3.0,152.0,277.0], 0.0) , prediction=0.0
([68.0,1.0,3.0,118.0,277.0], 0.0) , prediction=1.0
([42.0,1.0,4.0,136.0,315.0], 1.0) , prediction=1.0
([43.0,0.0,4.0,132.0,341.0], 1.0) , prediction=0.0
```

Chapter 6: Naive Bayes and Sentiment Analysis





$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

$$\frac{1}{2} = 0.5$$

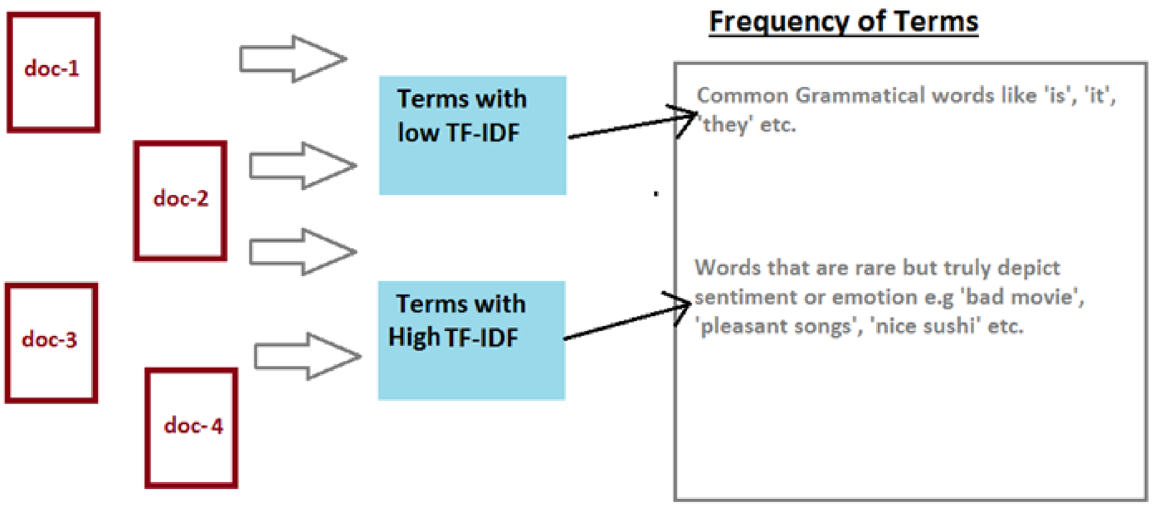
$$P(Triangle) = \frac{Number\ of\ Triangles}{Total\ Figures\ Count} = \frac{3}{7} = 0.43$$

$$P(Triangle \mid Set\_1) = \frac{P(\text{Triangle and Set}_1)}{P(\text{Set}_1)} = \frac{2}{5} = 0.4$$

$$P(Set\_1 \mid Triangle) = \frac{P(Triangle \mid Set\_1)\,P(Set\_1)}{P(Triangle)} = \frac{0.4 * 0.5}{0.43} = 0.47$$

$$P\left(Set\_2 \mid Triangle\right) = 1 - P\left(Set_1 \mid Triangle\right) = 1 - 0.47 = 0.53$$

$$Inverse\ Document\ Frequency = \log \frac{Total\ Number\ of\ Documents}{Number\ of\ Documents\ containing\ the\ Term}$$

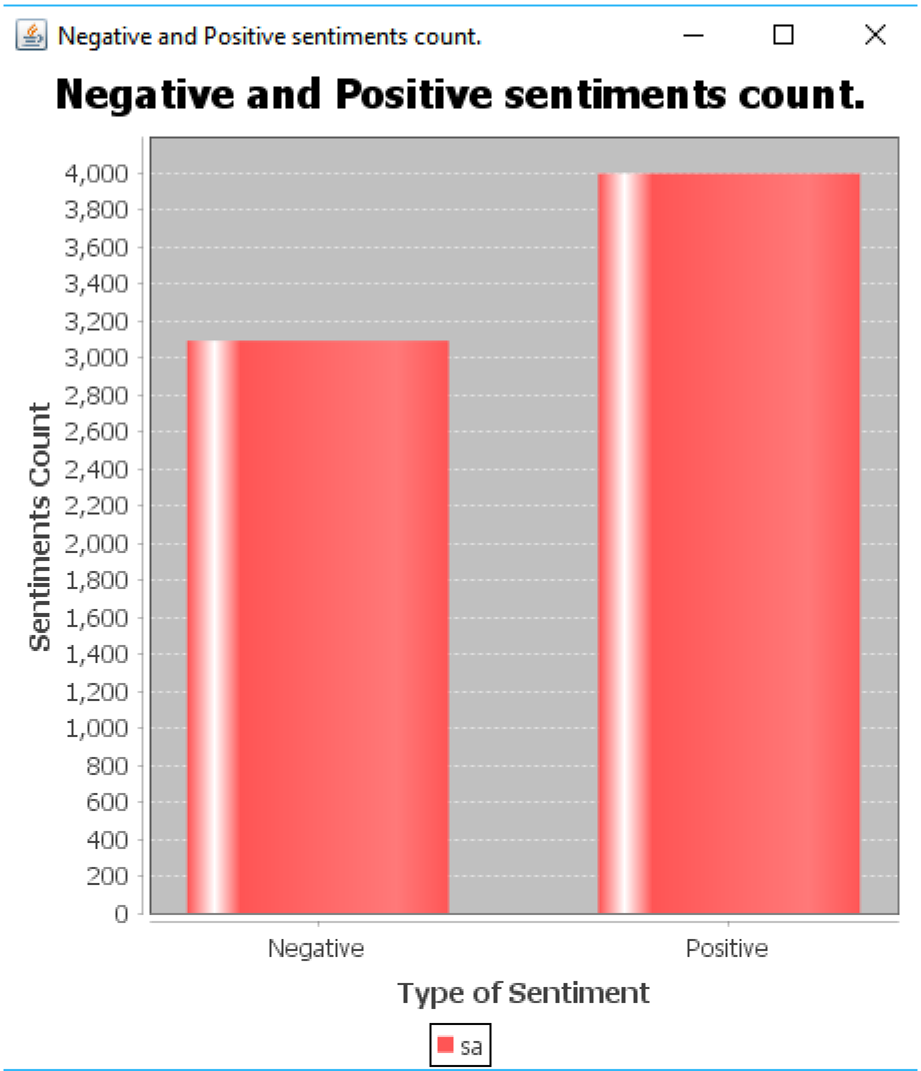


$$Inverse\ Document\ Frequency = \log \frac{Total\ Number\ of\ Documents}{Number\ of\ Documents\ containing\ the\ Term}$$

```
1    The Da Vinci Code book is just awesome.
1    i liked the Da Vinci Code a lot.
1    I loved the Da Vinci Code, but now I want something better and different!..
1    i liked the Da Vinci Code a lot.
1    I liked the Da Vinci Code but it ultimatly didn't seem to hold it's own.
```

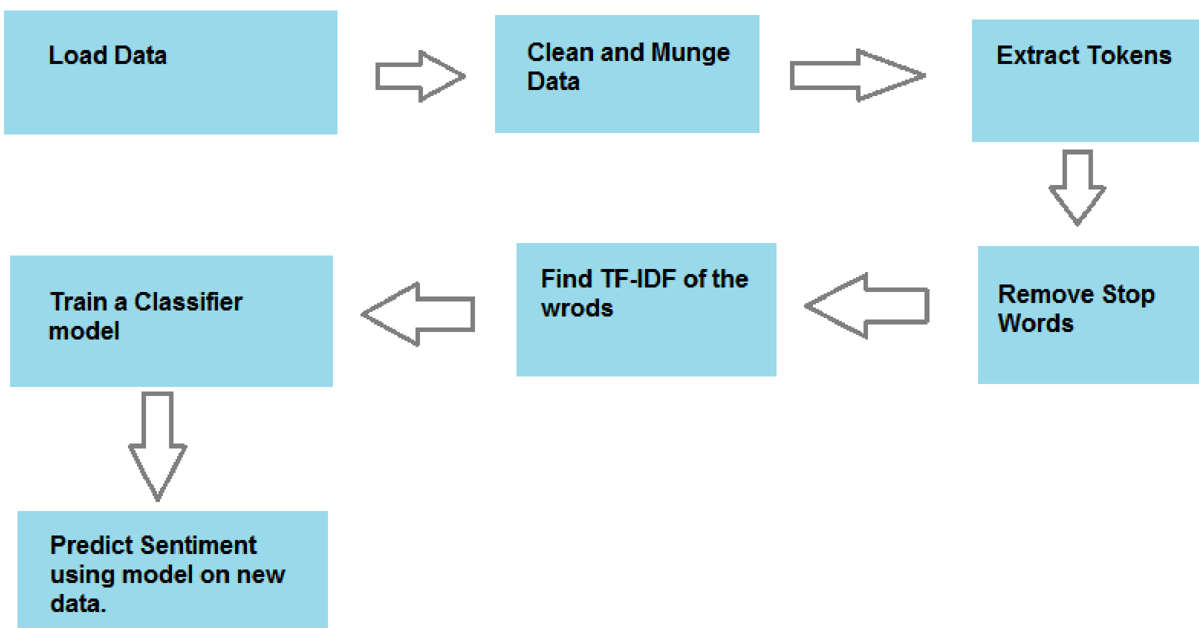
```
17/07/12 09:00:30 INFO CodeGenerator:
+-----+--------------------+
|label|               tweet|
+-----+--------------------+
|  1.0|The Da Vinci Code...|
|  1.0|this was the firs...|
|  1.0|i liked the Da Vi...|
|  1.0|i liked the Da Vi...|
|  1.0|I liked the Da Vi...|
+-----+--------------------+
only showing top 5 rows

17/07/12 09:00:30 INFO SparkContext:
```

```
17/07/12 09:07:03 INFO CodeGenerator:
+---------+--------+
|sentiment|count(1)|
+---------+--------+
|      0.0|    3091|
|      1.0|    3995|
+---------+--------+

17/07/12 09:07:03 INFO SparkContext:
```



word	count
i	4639
the	3221
and	2150
harry	2088
potter	2082
vinci	2001
da	1998
brokeback	1996
code	1895
mountain	1803
love	1542
is	1517
a	1305
was	1175

Stop words shown between ellipses.



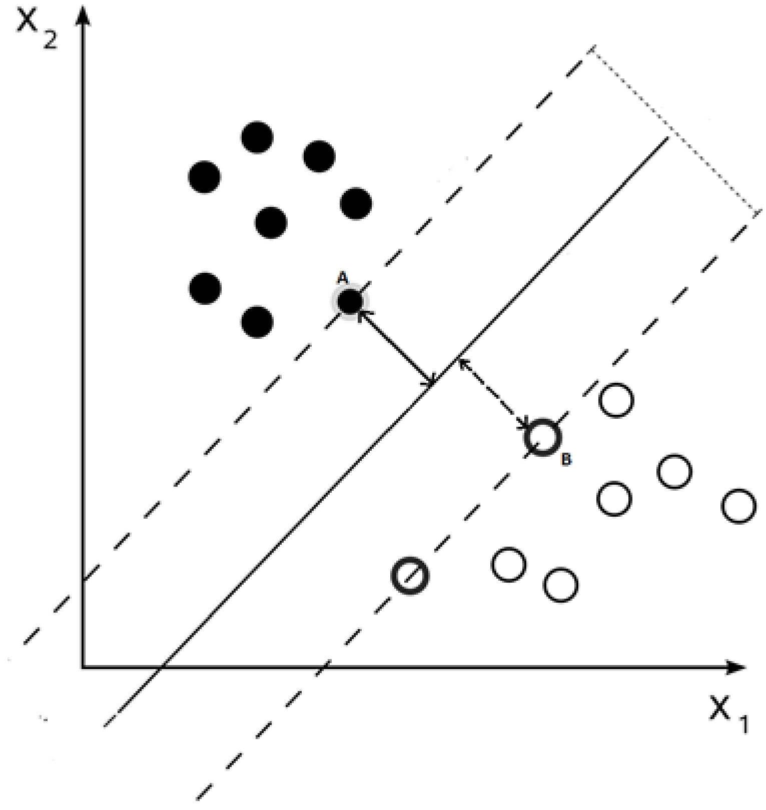
```
##############---- START ---- ##############
Opinion --> 0.0
Sentence --> 00 we rode bikes to hollywood and rented brokeback mountain which was also stupid
Tokens --> [00, we, rode, bikes, to, hollywood, and, rented, brokeback, mountain, which, was, also, stupid]
##############---- END ---- ##############

##############---- START ---- ##############
Opinion --> 0.0
Sentence --> 1-BROKEBACK MOUNTAIN IS A STUPID MOVIE
Tokens --> [1-brokeback, mountain, is, a, stupid, movie]
##############---- END ---- ##############

##############---- START ---- ##############
Opinion --> 0.0
Sentence --> 10 Things I Hate About You + A Knight's Tale  Brokeback Mountain =
Tokens --> [10, things, i, hate, about, you, +, a, knight's, tale, , brokeback, mountain, =]
##############---- END ---- ##############
```

Tokens

```
##############---- START ---- ##############
Opinion --> 0.0
Sentence --> 00 we rode bikes to hollywood and rented brokeback mountain which was also stupid
Tokens --> [00, we, rode, bikes, to, hollywood, and, rented, brokeback, mountain, which, was, also, stupid]
After removing Stop Words --> [00, rode, bikes, hollywood, rented, brokeback, mountain, also, stupid]
##############---- END ---- ##############

##############---- START ---- ##############
Opinion --> 0.0
Sentence --> 1-BROKEBACK MOUNTAIN IS A STUPID MOVIE
Tokens --> [1-brokeback, mountain, is, a, stupid, movie]
After removing Stop Words --> [1-brokeback, mountain, stupid, movie]
##############---- END ---- ##############

##############---- START ---- ##############
Opinion --> 0.0
Sentence --> 10 Things I Hate About You + A Knight's Tale  Brokeback Mountain =
Tokens --> [10, things, i, hate, about, you, +, a, knight's, tale, , brokeback, mountain, =]
After removing Stop Words --> [10, things, hate, +, knight's, tale, , brokeback, mountain, =]
##############---- END ---- ##############
```

features	rawPrediction	probability	predictions
(10000,[1702,2007...	[-460.73982498224...	[1.0,2.8831832778...	0.0
(10000,[493,2007,...	[-664.19237642734...	[0.99999999970660...	0.0
(10000,[1871,4420...	[-301.86338122585...	[0.99999999984274...	0.0
(10000,[2007,2069...	[-520.19457361779...	[2.80856472431555...	1.0
(10000,[1083,1402...	[-154.19765901965...	[1.0,1.7444214330...	0.0
(10000,[1083,1402...	[-154.19765901965...	[1.0,1.7444214330...	0.0
(10000,[1083,1402...	[-154.19765901965...	[1.0,1.7444214330...	0.0

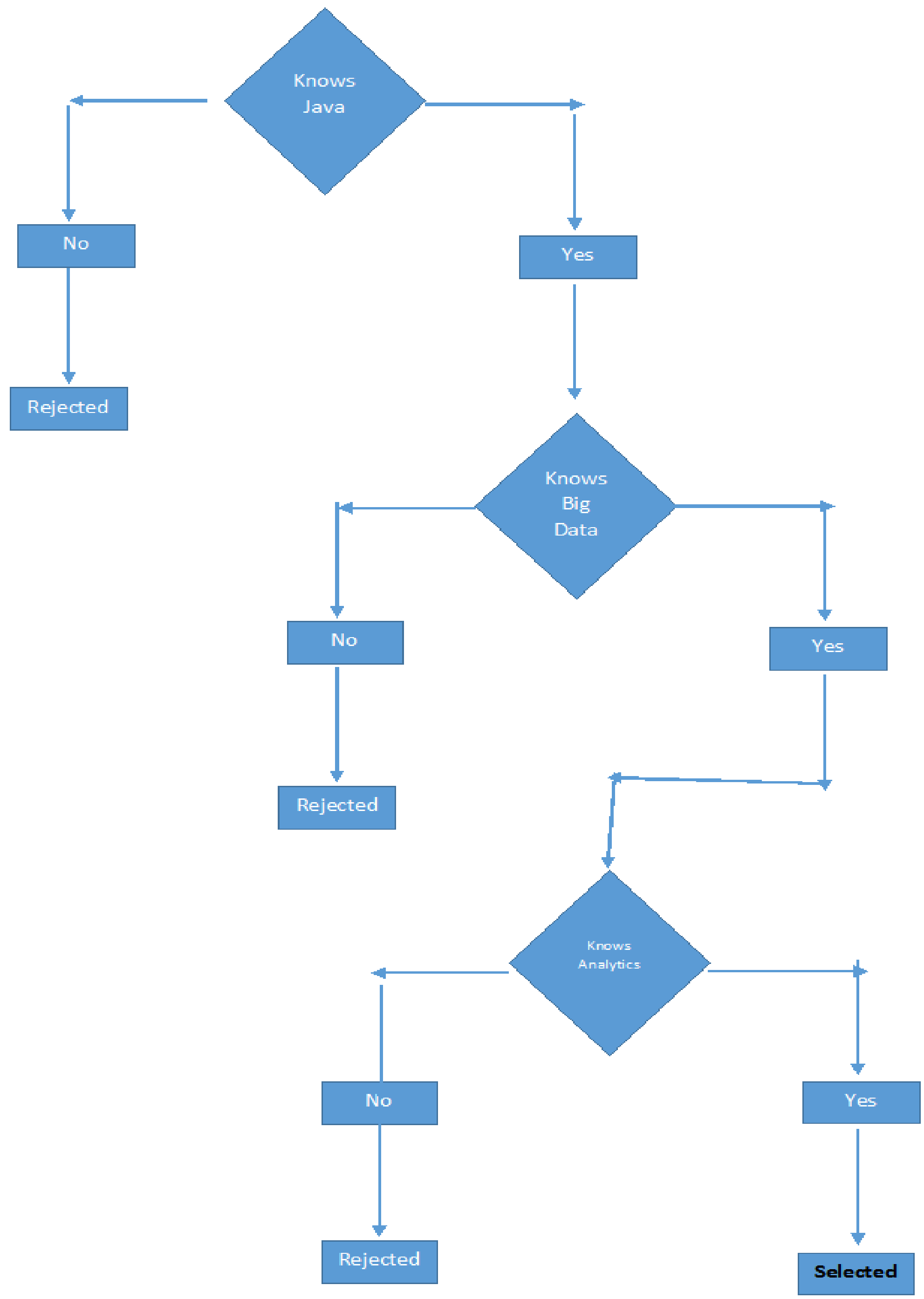
```
Accuracy = 0.9753086419753086
Test Error = 0.024691358024691357
```

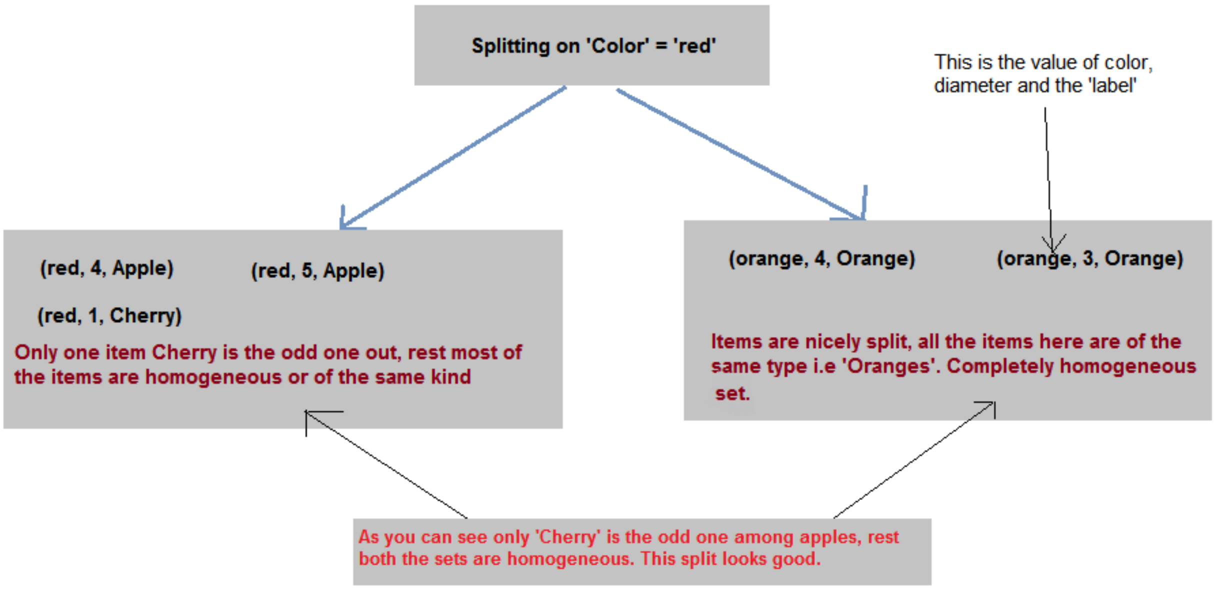


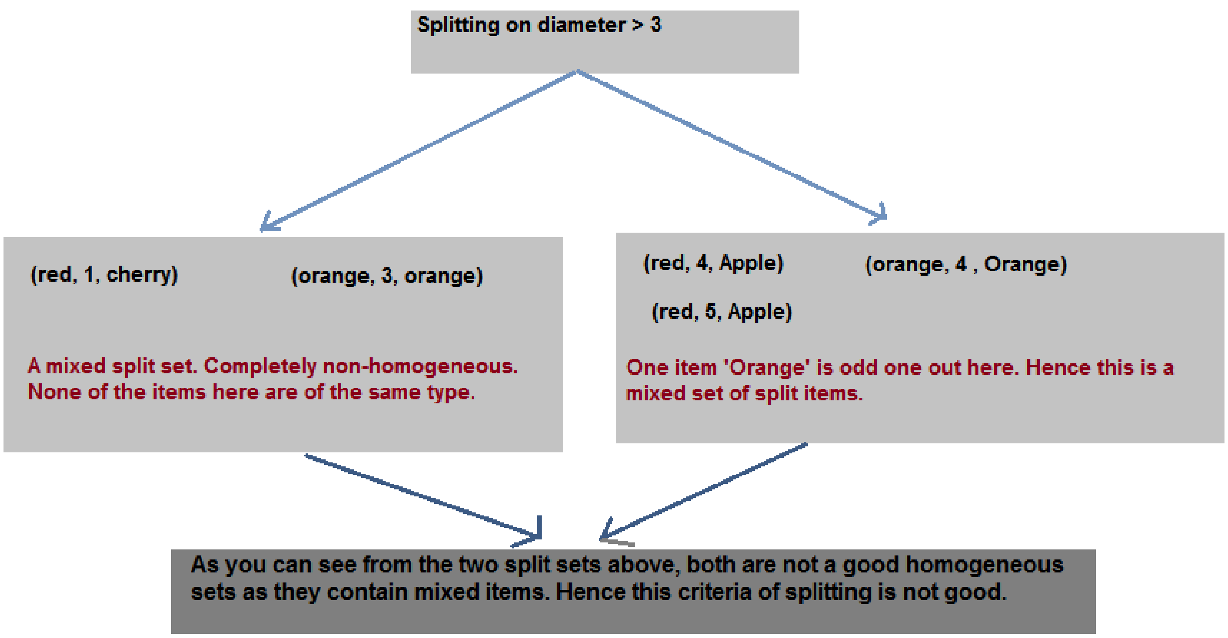

