


PG Department of Computer Science  
Sacred Heart College (Tirupattur)

In Plant Training

Boscsoft, Yelagiri - April 29<sup>th</sup> to May 28<sup>th</sup>  
Friendzion Technologies- May 4<sup>th</sup> to May 30<sup>th</sup>

S.NO	REGNO	NAME	COMPANY
1	BP180501	PRASANTH RAO S	Boscsoft, Yelagiri Hills
2	BP180502	PRIYA S	Boscsoft, Yelagiri Hills
3	BP180503	ARIVAZHAGAN M	Friendzion Technologies
4	BP180504	NISHA M	Boscsoft, Yelagiri Hills
5	BP180505	MANGALA MARY A	Boscsoft, Yelagiri Hills
6	BP180506	VINODHINI V	Boscsoft, Yelagiri Hills
7	BP180507	PUSHPALATHA P	Boscsoft, Yelagiri Hills
8	BP180508	PRABHU S	Boscsoft, Yelagiri Hills
9	BP180509	KOWSALYA M	Boscsoft, Yelagiri Hills
10	BP180510	ANTONY XAVIER S	Friendzion Technologies
11	BP180511	POOJA E	Boscsoft, Yelagiri Hills
12	BP180512	RENIRATHINAM R	Lesoftek, Salem
13	BP180513	IVISALLWYN A	Friendzion Technologies
14	BP180514	WILLIAM A	Friendzion Technologies
15	BP180515	SHALINI K	Boscsoft, Yelagiri Hills
16	BP180516	JOAN GEORGE T	Friendzion Technologies
17	BP180517	JOHNSON JERO S	Friendzion Technologies
18	BP180518	DHANALAKSHMI A	Boscsoft, Yelagiri Hills

19	BP180519	MASEEHA TASNEEM K	Boscsoft, Yelagiri Hills
20	BP180520	SOWMIYA A	Boscsoft, Yelagiri Hills
21	BP180521	JABEER S	Boscsoft, Yelagiri Hills
22	BP180522	GEETHAPRIYA R	Boscsoft, Yelagiri Hills
23	BP180523	BHARATHKUMAR S	Boscsoft, Yelagiri Hills
24	BP180524	KOTHANDARAMAN S	Friendzion Technologies
25	BP180525	DHANALAKSHMI H	Boscsoft, Yelagiri Hills
26	BP180526	ASHWINIPRIYA G	Boscsoft, Yelagiri Hills

  
Head  
PG Department of Computer Science  
Sacred Heart College (Autonomous)  
Tirupattur, Vellore Dt. - 635 601.

# CERTIFICATE

## OF COMPLETION

---

This certificate is proudly presented to

**S KOTHANDARAMAN**

For successfully completing his industrial inplant training in  
Adobe Photoshop and Android from  
**04-05-2019 to 30-05-2019**

During the training period the performance of trainee was found to be **Good**.

**30-05-2019**

---

**DATE TODAY**



**Friendzion**  
Technologies

[www.friendzion.com](http://www.friendzion.com)




---

**PRASANTH R**  
Program Coordinator

PG Department of Computer Science  
Sacred Heart College (Tirupattur)

Software project- Name List :: April 2019

S.No	REGNO	NAME	GENDER
1	BP170501	VASANTHA KUMARI G	Female
2	BP170502	SUGASHINI V	Female
3	BP170503	DEEPIKA J	Female
4	BP170504	MALINI A	Female
5	BP170505	LOKESH C	Male
6	BP170506	MURALIDHARAN D	Male
7	BP170507	CARMEL RAJ A	Male
8	BP170508	PAVITHRA S	Female
9	BP170509	CHINNAPPA RAJ J	Male
10	BP170510	PAVITHRA K	Female
11	BP170511	STEPHY A	Female
12	BP170512	SHINY MOULISHA R	Female
13	BP170513	JASMINE SHILPA G	Female
14	BP170514	SWETHA C	Female
15	BP170515	SHARADHA S	Female
16	BP170516	VIDYA E	Female
17	BP170517	GOKULASELVAN T	Male
18	BP170518	AMMU S	Female
19	BP170520	SUGANTHI S	Female
20	BP170521	MURUGAN R	Male
21	BP170522	PRABAVATHI P	Female
22	BP170523	PAVITHRA V	Female
23	BP170524	AISWARYA S	Female
24	BP170525	SRIDHAR R	Male
25	BP170526	IMMACULATE MARY J	Female

  
Head  
PG Department of Computer Science  
Sacred Heart College (Autonomous)  
Tirupattur, Vellore Dt. - 635 601.



# **ANALYSIS OF STUDENT'S MARKS AND PREDICTING THEM BY LOGISTIC REGRESSION**

A Report

Submitted in partial fulfillment for the award of the degree of

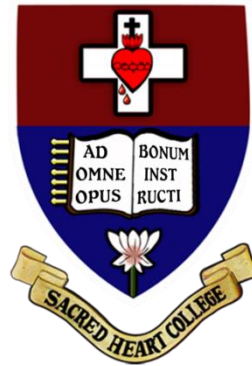
MASTER OF COMPUTER SCIENCE

**MURALIDHARAN D**

**(REG.NO:BP170506)**

**Under the Guidance of**

**Miss. D.Gajalakshmi M.C.A.,**



**PG DEPARTMENT OF COMPUTER SCIENCE**

**SACRED HEART COLLEGE (AUTONOMOUS)**

**TIRUPATTUR, VELLORE DT – 635 601**

**April - 2019**

## Contents

CHAPTER – I INTRODUCTION .....	4
1.2 Project Overview .....	4
CHAPTER – II SYSTEM ANALYSIS.....	8
2.1 System Study.....	8
2.1.5. 1 ABOUT CLASSIFICATION DETAILS OF THE PROJECTS .....	9
2.2 Feasibility Study .....	10
2.2.1.1 Technical Feasibility: .....	10
2.2.1.2 Economical Feasibility.....	10
CHAPTER – III System Configuration .....	12
3.1 Requirement Specification.....	12
3.3.2.1 SUPP18 Availability .....	13
3.3.2.2 Robustness .....	13
3.3.2.3 SUPP20 Accuracy .....	13
3.3.2.4 Safety .....	13
3.3.3.1 Response Time .....	14
3.3.3.2 Capacity.....	14
3.3.3.3 Utilization of resources .....	14
3.3.4.1 Adaptability.....	14
3.3.4.2 Maintainability .....	14
3.3.4.3 Reusability.....	14
CHAPTER – IV SCREENSHOT .....	15
4.7 User Interact Page.....	16
CHAPTER – V CONCLUSION .....	32
6.1 Conclusion.....	32

# INTRODUCTION

## 1.1 Abstract

**“Analysis of students marks and predicting them by logistic regression “this** is an predicting applications of the students mark details, by using the python language. In this the analysis are made of every student of the particular class and predicting the percentage of passing students in each subject and

overall passing percentage of the students in the class. We can also represent the predicted results of the students in graphical representation by plotting in bar diagram, pie chart and different kind of plotting method. The main purpose of the project is to identify whether the students result is pass or fail in the class by fixing the threshold value. We use logistic regression for this operation because the algorithms perfectly suits for our project by which it can be predicted by “pass” or “fail” /”0” or “1” /” yes”/”no”.

## **CHAPTER – I INRODUCTION**

### **1.2 Project Overview**

#### **1.2.1 Introduction**

**“Analysis of students marks and predicting them by logistic regression “this is an predicting applications of the students mark details, by using the python language. In this the analysis are made of every student of the particular class and predicting the percentage of passing students in**

each subject and overall passing percentage of the students in the class. We can also represent the predicted results of the students in graphical representation by plotting in bar diagram, pie chart and different kind of plotting method. The main purpose of the project is to identify whether the students result is pass or fail in the class by fixing the threshold value. We use logistic regression for this operation because the algorithms perfectly suits for our project by which it can be predicted by “pass” or “fail” /”0” or “1” /” yes”/”no”.

### **1.2.2 Scope**

The scope of the project is used to bring the outcome of the students details regarding the academic of the candidates. We can classify them into different norms and category by classifying the dataset, we can predict different type of possibilities of in various subjects. We predict outcome like passing percentage of the students in each subject and overall subject of each students. Finally, we can also plot the outcome predicted values in histogram and we can perform them in the graphical representations.

### **1.2.3 Problem Statement**

In this project the major problem was all the data are manually for each students and employees for the database due to is the time duration is to long huge man power is involved so that the new system in this project evaluate out comes of search problem in this project

### 1.2.4 Product position statement

For	For COE of College
Who	Staff members.
That	It is more flexible.
Our Product	Provide up-to-date information and easy to maintain all the details. An classify them and make prediction and represent them in graphical representation manner



# **SYSTEM ANALYSIS**

## **CHAPTER – II SYSTEM ANALYSIS**

### **2.1 System Study**

#### **2.1.1. Introduction**

The system study of “**Analysis of students marks and predicting them by logistic regression**” is that since the data of the students were stored in excel data format due to which the staff members of the class are facing problems in getting the proper results .so in this I have proposed the new way representing the data into graphical representations by which the staff members are able to identify the results easily.

#### **2.1.2. Over view**

Analysis of students marks and predicting them by logistic regression, this is an predicting applications of the students mark details, by using the python language. In this the analysis are made of every student of the particular class and predicting the percentage of passing students in each subject and overall passing percentage of the students in the class. the data of the students were stored in excel data format due to which the staff members of the class are facing problems in getting the proper results .so in this I have proposed the new way representing the data into graphical representations by which the staff members are able to identify the results easily. We can also represent the predicted results of the students in graphical representation by plotting in bar diagram, pie chart and different kind of plotting method. The main purpose of the project is to identify whether the students result is pass or fail in the class by fixing the threshold value.

#### **2.1.3 Proposed System**

- Representing the new way of present the data into the graphical reorientations.

## **2.1.5. 1 ABOUT CLASSIFICATION DETAILS OF THE PROJECTS**

### **Purpose**

To produce the complete details about the classifications and prediction.

### **Overview**

The user, who wants to know about the prediction of any case, can easily get the details in graphical representations.

### **Entry Criteria**

The users can directly import the data set, train and test them to get predictions.

### **Input**

Here the input is test data from the data set.

### **Steps involved**

Import .csv file into the python and perform the code form train and test using classification algorithms and make predict the desired output.

### **Output**

The required detail is shown to the knowledge seekers.

## **2.2 Feasibility Study**

### **2.2.1 Introduction**

The scope of the document is to define the system requirements “**Analysis of students marks and predicting them by logistic regression**” software. The impact can be either positive or negative. When the positives nominate the negatives, then the system is considered feasible. Here the feasibility study can be performed in two ways such as technical feasibility and Economical Feasibility.

#### **2.2.1.1 Technical Feasibility:**

We can strongly say that it is technically feasible, since there will not be much difficulty in getting required resources for the development and maintaining the system as well. All the resources needed for the development of the software as well as the maintenance of the same is available in the organization here we are utilizing the resources which are available already.

#### **2.2.1.2 Economical Feasibility**

Development of this application is highly economically feasible. The organization needed not spend much money for the development of the system already available. The only thing is to be done is making an environment for the development with an effective supervision. If we are doing so, we can attain the maximum usability of the corresponding resources. Even after the development, the organization will not be in condition to invest more in the organization. Therefore, the system is economically feasible.

# **SYSTEM CONFIGURATION**

## **CHAPTER – III System Configuration**

### **3.1 Requirement Specification**

#### **3.1.1 Hardware Requirements**

**Processor : Intel i3 processor**

**RAM : 4 MB**

**Hard Disk : 500 GB**

**Input/Output : Keyboard, mouse, monitor**

#### **3.1.2 Software Requirements**

**Front End : python 3.5 (jupyter)**

**Back End : Excel file (CSV FORMAT)**

**Documentation : Microsoft Word 2013**



## **3.3.2 Reliability**

### **3.3.2.1 SUPP18 Availability**

- ❖ This system will be configured as an online website, hence it will be accessible at 24\*7 days.

### **3.3.2.2 Robustness**

- ❖ For every invalid input from the user, the system should display a meaningful error message explaining in what format the input is to be fed into the system.

Ex: Error message for invalid Username or Password in login screen

“Invalid Username”, “Incorrect Password”

- ❖ Session should not expire in between the operations.

### **3.3.2.3 SUPP20 Accuracy**

- ❖ The result should be accurate.
- ❖ The reports should give accurate data.
- ❖ Transactions should not be collapsed in between operation.

### **3.3.2.4 Safety**

- ❖ Information maintained in the system should be kept confidential especially the user details.

### **3.3.3 Performance**

#### **3.3.3.1 Response Time**

- ❖ Average system response time should be less than five seconds to open the forms, to do the events and to display the result and to generate the report etc.

#### **3.3.3.2 Capacity**

- ❖ It should allow maximum number of users to access the system at the same time.

#### **3.3.3.3 Utilization of resources**

- ❖ CPU utilization should not exceed 25% when the application is functioning.

### **3.3.4 Supportability**

#### **3.3.4.1 Adaptability**

- ❖ Deployment time for new version should not take much time.
- ❖ System should support all web browsers.
- ❖ Application should run in windows family such as Windows XP, Windows 7.

#### **3.3.4.2 Maintainability**

- ❖ Code should be developed in such a way that it should be maintainable, so define the coding standard and follow it strictly.

#### **3.3.4.3 Reusability**

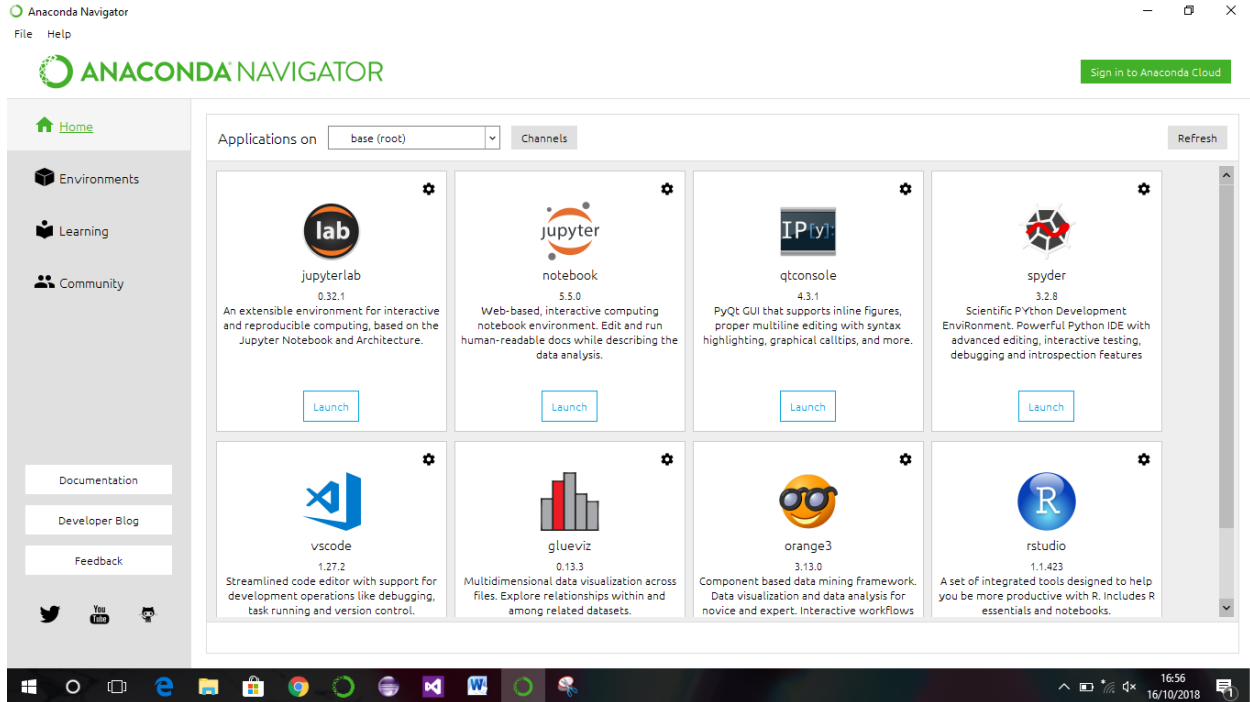
- ❖ Create some components as reusable components, so that it can be integrated with other system.

**CONCLUSION  
AND  
SAMPLE SCREEN SHOTS**

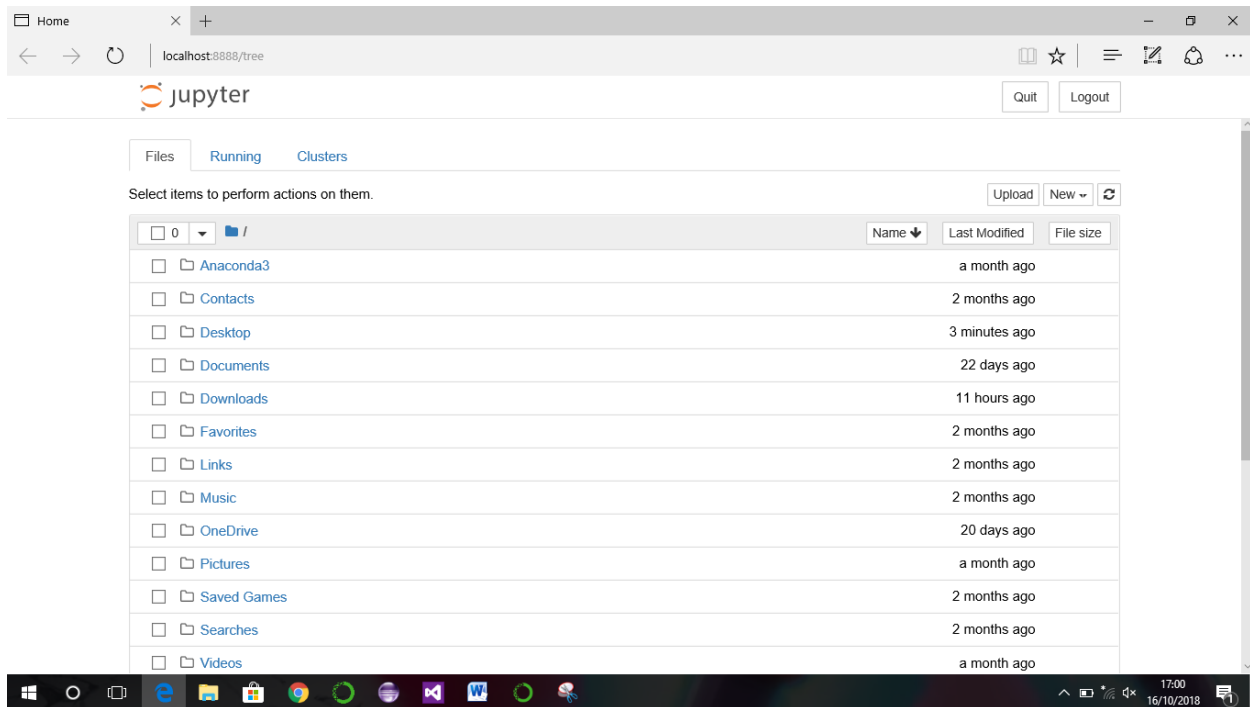
**CHAPTER – IV SCREENSHOT**

## 4.7 User Interact Page

### The anaconda IDE



The main window of jupyter IDE



## The editing window in jupyter

```
In [ ]: 1  
        2  
        3  
        4  
        5  
        6  
        7
```

## Importing the packages and basic file to perform the prediction.

```
In [1]: import pandas as pd  
        import numpy as np  
        #import cufflinks as cf  
        %matplotlib inline  
        import matplotlib.pyplot as plt  
        import random
```

## Reading the .csv file(raw data)

```
In [2]: cd C:\\Users\\madhu\\Downloads\\CSV files  
        C:\\Users\\madhu\\Downloads\\CSV files
```

## Extracting the data in the csv file

```
In [3]: pd.read_csv('student.csv')
```

Out[3]:

	Reg.No	Student_Name	Gender	IOT	CC	DOS	ASP	MA	Practical 1	Practical 2
0	Bp170501	Vasantha kumari	Female	70	69	87	68	56	89	86
1	Bp170502	Sugashini	Female	48	79	89	78	72	78	87
2	Bp170503	Deepika	Female	50	98	78	86	89	89	90
3	Bp170504	Malini	Female	60	86	98	94	92	90	90
4	Bp170505	Lokesh	Male	78	89	74	74	87	81	87
5	Bp170506	Muralidharan	Male	40	90	67	54	90	90	90
6	Bp170507	Carmel	Male	33	89	45	89	87	76	87
7	Bp170508	Pavithra 1	Female	87	78	43	85	78	87	88
8	Bp170509	Chinnapa raj	Male	20	78	23	23	81	77	80
9	Bp170510	Pavithra 2	Female	70	69	87	68	56	89	86
10	Bp170511	Stephy	Female	60	79	89	78	72	78	87
11	Bp170512	Shilpa	Female	78	98	78	86	89	89	90

## Creating the object for the csv file

```
In [4]: dd=pd.read_csv('student.csv')
```

## Showing the first 5 values

```
In [5]: dd.head()
```

Out[5]:

	Reg.No	Student_Name	Gender	IOT	CC	DOS	ASP	MA	Practical 1	Practical 2
0	Bp170501	Vasantha kumari	Female	70	69	87	68	56	89	86
1	Bp170502	Sugashini	Female	48	79	89	78	72	78	87
2	Bp170503	Deepika	Female	50	98	78	86	89	89	90
3	Bp170504	Malini	Female	60	86	98	94	92	90	90
4	Bp170505	Lokesh	Male	78	89	74	74	87	81	87

## Showing the last 5 values

```
In [6]: dd.tail()
```

Out[6]:

	Reg.No	Student_Name	Gender	IOT	CC	DOS	ASP	MA	Practical 1	Practical 2
21	Bp170522	Prabavathi	Female	89	86	98	56	92	90	90
22	Bp170523	Swetha	Female	78	89	74	74	87	81	87
23	Bp170524	Ishvariya	Female	56	90	34	23	90	90	90
24	Bp170525	Sridhar	Male	78	23	67	89	87	76	87
25	Bp170526	Mary	Female	89	87	85	89	89	89	90

## Describing the maximum marks in the each subject.



```
In [9]: dd.describe()
```

```
Out[9]:
```

	IOT	CC	DOS	ASP	MA	Practical 1	Practical 2
<b>count</b>	26.000000	26.000000	26.000000	26.000000	26.000000	26.000000	26.000000
<b>mean</b>	65.346154	81.076923	70.769231	74.307692	76.538462	84.461538	87.576923
<b>std</b>	20.834476	14.912875	21.345365	18.561830	19.206730	5.756602	2.715483
<b>min</b>	20.000000	23.000000	23.000000	23.000000	23.000000	76.000000	80.000000
<b>25%</b>	51.500000	78.000000	67.000000	68.000000	72.000000	78.000000	87.000000
<b>50%</b>	70.000000	86.000000	76.000000	78.000000	87.000000	88.000000	87.000000
<b>75%</b>	78.000000	89.000000	87.000000	87.500000	89.000000	89.000000	90.000000
<b>max</b>	89.000000	98.000000	98.000000	94.000000	92.000000	90.000000	90.000000

## Showing the attribute in the vertical order.

```
In [10]: dd.T
```

```
Out[10]:
```

	0	1	2	3	4	5	6	7	8	9	..
<b>Reg.No</b>	Bp170501	Bp170502	Bp170503	Bp170504	Bp170505	Bp170506	Bp170507	Bp170508	Bp170509	Bp170510	..
<b>Student_Name</b>	Vasanthakumari	Sugashini	Deepika	Malini	Lokesh	Muralidharan	Carmel	Pavithra 1	Chinnaparaaj	Pavithra 2	..
<b>Gender</b>	Female	Female	Female	Female	Male	Male	Male	Female	Male	Female	..
<b>IOT</b>	70	48	50	60	78	40	33	87	20	70	..
<b>CC</b>	69	79	98	86	89	90	89	78	78	69	..
<b>DOS</b>	87	89	78	98	74	67	45	43	23	87	..
<b>ASP</b>	68	78	86	94	74	54	89	85	23	68	..
<b>MA</b>	56	72	89	92	87	90	87	78	81	56	..
<b>Practical 1</b>	89	78	89	90	81	90	76	87	77	89	..
<b>Practical 2</b>	86	87	90	90	87	90	87	88	80	86	..

10 rows × 26 columns



## Showing all the register numbers

```
In [11]: dd['Reg.No']
```

```
Out[11]: 0    Bp170501
          1    Bp170502
          2    Bp170503
          3    Bp170504
          4    Bp170505
          5    Bp170506
          6    Bp170507
          7    Bp170508
          8    Bp170509
          9    Bp170510
         10    Bp170511
         11    Bp170512
         12    Bp170513
         13    Bp170514
         14    Bp170515
         15    Bp170516
         16    Bp170517
         17    Bp170518
         18    Bp170519
         19    Bp170520
         20    Bp170521
         21    Bp170522
```

## Showing the 5<sup>th</sup> record of the dataset

```
In [16]: dd.iloc[5]
```

```
Out[16]: Reg.No      Bp170506
Student_Name  Muralidharan
Gender        Male
IOT           40
CC            90
DOS           67
ASP           54
MA            90
Practical 1   90
Practical 2   90
Name: 5, dtype: object
```

**Selecting the particular attributes and there corresponding values:**

```
In [20]: dd.iloc[[2,3,5,25],[0,1,2]]
```

```
Out[20]:
```

	Reg.No	Student_Name	Gender
2	Bp170503	Deepika	Female
3	Bp170504	Malini	Female
5	Bp170506	Muralidharan	Male
25	Bp170526	Mary	Female

## Selecting and showing first 5 records:

```
In [17]: dd[0:5]
```

```
Out[17]:
```

	Reg.No	Student_Name	Gender	IOT	CC	DOS	ASP	MA	Practical 1	Practical 2
0	Bp170501	Vasantha kumari	Female	70	69	87	68	56	89	86
1	Bp170502	Sugashini	Female	48	79	89	78	72	78	87
2	Bp170503	Deepika	Female	50	98	78	86	89	89	90
3	Bp170504	Malini	Female	60	86	98	94	92	90	90
4	Bp170505	Lokesh	Male	78	89	74	74	87	81	87

## Selecting the 25<sup>th</sup> record and showing the 1<sup>st</sup> values of the record:

```
In [44]: dd.iat[25,1]
```

```
Out[44]: 'Mary'
```

```
In [27]: dd.index
```

```
Out[27]: RangeIndex(start=0, stop=26, step=1)
```

## Selecting and showing the columns of the trained records.

```
In [28]: dd.columns
```

```
Out[28]: Index(['Reg.No', 'Student Name', 'Gender', 'IOT', 'CC', 'DOS', 'ASP', 'MA',  
              'Practical 1', 'Practical 2'],  
              dtype='object')
```

## Showing the entire data of the whole record without the attribute:

```
In [29]: dd.values
```

```
Out[29]: array([[ 'Bp170501', 'Vasantha kumari', 'Female', 70, 69, 87, 68, 56, 89,  
                 86],  
               [ 'Bp170502', 'Sugashini', 'Female', 48, 79, 89, 78, 72, 78, 87],  
               [ 'Bp170503', 'Deepika', 'Female', 50, 98, 78, 86, 89, 89, 90],  
               [ 'Bp170504', 'Malini', 'Female', 60, 86, 98, 94, 92, 90, 90],  
               [ 'Bp170505', 'Lokesh', 'Male', 78, 89, 74, 74, 87, 81, 87],  
               [ 'Bp170506', 'Muralidharan', 'Male', 40, 90, 67, 54, 90, 90, 90],  
               [ 'Bp170507', 'Carmel', 'Male', 33, 89, 45, 89, 87, 76, 87],  
               [ 'Bp170508', 'Pavithra 1', 'Female', 87, 78, 43, 85, 78, 87, 88],  
               [ 'Bp170509', 'Chinnapa raj', 'Male', 20, 78, 23, 23, 81, 77, 80],  
               [ 'Bp170510', 'Pavithra 2', 'Female', 70, 69, 87, 68, 56, 89, 86],  
               [ 'Bp170511', 'Stephy', 'Female', 60, 79, 89, 78, 72, 78, 87],  
               [ 'Bp170512', 'Shilpa', 'Female', 78, 98, 78, 86, 89, 89, 90],  
               [ 'Bp170513', 'Jasmin', 'Female', 89, 86, 23, 94, 23, 90, 90],  
               [ 'Bp170514', 'Ammu', 'Female', 78, 89, 74, 74, 87, 81, 87],  
               [ 'Bp170515', 'Vidhiya', 'Female', 89, 65, 54, 76, 23, 90, 90],  
               [ 'Bp170516', 'Sharadha', 'Female', 76, 89, 67, 89, 87, 76, 87],  
               [ 'Bp170517', 'Gokul', 'Male', 65, 78, 76, 65, 78, 87, 88],  
               [ 'Bp170518', 'Lawrance', 'Male', 43, 78, 76, 88, 81, 77, 80],  
               [ 'Bp170519', 'Suganthi', 'Female', 23, 69, 87, 68, 56, 89, 86],  
               [ 'Bp170520', 'Pavithra 3', 'Female', 65, 79, 89, 78, 72, 78, 87],  
               [ 'Bp170521', 'Murugan', 'Male', 87, 98, 78, 86, 89, 89, 90],  
               [ 'Bp170522', 'Prabavathi', 'Female', 89, 86, 98, 56, 92, 90, 90],  
               [ 'Bp170523', 'Swetha', 'Female', 78, 89, 74, 74, 87, 81, 87],  
               [ 'Bp170524', 'Ishvariya', 'Female', 56, 90, 34, 23, 90, 90, 90],  
               [ 'Bp170525', 'Sridhar', 'Male', 78, 23, 67, 89, 87, 76, 87],  
               [ 'Bp170526', 'Mary', 'Female', 89, 87, 85, 89, 89, 89, 90]],  
              dtype=object)
```

## Selecting the 26 rows and 10 columns:

```
In [30]: dd.shape
```

```
Out[30]: (26, 10)
```

## Counting the number of records:

```
In [31]: dd.count
```

```
Out[31]: <bound method DataFrame.count of
0  Bp170501  Vasantha kumari  Female  70  69  87  68  56  89
1  Bp170502  Sugashini  Female  48  79  89  78  72  78
2  Bp170503  Deepika  Female  50  98  78  86  89  89
3  Bp170504  Malini  Female  60  86  98  94  92  90
4  Bp170505  Lokesh  Male  78  89  74  74  87  81
5  Bp170506  Muralidharan  Male  40  90  67  54  90  90
6  Bp170507  Carmel  Male  33  89  45  89  87  76
7  Bp170508  Pavithra 1  Female  87  78  43  85  78  87
8  Bp170509  Chinnapa raj  Male  20  78  23  23  81  77
9  Bp170510  Pavithra 2  Female  70  69  87  68  56  89
10 Bp170511  Stephy  Female  60  79  89  78  72  78
11 Bp170512  Shilpa  Female  78  98  78  86  89  89
12 Bp170513  Jasmin  Female  89  86  23  94  23  90
13 Bp170514  Ammu  Female  78  89  74  74  87  81
14 Bp170515  Vidhiya  Female  89  65  54  76  23  90
15 Bp170516  Sharadha  Female  76  89  67  89  87  76
16 Bp170517  Gokul  Male  65  78  76  65  78  87
17 Bp170518  Lawrance  Male  43  78  76  88  81  77
18 Bp170519  Suganthi  Female  23  69  87  68  56  89
19 Bp170520  Pavithra 3  Female  65  79  89  78  72  78
20 Bp170521  Murugan  Male  87  98  78  86  89  89
21 Bp170522  Prabavathi  Female  89  86  98  56  92  90
22 Bp170523  Swetha  Female  78  89  74  74  87  81
23 Bp170524  Ishvariya  Female  56  90  34  23  90  90
24 Bp170525  Sridhar  Male  78  23  67  89  87  76
25 Bp170526  Mary  Female  89  87  85  89  89  89
```

```
In [32]: dd.size
```

```
Out[32]: 260
```