```
17/07/13 00:12:44 INFO Code
+-----+--------------------+--------------------+--------------------+-----------+
|label|               tweet|            features|       rawPrediction|predictions|
+-----+--------------------+--------------------+--------------------+-----------+
|  0.0|10 Things I Hate ...|(10000,[2007,2069...|[1.73011104770074...|        0.0|
|  0.0|A couple of very ...|(10000,[2007,3189...|[1.24141630788482...|        0.0|
|  0.0|A futile mission ...|(10000,[2345,3326...|[-0.2558815088181...|        1.0|
|  0.0|A mother in Georg...|(10000,[161,294,4...|[2.93919879710469...|        0.0|
|  0.0|AND BROKEBACK MOU...|(10000,[493,2007,...|[1.51404977982603...|        0.0|
|  0.0|After school I we...|(10000,[1871,4420...|[1.90088379349260...|        0.0|
|  0.0|After the festivi...|(10000,[437,2007,...|[1.99976959365757...|        0.0|
```

```
17/07/13 00:12:45 INFO DAGScheduler: Job 87
Accuracy = 0.9846796657381616
Test Error = 0.01532033426183843
17/07/13 00:12:45 INFO SparkContext: Invoki
```

Chapter 7: Decision Trees


Knows Java
No
Yes
Rejected
Knows Big Data
No
Yes
Rejected
Knows Analytics
No
Yes
Rejected
Selected





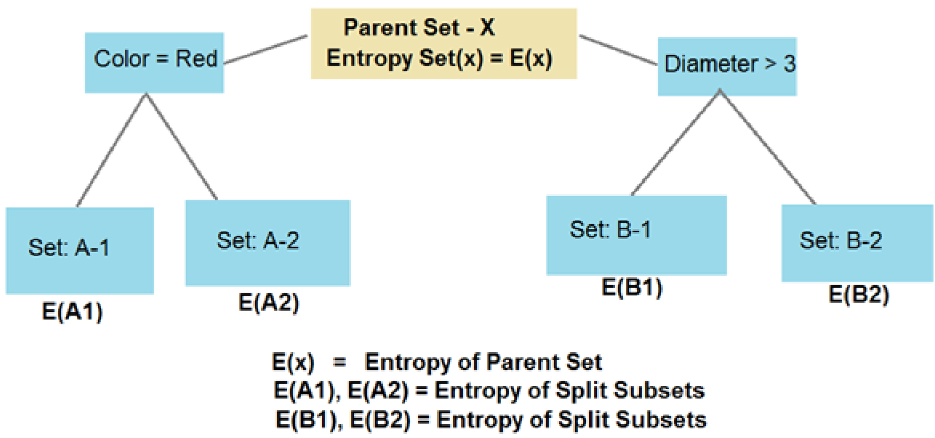
$$Gini\ Impurity = 1 - \sum_{i=1}^{k} P_i^2$$

$$Gini\ Impurity = 1 - \left[\left(\frac{3}{5}\right)^2 + \left(\frac{2}{5}\right)^2\right] = 0.48$$

$$\text{Net Gini of Split Sets} = \sum_{i}^{n} \frac{\textit{elements in set}}{\textit{Total Elements}} * \textit{Gini of Set}$$

$$\textbf{Entropy(X)} = -\sum_{i=1}^{n} \mathrm{P}(x_i) \log_b \mathrm{P}(x_i)$$

$$Entropy(fruits) = -\left[\left(\frac{3}{8}\right)\log_2\frac{3}{8} + \left(\frac{5}{8}\right)\log_2\frac{5}{8}\right] = 0.95$$



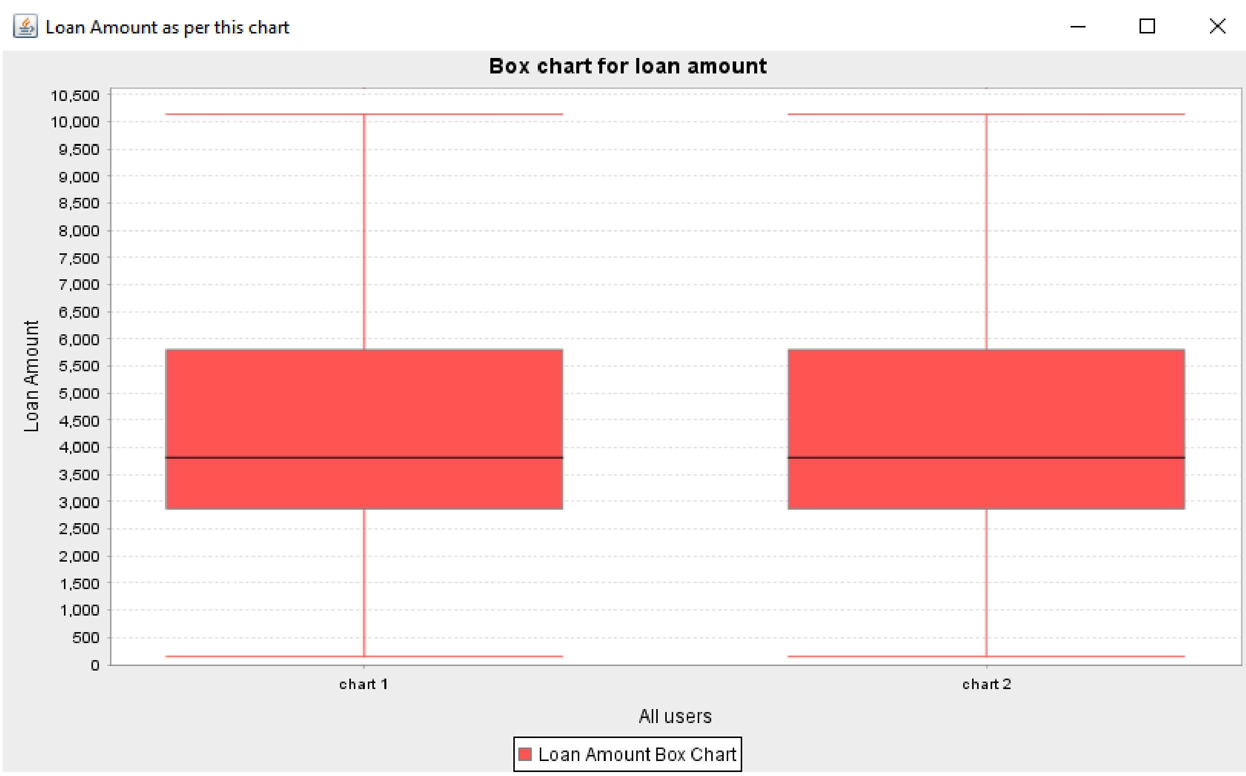
$$\text{Net Gini of Split Sets} = \sum_{i}^{n} \frac{\textit{elements in set}}{\textit{Total Elements}} * \textit{Gini of Set}$$

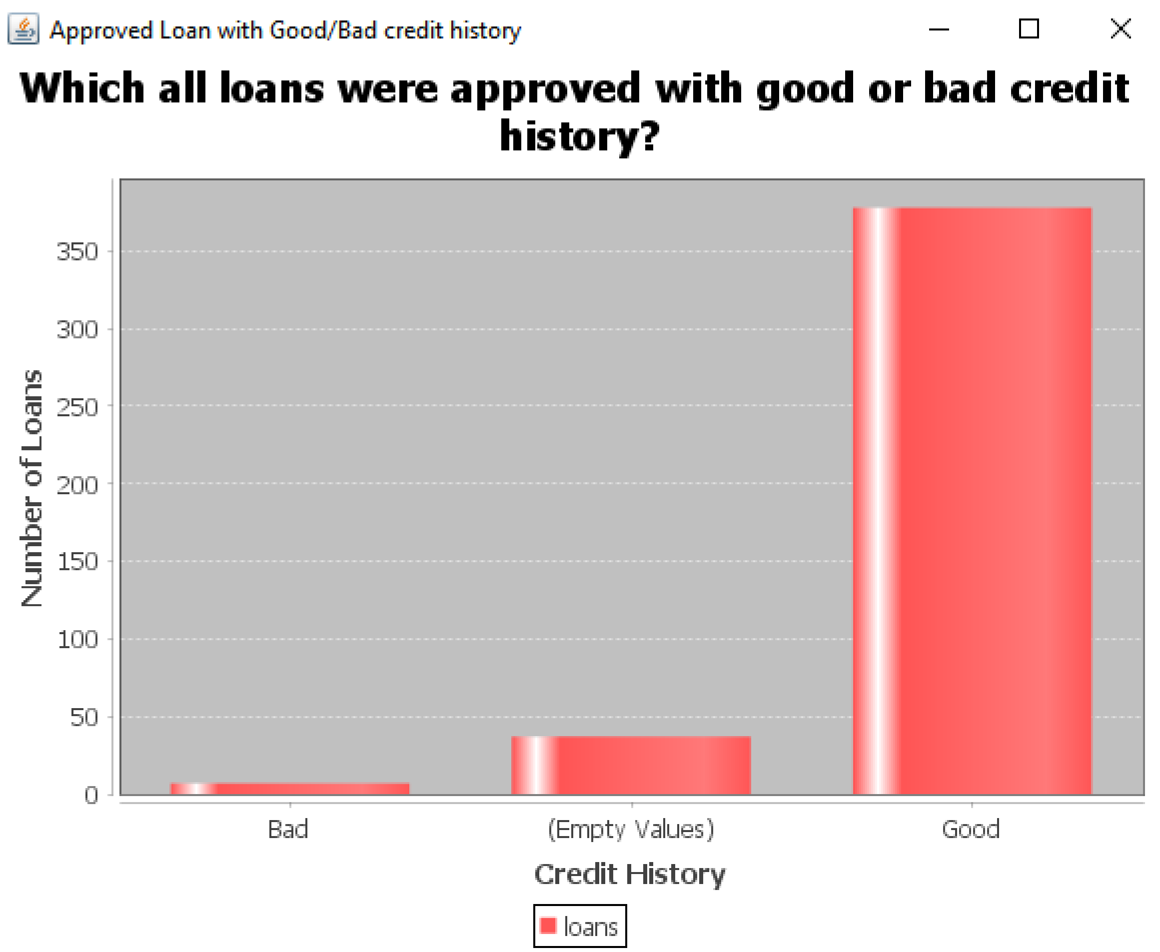
_c0	_c1	_c2	_c3	_c4	_c5	_c6	_c7	_c8	_c9	_c10	_c11	_c12
LP001002	Male	No	0	Graduate	No	5849	0	null	360	1	Urban	Y
LP001003	Male	Yes	1	Graduate	No	4583	1508	128	360	1	Rural	N
LP001005	Male	Yes	0	Graduate	Yes	3000	0	66	360	1	Urban	Y
LP001006	Male	Yes	0	Not Graduate	No	2583	2358	120	360	1	Urban	Y
LP001008	Male	No	0	Graduate	No	6000	0	141	360	1	Urban	Y
LP001011	Male	Yes	2	Graduate	Yes	5417	4196	267	360	1	Urban	Y
LP001013	Male	Yes	0	Not Graduate	No	2333	1516	95	360	1	Urban	Y
LP001014	Male	Yes	3+	Graduate	No	3036	2504	158	360	0	Semiurban	N
LP001018	Male	Yes	2	Graduate	No	4006	1526	168	360	1	Urban	Y

null	13
Female	112
Male	489

Count is much lesser than the total values, hence there are lots of missing values or null values here.

summary	_c0	_c1	_c2	_c3	_c4
count	614	601	611	599	614
mean	null	null	null	0.5547445255474452	null
stddev	null	null	null	0.7853289861674311	null
min	LP001002	Female	No	0	Graduate
max	LP002990	Male	Yes	3+	Not Graduate





_c10	count(1)
0	89
1	475

avgLoanAmount
146.41216216216216

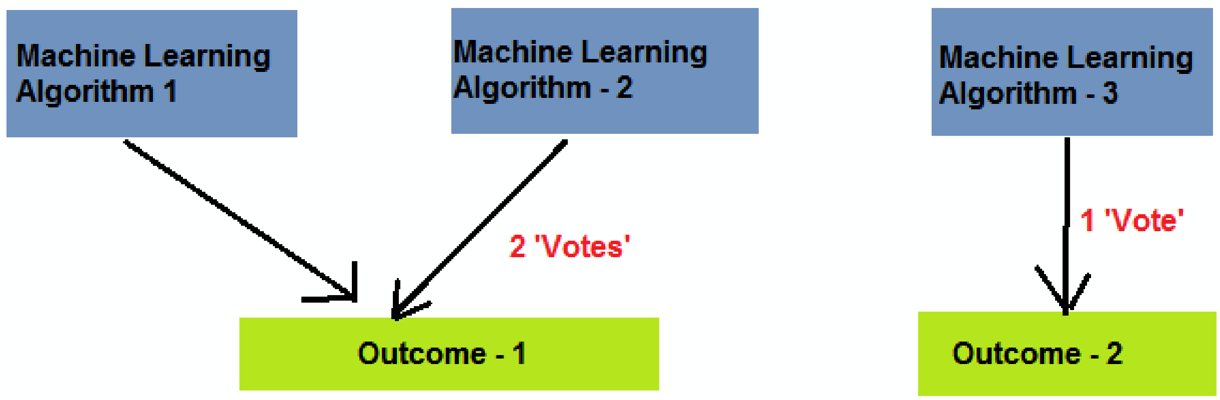
```
+-----------------+
|        avgIncome|
+-----------------+
|5403.459283387622|
+-----------------+
```

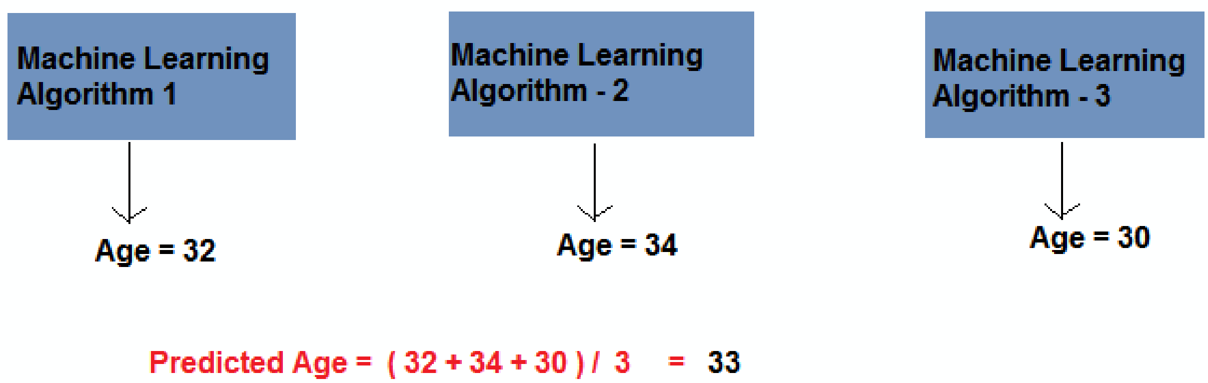
p_i

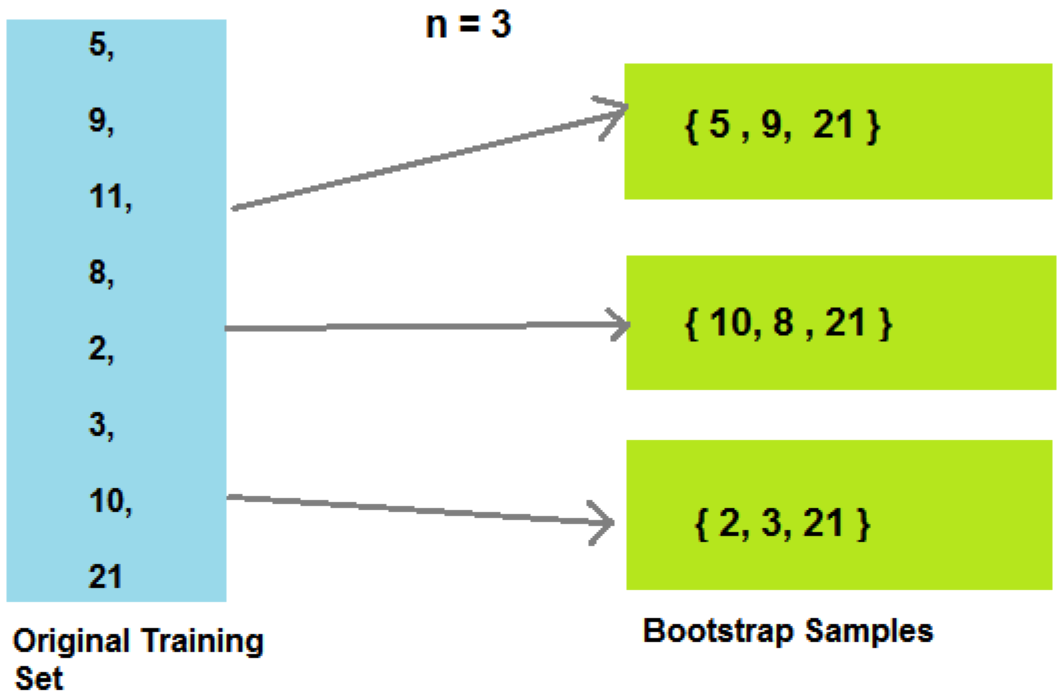
```
17/03/30 08:54:33 INFO CodeGenerator: Code gene
+--------------+------+--------------------+
|predictedLabel|result|            features|
+--------------+------+--------------------+
|           1.0|   0.0|    [98.0,210.0,1.0]|
|           0.0|   0.0|  [216.0,1025.0,1.0]|
|           1.0|   1.0|   [35.0,1442.0,1.0]|
|           0.0|   1.0|   [93.0,1800.0,0.0]|
|           1.0|   1.0|  [114.0,1853.0,1.0]|
|           1.0|   0.0|  [100.0,1928.0,1.0]|
|           1.0|   1.0|[146.412162162162...|
|           1.0|   0.0|  [113.0,2031.0,1.0]|
|           1.0|   0.0|  [101.0,2045.0,1.0]|
|           1.0|   0.0|   [88.0,2058.0,1.0]|
+--------------+------+--------------------+
only showing top 10 rows
```