**Showing the records in descending order:**

```
In [33]: dd.iloc[:-1]
```

```
Out[33]:
```

	Reg.No	Student_Name	Gender	IOT	CC	DOS	ASP	MA	Practical 1	Practical 2
25	Bp170528	Mary	Female	89	87	85	89	89	89	90
24	Bp170525	Sridhar	Male	78	23	67	89	87	76	87
23	Bp170524	Ishwariya	Female	56	90	34	23	90	90	90
22	Bp170523	Swetha	Female	78	89	74	74	87	81	87
21	Bp170522	Prabavathi	Female	89	86	98	56	92	90	90
20	Bp170521	Murugan	Male	87	98	78	86	89	89	90
19	Bp170520	Pavithra 3	Female	65	79	89	78	72	78	87
18	Bp170519	Suganthi	Female	23	69	87	68	56	89	86
17	Bp170518	Lawrance	Male	43	78	76	88	81	77	80
16	Bp170517	Gokul	Male	65	78	76	65	78	87	88
15	Bp170516	Sharadha	Female	78	89	67	89	87	76	87
14	Bp170515	Vidhiya	Female	89	65	54	76	23	90	90
13	Bp170514	Ammu	Female	78	89	74	74	87	81	87
12	Bp170513	Jasmin	Female	89	86	23	94	23	90	90
11	Bp170512	Shilpa	Female	78	98	78	86	89	89	90
10	Bp170511	Stephy	Female	60	79	89	78	72	78	87
9	Bp170510	Pavithra 2	Female	70	69	87	68	56	89	86
8	Bp170509	Chinnapa raj	Male	20	78	23	23	81	77	80
7	Bp170508	Pavithra 1	Female	87	78	43	85	78	87	88
6	Bp170507	Carmel	Male	33	89	45	89	87	76	87
5	Bp170506	Muralidharan	Male	40	90	67	54	90	90	90
4	Bp170505	Lokesh	Male	78	89	74	74	87	81	87
3	Bp170504	Malini	Female	60	86	98	94	92	90	90
2	Bp170503	Deepika	Female	50	98	78	86	89	89	90
1	Bp170502	Sugashini	Female	48	79	89	78	72	78	87

```
In [34]: dd.iloc[:-1].cummax()
```

```
Out[34]:
```

	Reg.No	Student_Name	Gender	IOT	CC	DOS	ASP	MA	Practical 1	Practical 2
25	Bp170528	Mary	Female	89	87	85	89	89	89	90
24	Bp170528	Sridhar	Male	89	87	85	89	89	89	90
23	Bp170528	Sridhar	Male	89	90	85	89	90	90	90
22	Bp170528	Swetha	Male	89	90	85	89	90	90	90
21	Bp170528	Swetha	Male	89	90	98	89	92	90	90
20	Bp170528	Swetha	Male	89	98	98	89	92	90	90
19	Bp170528	Swetha	Male	89	98	98	89	92	90	90
18	Bp170528	Swetha	Male	89	98	98	89	92	90	90
17	Bp170528	Swetha	Male	89	98	98	89	92	90	90
16	Bp170528	Swetha	Male	89	98	98	89	92	90	90
15	Bp170528	Swetha	Male	89	98	98	89	92	90	90
14	Bp170528	Vidhiya	Male	89	98	98	89	92	90	90
13	Bp170528	Vidhiya	Male	89	98	98	89	92	90	90
12	Bp170528	Vidhiya	Male	89	98	98	94	92	90	90
11	Bp170528	Vidhiya	Male	89	98	98	94	92	90	90
10	Bp170528	Vidhiya	Male	89	98	98	94	92	90	90
9	Bp170528	Vidhiya	Male	89	98	98	94	92	90	90
8	Bp170528	Vidhiya	Male	89	98	98	94	92	90	90
7	Bp170528	Vidhiya	Male	89	98	98	94	92	90	90
6	Bp170528	Vidhiya	Male	89	98	98	94	92	90	90
5	Bp170528	Vidhiya	Male	89	98	98	94	92	90	90
4	Bp170528	Vidhiya	Male	89	98	98	94	92	90	90
3	Bp170528	Vidhiya	Male	89	98	98	94	92	90	90
2	Bp170528	Vidhiya	Male	89	98	98	94	92	90	90
1	Bp170528	Vidhiya	Male	89	98	98	94	92	90	90



```
In [35]: dd.iloc[:,-1].cummin()
```

```
Out[35]:
```

	Reg.No	Student_Name	Gender	IOT	CC	DOS	ASP	MA	Practical 1	Practical 2
25	Bp170526	Mary	Female	89	87	85	89	89	89	90
24	Bp170525	Mary	Female	78	23	67	89	87	76	87
23	Bp170524	Ishvariya	Female	56	23	34	23	87	76	87
22	Bp170523	Ishvariya	Female	56	23	34	23	87	76	87
21	Bp170522	Ishvariya	Female	56	23	34	23	87	76	87
20	Bp170521	Ishvariya	Female	56	23	34	23	87	76	87
19	Bp170520	Ishvariya	Female	56	23	34	23	72	76	87
18	Bp170519	Ishvariya	Female	23	23	34	23	56	76	86
17	Bp170518	Ishvariya	Female	23	23	34	23	56	76	80
16	Bp170517	Gokul	Female	23	23	34	23	56	76	80
15	Bp170516	Gokul	Female	23	23	34	23	56	76	80
14	Bp170515	Gokul	Female	23	23	34	23	23	76	80
13	Bp170514	Ammu	Female	23	23	34	23	23	76	80
12	Bp170513	Ammu	Female	23	23	23	23	23	76	80
11	Bp170512	Ammu	Female	23	23	23	23	23	76	80
10	Bp170511	Ammu	Female	23	23	23	23	23	76	80
9	Bp170510	Ammu	Female	23	23	23	23	23	76	80
8	Bp170509	Ammu	Female	20	23	23	23	23	76	80
7	Bp170508	Ammu	Female	20	23	23	23	23	76	80
6	Bp170507	Ammu	Female	20	23	23	23	23	76	80
5	Bp170506	Ammu	Female	20	23	23	23	23	76	80
4	Bp170505	Ammu	Female	20	23	23	23	23	76	80
3	Bp170504	Ammu	Female	20	23	23	23	23	76	80
2	Bp170503	Ammu	Female	20	23	23	23	23	76	80
1	Bp170502	Ammu	Female	20	23	23	23	23	76	80

```
In [37]: print(dd.shape)
```

```
(26, 10)
```

## Showing the values of the particular attribute of the dataset:

```
In [38]: print(dd['Reg.No'].unique())
```

```
['Bp170501' 'Bp170502' 'Bp170503' 'Bp170504' 'Bp170505' 'Bp170506'  
'Bp170507' 'Bp170508' 'Bp170509' 'Bp170510' 'Bp170511' 'Bp170512'  
'Bp170513' 'Bp170514' 'Bp170515' 'Bp170516' 'Bp170517' 'Bp170518'  
'Bp170519' 'Bp170520' 'Bp170521' 'Bp170522' 'Bp170523' 'Bp170524'  
'Bp170525' 'Bp170526']
```

## Showing the 5 fields of the attributes:

```
In [19]: dd.iloc[:,0:5]
```

```
Out [19]:
```

	<b>Reg.No</b>	<b>Student_Name</b>	<b>Gender</b>	<b>IOT</b>	<b>CC</b>
<b>0</b>	Bp170501	Vasantha kumari	Female	70	69
<b>1</b>	Bp170502	Sugashini	Female	48	79
<b>2</b>	Bp170503	Deepika	Female	50	98
<b>3</b>	Bp170504	Malini	Female	60	86
<b>4</b>	Bp170505	Lokesh	Male	78	89
<b>5</b>	Bp170506	Muralidharan	Male	40	90
<b>6</b>	Bp170507	Carmel	Male	33	89
<b>7</b>	Bp170508	Pavithra 1	Female	87	78
<b>8</b>	Bp170509	Chinnapa raj	Male	20	78
<b>9</b>	Bp170510	Pavithra 2	Female	70	69
<b>10</b>	Bp170511	Stephy	Female	60	79
<b>11</b>	Bp170512	Shilpa	Female	78	98
<b>12</b>	Bp170513	Jasmin	Female	89	86
<b>13</b>	Bp170514	Ammu	Female	78	89

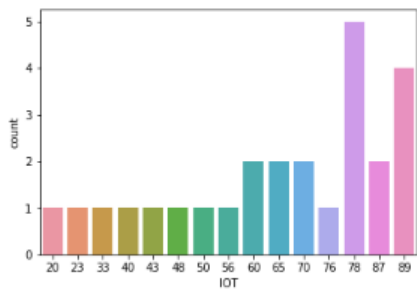
---

**Showing the IOT subject students mark details:**

```
In [40]: print(dd.groupby('IOT').size())

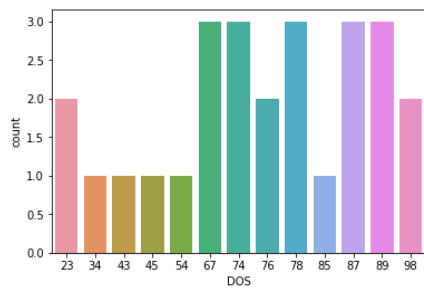
import seaborn as sns
sns.countplot(dd['IOT'], label="Count")
plt.show()
```

```
IOT
20  1
23  1
33  1
40  1
43  1
48  1
50  1
56  1
60  2
65  2
70  2
76  1
78  5
87  2
89  4
dtype: int64
```

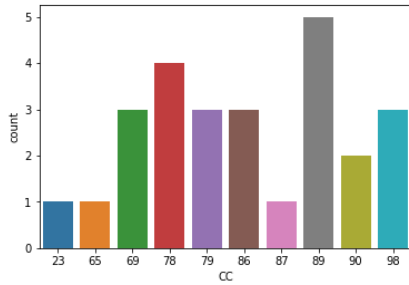


## Showing the DOS subject students mark details:

```
In [52]: 1 import seaborn as sns
2         sns.countplot(dd['DOS'], label="Count")
3         plt.show()
```



```
In [50]: 1 import seaborn as sns
2         sns.countplot(dd['CC'], label="Count")
3         plt.show()
```

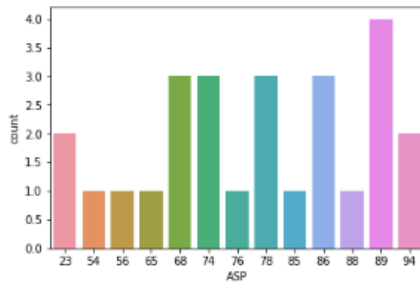


## Showing the ASP.NET subject students mark details:

```
In [53]: 1 print(dd.groupby('ASP').size())
```

```
ASP
23    2
54    1
56    1
65    1
68    3
74    3
76    1
78    3
85    1
86    3
88    1
89    4
94    2
dtype: int64
```

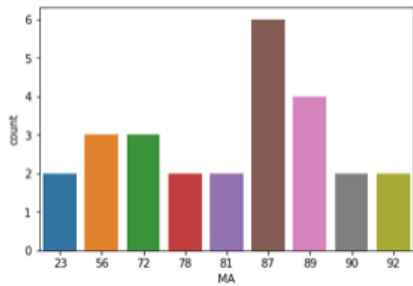
```
In [54]: 1 import seaborn as sns
2         sns.countplot(dd['ASP'], label="Count")
3         plt.show()
```



## Showing the MOBILE APPLICATION subject students mark details:

```
In [56]: 1 print(dd.groupby('MA').size())
2
3 import seaborn as sns
4 sns.countplot(dd['MA'],label="Count")
5 plt.show()
```

```
MA
23    2
56    3
72    3
78    2
81    2
87    6
89    4
90    2
92    2
dtype: int64
```



## Predicting the null values:

```
In [57]: 1 dd.isnull().sum()
```

```
Out[57]: Reg.No      0
Student_Name  0
Gender        0
IOT          0
CC           0
DOS          0
ASP          0
MA           0
Practical 1  0
Practical 2  0
dtype: int64
```

## Making the cross table function:

```
In [58]: 1 pd.crosstab(dd['Gender'],dd['Reg.No'])
```

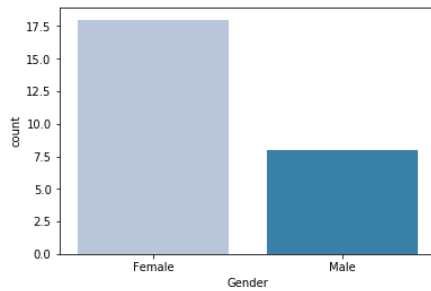
```
Out[58]: Reg.No Bp170501 Bp170502 Bp170503 Bp170504 Bp170505 Bp170506 Bp170507 Bp170508 Bp170509 Bp170510 ... Bp170517 Bp170518 Bp170519 Bp170520
Gender
Female      1      1      1      1      0      0      0      1      0      1 ...      0      0      1
Male        0      0      0      0      1      1      1      0      1      0 ...      1      1      0
```

2 rows x 26 columns



## Representing the number of male and female in bar chart:

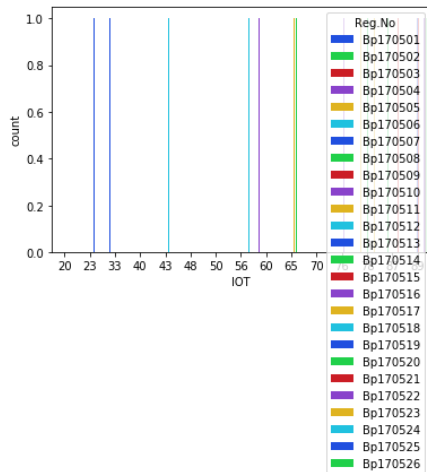
```
In [59]: 1 sns.countplot(x='Gender', data=dd, palette='PuBu')
        2 plt.show()
```



```
In [60]: 1 dd.Gender.value_counts()
```

```
Out[60]: Female    18
         Male      8
         Name: Gender, dtype: int64
```

```
In [61]: 1 sns.countplot(x='IOT', data=dd, hue='Reg.No', palette='bright')
        2 plt.show()
```



## Marks greater than 50 in IOT subject:

```
In [62]: 1 dd[dd.IOT > 50]
```

```
Out [62]:
```

	Reg.No	Student_Name	Gender	IOT	CC	DOS	ASP	MA	Practical 1	Practical 2
0	Bp170501	Vasantha kumari	Female	70	69	87	68	56	89	88
3	Bp170504	Malini	Female	60	86	98	94	92	90	90
4	Bp170505	Lokesh	Male	78	89	74	74	87	81	87
7	Bp170508	Pavithra 1	Female	87	78	43	85	78	87	88
9	Bp170510	Pavithra 2	Female	70	69	87	68	56	89	88
10	Bp170511	Stephy	Female	60	79	89	78	72	78	87
11	Bp170512	Shilpa	Female	78	98	78	86	89	89	90
12	Bp170513	Jasmin	Female	89	86	23	94	23	90	90
13	Bp170514	Ammu	Female	78	89	74	74	87	81	87
14	Bp170515	Vidhiya	Female	89	65	54	76	23	90	90
15	Bp170516	Sharadha	Female	76	89	67	89	87	76	87
16	Bp170517	Gokul	Male	65	78	76	65	78	87	88
19	Bp170520	Pavithra 3	Female	65	79	89	78	72	78	87
20	Bp170521	Murugan	Male	87	98	78	86	89	89	90
21	Bp170522	Prabavathi	Female	89	86	98	56	92	90	90
22	Bp170523	Swetha	Female	78	89	74	74	87	81	87
23	Bp170524	Ishvariya	Female	56	90	34	23	90	90	90
24	Bp170525	Sridhar	Male	78	23	67	89	87	76	87
25	Bp170526	Mary	Female	89	87	85	89	89	89	90