```
Accuracy = 0.9753086419753086
Test Error = 0.024691358024691357
```

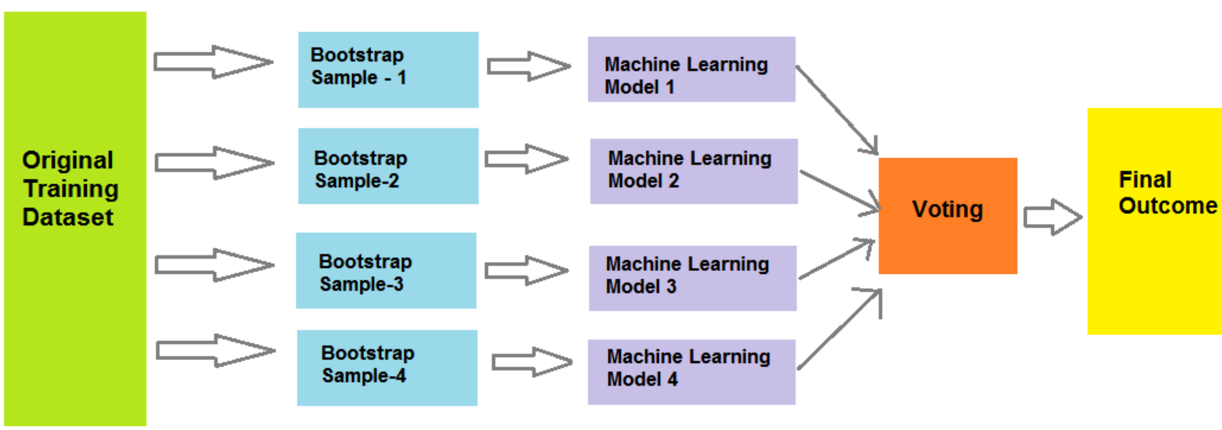
Chapter 8: Ensembling on Big Data



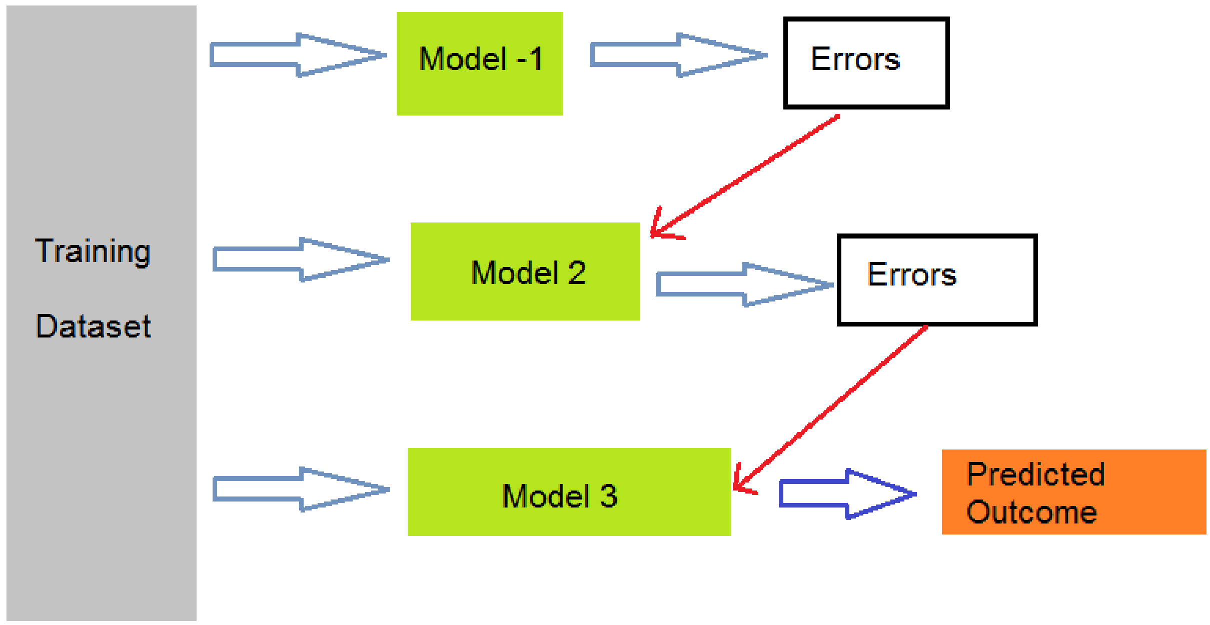




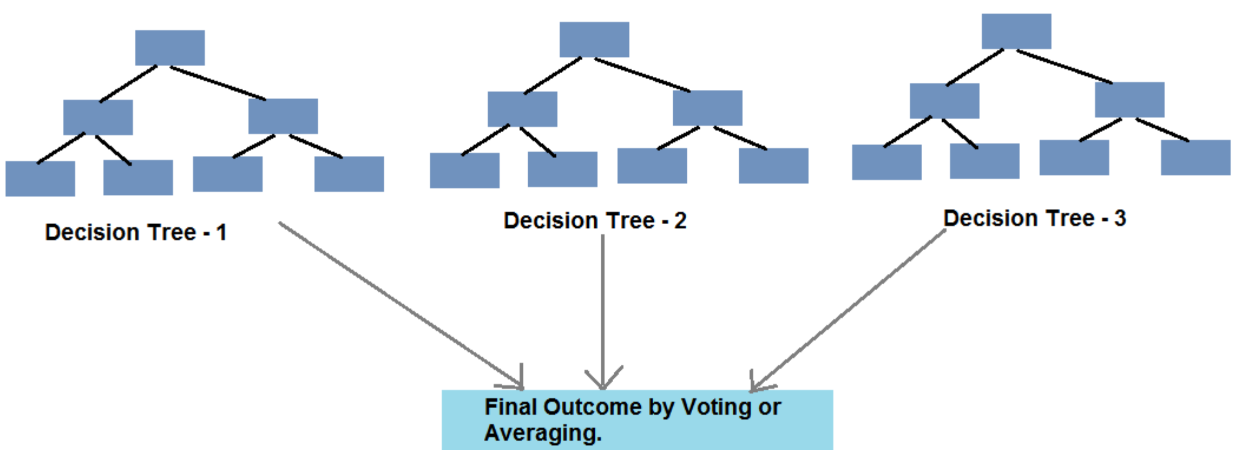
3 bootstrap samples from original training set.

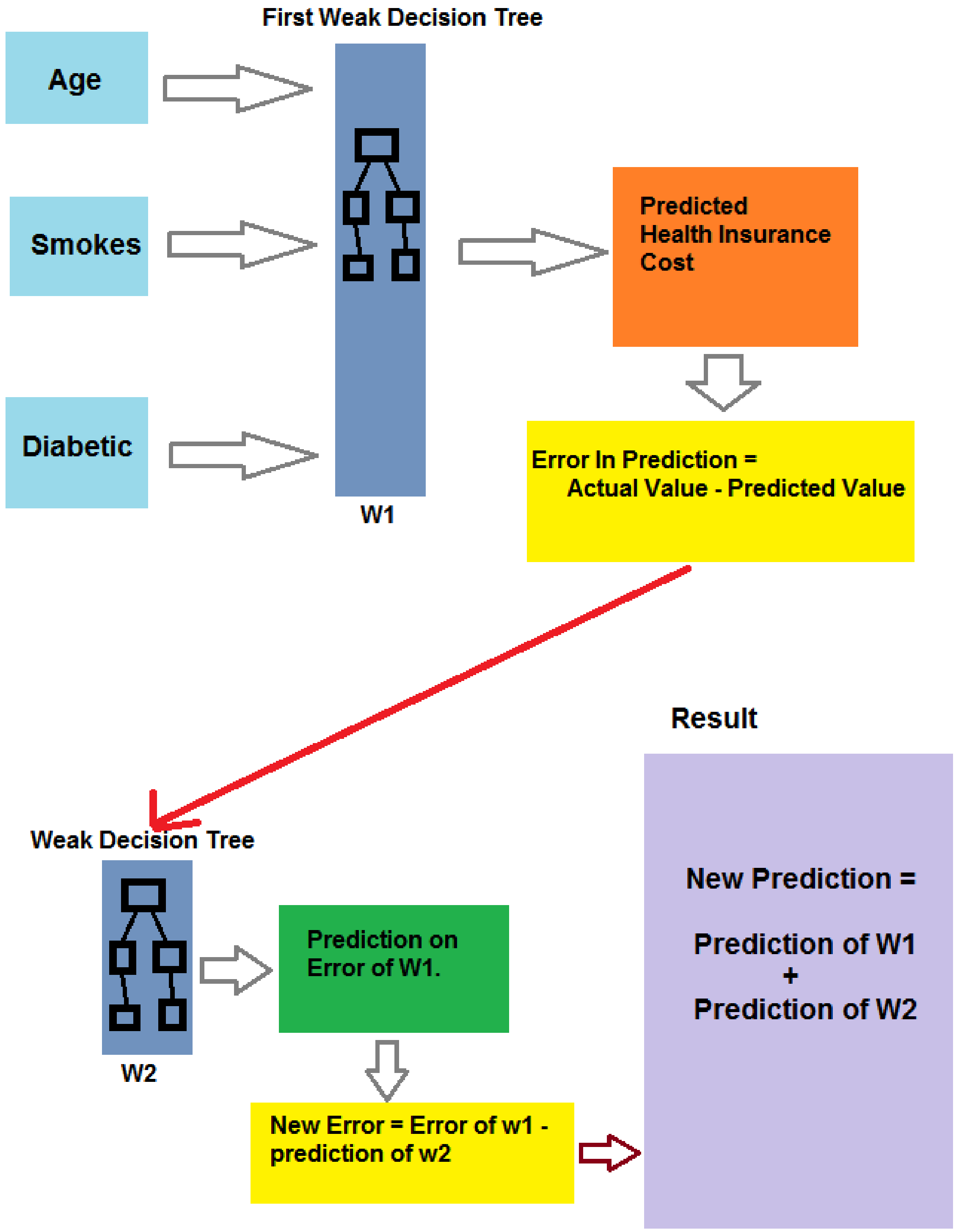


Bagging


Training
Dataset
Model -1
Errors
Model 2
Errors
Model 3
Predicted Outcome

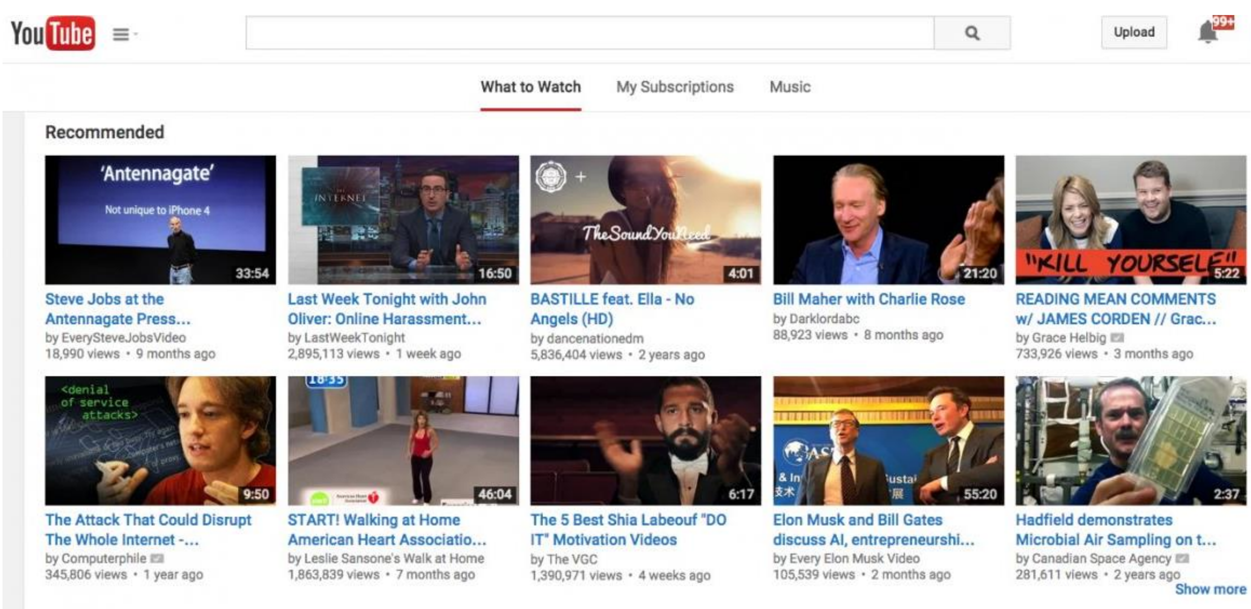
Boosting

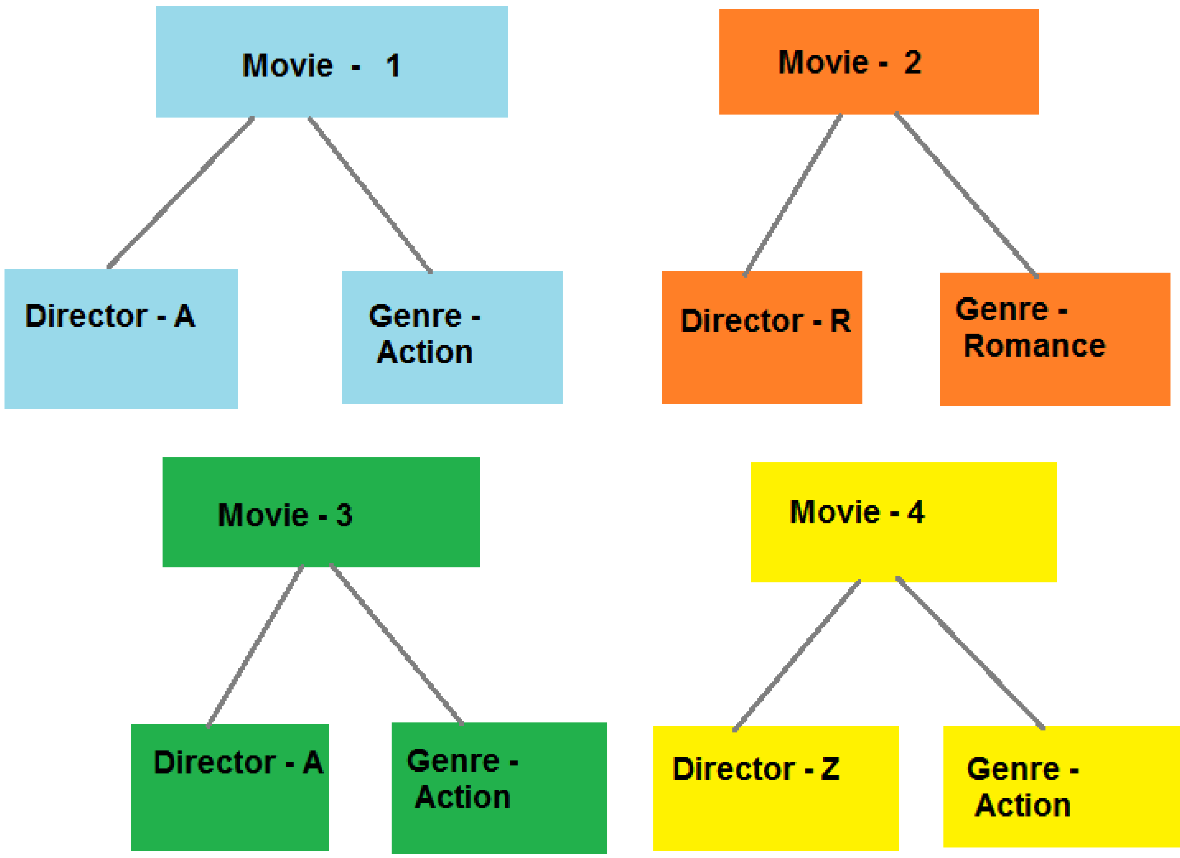

Decision Tree - 1
Decision Tree - 2
Decision Tree - 3
Final Outcome by Voting or Averaging.

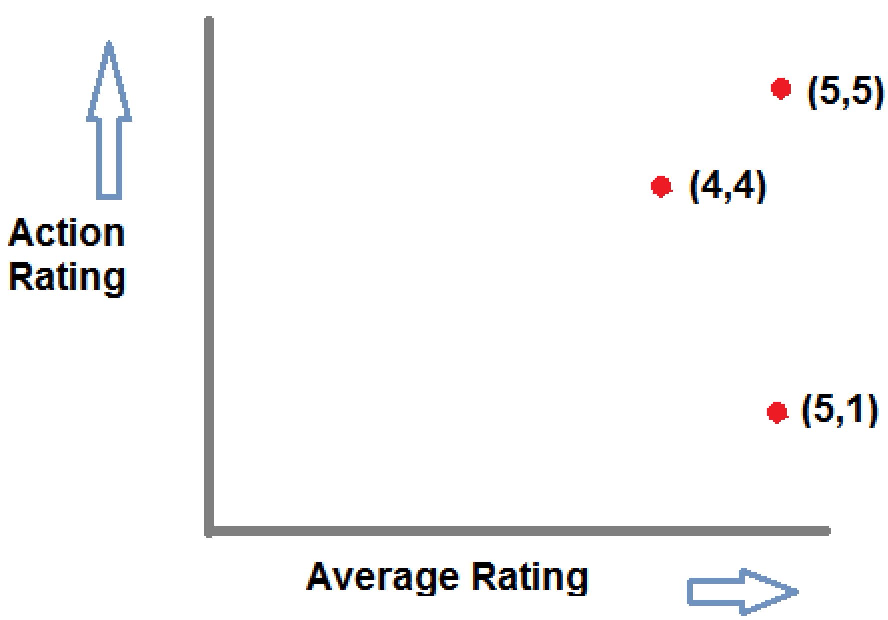

First Weak Decision Tree
Age
Smokes
Diabetic
W1
Predicted Health Insurance Cost
Error In Prediction = Actual Value - Predicted Value
Result
Weak Decision Tree
W2
Prediction on Error of W1.
New Error = Error of w1 - prediction of w2
New Prediction =
Prediction of W1
+
Prediction of W2

$$y = F(x) + error$$

Chapter 9: Recommendation Systems



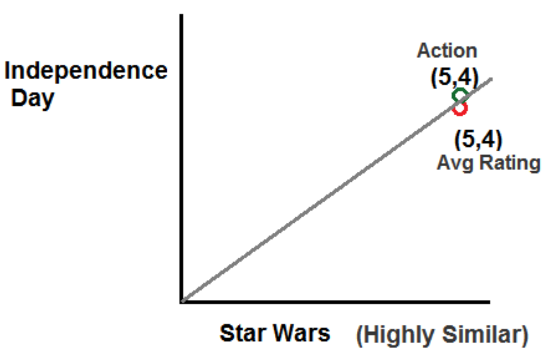

Movie - 1
Director - A
Genre - Action
Movie - 2
Director - R
Genre - Romance
Movie - 3
Director - A
Genre - Action
Movie - 4
Director - Z
Genre - Action

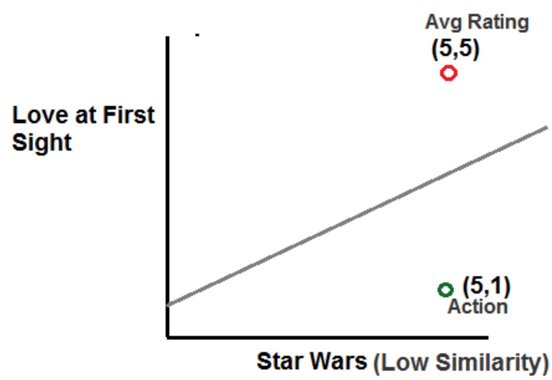

(5,5)
(4,4)
(5,1)
Action Rating
Average Rating

$$\text{Euclidean Distance between points } (x1, y1) \text{ and } (x2, y2)$$
$$= \sqrt{(x1 - x2)^2 + (y1 - y2)^2}$$

$$\sqrt{(5-4)^2 + (5-4)^2} = 1.4$$

$$\sqrt{(5-5)^2 + (5-1)^2} = 4$$





$$r = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{\sqrt{\left[n\sum x^2 - \left(\sum x\right)^2\right]\left[n\sum y^2 - \left(\sum y\right)^2\right]}}$$