## Marks greater than 50 in DOS subject:

```
In [63]: 1 dd[dd.DOS > 50]
```

```
Out [63]:
```

	Reg.No	Student_Name	Gender	IOT	CC	DOS	ASP	MA	Practical 1	Practical 2
0	Bp170501	Vasantha kumari	Female	70	69	87	68	56	89	88
1	Bp170502	Sugashini	Female	48	79	89	78	72	78	87
2	Bp170503	Deepika	Female	50	98	78	86	89	89	90
3	Bp170504	Malini	Female	60	86	98	94	92	90	90
4	Bp170505	Lokesh	Male	78	89	74	74	87	81	87
5	Bp170506	Muralidharan	Male	40	90	67	54	90	90	90
9	Bp170510	Pavithra 2	Female	70	69	87	68	56	89	88
10	Bp170511	Stephy	Female	60	79	89	78	72	78	87
11	Bp170512	Shilpa	Female	78	98	78	86	89	89	90
13	Bp170514	Ammu	Female	78	89	74	74	87	81	87
14	Bp170515	Vidhiya	Female	89	65	54	76	23	90	90
15	Bp170516	Sharadha	Female	76	89	67	89	87	76	87
16	Bp170517	Gokul	Male	65	78	76	65	78	87	88
17	Bp170518	Lawrance	Male	43	78	76	88	81	77	80
18	Bp170519	Suganthi	Female	23	69	87	68	56	89	88
19	Bp170520	Pavithra 3	Female	65	79	89	78	72	78	87
20	Bp170521	Murugan	Male	87	98	78	86	89	89	90
21	Bp170522	Prabavathi	Female	89	86	98	56	92	90	90
22	Bp170523	Swetha	Female	78	89	74	74	87	81	87
24	Bp170525	Sridhar	Male	78	23	67	89	87	76	87
25	Bp170526	Mary	Female	89	87	85	89	89	89	90

## **CHAPTER – V CONCLUSION**

### **5.1 Conclusion**

- This project has been successfully developed and interpreted and system was developed according to the user requirement.
- The system produces accurate result and it also reduces a lot of overheads, which the manual system faced.
- The system was thoroughly tested according to the testing methodologies.



# Natural Language Processing

*by* S Sagayaraj

---

**Submission date:** 02-Apr-2019 12:32PM (UTC+0530)

**Submission ID:** 1104361234

**File name:** BP170522\_Document\_classification.pdf (308.95K)

**Word count:** 3753

**Character count:** 20051

DOCUMENT CLASSIFICATION WITH NATURAL LANGUAGE PROCESSING

A Software Project

26

Submitted in partial fulfillment for the award of the degree of

MASTER OF COMPUTER SCIENCE

By

PRABAVATHI P

(REG. NO: BP170522)

5

Under the Guidance of

Mrs. M. POOVIZHI. M.C.A., M.Phil,



PG DEPARTMENT OF COMPUTER SCIENCE

SACRED HEART COLLEGE (AUTONOMOUS)

TIRUPATTUR, VELLORE DT – 635 601

APRIL-2019

## CERTIFICATE

This is to certify that the project work entitled “DOCUMENT CLASSIFICATION WITH NATURAL LANGUAGE PROCESSING” is submitted to Sacred Heart College (Autonomous), Tirupattur-635 601, Vellore District by PRABAVATHI P (REG.NO: BP170522) for the partial fulfillment for the award of the Degree of Master of Science in Computer Science is a bonafied record of the work carried out by him, under my guidance and supervision.

Signature of the Project Guide

Mrs. M.Poovizhi, M.C.A., M.Phil.

Submitted for Viva- Voice examination held on \_\_\_\_\_

Internal Examiner

1.

Internal Examiner

2.

## ACKNOWLEDGEMENT

I thank God Almighty for his blessings and graces by which I was able to complete project work successfully.

I sincerely thank my parents who have given the gift of life to me to attain many achievements.

I express my deep sense of gratitude to <sup>36</sup> **Dr. S. Sagayaraj**, Head of the **Department of Computer Science** and My Project Guide **Mrs. M.Poovizhi** who is the source of inspiration to pursuer with every efficient work. And I am heart fully honor her valuable guidance and encouragement of my every activities.

<sup>5</sup> I would like to thank the entire teaching and non-teaching staff members, Department of Computer Science for their help to complete the project successfully.

Finally, I thank <sup>5</sup> each and every one of my friends, especially who have assisted me in completing the project work.

## Document Classification with Natural Language Processing

### ABSTRACT:

Text Document classification is the process of assigning the class labels to the text documents in the training model it's used by the large enough collection. It's a more multiplex algorithm to deal with the known attributes and also a limited set of possible values for the each set of attribute. Preprocessing and the representation of the reduced data is the essential for obtaining the effective result of text classification. Text classification is the act of the assigning classes to the text documents. This project helps to analyze and test the dataset with trained dataset model. Natural Language analysis is followed by the statistical analysis process. It's the process to provide the size reduction of the document and also improve the performance of the classification without the compromising accuracy.

### INTRODUCTION:

In the world we have the huge amount of data on the web it's randomly increased day by day. All the huge information of the data is valuable and also it contain most of the information is text. So it becomes a very complex or a challenge for the humans to identify the relevant information. In this project document classification is helps to overcome these complex problems.

Text document classification is the dividing the set of inputs for the document into the two /more classes in the each document belong to the one / multiple classes. Text classification is maybe manual / automated. The manual classification helps to consume the time and high accuracy. The automated classification process is the fast and more efficient it's automatically classify the document.

## SCOPE:

The scope of this project is used to bring the outcome of the document details regarding the document count and running time. It can be classify them into different norms and category by the classifying the dataset. It can be predict the different type of possibilities of various documents. We can predict the document title and total number of text file will be calculate and generating the graph. Finally, it can also plot the outcome predicted values in histogram and can perform them into the graphical representations.

## PROBLEM STATEMENT:

In this project the major problem for all the data are manual document and calculating time for the database due to the time duration is to long huge man power is involved so that the new system in this project evaluates outcomes of search problem in this project.

## Product Position Statement:

For	IT Industry
Who	IT Members
That	It is more flexible
Our Product	Provide up-to-date information and easy to maintain all the data. An classify them and make prediction and represent them in graphical representation manner.

## SYSTEM ANALYSIS

System Study:

Introduction:

**Document Classification** this is the application for predicting the document details, by using the python language. In this project is to analysis made of every document to analyzing the data. It should be calculating the text files and also calculating the times of seconds. We can also represent the predicted results of the document in graphical representation by plotting in bar diagram, different kind of plotting method. The main purpose of this project is to identify the document and analyzing data of the text file and also calculating the time and also base time of the graph will be generate.

Overview:

Analysis of the text document files is analyzing and predicting by Naïve-Bayes in Natural Language Processing. This is a predicting application of the document file details is using python language. In this project is to analyze the every document and it should be calculating the data and also calculating the number of data files in the particular document and display the graph based on the running times. So that the proposed the new way is representing the data into graphical representations by which the graphs are able to identify the results easily.

It can also represent the predicted result of the document in graphical representation by plotting diagram and also different kind of plotting method. The main purpose of this project is to identify the document data and text file calculation and also generating the graph based on the text file format.

### Technical Feasibility:

In this project it can strongly says that it is technically feasible. So there will not be much difficulty in getting required resources for the development and maintaining the system as well. All the resources of the organization needed for the development of the software as well as the maintenance of the same is available in the organization here we are utilizing the resources which are available already.

### Economical Feasibility:

Development of this application is highly economically feasible. The resources of the organization needed not spend much money for the development of the system already available. The only thing is to be done is making an environment for the development with the effective supervision. If we are doing so, we can increase the maximum usability of the corresponding resources. Even after the development process, the organization will not be in condition to invest more in the organization. Therefore the system is economically feasible.

### Classification details of this project:

#### Purpose:

To produce the complete details about the classification and prediction

#### Overview:

The user, who wants to know about the prediction of any case, can easily get the details in graphical representations.

#### Entry Criteria:

The users can directly import the data set, train and test them to get predictions.



### Input:

Here, the input is test data from the data set.

### Steps Solved:

Import text file into the python and perform the code form train and test using classification algorithms and make predict the desired output.

### Output:

The required detail is shown to the knowledge seekers.

## SYSTEM CONFIGURATION

### REQUIREMENT SPECIFICATION:

29

#### Hardware Requirements:

Processor	Intel i3 processor
RAM	4 GB
Hard Disk	500 GB
Input/output	Keyboard, mouse, monitor

#### Software Requirements:

Front End	Python 3.7 (Jupyter Notebook)
Document	Microsoft Word 2007

## NATURAL LANGUAGE PROCESSING:

Natural Language Processing is the subfield of the computer science and the artificial intelligence. It's alarmed with the communications between the computer and creature language. It's to process and analyze the large amount of data in Natural Language.

Natural Language Processing is used to apply the machine learning algorithms to text and the speech recognition/categorization. The development of the Natural Language Processing is the challenging process. Because computers are traditionally require the humans to 'speak' to them in to a programming language and it is highly structured or a limited number of voice commands.

NLP is used to interpret the free text and its make analyzable. So there is a large amount of information is stored in the free text files, like a business records and the medical records. NLP allows the analyst to shift the massive troves of free text to find the relevant information of the files.

## WORK FLOW OF NATURAL LANGUAGE PROCESSING:

Current approach of natural language processing is based on the Deep Learning. This approach is a type of Artificial Intelligence and its uses the pattern in the data to improve the programs to understanding. Deep Learning model is required a massive amount of labeled data is to train and identify the relevant categorization Earlier NLP is involved a more rules based approaches , where it's a simpler machine learning algorithms.

## IMPORTANCE OF NLP:

The benefit of the natural language processing can be considering the following statement that is cloud computing insurance should be the part of every service level agreement. These types of agreements that appear in frequently in the human language and the machine learning an algorithm is historically have bad interpreting. But the improvement of Deep Learning and Artificial Intelligence algorithms is effectively interpreted. Advances of Natural Language processing make it is easy to analyze and learn a greater range of the data sources.

## TEXT CLASSIFICATION PROCESS:

Text classification process it contains various sub phases, each phase has its own importance and needs.

Data collection

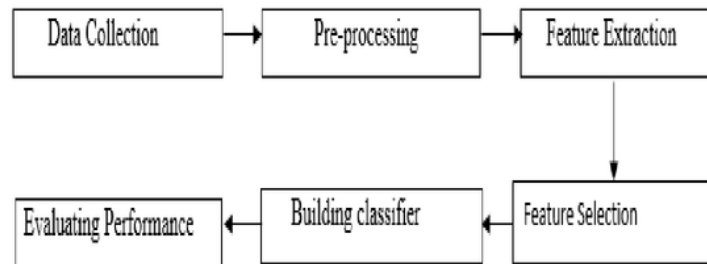
Pre-processing phase

Feature Extraction

Feature Selection

Building a classifier

Performance Evaluation



## DATA COLLECTION:

The first step of classification process is the corpus building it consist of collect the dissimilar formats of the documents like a **HTML, PDF, DOC, WEB** substance and so on. so these documents are used to during the training and testing classifier.

## PRE-PROCESSING:

Pre-processing is used to represent the text documents into the clear word format. The documents are prepared for the next step in the classification processes are represented by the great amount of features.

FEATURES:

### Tokenization:

A document is treated as the string and then it is partitioned into the list of tokens.

### Removing stop words:

Stopping words are 'the', 'a', 'and', and etc. these are regularly occur. So that the insignificant words are need to be removed.

### Stemming word:

Stemming Algorithm it converts the different words from into the similar canonical forms. So these steps are the process of conflating tokens into their root form. Eg: Connection to connect and computing to compute.

## FEATURE EXTRACTION:

The document classification used in the individual of the pre-processing performance that's used to decrease the convolution of the documents and it construct them into easier to switch, the documents have to be distorted into the full text edition to a certificate vector. The unaniously using document exhibition is called the vector space model. So the credentials are representing by the vectors of terms.

### Restrictions:

1. Towering dimensionality of the document demonstration.
2. Failure of association with the flanking of the terms.
3. Thrashing of semantic relationships to facilitate is exists along with the expressions into a certificate.

### FEATURE SELECTION:

After completing pre-processing and the feature extraction that is significant march of the transcript taxonomy is aspect variety to assemble the vector liberty, which is improve the scalability, competence and precision of the transcript organization. The foremost objective of the aspect variety is to decide on the separation of features from the creative documents. Feature selection is perform to maintenance the terms into the utmost achieve according to the pre-determined measurement substance of the words.

Major problem of the content categorization is the towering dimensionality of the aspect liberty. So the several aspect assessment metrics cover be noticed into the <sup>3</sup> information gain, term frequency and the chi-square, expected cross entropy and odds ratio, mutual information, the weight of the evidence of the text and finally Gini index.

### CLASSIFICATION:

The automatic classification of the documents is the predefined categories has been observed an active attention of the documents can be classified into three ways.

Unsupervised classification

Supervised classification

Semi supervised categorization method.

<sup>3</sup> The task of the automatic classification is extensively studied and the rapid progress in this area, including the approaches of the machine learning.

Naïve Bayes Classifier

Decision Tree

K-Nearest Neighbor

Support vector machines

Neural Networks

## PERFORMANCE EVALUATION:

The performance of the text classification system is evaluated using four metrics

Accuracy

Precision

Recall

F1 measure

Precision is measure the exactness of the classifier. The higher precision is less false positives, while the lower precision is more false positives.

$$\text{Precision} = \frac{\text{Number of correct extracted text}}{\text{Total number of extracted texts}}$$

Recall is measures the completeness or the sensitivity of the classifier. Higher recall is the lass false negatives and lower recall is more false negatives

$$\text{Recall} = \frac{\text{Number of correct extracted texts}}{\text{Total no of annotated texts}}$$

10

Precision and recall is combined to produce the single metrics that is known as F-measure and which is represents the weighted harmonic of precision and recall. The advantage of the F-measure is able to rate the system with a one unique rating.

$$\text{F-measure} = \frac{2 * \text{precision} * \text{recall}}{\text{Precision} + \text{recall}}$$

Precision + recall

18

Accuracy is the overall degree which is instances have been correctly classified into the formula,

18

$$\text{Accuracy} = \frac{\text{Number of correctly classified instances}}{\text{Total number of instances}}$$

Total number of instances

### Approaches of Text Classification Task:

In the literature survey is done in the two types of techniques.

- Supervised learning
- Unsupervised learning

24

### Naïve Bayes Classifier:

Naïve Bayes Classifier is the simple and probabilistic classifier is fully based on the applying Bayes theorem with the strong independent assumption. One of the main Naïve Bayes model is work well for the text domain because it is the evidences of the 'Vocabularies' and 'words' are appearing in the text and the size of the words is typically used in the range of thousand.

The huge size of the vocabularies makes the Naïve Bayes Model is works well for the text document classification problem. The Naïve Bayes Classifier model is widely used algorithm for the purpose of the text document classification.

### A Basic Naïve Bayes Model:

The most basic models for the text classification is Naïve Bayes Model. The Naïve Bayes classification model will predicts the document topic,



$$Y = \{c_1, c_2, \dots, c_{20}\}$$

Where  $C_k$  is the class otherwise topic based on the document features  $x \in \mathbb{N}^p$ , and  $p$  denotes the number of items in bag of words list.

The feature vector contains count or length of  $x_i$  for the tf-idf value of the  $i$ -th term in the bag of words list.

$$\mathbf{x} = [x_1, x_2, \dots, x_p]$$

Using Bayes theorem can develop a model to predict the topic class ( $c_k$ ) of a document its feature vector  $x$ .

The Naïve Bayes model makes the “Naive” assumption the probability of each term’s tf-idf is **conditionally independent of every other** term. This reduces **the conditional** probability function to **the product**

$$P(x_1, \dots, x_p | C_k) = \prod_{i=1}^p P(x_i | C_k)$$

Moreover Bayes theorem for classification of problem becomes,

$$P(C_k | x_1, \dots, x_p) = \frac{P(C_k) \prod_{i=1}^p P(x_i | C_k)}{P(x_1, \dots, x_p)}$$

Also, Multi-nominal Naïve Bayes classifier using scikit learn commands are:

$$P(x_1, \dots, x_p | C_k) = \frac{(\sum_{i=1}^p x_i)!}{\prod_{i=1}^p x_i!} \prod_{i=1}^p p_{k,i}^{x_i}$$

Accuracy: 0.77389803505

**Scikit learn Pipelines:**

In Scikit-learn pipelines are a sequence of transforms followed by a final estimator. Intermediate steps within the pipeline must be 'transforms', they must implement fit and transform methods. The Count Vectorizer and tfidf Transformer are used as transformers in above example. The final estimator of the pipeline is only needs to implement the fit method. It sees the simplicity of pipelines by using it to re-implement our above analysis using the Naive Bayes model:

Accuracy: 0.77389803505

### **An Improved Naive Bayes Model:**

In this model it seems to improve by removing stop words which are common words in the english language and do not add any information into the text. These includes words such as, "the", "at", "is", etc. Finally can remove them in the CountVectorizer constructor call,

Accuracy: 0.831651619756

A less improvement, but an improvement is none-the-less. In the Naive Bayes classifier can be fast compared to more sophisticated methods due to the decoupling of the class conditional feature distributions, i.e.

$$P(x_1, \dots, x_p | C_k) = \prod_{i=1}^p P(x_i | C_k)$$

The decoupling is the class conditional distributions it allows for each distribution to be independently estimated as a one dimensional distribution and helps to alleviate problems with the curse of dimensionality.

### **Implementation of Navies-Bayes to an improved Navies-Bayes Algorithm:**

**To obtain the training and testing sets directly with the following commands:**

21

```
from sklearn.datasets import fetch_20newsgroups
```

```
twenty_train= fetch_20newsgroups(subset='train', shuffle=True)
```

```
twenty_test = fetch_20newsgroups(subset='test', shuffle=True)
```

**To view the total number of articles:**

```
len(twenty_train.data) + len(twenty_test.data)
```

**All the documents contain the data set below the following 20 topics,**

6

```
twenty_train.target_names
```

**Result:**

```
['alt.atheism','comp.graphics','comp.os.ms.`windows.misc`',  
'comp.sys.ibm.pc.hardware','comp.sys.mac.hardware','comp.windows.x','misc.forsale','rec.autos','  
rec.motorcycles','rec.sport.baseball','rec.sport.hockey','sci.crypt','sci.electronics','sci.med','sci.spac  
e','soc.religion.christian','talk.politics.guns','talk.politics.mideast','talk.politics.misc','talk.religion.  
misc']
```

**To particularly view at the actual message within the documents,**

9

```
print("\n".join(twenty_train.data[0].split("\n")))
```

**Result:**

8

```
From: leroxst@wam.umd.edu (where's my thing)
```

```
Subject: WHAT car is this!?
```

```
Nntp-Posting-Host: rac3.wam.umd.edu
```

```
Organization: University of Maryland, College Park
```

```
Lines: 15
```

It is wondering if anyone out there could enlighten on this car It seems

the other day. It was a 2-door sports car, 60s/ delayed, <sup>11</sup> early 70s. It was called a Bricklin. The doors were really small. In insertion, the front bumper was separate from the rest of the body.

<sup>11</sup>  
Thanks,

- IL

-brought to you by your neighborhood Lersxt -

**Using a categorical variable they are encoding the target classes**

```
twenty_train.target[0]
```

**Result: 7**

**A Basic Naive Bayes Model :**

To instantiate a multinomial Naive Bayes classifier using the Scikit-learn library and fit it to our tf-idf matrix using the commands,

```
27  
from sklearn.naive_bayes  
import MultinomialNB  
from sklearn.metrics  
import accuracy_score  
mod = MultinomialNB()  
mod.fit(X_train_tfidf, twenty_train.target)
```

**Result:**

```
30  
MultinomialNB(alpha=1.0, class_prior=None, fit_prior=True)
```

The word is `alpha=1` means it is using Laplace smoothing. Otherwise, now see the accuracy of classifier using Scikit-learn's `accuracy_score` function:

```

20 X_test_tf = count_vect.transform(twenty_test.data)
X_test_tfidf = tfidf_transformer.transform(X_test_tf)
predicted = mod.predict(X_test_tfidf)
print("Accuracy:", accuracy_score(twenty_test.target, predicted))

```

### Result:

Accuracy: 0.77389803505

### To plot a graph for Navies-Bayes Algorithm:

```

import matplotlib.pyplot as plt
4 x=[
    "alt.atheism",
    "comp.graphics", "comp.os.ms-
windows.misc", "comp.sys.ibm.pc.hardware", "comp.sys.mac.hardware", "comp.windows.x",
    "misc.forsale", "rec.autos", "rec.motorcycles", "rec.sport.baseball", "rec.sport.hockey", "sci.crypt", "
sci.electronics", "sci.med", "sci.space", "soc.religion.christian", "talk.politics.guns", "talk.politics.mi
deast", "talk.politics.misc", "talk.religion.misc"]

```

```

y=[0.80,0.78,0.79,0.68,0.86,0.87,0.87,0.88,0.93,0.91,0.88,0.75,0.84,0.92,0.82,0.62,0.66,0.95,0.9
4,0.95]

```

```

y1=[0.69,0.72,0.72,0.81,0.81,0.78,0.80,0.91,0.96,0.92,0.98,0.96,0.65,0.79,0.94,0.96,0.95,0.94,0.
52,0.24]

```

```

y2=[0.74,0.75,0.75,0.74,0.84,0.82,0.83,0.90,0.95,0.92,0.93,0.84,0.74,0.85,0.88,0.76,0.78,0.94,0.
67,0.38]

```

```

y3=[319,389,394,392,385,395,390,396,398,397,399,396,393,396,394,398,364,376,310,251]

```

```

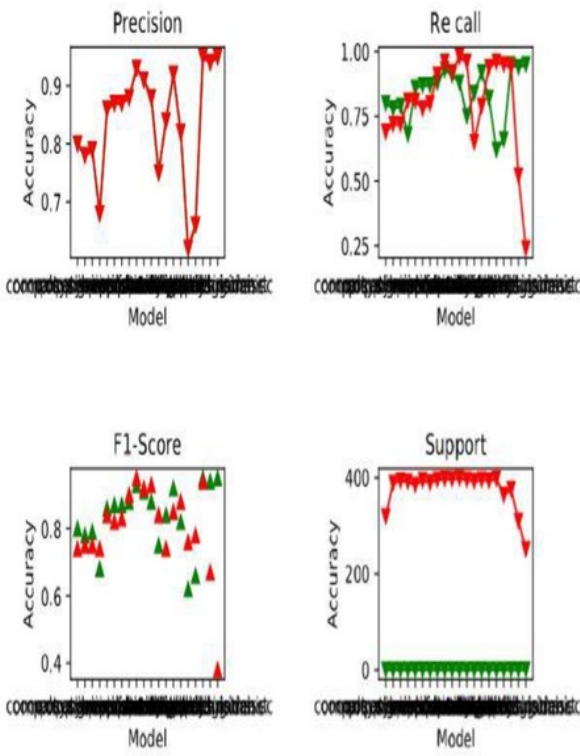
9 plt.subplot(2,2,1)
plt.subplots_adjust(hspace=5)
plt.subplots_adjust(wspace=10)
19 plt.plot(x,y,'gv-',x,y,'rv-')
plt.title("Precision")
9 plt.xlabel("Model")
plt.ylabel("Accuracy")

```

```
plt.subplot(2,2,2)
plt.subplots_adjust(hspace=1)
plt.subplots_adjust(wspace=1)
plt.plot(x,y,'gv-',x,y1,'rv-')
plt.title("Re call")
plt.xlabel("Model")
plt.ylabel("Accuracy")
plt.subplot(2,2,3)
plt.subplots_adjust(hspace=1)
plt.subplots_adjust(wspace=1)
plt.plot(x,y,'g^',x,y2,'r^')
plt.title("F1-Score")
plt.xlabel("Model")
plt.ylabel("Accuracy")

plt.subplot(2,2,4)
plt.subplots_adjust(hspace=1)
plt.subplots_adjust(wspace=1)
plt.plot(x,y,'gv-',x,y3,'rv-')
plt.title("Support")
plt.xlabel("Model")
plt.ylabel("Accuracy")
plt.savefig("C:/Users/malini/Desktop/graphnlp1.pdf")
plt.show()
```

**Result:**



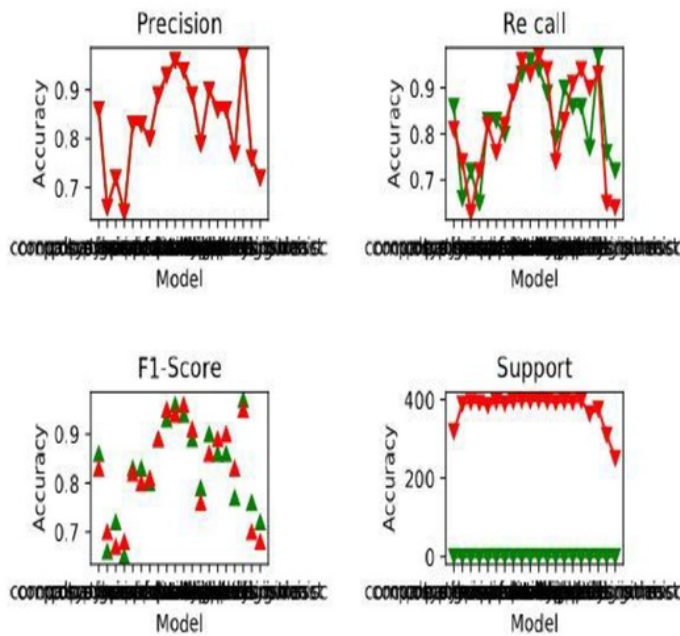
**An improved Navies-Bayes model:**

To look for an improved model by removing stop words which are common words in the english language and do not add any information into the text. These includes words such as, "the", "at", "is", etc. Also can remove them in the CountVectorizer constructor call,

```

32 pipe = Pipeline([('vect', CountVectorizer(stop_words='english')),
('tfidf', TfidfTransformer()),
('model', MultinomialNB())])
14 mod = pipe.fit(twenty_train.data, twenty_train.target)
14 predicted = mod.predict(twenty_test.data)
print(classification_report(twenty_test.target,
predicted, target_names=twenty_test.target_names))
print("Accuracy:", accuracy_score(twenty_test.target, predicted))

```





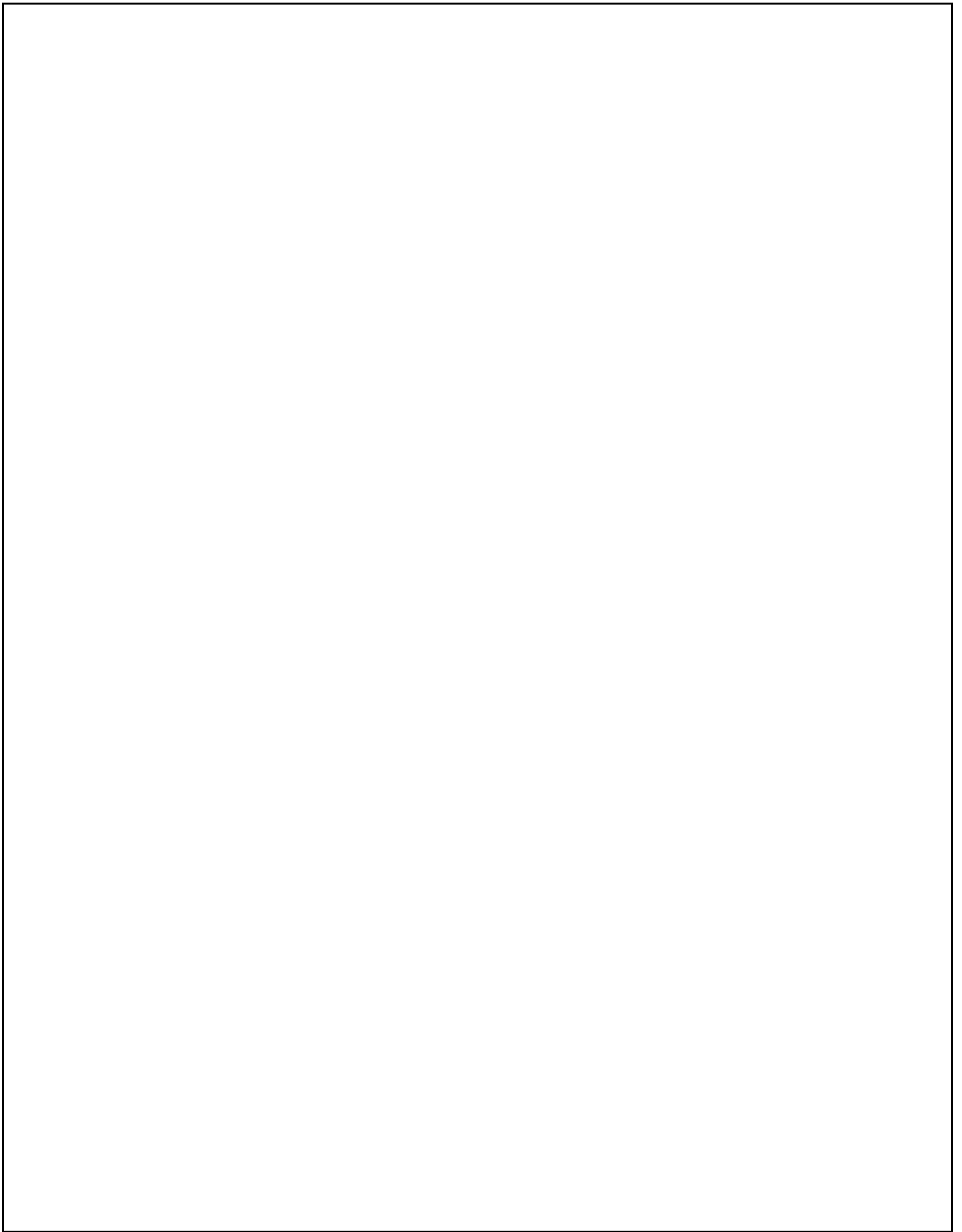
**Future Scope:**

- One thing that did not address was the topic of stemming and lemmatisation, both do with reducing a word down to its base form and also can be used to improve the performance of text classification models.
- Lemmatisation differs from stemming topic because it depends only identifying the intended part of speech and produce meaning of a word in a sentence.
- Natural Language Tool Kit or NLTK and spaCy libraries and both provided by the Stemmers and lemmatizers.