```
17/05/05 08:50:54 INFO CodeGenerator: Code generat
+----+-------+------+---------+------+
|like|movieId|rating|timestamp|userId|
+----+-------+------+---------+------+
|   1|     50|   5.0|881250949|     0|
|   1|    172|   5.0|881250949|     0|
|   0|    133|   1.0|881250949|     0|
|   0|    242|   3.0|881250949|   196|
|   0|    302|   3.0|891717742|   186|
|   0|    377|   1.0|878887116|    22|
|   0|     51|   2.0|880606923|   244|
|   0|    346|   1.0|886397596|   166|
```

l	movieTitle	musical	mystery	releaseDate	romance	sciFi	thriller	war	western	averageRating
.	Hoodlum (1997)	0	0	22-Aug-1997	0	0	0	0	0	2.9315068493150687
.	Ice Storm, The (1...	0	0	01-Jan-1997	0	0	0	0	0	3.643678160919540З
.	It's a Wonderful ...	0	0	01-Jan-1946	0	0	0	0	0	4.121212121212121
.	Heavenly Creature...	0	0	01-Jan-1994	0	0	1	0	0	3.671428571428571З
.	Hunchback of Notr...	1	0	21-Jun-1996	0	0	0	0	0	3.377952755905512
.	American Presiden...	0	0	01-Jan-1995	1	0	0	0	0	3.6280487804878048
.	Congo (1995)	0	1	01-Jan-1995	0	1	0	0	0	2.4523809523809526
.	Preacher's Wife, ...	0	0	13-Dec-1996	0	0	0	0	0	2.926470588235294

movieId1	movieTitle1	action1	adventure1
299	Hoodlum (1997)	0	0
299	Hoodlum (1997)	0	0
299	Hoodlum (1997)	0	0
299	Hoodlum (1997)	0	0
299	Hoodlum (1997)	0	0
299	Hoodlum (1997)	0	0
299	Hoodlum (1997)	0	0
299	Hoodlum (1997)	0	0
305	Ice Storm, The (1...	0	0
305	Ice Storm, The (1...	0	0
305	Ice Storm, The (1...	0	0
305	Ice Storm, The (1...	0	0

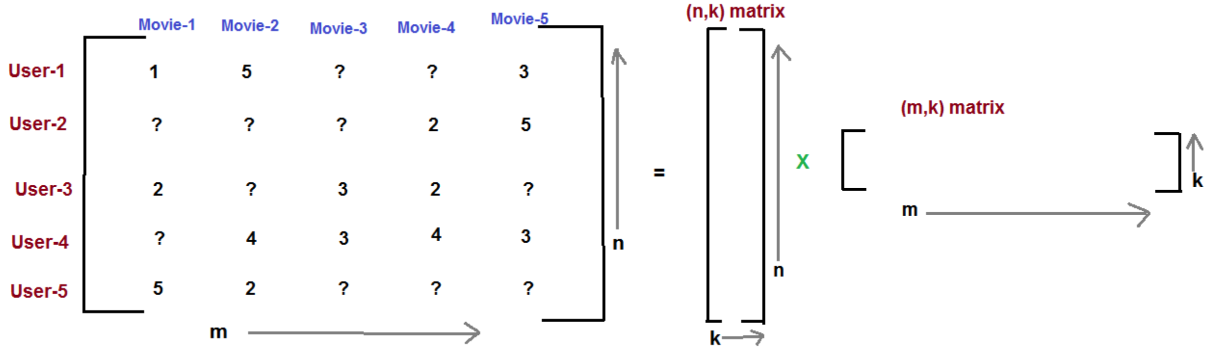
movieTitle2	action2	adventure2	animation2	c
Ice Storm, The (1...	0	0	0	
It's a Wonderful ...	0	0	0	
Heavenly Creature...	0	0	0	
Hunchback of Notr...	0	0	1	
American Presiden...	0	0	0	
Congo (1995)	1	1	0	
Preacher's Wife, ...	0	0	0	
Associate, The (1...	0	0	0	
Hoodlum (1997)	0	0	0	
It's a Wonderful ...	0	0	0	
Heavenly Creature...	0	0	0	

euclidDist	movieId1	movieId2	movieTitle1	movieTitle2
1.0021449993740907	1	95	Toy Story (1995)	Aladdin (1992)
1.003062211735072	1	969	Toy Story (1995)	Winnie the Pooh a...
1.0208065527493175	1	404	Toy Story (1995)	Pinocchio (1940)
1.0256053961344038	1	189	Toy Story (1995)	Grand Day Out, A ...
1.03216473791695	1	422	Toy Story (1995)	Aladdin and the K...
1.140813842164051	1	625	Toy Story (1995)	Sword in the Ston...
1.159952147889726	1	169	Toy Story (1995)	Wrong Trousers, T...
1.2024751613136604	1	477	Toy Story (1995)	Matilda (1996)
1.2385620296700004	1	946	Toy Story (1995)	Fox and the Hound...

<u>Similar Users, User-A and User-B with the moviews they have watched</u>

$$\begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 3 & 4 & 3 \\ 2 & 4 & 2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 2 \end{bmatrix} * \begin{bmatrix} 1 & 2 & 3 \end{bmatrix}$$

$$\begin{array}{cc} & \begin{array}{ccccc} \text{Movie-1} & \text{Movie-2} & \text{Movie-3} & \text{Movie-4} & \text{Movie-5} \end{array} \\ \begin{array}{c} \text{User-1} \\ \text{User-2} \\ \text{User-3} \\ \text{User-4} \\ \text{User-5} \end{array} & \begin{bmatrix} 1 & 5 & ? & ? & 3 \\ ? & ? & ? & 2 & 5 \\ 2 & ? & 3 & 2 & ? \\ ? & 4 & 3 & 4 & 3 \\ 5 & 2 & ? & ? & ? \end{bmatrix} \end{array}$$



movieId	rating	timestamp	userId	prediction
148	1.0	875326138	633	3.8584247
148	3.0	886106165	271	2.7496443
148	3.0	878150506	606	3.4842434
148	4.0	890117028	236	3.028388
148	3.0	876348140	601	1.1929333
148	3.0	888104154	224	4.195447
148	2.0	887160606	896	2.6896698
148	4.0	882824325	178	3.8263626
148	3.0	887740788	308	2.5620613
148	4.0	880387474	923	3.4326193
148	3.0	889490499	120	2.3387685
148	2.0	877226047	430	2.688941
148	2.0	877383934	92	2.5610256
148	4.0	878854729	447	3.031723

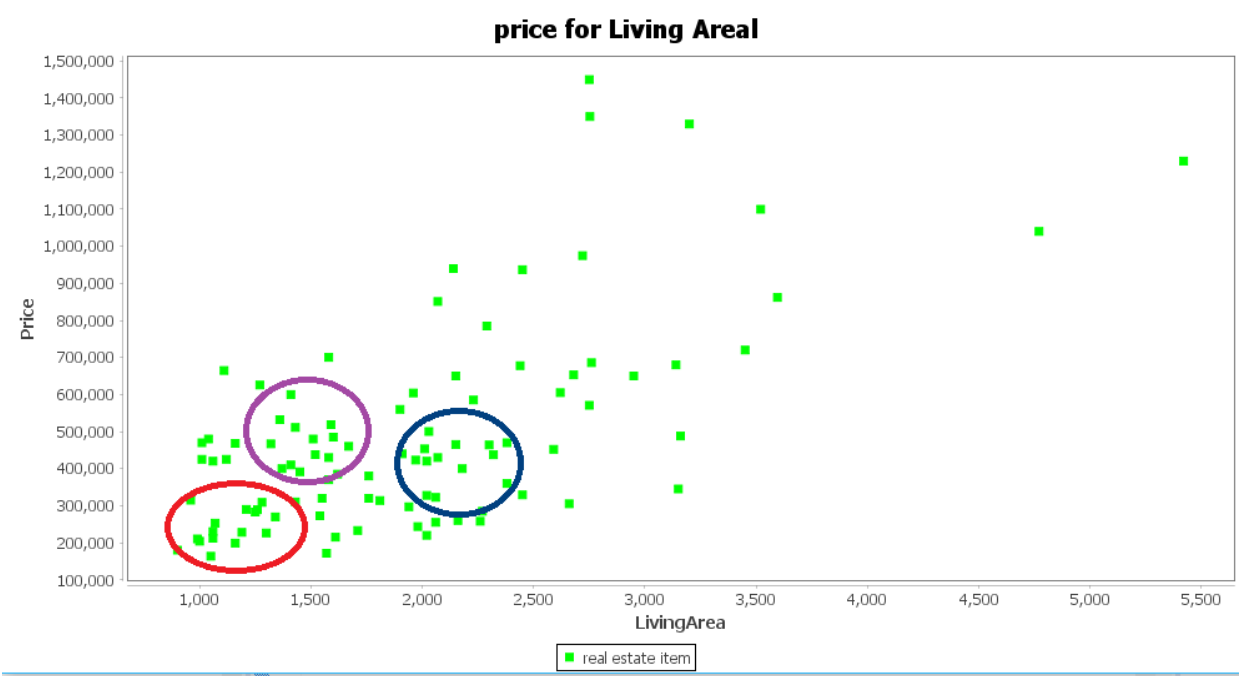
Predictions for User Ratings

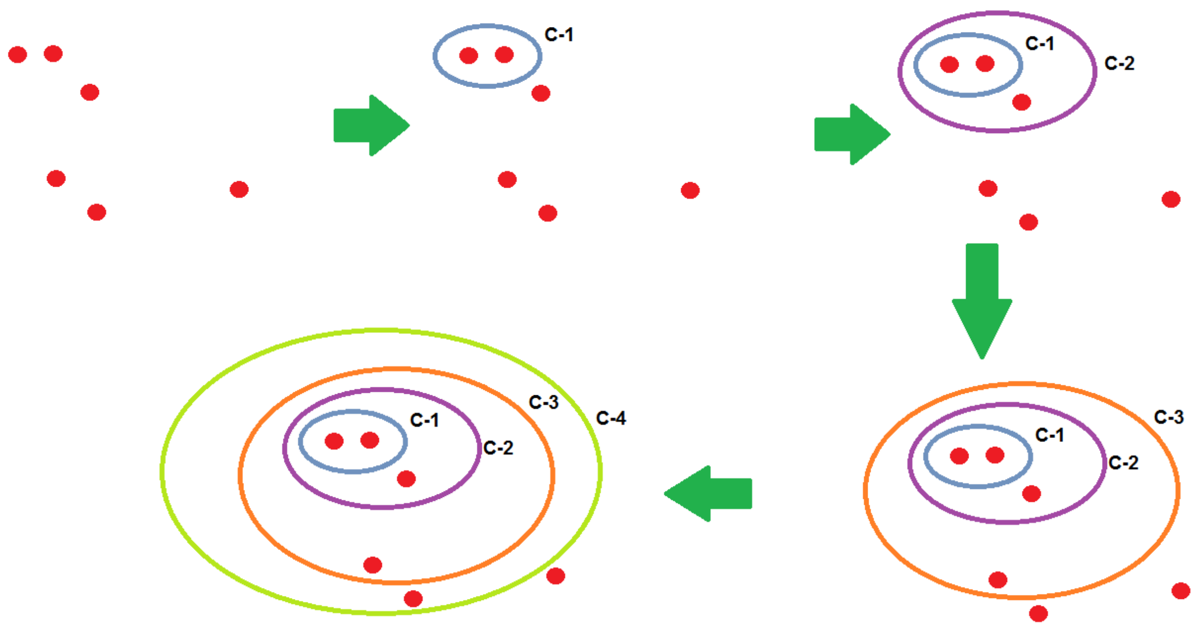
movieTitle	like	movieId	rating	timestamp	userId	prediction
Murder in the Fir...	0	939	4.0	877212045	633	4.1661263
Terminator 2: Jud...	0	96	4.0	875324997	633	4.10481
Home Alone (1990)	0	94	4.0	877211684	633	3.7664657
Fugitive, The (1993)	0	79	5.0	875325128	633	3.6715062
Schindler's List ...	0	318	4.0	875324813	633	3.6350746
Ben-Hur (1959)	0	526	4.0	877212250	633	3.5871043
Jaws (1975)	0	234	4.0	877212594	633	3.5124974
Good, The Bad and...	0	177	3.0	875325654	633	3.3321676
Scream (1996)	0	288	2.0	875324233	633	3.3247242
Supercop (1992)	0	128	3.0	875325225	633	3.2833974
Glory (1989)	0	651	3.0	877212283	633	2.9536574

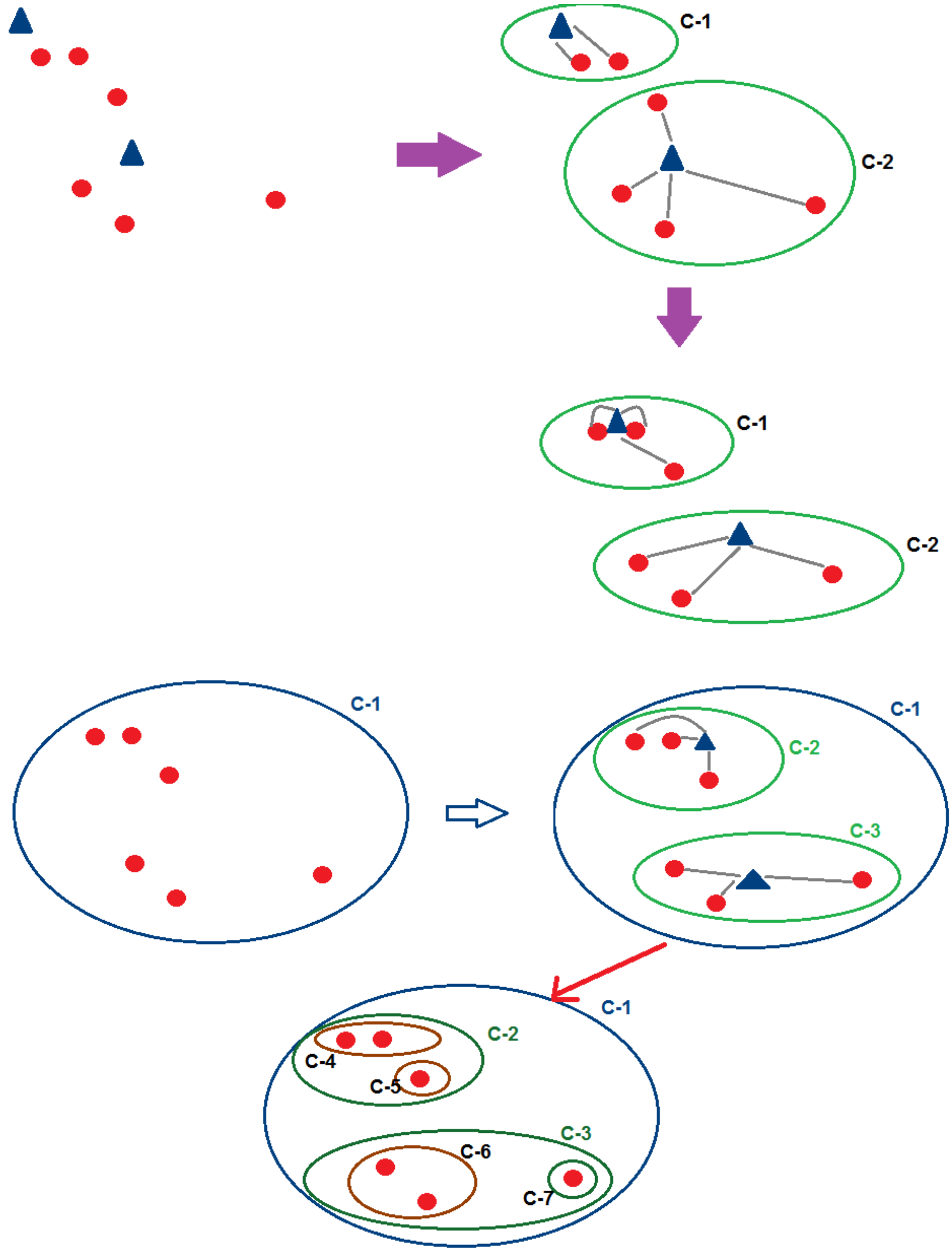
Predicted Ratings

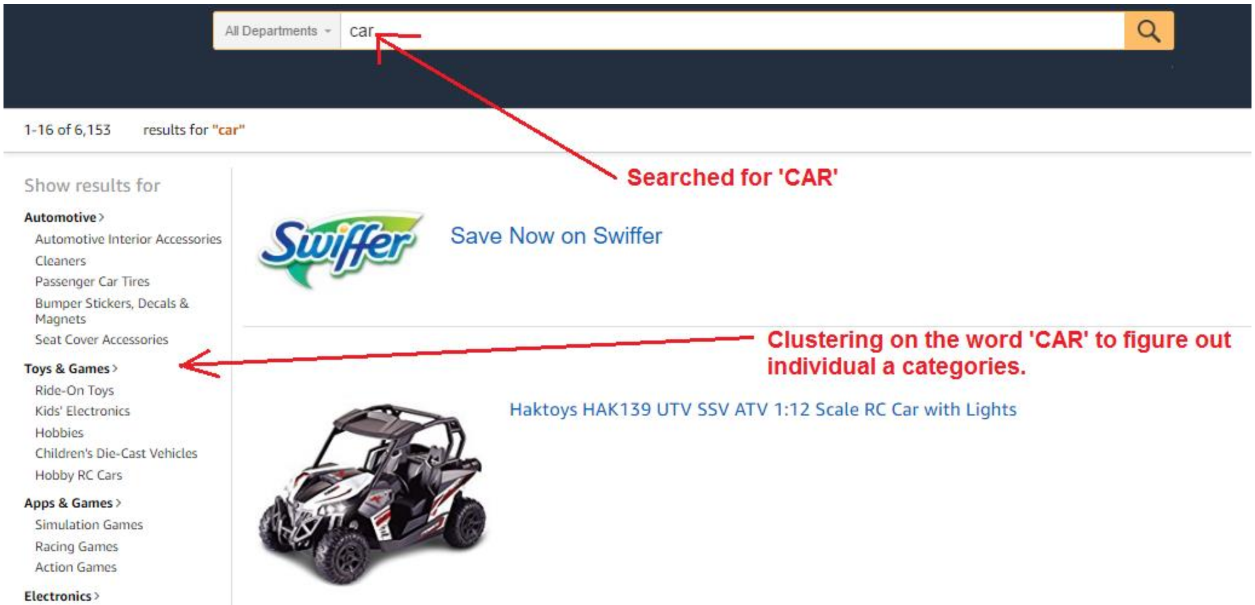
Predicted Movies

Chapter 10: Clustering and Customer Segmentation on Big Data





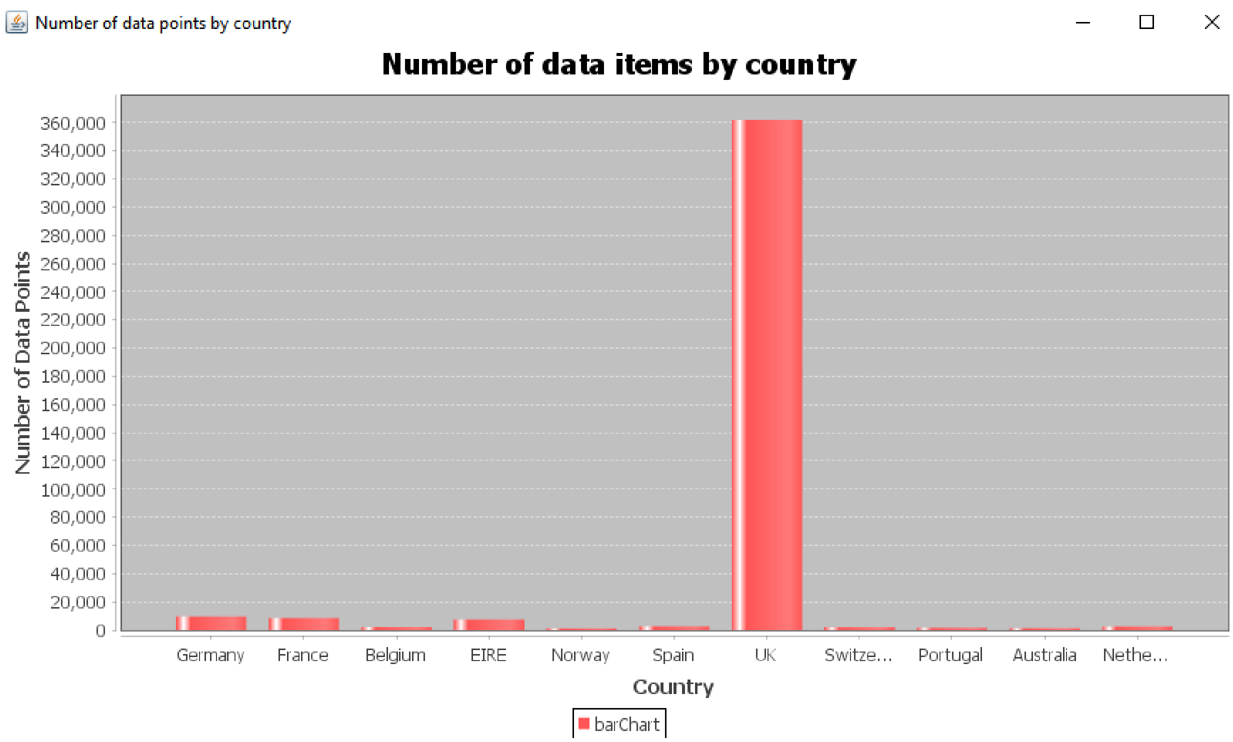

C-1
C-2
C-1
C-2
C-1
C-1
C-2
C-3
C-1
C-2
C-4
C-5
C-3
C-6
C-7



17/05/16 08:50:06 INFO CodeGenerator: Code generated in 32.169375 ms

_c0	_c1	_c2	_c3	_c4	_c5	_c6	_c7
536365	85123A	WHITE HANGING HEA...	6	12/1/2010 8:26	2.55	17850	United Kingdom
536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850	United Kingdom
536365	84406B	CREAM CUPID HEART...	8	12/1/2010 8:26	2.75	17850	United Kingdom
536365	84029G	KNITTED UNION FLA...	6	12/1/2010 8:26	3.39	17850	United Kingdom
536365	84029E	RED WOOLLY HOTTIE...	6	12/1/2010 8:26	3.39	17850	United Kingdom
536365	22752	SET 7 BABUSHKA NE...	2	12/1/2010 8:26	7.65	17850	United Kingdom
536365	21730	GLASS STAR FROSTE...	6	12/1/2010 8:26	4.25	17850	United Kingdom
536366	22633	HAND WARMER UNION...	6	12/1/2010 8:28	1.85	17850	United Kingdom
536366	22632	HAND WARMER RED P...	6	12/1/2010 8:28	1.85	17850	United Kingdom
536367	84879	ASSORTED COLOUR B...	32	12/1/2010 8:34	1.69	13047	United Kingdom
536367	22745	POPPY'S PLAYHOUSE...	6	12/1/2010 8:34	2.1	13047	United Kingdom
536367	22748	POPPY'S PLAYHOUSE...	6	12/1/2010 8:34	2.1	13047	United Kingdom
536367	22749	FELTCRAFT PRINCES...	8	12/1/2010 8:34	3.75	13047	United Kingdom

country	cnt
Germany	9495
France	8491
Belgium	2069
EIRE	7485
Norway	1086
Spain	2533
UK	361878
Switzerland	1877
Portugal	1480
Australia	1259
Netherlands	2371



summary	unitPrice	quantity
count	361878	361878
mean	3.256006897353581	11.077028722387103
stddev	70.6547311689391	263.1292655989739
min	0.0	-80995
max	38970.0	80995