**Conclusion:**

In this paper, it covers document classification using Scikit-learn and the 20 News Groups dataset. It went over the basics of term frequency-inverse document frequency, pipelines and the Naive Bayes classifier. While possibilities that it able to improve the accuracy of the basic Naive Bayes classifier from 77% to 83%, there are much more sophisticated models like a support vector machines it also achieve higher performance. One important thing that , it did not address the topic of stemming and lemmatisation, where both reducing a word down to its base form and it can be used to improve the performance of text classification models. In this model of Lemmatisation and stemming ,the stemming differs from it. Because it

Only depends on identifying the intended part of speech and meaning of a word in a sentence. Natural Language Tool Kit or NLTK and spaCy libraries and both provided by the Stemmers and lemmatizers



# Natural Language Processing

---

## ORIGINALITY REPORT

---

**31** %

SIMILARITY INDEX

**26** %

INTERNET SOURCES

**15** %

PUBLICATIONS

**24** %

STUDENT PAPERS

---

## PRIMARY SOURCES

---

**1** [de.slideshare.net](https://de.slideshare.net) Internet Source **4** %

---

**2** [searchbusinessanalytics.techtarget.com](https://searchbusinessanalytics.techtarget.com) Internet Source **3** %

---

**3** Submitted to Higher Education Commission Pakistan Student Paper **2** %

---

**4** [scidok.sulb.uni-saarland.de](https://scidok.sulb.uni-saarland.de) Internet Source **2** %

---

**5** Submitted to Sacred Heart College Student Paper **2** %

---

**6** [srome.github.io](https://srome.github.io) Internet Source **2** %

---

**7** Submitted to Eiffel Corporation Student Paper **2** %

---

**8** [fnl.es](https://fnl.es) Internet Source **1** %

---

**9** [documents.mx](https://documents.mx)

Internet Source

1%

10

Submitted to National Institute of Technology,  
Raipur

Student Paper

1%

11

[www.machinelearningplus.com](http://www.machinelearningplus.com)

Internet Source

1%

12

Submitted to IPMC Kumasi

Student Paper

1%

13

Mary Joy Canon, Arlene Satuito, Christian Sy.  
"Determining Disaster Risk Management  
Priorities through a Neural Network-Based Text  
Classifier", 2018 International Symposium on  
Computer, Consumer and Control (IS3C), 2018

Publication

1%

14

[khanrc.tistory.com](http://khanrc.tistory.com)

Internet Source

1%

15

Submitted to Fort Valley State Univeristy

Student Paper

1%

16

[scikit-learn.org](http://scikit-learn.org)

Internet Source

1%

17

[www.slideshare.net](http://www.slideshare.net)

Internet Source

1%

18

Submitted to University of Sheffield

Student Paper

1%

---

19

K. G. Srinivasa, Siddesh G. M., Srinidhi H..  
"Network Data Analytics", Springer Nature,  
2018

Publication

<1%

---

20

Submitted to University of Westminster

Student Paper

<1%

---

21

[www.antidot.net](http://www.antidot.net)

Internet Source

<1%

---

22

Submitted to Indian School of Business

Student Paper

<1%

---

23

[lijiancheng0614.github.io](https://lijiancheng0614.github.io)

Internet Source

<1%

---

24

[www.cse.ust.hk](http://www.cse.ust.hk)

Internet Source

<1%

---

25

Submitted to iGroup

Student Paper

<1%

---

26

[elearning.nic.in](http://elearning.nic.in)

Internet Source

<1%

---

27

Manohar Swamynathan. "Mastering Machine  
Learning with Python in Six Steps", Springer  
Nature, 2017

Publication

<1%

---

28

Submitted to St. Xavier's College

Student Paper

<1%

---

29

Submitted to Staffordshire University

Student Paper

<1%

---

30

Submitted to The Robert Gordon University

Student Paper

<1%

---

31

[www.leapshc.org:8080](http://www.leapshc.org:8080)

Internet Source

<1%

---

32

[carrefax.com](http://carrefax.com)

Internet Source

<1%

---

33

[ar.scribd.com](http://ar.scribd.com)

Internet Source

<1%

---

34

Chetana Janardan Kolte, Avinash Shrivasa.  
"Intelligent knowledge sharing for agricultural  
information", 2017 International Conference on  
Trends in Electronics and Informatics (ICEI),  
2017

Publication

<1%

---

35

Mohamed K. Elhadad, Khaled M. Badran,  
Gouda I. Salama. "A Novel Approach for  
Ontology-Based Dimensionality Reduction for  
Web Text Document Classification",  
International Journal of Software Innovation,  
2017

Publication

<1%

---

36

[shctpt.edu](http://shctpt.edu)

Internet Source

<1%

---

37

Andreas François Vermeulen. "Practical Data Science", Springer Nature, 2018

<1%

Publication

---

Exclude quotes Off

Exclude matches Off

Exclude bibliography Off

# K-Nearest classification

*by* S Sagayaraj

---

**Submission date:** 02-Apr-2019 12:35PM (UTC+0530)

**Submission ID:** 1104361942

**File name:** BP170523.pdf (884.99K)

**Word count:** 4465

**Character count:** 24149



**TEXT CATEGORIZATION USING K-NEAREST NEIGHBOR  
CLASSIFICATION**

A Software Project

16

Submitted in partial fulfillment for the award of the degree of

**MASTER OF COMPUTER SCIENCE**

by

**PAVITHRA.V  
(REG.NO:BP170523)**

9

Under the Guidance of

**Mr. R.DENIS. M.C.A., M.Phil,**



**PG DEPARTMENT OF COMPUTER SCIENCE**

**SACRED HEART COLLEGE (AUTONOMOUS)**

**TIRUPATTUR, VELLORE DT – 635 601**

**April- 2019**

**DECLARATION BY THE STUDENT**

**I hereby declare that the report entitled “TEXT CATEGORIZATION USING**

**K-NEAREST NEIGHBOR CLASSIFICATION ” submitted by me to PG Department of Computer Science, Sacred Heart College, Tirupattur in partial fulfillment of the requirement for the award of the degree of M.Sc. in Computer Science is an ‘Industrial Plant Training’ undertaken by me under the Supervision of Department faculty. I further declare that the work in this report has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university.**

## CERTIFICATE

This is to certify that the project work<sup>22</sup> entitled “<sup>2</sup> Text Categorization Using k-Nearest Neighbor Classification” is submitted to Sacred Heart College (Autonomous), Tirupattur-<sup>4</sup> 635601, Vellore District by PAVITHRA.V Reg. No: BP170523 for the partial fulfillment for the award of the Degree of Master of Science in Computer Science is a bona fide record of the work carried out by him, under my guidance and supervision.

Signature of the Project Guide

Mr. R. Denis, M.C.A., M.Phil.,

Submitted for Viva – Voice examination held on \_\_\_\_\_.

Internal Examiner

1.

Internal Examiner

2.

## **1.1 Abstract**

K-Nearest Neighbor is a popular algorithm for text categorization. In this project represents a document categorization using KNN. This method is to categorize the business, entertainment, politics, sports, and technology and training documents and it finds all the closest neighbors of the sample documents. Its focus the speed and quality of the document classification.

In this project shows the k-nearest neighbors enables to find the document categorization. In this project involves a solution of the similarity functions and previously finds the document categorization.

## **CHAPTER – I INRODUCTION**

### **1.2 Project Overview**

#### **1.2.1 Introduction**

Text Categorization can be classification of document different data set as the original documents dataset process a categories. As the volume of data information available, document classification on the intranets Text data and Text Classification is an different text document using.

1. K-Nearest Neighbor classification text categorization in an algorithm model. Document class approach level categorized Text.
2. Text Classification to business document. five Types of data are using this of consumer files.
3. Different Text Categorization many documents five categories are worked.
4. Text used in K-NN algorithm work below text classified.
5. K-NN Text Categorization is mainly concept of machine learning classifier document in java page.

### 1.2.2 Scope

The scope of the project is used to bring the outcome of the document details regarding the document count and running time. We can classify them into different norms and category by categories the dataset, we can predict different type of possibilities of various document. We predict document title total number of text file will be calculate and generating the document. Finally, we can also plot the outcome predicted values in histogram and we can perform them in the K-NN representations.

### 1.2.3 Problem Statement

A document classification, matrix formula are using in this projects Cosine Similarity. Non zero vectors assign to same document need to classified those data of this <sup>13</sup> an inner product space that measures the matrix formula are using in <sup>13</sup> cosine of the angle between them. The cosine similarity as a math function are calculated the seconds minimum time is particularly used in algorithm based to positive space calculate every second assign distance problem , where the document outcome is neatly bounded in this project.

This method is to categorize the business, entertainment, politics, sports, and technology and training documents. Using in K-Nearest Neighbor algorithm some examples can be classified documents in sports and politics same the test which are same in the content easily match in Text data, in our problem be eliminated for the same of efficiency. While categorizing documents, these terms in document categorization, Cosine Similarity both text document are KNN classify Test Document Thread in class using machine learning concepts. java Implementation of K-Nearest Neighbor Algorithm using in File cache.

Text Categorization data for BBC Dataset is under “data” folder. Here dataset import the data level in set of data in similarity between these documents.

#### 1.2.4 Product position statement

For	IT Industry
Who	IT members.
That	It is more flexible.
Our Product	Provide up-to-date information and easy to maintain all the data. An categories them and make prediction and represent them in K-NN representation manner

## SYSTEM ANALYSIS

### CHAPTER – II SYSTEM ANALYSIS

#### 2.1 System Study

##### 2.1.1. Introduction

Text Categorization used in K-NN algorithm text document project is an k-Nearest Neighbor predicting algorithm of the document text details, by using the java language. In this the analysis are made of every document to analyzing the data it should be calculating the text documents. And also calculating times of the seconds We can also represent the predicted results of the document in K-NN representation by Text in bar diagram, different kind of Text method. The main purpose of the project is to identify the document and categories data of the text file and also calculating time .base of the time the text will be generate.

##### 2.1.2 STATEMENT OF SOLUTION:-

<sup>12</sup> Text categorization, also known as Text classification, is the performance of automatically classifying a dataset of text documents into five categories from a dataset, in this project, we first categorize the algorithms using K-NN Based machine learning approach and then return's the most relevant documents. We will explain the solution method give a K-NN algorithm to classification and analysis.

##### 2.1.3. Overview

Analysis of document text documents and predicting them by K-Nearest Neighbor machine learning models, this is an predicting applications of the document file details, by using java language. In this the analysis of every document should be analyzed and calculating the data and calculating number of data files in the particular document and displayed Test based on the running times so that proposed the new way representing the data into Text representations by which the Text are able to identify the results easily. It can also represent the predicted results of the document in Text representation by K-NN diagram, and different kind of plotting method.

The main purpose of the project is to identify document data and text file calculation and running times and also generating Text based on the data file format.

### **2.1.3 Proposed System**

- Representing the new way of present the data into the categorizing.

### **2.1.5. 1 ABOUT CLASSIFICATION DETAILS OF THE PROJECTS**

#### **Purpose**

To produce the complete details about the classifications and Text categorizing.

#### **Overview**

The user, who wants to know about the prediction of any case, can easily get the details in classifications and Text categorizing.

#### **Entry Criteria**

The users can directly import the data set, train and test them to get predictions.

#### **Input**

Here the input is test data from the data set.

#### **Steps involved**

Import text data into the java and perform the code form train and test using classification algorithms and make predict the desired output.

#### **Output**

The required detail is shown to the knowledge seekers.



## 2.2 Feasibility Study

### 2.2.1 Introduction

Text Categorization can be classification of document <sup>2</sup> As the volume of data information available, document classification on the intranets Import text data into the java and perform the code form train and test using classification algorithms and make predict the desired output.

#### <sup>19</sup> 2.2.1.1 Technical Feasibility:

In this, one can strongly says that it is technically feasible, solution based review documents and assign those documents verify quickly access catches few documents development software hardware specify info text documents calculated feasibility solution, <sup>3</sup> since there will not be much difficulty in getting required resources for the development and maintaining the system as well. All the resources needed for the development of the software as well as the maintenance of the same is available in the organization here we are utilizing the resources which are available already.

#### 2.2.1.2 Economical Feasibility

Development of this application is highly economically feasible .The organization needed not spend much money for the development of the system already available. The only thing is to be done is making an environment for the <sup>8</sup> project application using available, and intranets volume of data default text development with an effective supervision. If we are doing so, we can attain the maximum usability of the corresponding resources .Even after the development, the organization will not be in condition to invest more in the organization. Therefore, the system is economically feasible.

## **SYSTEM CONFIGURATION**

### CHAPTER – III **System** Configuration

#### **3.1 Requirement Specification**

15

##### **3.1.1 Hardware Requirements**

<b>Processor</b>	<b>: Intel i3 processor</b>
<b>RAM</b>	<b>: 2 GB</b>
<b>Hard Disk</b>	<b>: 500 GB</b>
<b>Input/Output</b>	<b>: Keyboard, mouse, monitor</b>

##### **3.1.2 Software Requirements**

<b>Front End</b>	<b>: Java (NetBeans 8.2)</b>
<b>Back End</b>	<b>: PDF Format</b>
<b>Documentation</b>	<b>: Microsoft Word 2013</b>

##### **3.3.2.1 ALGOTITHM WORK'S**

The algorithm classified in text document where identifying the details in K-NN algorithm. Text representation by K-NN diagram, and different kind of plotting method. The main purpose of the project is to identify document data and text file calculation and running times and also generating Text based on the data file format. “K-Nearest Neighbor classification” was introduced. Model used both document cache file representation. Comparing text document business, sports and text will be same finally machine learning methods are using manual system not using only machine language using. The system produces accurate result and it also reduces a lot of overheads, which the manual system faced. The system was thoroughly tested

according to the testing methodologies. Dataset is including the project Similarity (Document test, Document train Document).

K-NN algorithm The text sample may be classified either to document classified Either to the File Document Factory of assigned or to the second class of file Comparators of tanning data if  $k=150$  it is assigned to the second document because there are more set of documents are available in this project of K-NN algorithm. The second file document class to compare the file cache document class are used. This is the inner circle. If  $k=150$  is assigned the first document and second machine leaning of classifiers in project.

Nearest Neighbor classifier:-

The popular algorithm in nearest neighbor classifier is those training dataset is available one of algorithm that assigns a point of data to the document class of its neighbor in the feature space, that is size of training data set Documents completely in this algorithm is infinity, the one nearest neighbor algorithms classifier sports document and training text below guarantees an error rate of no worse than twice the algorithm minimum achievable error rate given the distribution of the data.

### 3.3.2. TEXT CATEGORIZATION :-

The machine leaning concept including this project until and file cache page are including and three type of machine learning files are using be the Document similarity methods the usage in search engines. This methods, have been using manual categorization but k-Nearest Neighbor methods like a Text java main page in cluing the string variables and calculus the letters text categorizations router and check the words of data levels. Text categorizations, approaches in KNN Classification Comparator approach categorizes the documents. text categorization meaning of training dataset including the BBC dataset added adds this project. Text classification probabilistic on work to text categorization

Since them, it has been clarify to dataset used for a many numbers of different applications, one. which here briefly review of the most important one. Note that boarders between more different class of applications listed here fully and some artificial, these may be considered special cases of others. Applications we do not explicitly discuss are text categorization by mean of a combination of text classifier and text document categorization through the analysis of textual captions in text, author identifications for literary texts of unknown or disputed authorship

language identification for text identify for texts of unknown language automated identification of text genre of a project.