```
+----------+-------+---------+--------+
|customerID|recency|frequency|spending|
+----------+-------+---------+--------+
|     12847|    410|       91|     871|
|     13192|    483|       63|     922|
|     13282|    406|       40|    1055|
|     13610|    400|      228|    1129|
|     13772|    421|      177|    1145|
|     13865|    446|       30|     508|
|     14157|    407|       49|     401|
|     14204|    390|       44|     163|
|     14887|    467|        6|    1862|
|     15269|    411|        2|     409|
|     15271|    395|      275|    2507|
|     15555|    400|      925|    4784|
|     15574|    565|      168|     713|
|     15634|    405|       15|     246|
|     16250|    649|       24|     392|
|     16504|    413|       86|     486|
|     17427|    459|        2|     101|
|     17506|    458|       16|     297|
```

```
+----------+-------+---------+--------+--------------------+--------------------+
|customerID|recency|frequency|spending|            features|         normFeatures|
+----------+-------+---------+--------+--------------------+--------------------+
|     12847|    410|       91|     871|  [410.0,91.0,871.0]|[0.42400691750858...|
|     13192|    483|       63|     922|  [483.0,63.0,922.0]|[0.46319519866007...|
|     13282|    406|       40|    1055| [406.0,40.0,1055.0]|[0.35893227147767...|
|     13610|    400|      228|    1129|[400.0,228.0,1129.0]|[0.32806452843731...|
|     13772|    421|      177|    1145|[421.0,177.0,1145.0]|[0.34152159327630...|
|     13865|    446|       30|     508|  [446.0,30.0,508.0]|[0.65911170105826...|
|     14157|    407|       49|     401|  [407.0,49.0,401.0]|[0.70973271727211...|
```

```
17/05/17 17:02:24 INFO CodeGenerator: Code generated in 11.46881 ms
+----------+-------+---------+--------+--------------------+--------------------+----------+
|customerID|recency|frequency|spending|            features|         normFeatures|prediction|
+----------+-------+---------+--------+--------------------+--------------------+----------+
|     12847|    410|       91|     871|  [410.0,91.0,871.0]|[0.42400691750858...|         3|
|     13192|    483|       63|     922|  [483.0,63.0,922.0]|[0.46319519866007...|         3|
|     13282|    406|       40|    1055| [406.0,40.0,1055.0]|[0.35893227147767...|         3|
|     13610|    400|      228|    1129|[400.0,228.0,1129.0]|[0.32806452843731...|         3|
|     13772|    421|      177|    1145|[421.0,177.0,1145.0]|[0.34152159327630...|         3|
|     13865|    446|       30|     508|  [446.0,30.0,508.0]|[0.65911170105826...|         1|
|     14157|    407|       49|     401|  [407.0,49.0,401.0]|[0.70973271727211...|         1|
|     14204|    390|       44|     163|  [390.0,44.0,163.0]|[0.91769786703235...|         4|
|     14887|    467|        6|    1862|  [467.0,6.0,1862.0]|[0.24326978761296...|         2|
```

prediction	count(1)
1	668
3	822
4	716
2	1133
0	611

17/05/17 22:45:18 INFO DAGScheduler: Job 43 finished: show at KMeansClustering.java:139, took 0.127328 s

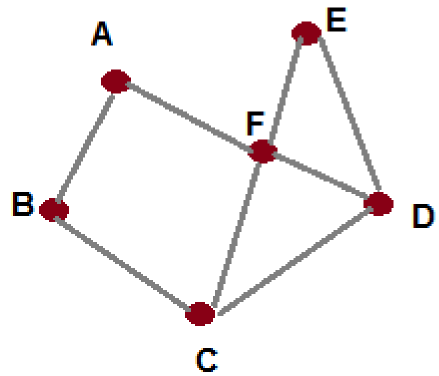
customerID	recency	frequency	spending	features	normFeatures
14887	467	6	1862	[467.0,6.0,1862.0]	[0.24326978761296...
15271	395	275	2507	[395.0,275.0,2507.0]	[0.15473311361757...
15555	400	925	4784	[400.0,925.0,4784.0]	[0.08181639081528...
17686	395	286	5786	[395.0,286.0,5786.0]	[0.06802703458963...
13985	392	353	7072	[392.0,353.0,7072.0]	[0.05527629909846...
15947	470	29	1709	[470.0,29.0,1709.0]	[0.26513411031107...
16549	398	981	4200	[398.0,981.0,4200.0]	[0.09188778871879...
17757	389	742	5645	[389.0,742.0,5645.0]	[0.06816393548275...
13107	432	60	1526	[432.0,60.0,1526.0]	[0.27219383443550...
14525	396	298	4242	[396.0,298.0,4242.0]	[0.09272152767339...
15478	428	46	1449	[428.0,46.0,1449.0]	[0.28314576816739...
16303	413	167	5313	[413.0,167.0,5313.0]	[0.07746203711191...
18283	391	756	2140	[391.0,756.0,2140.0]	[0.16977522070462...

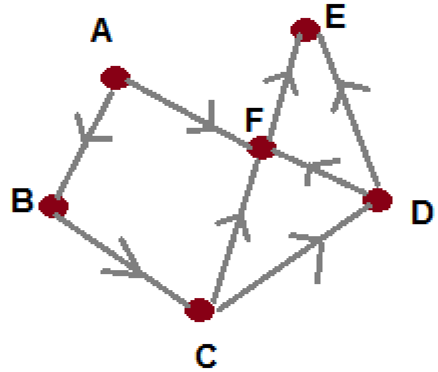
prediction	count(1)
1	601
3	833
4	926
2	917
0	673

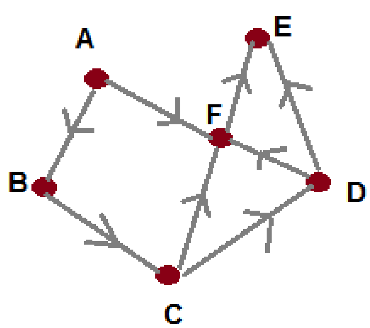
17/05/18 00:05:12 INFO DAGScheduler: Job 79 finished: show at BisectingKmeansClustering.java::

customerID	recency	frequency	spending	features	normFeatures
12847	410	91	871	[410.0,91.0,871.0]	[0.42400691750858...
13192	483	63	922	[483.0,63.0,922.0]	[0.46319519866007...
13282	406	40	1055	[406.0,40.0,1055.0]	[0.35893227147767...
13865	446	30	508	[446.0,30.0,508.0]	[0.65911170105826...
15574	565	168	713	[565.0,168.0,713.0]	[0.61074193335466...
16504	413	86	486	[413.0,86.0,486.0]	[0.64174937053506...
15539	395	41	539	[395.0,41.0,539.0]	[0.58999467069230...
16027	479	17	851	[479.0,17.0,851.0]	[0.49043000426939...
16340	495	153	563	[495.0,153.0,563.0]	[0.64696044239096...
13122	482	55	929	[482.0,55.0,929.0]	[0.45990565858249...

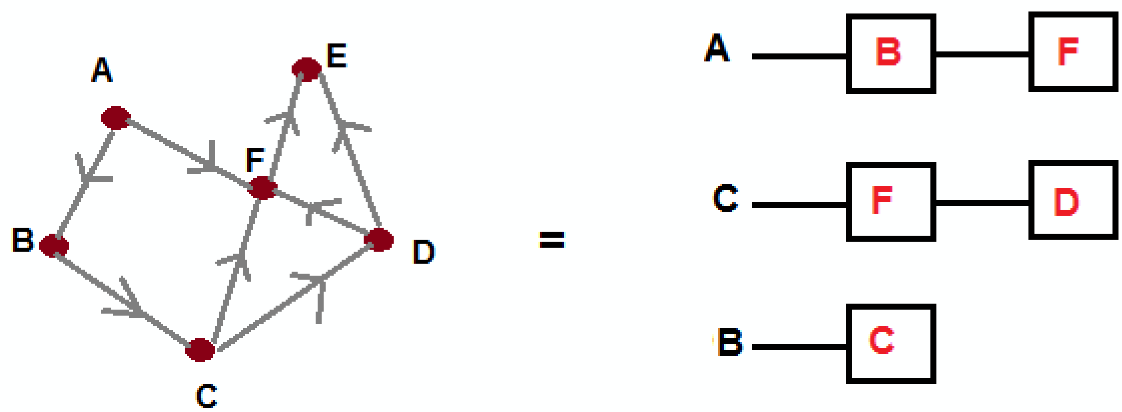
Chapter 11: Massive Graphs on Big Data

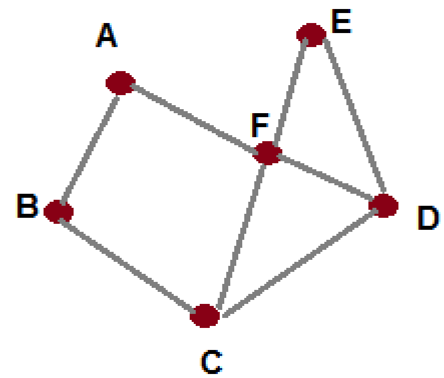


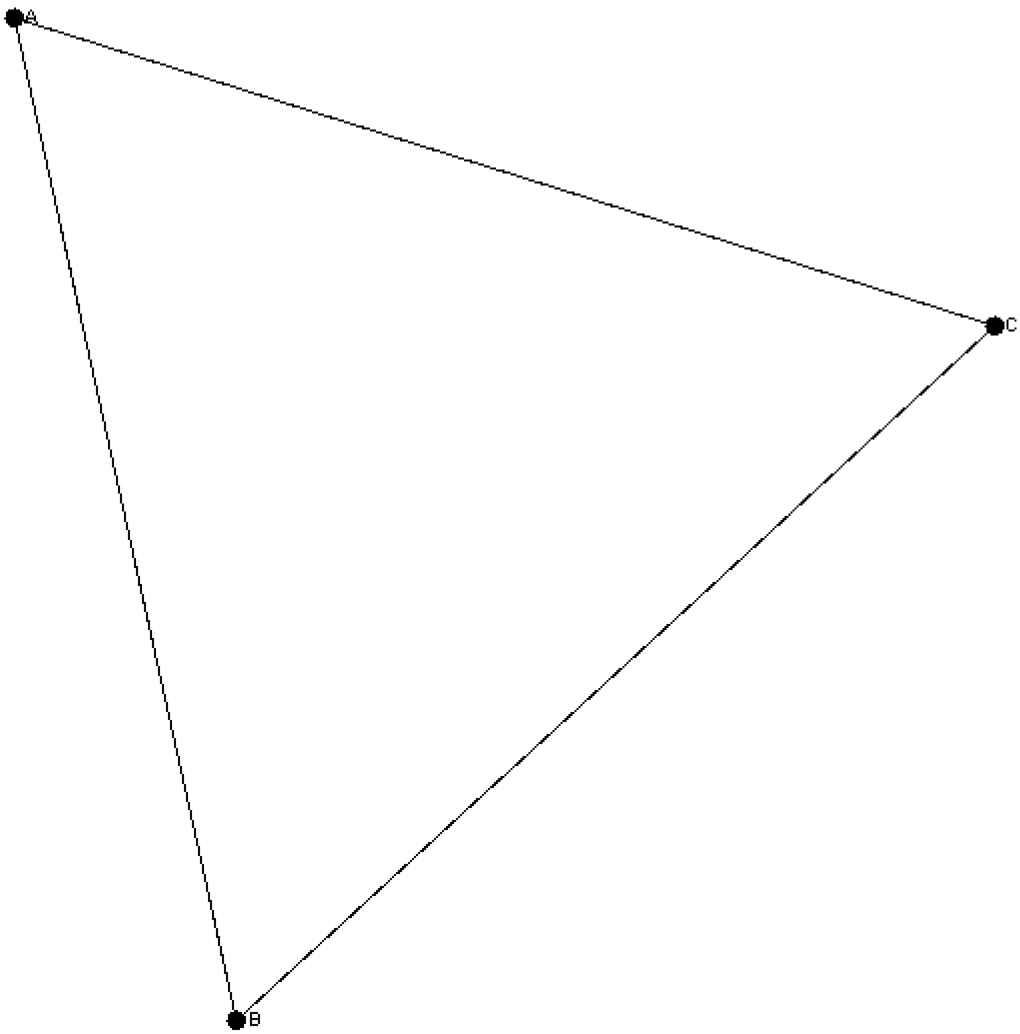


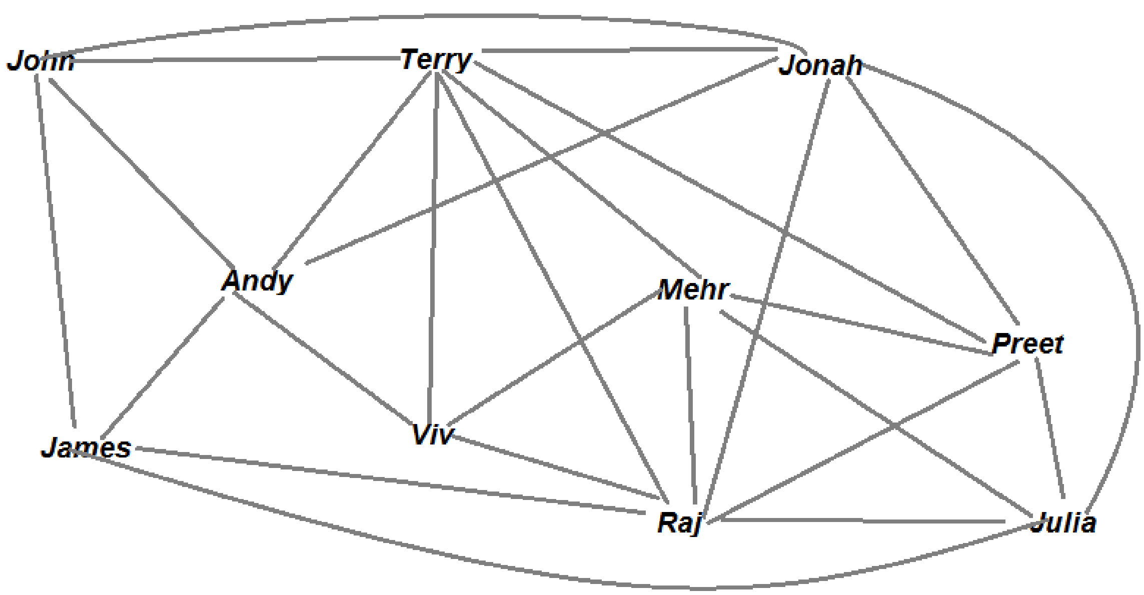


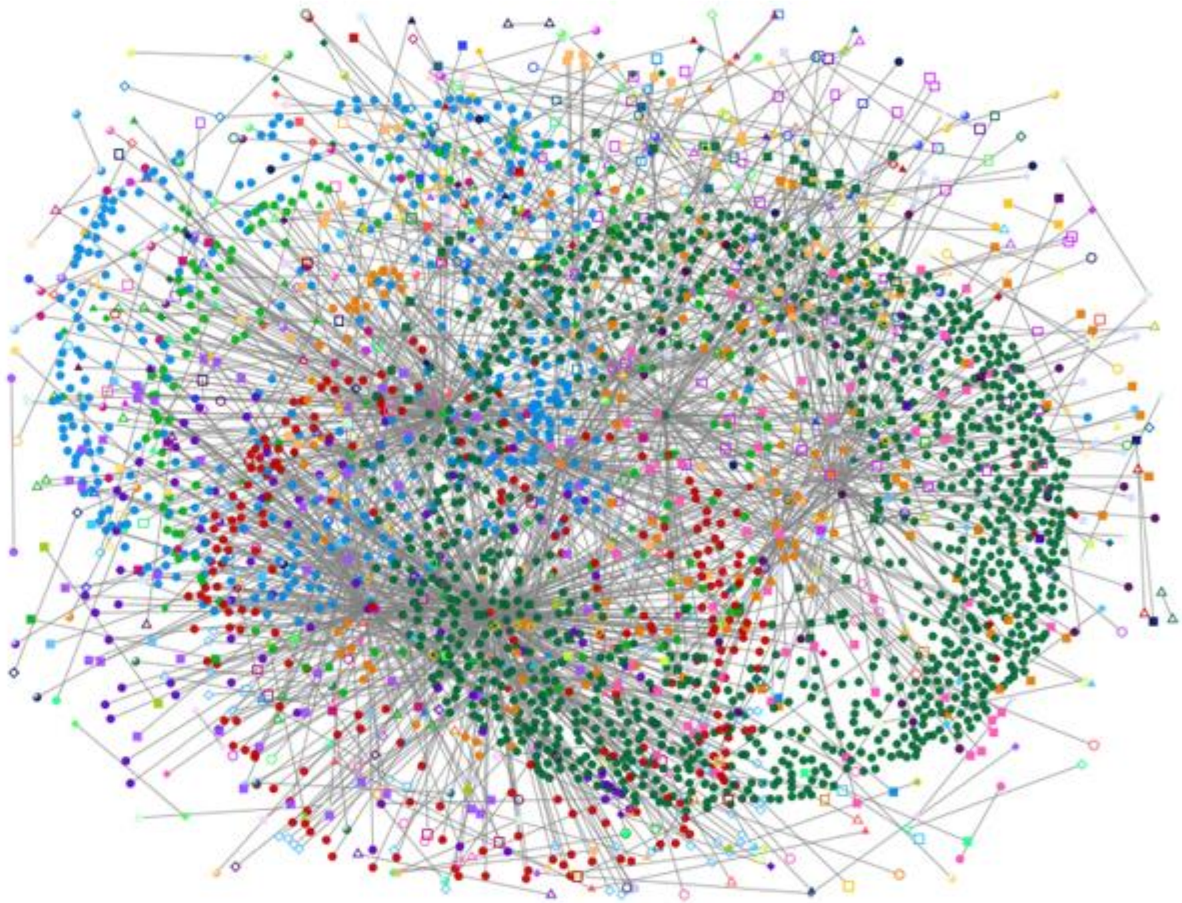
$$
= \quad \begin{array}{cc}
 & \begin{array}{cccccc} A & B & C & D & E & F \end{array} \\
\begin{array}{c} A \\ B \\ C \\ D \\ E \\ F \end{array} &
\begin{bmatrix}
0 & 1 & 0 & 0 & 0 & 1 \\
0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 1 \\
0 & 0 & 0 & 0 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0
\end{bmatrix}
\end{array}
$$

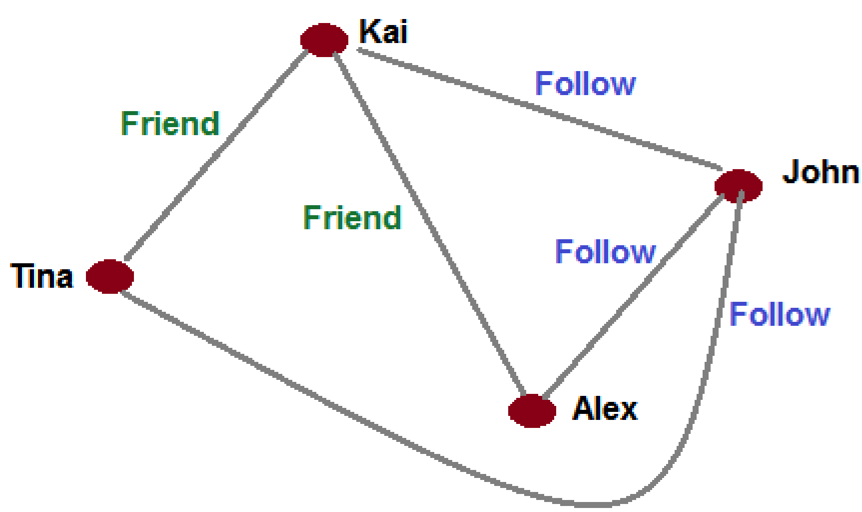

A
E
F
B
D
C
=
A
B
F
C
F
D
B
C


A
E
F
B
D
C


A
B
C


John
Terry
Jonah
Andy
Mehr
Preet
James
Viv
Raj
Julia




Kai
Follow
Friend
John
Friend
Follow
Tina
Follow
Alex

id	name	age
101	Kai	27
201	John	45
301	Alex	32
401	Tina	23

src	dst	relationType
101	301	Friends
101	401	Friends
401	201	Follow
301	201	Follow
201	101	Follow

id	inDegree
201	2
301	1
101	1
401	1

id	name	age
201	John	45
301	Alex	32

_c0	_c1	_c2	_c3	_c4	_c5	_c6	_c7	_c8
1	Goroka Airport	Goroka	Papua New Guinea	GKA	AYGA	-6.081689834590001	145.391998291	5282
2	Madang Airport	Madang	Papua New Guinea	MAG	AYMD	-5.20707988739	145.789001465	20
3	Mount Hagen Kagam...	Mount Hagen	Papua New Guinea	HGU	AYMH	-5.826789855957031	144.2960052490234 4	5388
4	Nadzab Airport	Nadzab	Papua New Guinea	LAE	AYNZ	-6.569803	146.725977	239
5	Port Moresby Jack...	Port Moresby	Papua New Guinea	POM	AYPY	-9.443380355834961	147.2200012207031 2	146
6	Wewak Internation...	Wewak	Papua New Guinea	WWK	AYWK	-3.58383011818	143.669006348	19
7	Narsarsuaq Airport	Narssarssuaq	Greenland	UAK	BGBW	61.1604995728	-45.4259986877	112
8	Godthaab / Nuuk A...	Godthaab	Greenland	GOH	BGGH	64.19090271	-51.6781005859	283

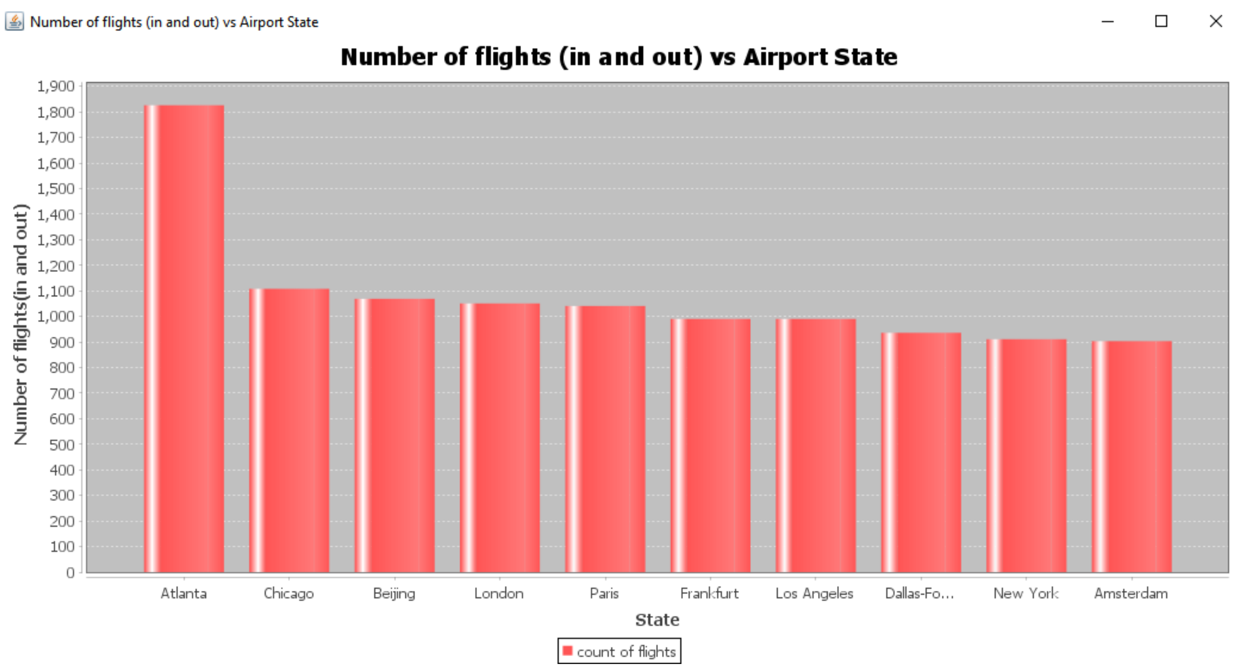
airLineCode	airlineId	dst	dstCode	dstId	relationType	src	srcCode	srcId
410	null	KZN	null	null	null	AER	null	null
410	null	KZN	null	null	null	ASF	null	null
410	null	MRV	null	null	null	ASF	null	null
410	null	KZN	null	null	null	CEK	null	null

airportIataCode	airportIcaoCode	airportId	airportName	country	id
GKA	AYGA	1	Goroka Airport	Papua New Guinea	GKA
MAG	AYMD	2	Madang Airport	Papua New Guinea	MAG
HGU	AYMH	3	Mount Hagen Kagam...	Papua New Guinea	HGU
LAE	AYNZ	4	Nadzab Airport	Papua New Guinea	LAE
POM	AYPY	5	Port Moresby Jack...	Papua New Guinea	POM
WWK	AYWK	6	Wewak Internation...	Papua New Guinea	WWK
UAK	BGBW	7	Narsarsuaq Airport	Greenland	UAK
GOH	BGGH	8	Godthaab / Nuuk A...	Greenland	GOH
SFJ	BGSF	9	Kangerlussuaq Air...	Greenland	SFJ
THU	BGTL	10	Thule Air Base	Greenland	THU
AEY	BIAR	11	Akureyri Airport	Iceland	AEY

airportIataCode	airportIcaoCode	airportId	airportName	country	id
BTI	PABA	3411	Barter Island LRR...	United States	BTI
LUR	PALU	3413	Cape Lisburne LRR...	United States	LUR
PIZ	PPIZ	3414	Point Lay LRRS Ai...	United States	PIZ
ITO	PHTO	3415	Hilo Internationa...	United States	ITO
ORL	KORL	3416	Orlando Executive...	United States	ORL
BTT	PABT	3417	Bettles Airport	United States	BTT
Z84	PACL	3418	Clear Airport	United States	Z84
UTO	PAIM	3419	Indian Mountain L...	United States	UTO
FYU	PFYU	3420	Fort Yukon Airport	United States	FYU

id	outDegree
EWR	253

airportName	State	Country	degree
Hartsfield Jackso...	Atlanta	United States	1826
Chicago O'Hare In...	Chicago	United States	1108
Beijing Capital I...	Beijing	China	1069
London Heathrow A...	London	United Kingdom	1051
Charles de Gaulle...	Paris	France	1041
Frankfurt am Main...	Frankfurt	Germany	990
Los Angeles Inter...	Los Angeles	United States	990
Dallas Fort Worth...	Dallas-Fort Worth	United States	936
John F Kennedy In...	New York	United States	911
Amsterdam Airport...	Amsterdam	Netherlands	903



source_city	destination_city	airlineName
Newark	Mumbai	Air India Limited
Chicago	Delhi	Air India Limited
New York	Delhi	Air India Limited
Newark	Mumbai	United Airlines
Newark	Delhi	United Airlines

src	edge	dst

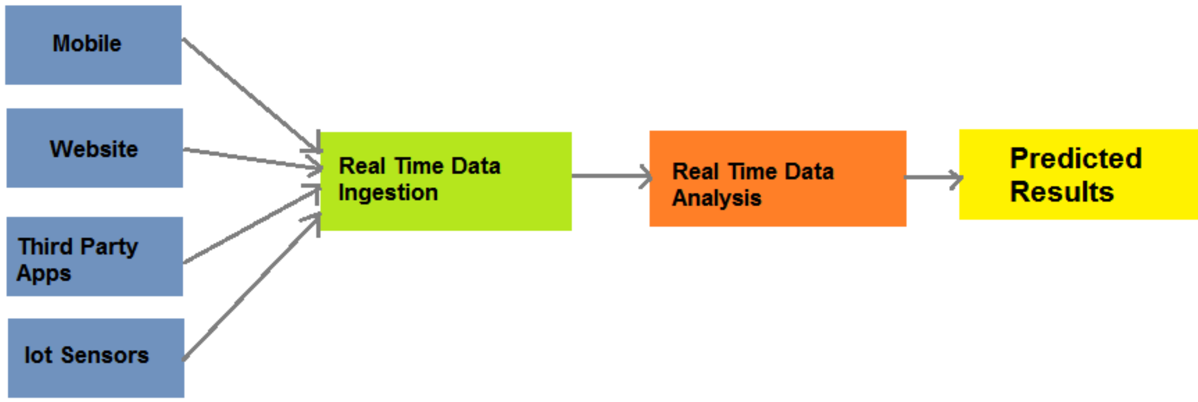
state	state	state
San Francisco	Minneapolis	Buffalo
San Francisco	Charlotte	Buffalo
San Francisco	Orlando	Buffalo
San Francisco	Las Vegas	Buffalo
San Francisco	Newark	Buffalo
San Francisco	Philadelphia	Buffalo
San Francisco	Atlanta	Buffalo
San Francisco	Phoenix	Buffalo
San Francisco	Washington	Buffalo
San Francisco	Detroit	Buffalo
San Francisco	Boston	Buffalo
San Francisco	Baltimore	Buffalo
San Francisco	Chicago	Buffalo
San Francisco	Fort Lauderdale	Buffalo
San Francisco	New York	Buffalo
San Francisco	Cleveland	Buffalo

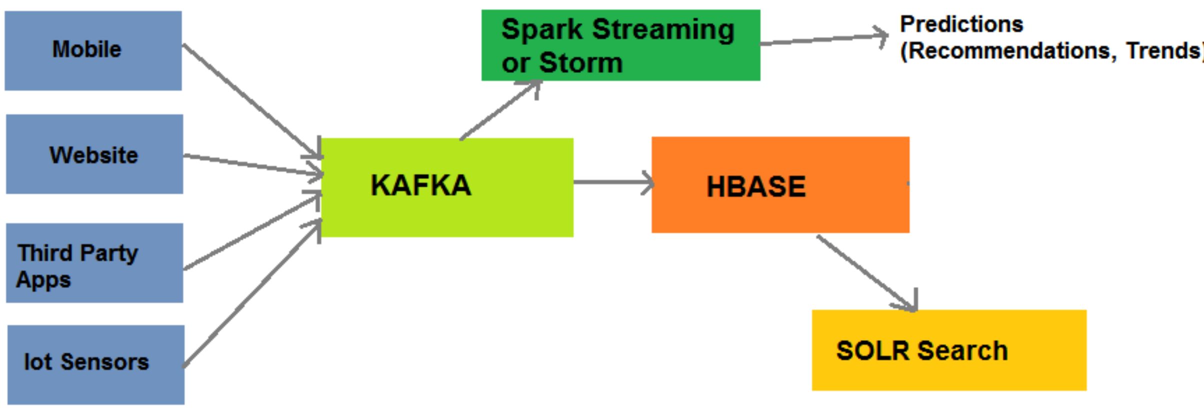
airportId	airportName	country	id	latitude	longitude	state	pagerank
2989	Syktyvkar Airport	Russia	SCW	61.64699935913086	50.84510040283203	Syktyvkar	0.41237783214480783
1686	?ncirlik Air Base	Turkey	UAB	37.002101898199996	35.4258995056	Adana	0.15
2548	Marechal Rondon A...	Brazil	CGB	-15.6528997421	-56.1166992188	Cuiaba	1.5706549071102351
8284	Capital City Airport	United States	CXY	40.2170982361	-76.85150146480001	Harrisburg	0.15
3368	Tianjin Binhai In...	China	TSN	39.124401092499994	117.346000671	Tianjin	1.0415126053412012
1837	Puerto Escondido ...	Mexico	PXM	15.8768997192	-97.08910369870001	Puerto Escondido	0.19125829670841465
8793	Hardwick Field	United States	HDI	35.22010040283203	-84.8323974609375	Cleveland	0.15
868	Virginia Airport	South Africa	VIR	-29.770599365234375	31.058399200439453	Durban	0.15
6917	Antonio Nery Juar...	Puerto Rico	ARE	18.4500007629	-66.6753005981	Arecibo	0.15
6013	Jolo Airport	Philippines	JOL	6.0536699295043945	121.01100158691406	Jolo	0.20276920684098537
128	Resolute Bay Airport	Canada	YRB	74.7169036865	-94.9693984985	Resolute	0.5402124513888888
413	Kuressaare Airport	Estonia	URE	58.22990036010742	22.50950050354004	Kuressaare	0.16411268502907458

airportIataCode	airportIcaoCode	airportId	airportName	country	id	state	pagerank
ATL	KATL	3682	Hartsfield Jackso...	United States	ATL	Atlanta	48.866507796851174
ORD	KORD	3830	Chicago O'Hare In...	United States	ORD	Chicago	30.85875567259695
LAX	KLAX	3484	Los Angeles Inter...	United States	LAX	Los Angeles	29.767643411164528
DFW	KDFW	3670	Dallas Fort Worth...	United States	DFW	Dallas-Fort Worth	28.614101408262563
SIN	WSSS	3316	Singapore Changi ...	Singapore	SIN	Singapore	25.836970924753565
CDG	LFPG	1382	Charles de Gaulle...	France	CDG	Paris	25.494167226690674
LHR	EGLL	507	London Heathrow A...	United Kingdom	LHR	London	25.258003242524943
DEN	KDEN	3751	Denver Internatio...	United States	DEN	Denver	24.906201842614077
DME	UUDD	4029	Domodedovo Intern...	Russia	DME	Moscow	23.29190755405923
JFK	KJFK	3797	John F Kennedy In...	United States	JFK	New York	23.064144993605314
FRA	EDDF	340	Frankfurt am Main...	Germany	FRA	Frankfurt	22.714716381995252
PEK	ZBAA	3364	Beijing Capital I...	China	PEK	Beijing	22.654135319752914
SYD	YSSY	3361	Sydney Kingsford ...	Australia	SYD	Sydney	22.603944128539965
MIA	KMIA	3576	Miami Internation...	United States	MIA	Miami	21.973498718530653
AMS	EHAM	580	Amsterdam Airport...	Netherlands	AMS	Amsterdam	21.067548191034817
IST	LTBA	1701	Atatürk Internati...	Turkey	IST	Istanbul	20.923165962750826
DXB	OMDB	2188	Dubai Internation...	United Arab Emirates	DXB	Dubai	20.764004952251774
BKK	VTBS	3885	Suvarnabhumi Airport	Thailand	BKK	Bangkok	19.537859448118365
BOG	SKBO	2709	El Dorado Interna...	Colombia	BOG	Bogota	19.471296723900714
ICN	RKSI	3930	Incheon Internati...	South Korea	ICN	Seoul	18.86590163053047

only showing top 20 rows

Chapter 12: Real-Time Analytics on Big Data





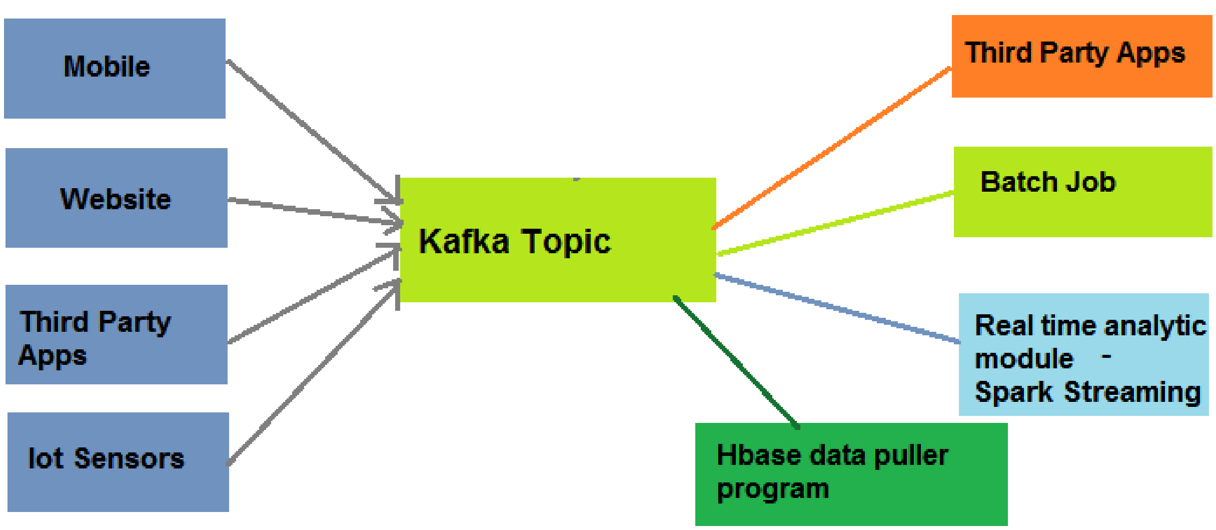