Automatic indexing for Text Categorization systems:-

The Text categorization system applications that has K-NN most of the early research in the field is that algorithm automatic document indexing classifier systems relying on a controlled dictionary, the most prominent example of which is that of document systems. In these latter each document is assigned one or more keywords or key phrases describing of content, where these document keywords and key phrases belong to a finite set called controlled dictionary, after consisting a cosine matrix hierarchical and text document belong to a finite set called controlled dictionary. Usually, this project assignment is done by trained Text documents and is thus a costly activity. If the entries in the controlled vocabulary are viewed as categories, text document is an instance of multi document may thus be Business, sports, politics by the automatic techniques described in this algorithm. Recalling note the text document note that this application may typically require that  $k_1 < x < k_2$  value are assigned to each document, for given  $k_1, k_2$ . Document classifiers are assigned to each document-pivoted is probably the best option, so that new documents may be classified as they become available. Various text classifiers and explicitly to conceived for document indexing have been some text described in the literature five kinds of documents Automatic text categories with controlled and dictionaries is matrix functions related to automated data generation. In digital libraries one is usually interested in tagging documents by data this describe then under a variety of aspects data model, some of the data are in matrix functions their role is to describe that semantics of the document by means of K-Nearest Neighbor algorithm, keywords of key phrases. The generation of these data may thus be viewed as a problem identify of document indexing with controlled dictionary, and thus take in k-NN algorithm techniques.

### 3.3.2.3 DATASET:-

K-Nearest Neighbor is a popular algorithm for text categorization. In this project represents a document categorization using KNN. This method is to categorize the business, entertainment, politics, sports, and technology and training documents and it finds all the closest neighbors of the sample documents. Its focus the speed and quality of the document classification.

The system project and accurate result and it also reduces less of time saving and a lot of overheads, which the manual system faced. The system was high-level of performance need this project model thoroughly tested according to the testing methodologies.

In this project shows the k-nearest neighbors enables to find the document categorization. In this project involves a solution of the similarity functions and previously finds the document categorization. Filtering this document according to

#### 3.3.2.4 Document organization:-

This organization of documents is critical to each space handle with area in business documents severally however is a lot of sequences but efface more project particularly important for corporations, firms and also the classifiers. It makes search on the categorised documents simple easy and it permits individuals to pay time. Some magazines and newspapers are exemplar for the firms in since they received several advertisements and most need and automatic system those to arrange these advertisements to classes in order that thousands of document similarity advertisements concerning sports, business, politics, educations etc. won't need to be divided into documents manually. Newspapers aren't solely printed this paper however they some document exits conjointly within the document electronic surroundings and create use of text categorization throughout some deciding the places of file caches places of article consistent with to the topic cover. This project to find out details see.

#### 3.3.2.5 Finding disambiguos words:-

Applications of text categorization square measure doable to use fine text documents with in an exceedingly text, that have over one document which means and confirm the means to which means of its that category's finding time calculate disambiguos words for instance the document base classifiers base in an categorization article it's presumably accustomed offer the same which means, used to give the some meaning. Text categorization finds the meant thatmeans. That is often referred to as File caches documents application and is usually used in K-Nearest Neighbor algorithm rule.



### 3.3.2.6 Different Approaches:-

We have mentioned about the decision Tree method, K nearest neighbor classification(k-NN), and lexical chains are generally not used in this text categorization but other research topics, so will not clearly explain them here very briefly in our proposal. But are many other applications used in this text categorization. We will try to explain all approaches including the K-NN classification method we use in our implementation.

### 3.3.2.7 K -nearest neighbor bayes probabilistic classifiers:-

The K -nearest neighbor probabilistic classifiers are employed in machine learning applications, to shortly make a case for it joint chances of words and text classes to estimate that of follow predefined classes are appropriate for a text. Then. K- nearest neighbor simplifies the computation and will increases to potency with the belief text document. In line with this assumption, text document in one class have freelance of this following additional blessings this document's conditional chances.

According to this assumption, text document in one category have freelance of this following a lot of blessings this document's conditional chances.

Let as justify exploitation mathematics:

$P(c/d)$ =probability for a document text to belong to class File caches document.

$P(c/d)=[P(c)*p(d/c)]/p(d)$  wherever matrix perform

$P(c)$ =probability that any document chosen haphazardly belongs to text class

$P(d)$ =probability that any document chosen haphazardly features a vector illustration.

$P(d/c)$ =probability that text class includes conjointly documents.

There are a lot of formulas to calculate  $p(d/c)$  and also the earlier mentioned k nearest neighbor Algorithm assumption permit this calculation of  $p(d/c)$  to be done easier.

### 3.3.2.8 Decision Tree method:-

The K-Nearest Neighbor rule is employed in machine learning, a choice tree classifier uses a tree structure in some internal nodes stands for things and terms and edges holds relevant weights associated with the terms in nodes.

Finally the lead nodes represent text classes. These classifiers reason documents that following a path to the connected lead node. This path is set consistent with the algorithmic partitioning specific rule and a some rule referred to as info gain algorithm rule.

The algorithm program is given below:

- i. Class within the set of coaching documents are set to 3 or 5 if containing or not.
- ii. The coaching documents are appointed to a root node and a perform be enforced are assigned to a root node and text document a info retrieved from training documents is that the root of the tree.
- iii. The Recursive algorithm partitioning algorithm program is enforced during this a part of document since the perform is recursive and takes nodes as parameters.
  - i. if this node has some specific documents stopping conditions, flip this node into a let in order that is appointed to the category most ordinarily used among coaching documents known as text catteries
  - ii. if there's chance to make a brand new text class, a word from the coaching documents, this could be word that holds most in an exceedingly document of not. According crate to kid documents current node and call and decision this project.

Document sense Disambiguation:-

Document sense clarification is that the activity of finding text classes, prevalence associate exceedingly in a very text of an ambiguous document, the sense of this explicit word prevalence file document. As an example, bank could have a lot of totally different check files are offered as within the establishment or the document. DSD is extremely vital for several applications, together with tongue process, the text documents by word senses rather of then by

words for algorithmic program functions. Once we have tendency to read document prevalence contexts as document and text senses as classes.

### 3.3.2.9 Hierarchical categorization of web pages:-

Text has recently or used a lot of interest also for its possible applications to belong very much of document automatically classifying web pages, or web sites, under the hierarchical section catalogues hosted on popular internet portals. Where text documents are catalogued in this way, rather than issuing letters in categories an algorithm to general-purpose web search engine model a searcher may find it easier to first navigate in the hierarchy of categories and then restrict her search to a given document of particular category of interest.

Classifying web pages automatically has more advantages, this project and algorithm since the manual categorization of a large size of document enough subset of this web is infeasible. In more unlike in the previous applications, it is file text modeling typically the case of study that each category must be populated the set of  $k_1 < x < k_2$  documents. Should be chosen so as to allow new categories to be added and document to be deleted. With respect to previously discussed text applications, automatic web page categorization

Two essential peculiarities:-

- i. The hypertext nature of this documents like a more links are a rich file structure source of information, as they may be understood as training data stating the relevance of the linked based pages to the linking pages. Techniques as exploiting this intuition in a context have been logics.
- ii. The hierarchical structure of file documents category sets this may be used on decomposing documents the classification in to regular experience problem into most a number of smaller classification problems, each text corresponding to a branching decision at an internal node techniques documents.

This organization of documents is important to each in business documents one by one however is additional project's particularly important for corporations, firms and also the classifiers. It makes search on the categorised documents straightforward and it allows folks to pay time.



Some magazines and classifiers documents most are model for the firms in since they received several advertisements and most need and automatic system those to arrange these advertisements to classes in order that thousands of document advertisements concerning sports, business, politics, educations etc. won't must be divided into documents manually. Newspapers aren't solely revealed this paper however they some document exits conjointly within the electronic setting and build use of text categorization throughout some crucial the places of file caches places of article per the subject.

<sup>1</sup> The machine learning approach to text categorization:-

The most popular approach for the document creation of text will be <sup>1</sup> automatic document classifiers consisted in to manually some building, by means of different documents knowledge techniques, an machine learning expert system capable of length text decisions. Such an expert system would like typically consist of a set of manually process defined logical rules, one text document per category, of type if the formula then category more types if formula using to identify the category <sup>14</sup> disjunctive normal formula is a disjunction of conjunctive class to added the document is classified under category if u want to more document add satisfies the <sup>1</sup> formula is a disjunction of conjunctive clauses, the document almost classified under category formula in using machine learning in automated text categorization. <sup>23</sup>

<sup>20</sup> A sample rule of the type used in construe is illustrated

<sup>6</sup> The drawback this project approach is the knowledge is question well-known from the expert system identify literature. That is some rules must be manually defined by a knowledge engineer with the text id concepts of a domain expert if the set of categories is updated, then those add two professionals must intervene again, and if the classifier is ported to a completely different domain, <sup>1</sup> expert user needs to intervene and the work has to be repeated both side from scratch. In the others hand, it was originally suggested that this approach can give very good effectiveness to document <sup>1</sup> intervene and the work has to be ready to repeated from scratch. On the other hand, its more specify originally suggested that this approach documents can give very normal effectiveness results, on a subset of the results <sup>4</sup> test collections that output performs even those best classifiers built in this 90 by state-of-art machine learning techniques. However, other classifier has been tested on this same dataset as construe, project not clear whether this was a randomly chosen or a favorable subsets of entire results collection. As project the results above

do not clear randomly chosen or a results again may be obtained in general methods. Since the early machine learning approach to documents gained popularity and has eventually become the dominant one, at least in the research comments a comprehensive introduction to machine learning approach a general inductive process also called the machine learning, automatically builds a classifier for a category files domain expert from these characteristics, the inductive process gleans the characteristics that inductive process characteristics that a new unseen document should have in order to be classified document .

## CONCLUSION AND SAMPLE SCREEN SHOTS

### CHAPTER – IV SCREENSHOT

#### 4.7 User Interact Page

```
package com.machinelearning.doc_classifier.knn;

import java.io.File;
import java.io.FileNotFoundException;

import com.machinelearning.doc_classifier.document.DocumentCache;
import com.machinelearning.doc_classifier.document.FileDocumentCache;
import com.machinelearning.doc_classifier.knn.core.KNNClassifier;

public class Main {

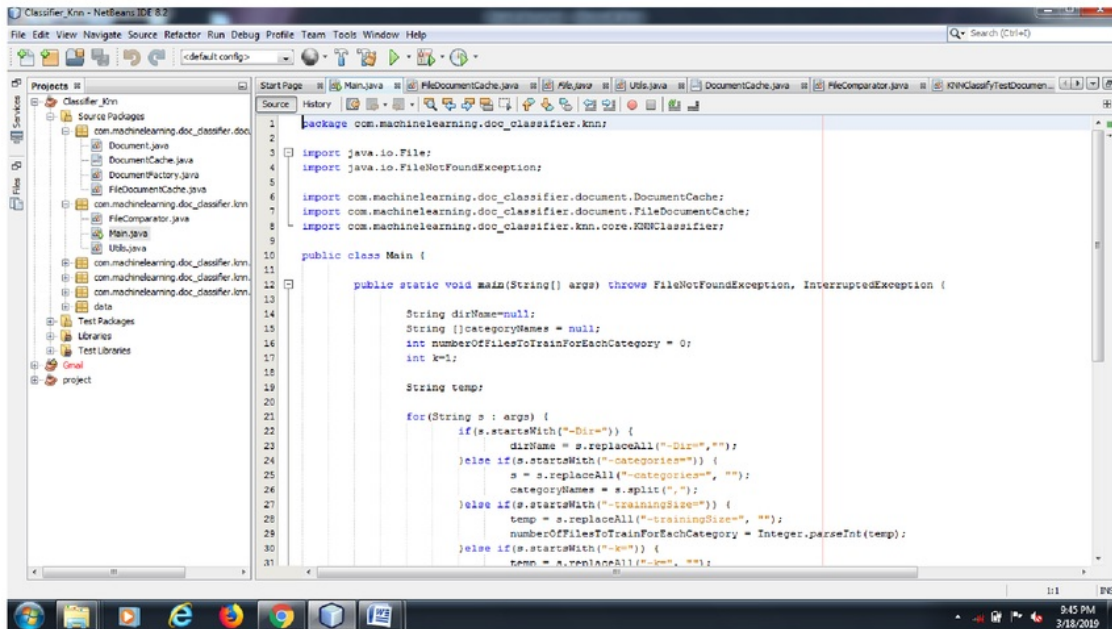
    public static void main(String[] args) throws FileNotFoundException, InterruptedException {

        String dirName=null;
        String []categoryNames = null;
        int numberOfFilesToTrainForEachCategory = 0;
        int k=1;
        String temp;

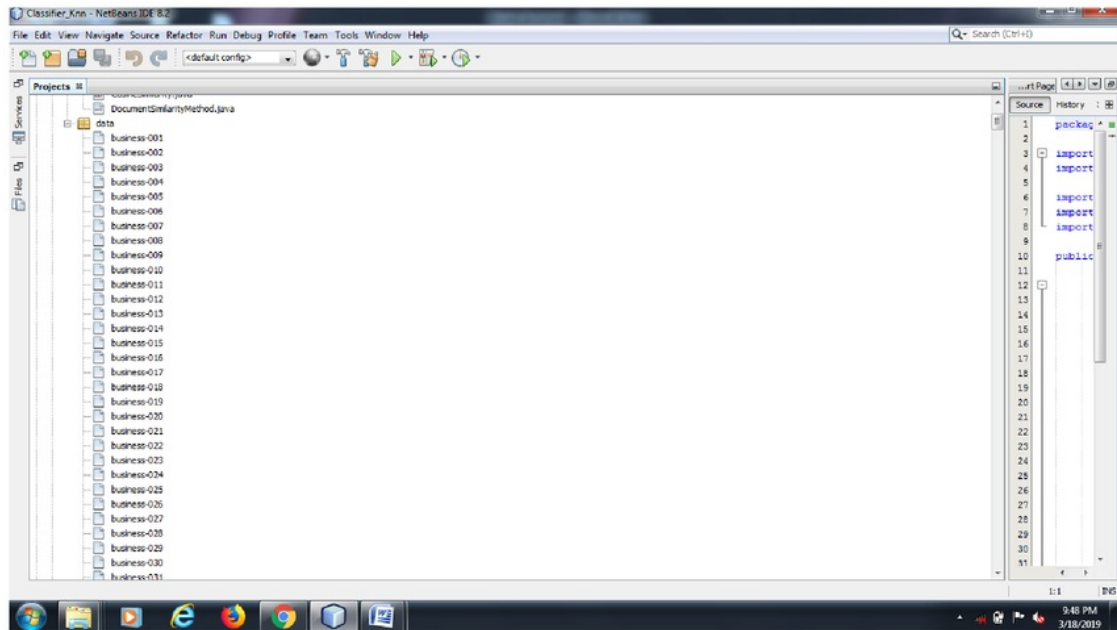
        for(String s : args) {

            if(s.startsWith("-Dir=")) {
                dirName = s.replaceAll("-Dir=", "");
            } else if(s.startsWith("-categories=")) {
                s = s.replaceAll("-categories=", "");
                categoryNames = s.split(",");
            } else if(s.startsWith("-trainingSize=")) {
                temp = s.replaceAll("-trainingSize=", "");
                numberOfFilesToTrainForEachCategory = Integer.parseInt(temp);
            } else if(s.startsWith("-k=")) {
                k = s.replaceAll("-k=", "");
            }
        }
    }
}
```