```
17/06/01 00:21:55 INFO
+--------+
|count(1)|
+--------+
|    4042|
+--------+

17/06/01 00:21:55 INFO
```

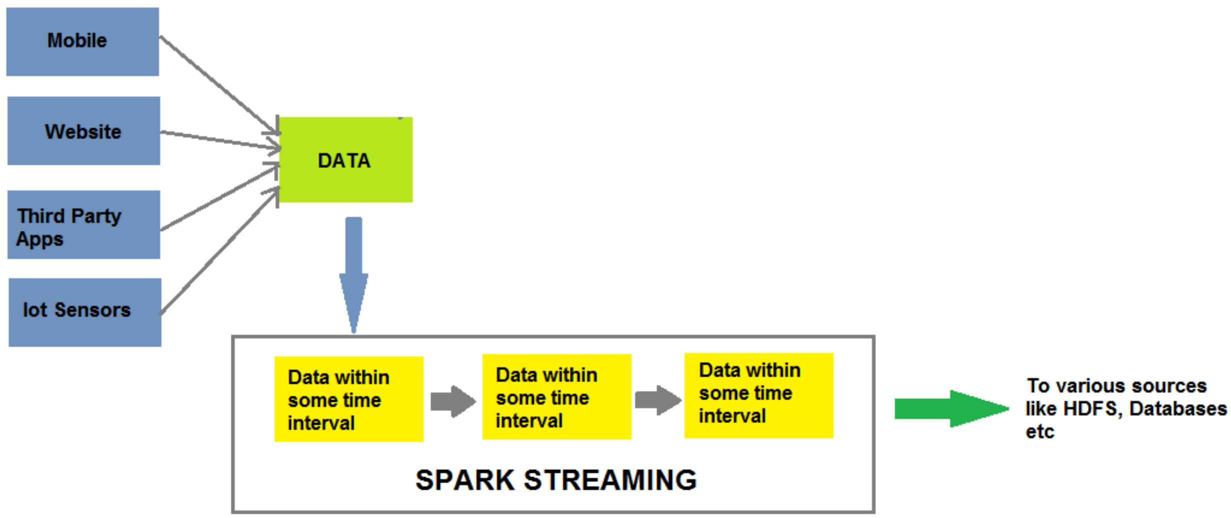
17/06/01 00:55:45 INFO DAGScheduler: Job 3 finished: show at CsvToParquet.java:32, took 0.471530 s

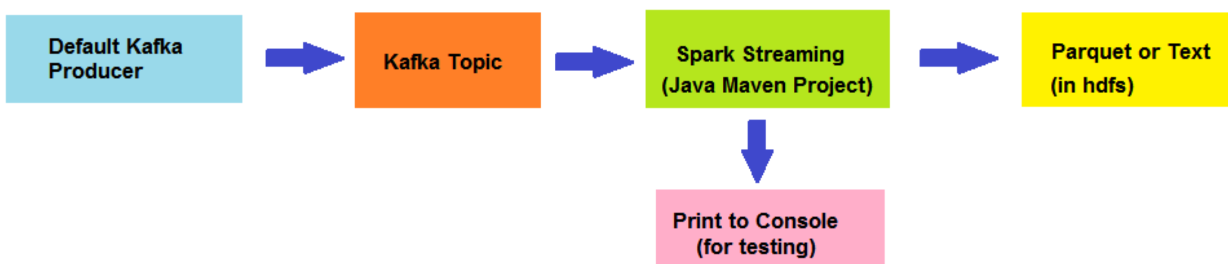
airline_id	origin	origin_state	dep_delay	dest	dest_state	arr_delay	distance	weather_delay	cancelled
19805	SFO	CA	53.00	PHX	AZ	50.00	651.00	0.00	0.00
19805	SFO	CA	38.00	PHX	AZ	24.00	651.00	0.00	0.00
19805	SFO	CA	142.00	PHX	AZ	156.00	651.00	2.00	0.00
19805	SFO	CA	52.00	DFW	TX	46.00	1464.00	0.00	0.00
19805	SFO	CA	100.00	DFW	TX	103.00	1464.00	0.00	0.00
19805	SFO	CA	46.00	DFW	TX	31.00	1464.00	0.00	0.00
19805	SFO	CA	85.00	DFW	TX	99.00	1464.00	0.00	0.00
19805	SFO	CA	31.00	DFW	TX	18.00	1464.00	0.00	0.00
19805	SFO	CA	294.00	PHX	AZ	291.00	651.00	0.00	0.00
19805	SFO	CA	46.00	ORD	IL	8.00	1846.00	null	0.00
19805	SFO	CA	54.00	ORD	IL	53.00	1846.00	0.00	0.00
19805	SFO	CA	56.00	MIA	FL	32.00	2585.00	0.00	0.00
19805	SFO	CA	33.00	MIA	FL	7.00	2585.00	null	0.00
19805	SFO	CA	49.00	MIA	FL	58.00	2585.00	0.00	0.00

17/06/01 01:10:11 INI

origin_state	cnt
CA	1425
FL	709
GA	676
TX	617
IL	554
CO	458
NY	438
MI	342
NC	311
OR	272







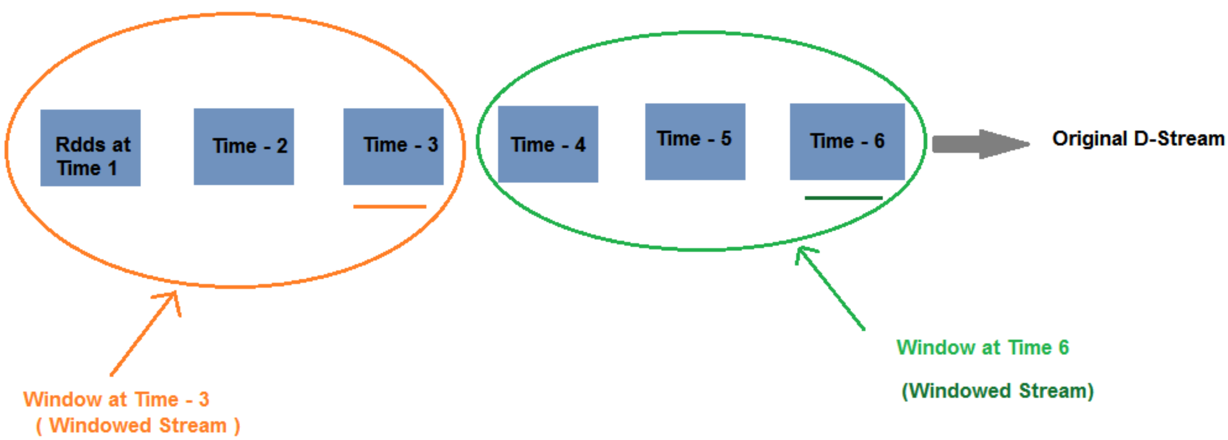
Anaconda - kafka-console-producer.bat --broker-list localhost:9092 --topic test11

```
C:\kafka_2.11-0.9.0.0\bin\windows>
C:\kafka_2.11-0.9.0.0\bin\windows>
C:\kafka_2.11-0.9.0.0\bin\windows>
C:\kafka_2.11-0.9.0.0\bin\windows>kafka-console-producer.bat --broker-list localhost:9092 --topic test11
Test Data on the topic test11
Some more data on the topic test11
Real time analytics is cool
```

Anaconda - kafka-console-producer.bat --broker-list localhost:9092 --topic test11

```
C:\kafka_2.11-0.9.0.0\bin\windows>kafka-console-producer.bat --broker-list localhost:9092 --topic test11
Cache is cleared
Hit count on the server is 50
Nullpointer exception in class BBCTestService at line 140
```

```
2017-06-08 08:44:06 INFO  JobScheduler:54 - Added jobs for time 1496925846000 ms
2017-06-08 08:44:06 INFO  JobScheduler:54 - Starting job streaming job 1496925846
--- New RDD with 1 partitions and 1 records
2017-06-08 08:44:06 INFO  SparkContext:54 - Starting job: foreach at SparkStreami
2017-06-08 08:44:06 INFO  DAGScheduler:54 - Got job 23 (foreach at SparkStreaming
2017-06-08 08:44:06 INFO  DAGScheduler:54 - Final stage: ResultStage 23 (foreach
2017-06-08 08:44:06 INFO  DAGScheduler:54 - Parents of final stage: List()
2017-06-08 08:44:06 INFO  DAGScheduler:54 - Missing parents: List()
2017-06-08 08:44:06 INFO  DAGScheduler:54 - Submitting ResultStage 23 (MapPartiti
2017-06-08 08:44:06 INFO  MemoryStore:54 - Block broadcast_23 stored as values in
2017-06-08 08:44:06 INFO  MemoryStore:54 - Block broadcast_23_piece0 stored as by
2017-06-08 08:44:06 INFO  BlockManagerInfo:54 - Added broadcast_23_piece0 in memo
2017-06-08 08:44:06 INFO  SparkContext:54 - Created broadcast 23 from broadcast a
2017-06-08 08:44:06 INFO  DAGScheduler:54 - Submitting 1 missing tasks from Resul
2017-06-08 08:44:06 INFO  TaskSchedulerImpl:54 - Adding task set 23.0 with 1 task
2017-06-08 08:44:06 INFO  TaskSetManager:54 - Starting task 0.0 in stage 23.0 (TI
2017-06-08 08:44:06 INFO  Executor:54 - Running task 0.0 in stage 23.0 (TID 23)
2017-06-08 08:44:06 INFO  KafkaRDD:145 - Computing topic test11, partition 0 offs
2017-06-08 08:44:06 INFO  VerifiableProperties:68 - Verifying properties
2017-06-08 08:44:06 INFO  VerifiableProperties:68 - Property group.id is overridd
2017-06-08 08:44:06 INFO  VerifiableProperties:68 - Property zookeeper.connect is
rowLine ------> Nullpointer exception in class BBCTestService at line 140
2017-06-08 08:44:06 INFO  Executor:54 - Finished task 0.0 in stage 23.0 (TID 23)
```



```
C:\kafka_2.11-0.9.0.0\bin\windows>kafka-console-producer.bat --broker-list localhost:9092 --topic test11
video-1,3
video-2,1
video-3,5
video-4,1
video-5,2
video-6,7
video-3,2
video-6,1
video-4,2
video-4,1
video-3,5
video-2,1
video-1,3
video-5,2
video-3,2
video-6,7
```

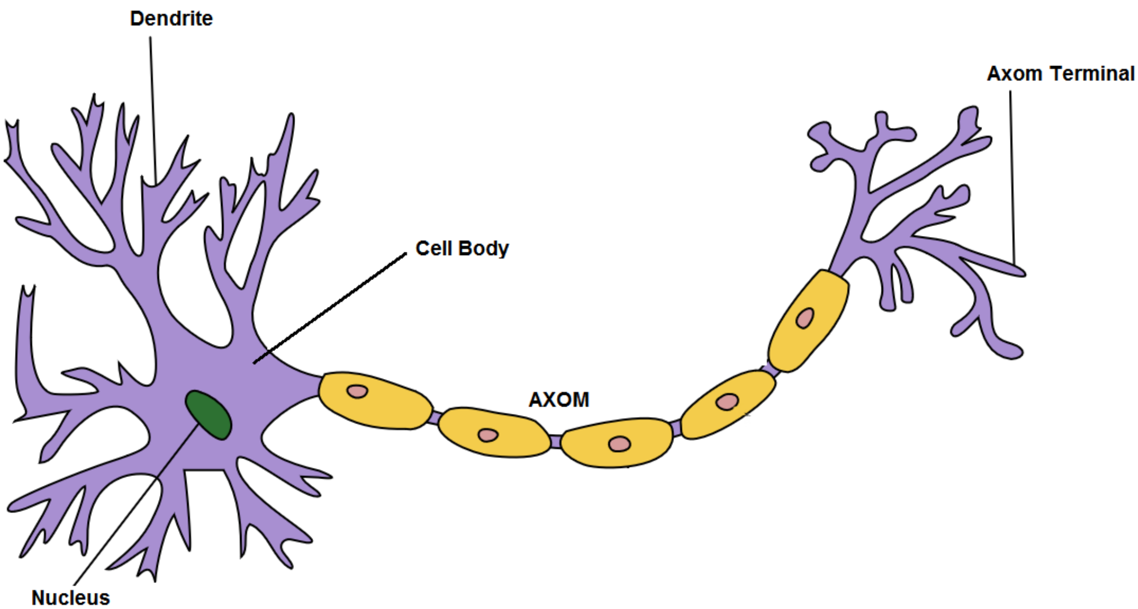
```
--- WINDOW ID --- 49
+-------+-------------+
|videoID|videoHitsCount|
+-------+-------------+
+-------+-------------+

--- WINDOW ID --- 61
+-------+-------------+
|videoID|videoHitsCount|
+-------+-------------+
|video-3|            5|
|video-1|            3|
|video-4|            1|
|video-2|            1|
+-------+-------------+

--- WINDOW ID --- 73
+-------+-------------+
|videoID|videoHitsCount|
+-------+-------------+
|video-6|            7|
|video-3|            7|
|video-5|            4|
|video-1|            3|
|video-2|            1|
|video-4|            1|
+-------+-------------+
```

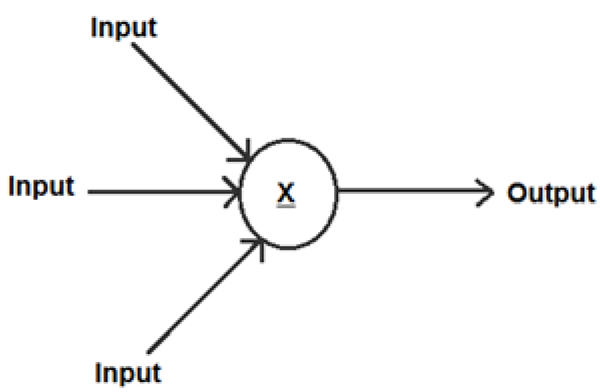
```
2017-06-09 19:50:40 INFO  DAGSc
+-----+--------------------+
|label|               tweet|
+-----+--------------------+
|  0.0|Da vinci code is ...|
|  0.0|da vinci book is ...|
+-----+--------------------+

######################## Predict
```

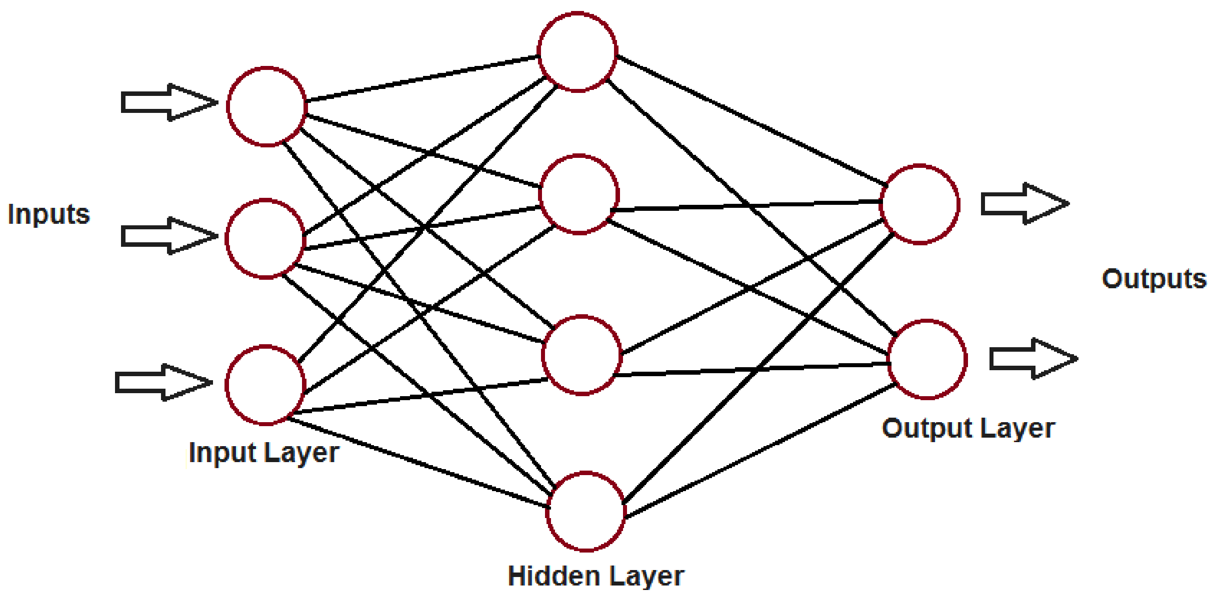
```
-09 20:02:30 INFO  Co
--------------------+      +--------------------+--------------------+-----------
               tweet|      |       rawPrediction|         probability|predictions
--------------------+      +--------------------+--------------------+-----------
 i love da vinci code      |[-30.580222214999...|[5.00686967717962...|        1.0
--------------------+      +--------------------+--------------------+-----------
```

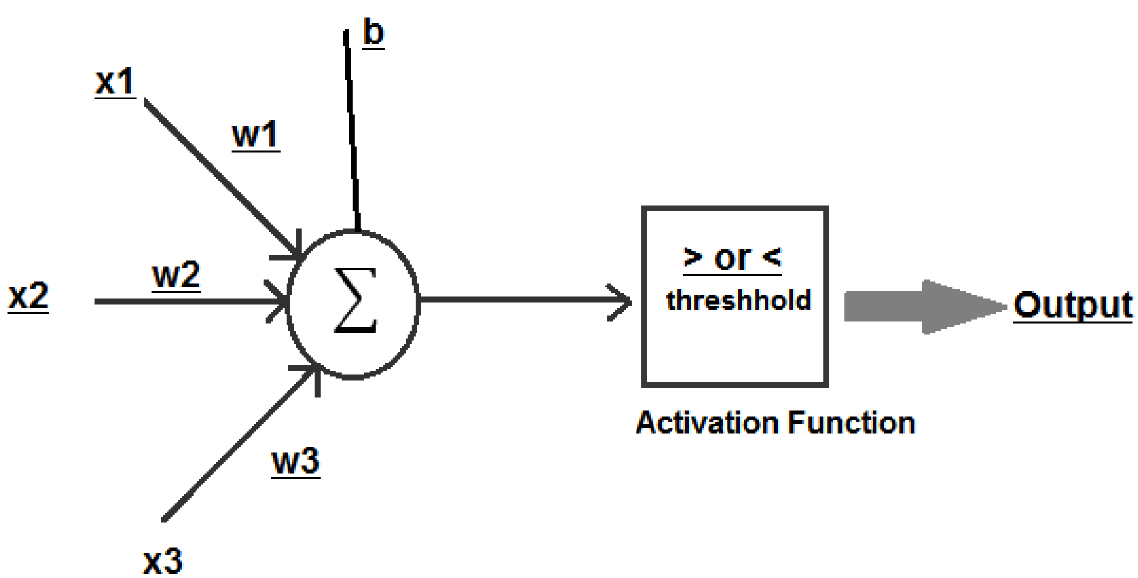
```
-09 20:02:40 INFO  CodeGenerator:54 - Code
---------------------+--------------------+        +--------------------+--------------------+-----------
                tweet|               words|        |       rawPrediction|         probability|predictions
---------------------+--------------------+        +--------------------+--------------------+-----------
 da vinci code is bad|[da, vinci, code,...|        |[-65.116600527323...|[0.99952330472415...|        0.0
 da vinci code is ...|[da, vinci, code,...|        |[-113.18459052402...|[0.99999970715803...|        0.0
---------------------+--------------------+        +--------------------+--------------------+-----------
```

Chapter 13: Deep Learning Using Big Data





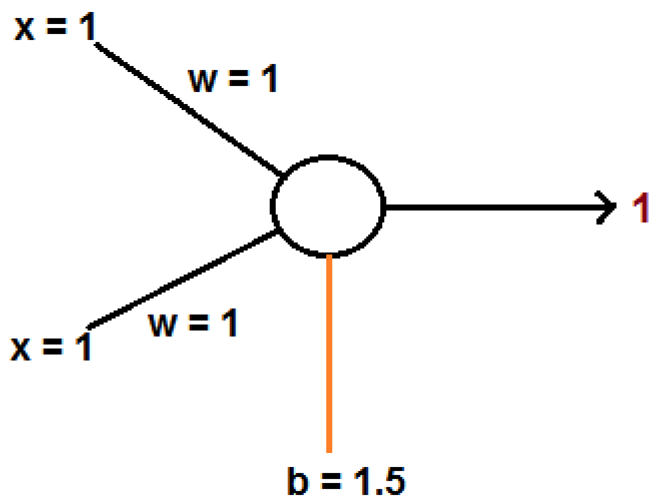


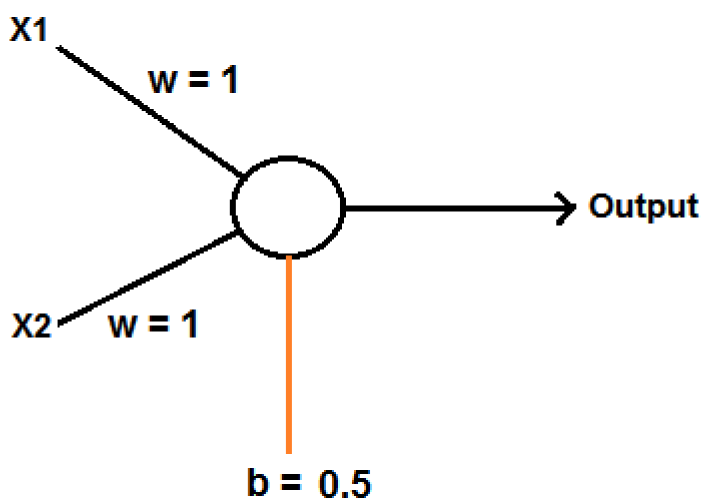


$$\text{Calculation Output} = b + \sum_{k=0}^{n} w * x$$

$$\text{Perception result} = 0 \ , \mathit{if\ Calculation\ Output} \leq 0$$

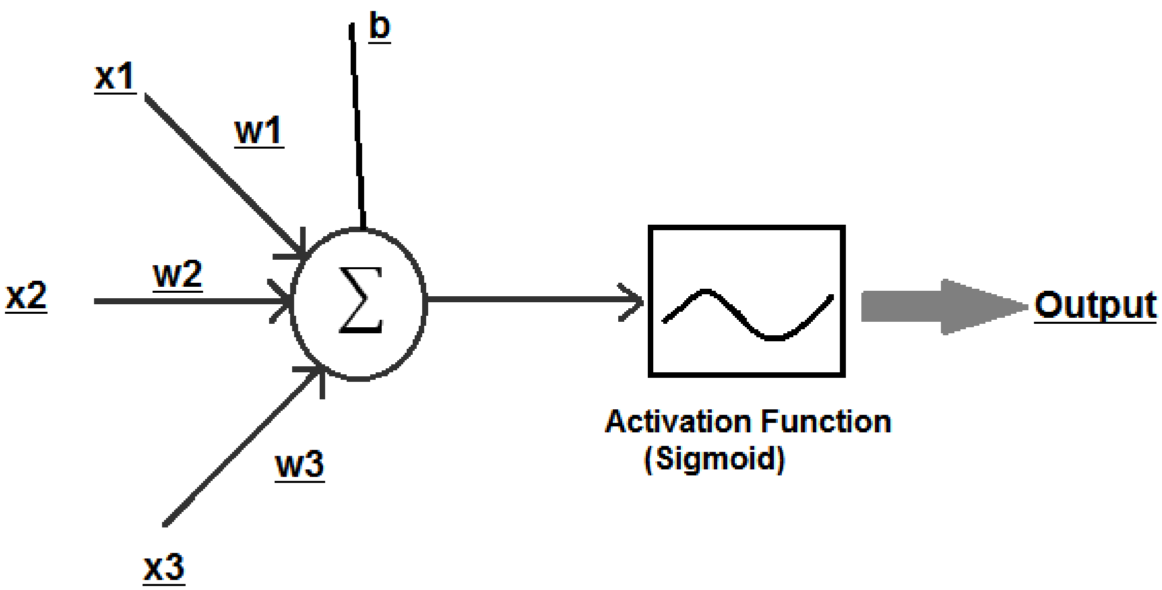
$$\text{result} = 1 \ , \mathit{if\ Calculation}$$

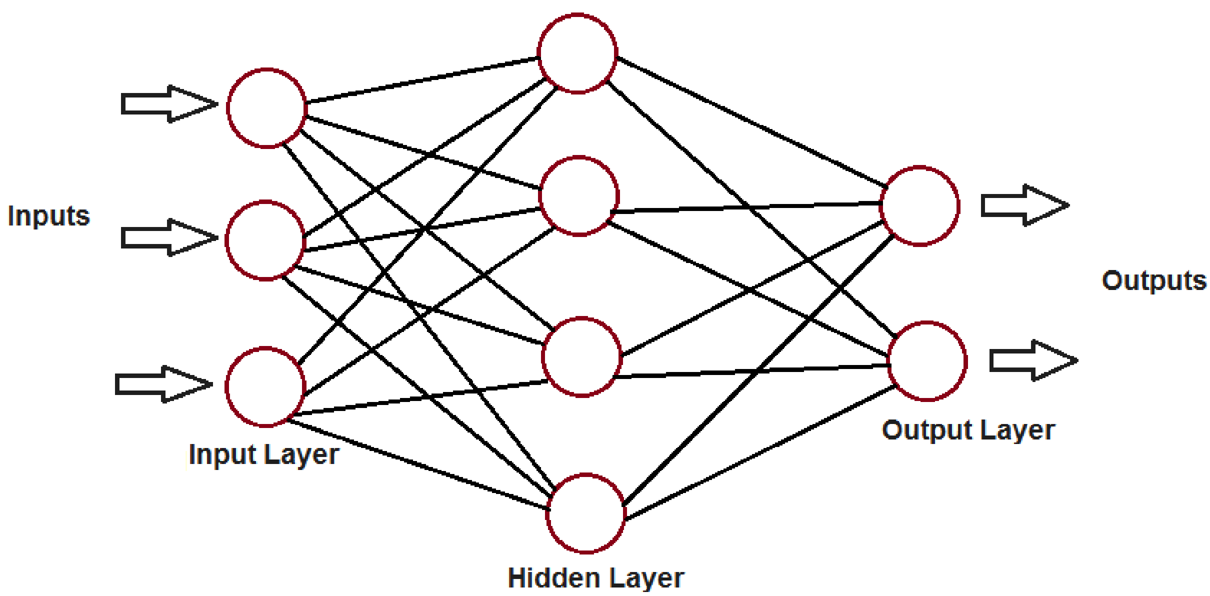




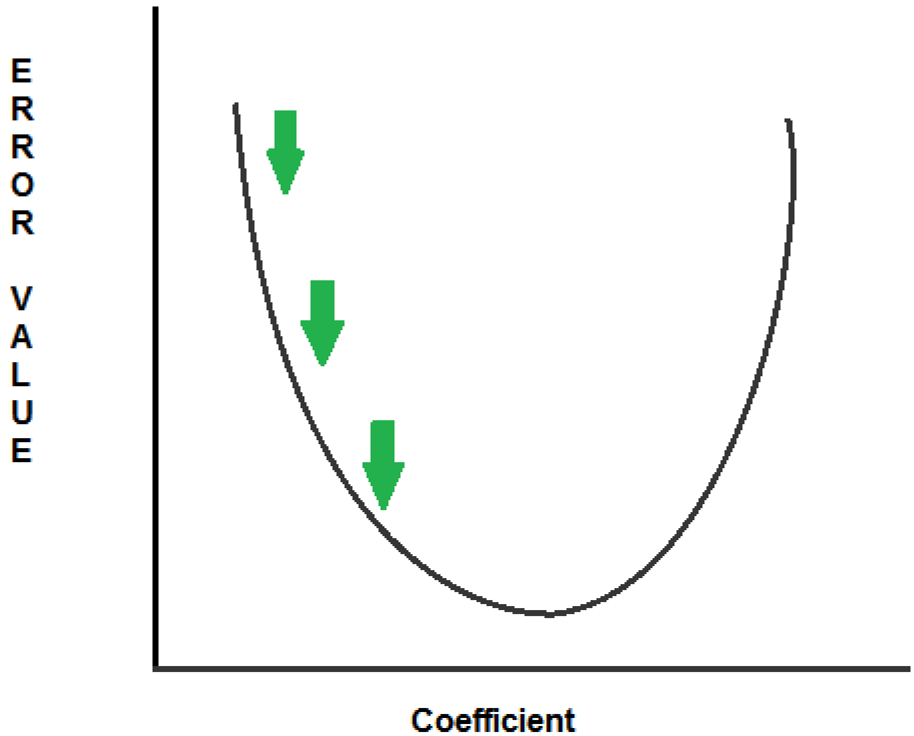
$$f(x) = \frac{1}{1+e^{-x}}$$

$$b + \sum_{k=0}^{n} w * x$$





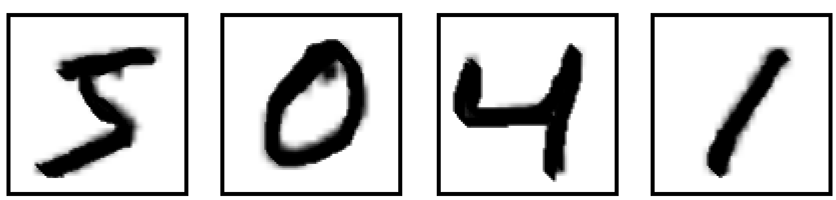
$$Mean\,Squard\,Error = \frac{1}{2n}\sum_{k=0}^{n}\left(y(x)-a\right)^2$$




```
17/06/27 19:17:24 INFO DAGSchedule
Test set accuracy = 0.95
17/06/27 19:17:24 INFO SparkUI: St
```

labelString	petalLength	petalWidth	sepalLength	sepalWidth	label	features	prediction
setosa	1.0	0.2	4.6	3.6	1.0	[4.6,3.6,1.0,0.2]	1.0
setosa	1.1	0.1	4.3	3.0	1.0	[4.3,3.0,1.1,0.1]	1.0
setosa	1.2	0.2	5.0	3.2	1.0	[5.0,3.2,1.2,0.2]	1.0
setosa	1.4	0.2	4.4	2.9	1.0	[4.4,2.9,1.4,0.2]	1.0
setosa	1.5	0.2	4.6	3.1	1.0	[4.6,3.1,1.5,0.2]	1.0
setosa	1.5	0.2	4.9	3.1	1.0	[4.9,3.1,1.5,0.2]	1.0
setosa	1.5	0.2	5.3	3.7	1.0	[5.3,3.7,1.5,0.2]	1.0
setosa	1.5	0.4	5.1	3.7	1.0	[5.1,3.7,1.5,0.4]	1.0
setosa	1.6	0.2	4.8	3.1	1.0	[4.8,3.1,1.6,0.2]	1.0
setosa	1.6	0.2	5.1	3.8	1.0	[5.1,3.8,1.6,0.2]	1.0
setosa	1.7	0.2	5.4	3.4	1.0	[5.4,3.4,1.7,0.2]	1.0
setosa	1.7	0.3	5.7	3.8	1.0	[5.7,3.8,1.7,0.3]	1.0
setosa	1.7	0.5	5.1	3.3	1.0	[5.1,3.3,1.7,0.5]	1.0
versicolor	3.3	1.0	5.0	2.3	2.0	[5.0,2.3,3.3,1.0]	2.0
versicolor	4.0	1.2	5.8	2.6	2.0	[5.8,2.6,4.0,1.2]	2.0
versicolor	4.0	1.3	6.1	2.8	2.0	[6.1,2.8,4.0,1.3]	2.0
versicolor	4.3	1.3	6.4	2.9	2.0	[6.4,2.9,4.3,1.3]	2.0
versicolor	4.4	1.4	6.7	3.1	2.0	[6.7,3.1,4.4,1.4]	2.0
versicolor	4.5	1.3	5.7	2.8	2.0	[5.7,2.8,4.5,1.3]	2.0
versicolor	4.5	1.5	5.6	3.0	2.0	[5.6,3.0,4.5,1.5]	2.0

```
17/06/27 19:17:24 INFO DAGSchedule
Test set accuracy = 0.95
17/06/27 19:17:24 INFO SparkUI: St
```



```
***** Evaluation *****

Examples labeled as 0 classified by model as 0: 5922 times
Examples labeled as 0 classified by model as 8: 1 times
Examples labeled as 1 classified by model as 1: 6715 times
Examples labeled as 1 classified by model as 2: 3 times
Examples labeled as 1 classified by model as 4: 1 times
Examples labeled as 1 classified by model as 7: 19 times
Examples labeled as 1 classified by model as 8: 4 times
Examples labeled as 2 classified by model as 1: 1 times
Examples labeled as 2 classified by model as 2: 5945 times
Examples labeled as 2 classified by model as 4: 1 times
Examples labeled as 2 classified by model as 6: 1 times
Examples labeled as 2 classified by model as 7: 6 times
Examples labeled as 2 classified by model as 8: 4 times
Examples labeled as 3 classified by model as 2: 2 times


==========================Scores========================================
 Accuracy:        0.9973
 Precision:       0.9974
 Recall:          0.9974
 F1 Score:        0.9974
========================================================================
```