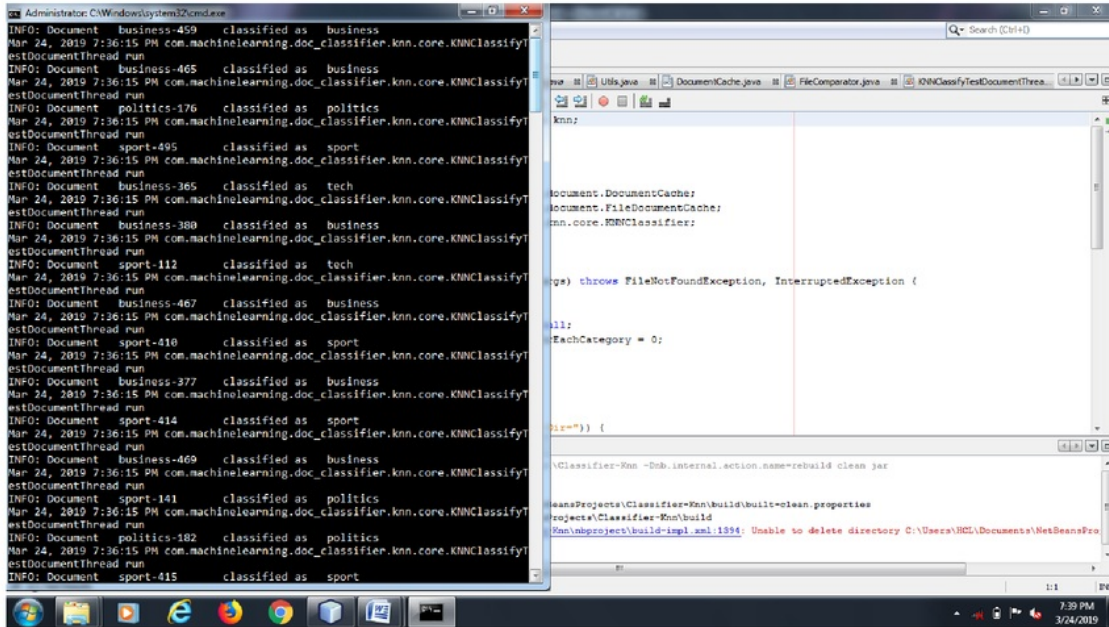
The main page:-



## Importing the Data set:-



## Output:-



```
Administrator: C:\Windows\system32\cmd.exe
INFO: Document business-459 classified as business
Mar 24, 2019 7:36:15 PM com.machinelearning.doc_classifier.knn.core.KNNClassifier
estDocumentThread run
INFO: Document business-465 classified as business
Mar 24, 2019 7:36:15 PM com.machinelearning.doc_classifier.knn.core.KNNClassifier
estDocumentThread run
INFO: Document politics-176 classified as politics
Mar 24, 2019 7:36:15 PM com.machinelearning.doc_classifier.knn.core.KNNClassifier
estDocumentThread run
INFO: Document sport-495 classified as sport
Mar 24, 2019 7:36:15 PM com.machinelearning.doc_classifier.knn.core.KNNClassifier
estDocumentThread run
INFO: Document business-365 classified as tech
Mar 24, 2019 7:36:15 PM com.machinelearning.doc_classifier.knn.core.KNNClassifier
estDocumentThread run
INFO: Document business-380 classified as business
Mar 24, 2019 7:36:15 PM com.machinelearning.doc_classifier.knn.core.KNNClassifier
estDocumentThread run
INFO: Document sport-112 classified as tech
Mar 24, 2019 7:36:15 PM com.machinelearning.doc_classifier.knn.core.KNNClassifier
estDocumentThread run
INFO: Document business-467 classified as business
Mar 24, 2019 7:36:15 PM com.machinelearning.doc_classifier.knn.core.KNNClassifier
estDocumentThread run
INFO: Document sport-410 classified as sport
Mar 24, 2019 7:36:15 PM com.machinelearning.doc_classifier.knn.core.KNNClassifier
estDocumentThread run
INFO: Document business-377 classified as business
Mar 24, 2019 7:36:15 PM com.machinelearning.doc_classifier.knn.core.KNNClassifier
estDocumentThread run
INFO: Document sport-414 classified as sport
Mar 24, 2019 7:36:15 PM com.machinelearning.doc_classifier.knn.core.KNNClassifier
estDocumentThread run
INFO: Document business-469 classified as business
Mar 24, 2019 7:36:15 PM com.machinelearning.doc_classifier.knn.core.KNNClassifier
estDocumentThread run
INFO: Document sport-141 classified as politics
Mar 24, 2019 7:36:15 PM com.machinelearning.doc_classifier.knn.core.KNNClassifier
estDocumentThread run
INFO: Document politics-182 classified as politics
Mar 24, 2019 7:36:15 PM com.machinelearning.doc_classifier.knn.core.KNNClassifier
estDocumentThread run
INFO: Document sport-415 classified as sport

knn:
document.DocumentCache;
document.FileDocumentCache;
knn.core.KNNClassifier;

) throws FileNotFoundException, InterruptedException {

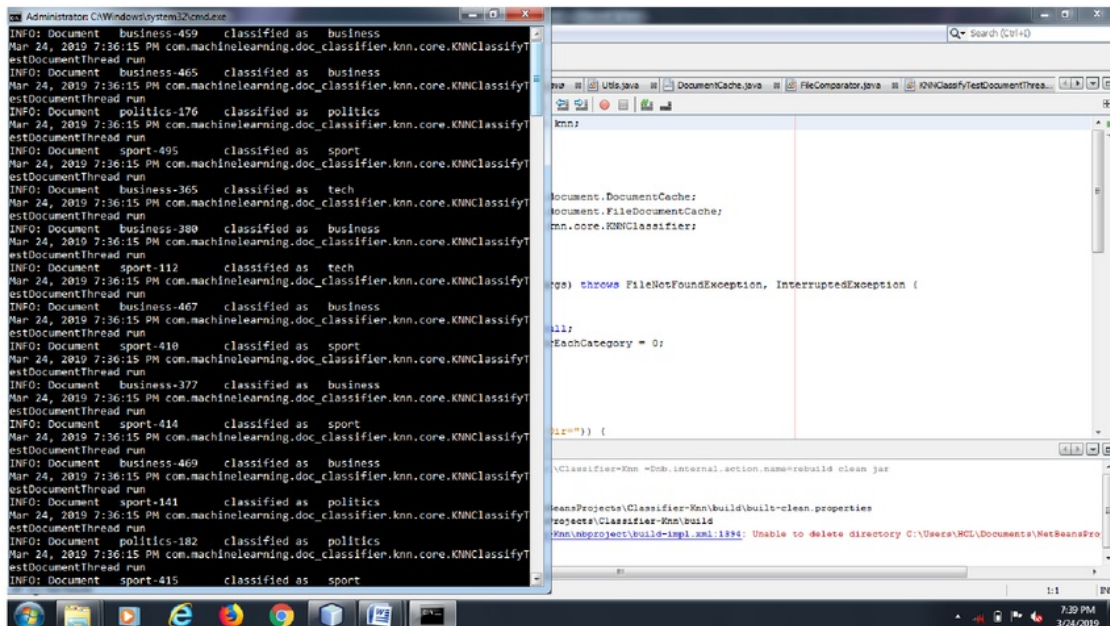
all:
EachCategory = 0;

}

Classifier-Knn -Ddb.internal.action.name=rebuild clean jar

beansProjects\Classifier-Knn\build\build-clean.properties
projects\Classifier-Knn\build
Knn\project\build-impl.xml:1394: Unable to delete directory C:\Users\RGJ\Documents\NetBeansPro
```





#### 4.8 FUTURE WORK:-

The k-Nearest Neighbor formula is a simple to modifiable algorithm and is filmable to totally different issues. This formula permits north American country to develop code any to extend potency and accuracy. We tend to assign specifically one classes to at least one article in our implementation for safe of potency. However there exists article containing several terms concerning many classes are offered. Articles don't seem to be perpetually written concerning every topic, however jump from topic to topic in way of life. This means, articles might have multiple classes. This version of our implementation seems to offer less info than it ought to, at that time not be tough to point out different any text documents classes since the k-NN formula already calculates the required info for distribution multiple categories. Keep in mind that we tend to calculates all similarity totals and confidences of every class document and select the best worth to assign our category.

## CHAPTER – V CONCLUSION

### **5.1 Conclusion**

- This project has been successfully developed and interpreted and system was developed according to the user requirement. The system produces accurate result and it also reduces a lot of overheads, which the manual system faced. The system was thoroughly tested according to the testing methodologies. We several documents are text strategies are exploitation text categorization, thanks to strategies and it's document that text categorization may be a helpful application in standard of living wherever the requirement of automatic system will increase information is automatic calculated the information.

# K-Nearest classification

## ORIGINALITY REPORT

35%

SIMILARITY INDEX

33%

INTERNET SOURCES

20%

PUBLICATIONS

18%

STUDENT PAPERS

## PRIMARY SOURCES

1	<a href="http://www.clips.ua.ac.be">www.clips.ua.ac.be</a> Internet Source	13%
2	<a href="http://www.ceng.metu.edu.tr">www.ceng.metu.edu.tr</a> Internet Source	9%
3	<a href="http://de.slideshare.net">de.slideshare.net</a> Internet Source	2%
4	<a href="http://www.slideshare.net">www.slideshare.net</a> Internet Source	2%
5	<a href="http://elvis.slis.indiana.edu">elvis.slis.indiana.edu</a> Internet Source	1%
6	Submitted to CSU, San Jose State University Student Paper	1%
7	<a href="http://www.lwfree.cn">www.lwfree.cn</a> Internet Source	1%
8	Submitted to Info Myanmar College Student Paper	1%
9	Submitted to Sacred Heart College Student Paper	1%



10

Submitted to University of Warwick

Student Paper

&lt;1%

11

Submitted to Selçuk Üniversitesi

Student Paper

&lt;1%

12

Célia Valente. "A Tool for Text Mining in Molecular Biology Domains", Repositório Aberto da Universidade do Porto, 2014.

Publication

&lt;1%

13

Vulić, Ivan, Wim De Smet, Jie Tang, and Marie-Francine Moens. "Probabilistic topic modeling in multilingual settings: An overview of its methodology and applications", Information Processing &amp; Management, 2015.

Publication

&lt;1%

14

Laurence Hirsch, Masoud Saeedi, Robin Hirsch. "Chapter 8 Evolving Rules for Document Classification", Springer Nature, 2005

Publication

&lt;1%

15

Submitted to Staffordshire University

Student Paper

&lt;1%

16

[elearning.nic.in](http://elearning.nic.in)

Internet Source

&lt;1%

17

Ali, Amal Seralkhatem Osman, Vijanth Sagayan, Aamir Saeed Malik, Mohamed Meselhy Eltoukhy, and Azrina Aziz. "Age-invariant Face Recognition Using Triangle

&lt;1%

# Geometric Features", International Journal of Pattern Recognition and Artificial Intelligence, 2015.

Publication

18

[nmis.isti.cnr.it](http://nmis.isti.cnr.it)

Internet Source

<1%

19

Submitted to Lovely Professional University

Student Paper

<1%

20

[reference.kfupm.edu.sa](http://reference.kfupm.edu.sa)

Internet Source

<1%

21

[tu-dresden.de](http://tu-dresden.de)

Internet Source

<1%

22

[www.leapshc.org:8080](http://www.leapshc.org:8080)

Internet Source

<1%

23

Fabrizio Sebastiani. "Machine learning in automated text categorization", ACM Computing Surveys, 3/1/2002

Publication

<1%

Exclude quotes Off

Exclude matches Off

Exclude bibliography Off

# Image Retrieval

*by* S Sagayaraj

---

**Submission date:** 02-Apr-2019 12:37PM (UTC+0530)

**Submission ID:** 1104362585

**File name:** BP170524\_Tag\_Based\_Image\_Retrival.pdf (403.19K)

**Word count:** 2196

**Character count:** 10971

**JOINT HYPERGRAPH LEARNING FOR TAG-BASED IMAGE RETRIEVAL**

**5**  
**A PROJECT REPORT**

*Submitted in partial fulfillment for*

*The award of the degree of*

**MASTER OF SCIENCE IN COMPUTER SCIENCE**

*By*

**AISWARYA.S**

Reg. No: BP170524

*Under the guidance of*

**Prof . D. Gajalakshmi, MCA.,**



**2**  
**PG DEPARTMENT OF COMPUTER SCIENCE**

**SACRED HEART COLLEGE (AUTONOMOUS)**

**Tirupattur, Vellore Dist-635 601**

**APRIL - 2019**

## **BONAFIDE CERTIFICATE**

This is to certify that the project work entitled **JOINT HYPERGRAPH LEARNING FOR TAG-BASED IMAGE RETRIEVAL** is submitted to Sacred Heart College (Autonomous) Tirupattur by **AISWARYA S (Reg.No. BP170524)** for the partial fulfillment for the award of the Degree of Master of Science in Computer Science is a bonafide record of the work carried out by her, under my guidance and supervision.

Date:

Signature of the guide

Submitted for <sup>7</sup>Viva –Voce examination held on: \_\_\_\_\_

**Internal Examiners**

**External Examiner**

## ACKNOWLEDGEMENT

I thank God Almighty for his blessings and graces by which I was able to complete project work successfully.

I sincerely thank my parents who have given the gift of life to me to attain many achievements.

I express my deep sense of gratitude to <sup>8</sup> **Dr.S.Sagayaraj**, Head of the Department of Computer Science and my Project Guide **Ms.D.Gajalakshmi** who is the source of inspiration to pursuer with every efficient work. And I am heart fully honor her valuable guidance and encouragement of my every activities.

<sup>2</sup> I would like to thank the entire teaching and non-teaching staff members, Department of Computer Science for their help to complete the project successfully.

<sup>2</sup> Finally, I thank each and every one of my friends, especially who have assisted me in completing the project work.

**AISWARYA.S**

# 1. INTRODUCTION

## 1.1 Project Summary

**1** **Joint Hypergraph Learning for Tag-based Image Retrieval** is a web application. It's helpful for social network users. The social network users can easily create their own account, and use this account. Through this web application, the users can share the images approach and Search for the image in efficient way using tag based image retrieval. While the user sharing the images, the images are dumped into search option and the same can be added into timeline news feed. This application not only limited to image sharing but also the users can able to follow or following other users in the friends list.

## **1.2 List of Modules**

- Home
- Admin
  - User Details
  - View Images
- User
  - User Home
  - Profile
  - Share
  - TimeLine
  - Search Image
  - Followers
  - Followings
  - Logout
- User Registration

## **1.3 Conclusion**

- This chapter says about the main concept of the project and motive of the Application. This chapter concludes with motive of the project.



## 2. REQUIREMENTS ANALYSIS

### 2.1 INTRODUCTION

Requirement analysis brings the system requirements of the project. This chapter gives the analysis of the application, E-R diagram, DFD diagram, Product vision, Use case specification of the application. This brings to get an idea about the scope of the project.

### 2.2 SYSTEM STUDY

#### 2.2.1 Overview

The public networking users anywhere and anytime to share and view the image details through offline.

#### 2.2.3 Proposed System

- ❖ This proposed system helpful for the public social networking users. Through this application performance is sharing the images, search and view the images.
- ❖ Through this application approach to automatically use different visual features and tags for images.
- ❖ This application is offline based approach and can capture more reliable relationships between the images.

#### Features

- Easily share unlimited images
- Easily view the images and image details

#### Advantages of proposed system:

- ❖ This application not only limited to image sharing but also the users can able to follow or following other users in the friends list.

#### 2.2.4 Responsibilities of Admin

- ✓ User Details
- ✓ View Images

#### 2.2.5 Modules and their Description

##### List of modules:

- Home
- Admin
- User
- User Registration

**Module No: 01**

**Module Name: Home**

**Overview**

To view about the website

**Module No: 02**

**Module Name: Admin**

**Overview:**

Securing administrator tools from the users

**Activities: 2.1**

**Activity Name: Admin Home**

**Purpose**

To ensure that the login person is admin

**Entry Criteria**

Valid admin name and password

**Input**

Admin name

Password

**Output**

Now the admin can use his tools

**Edit Criteria**

Admin exit the login of the moment he/she press the proceed or back button.

**Activities: 2.2**

**Activity Name: User Details**

**Purpose**

To see the how many users are registered

**Entry Criteria**

Valid admin name and password

**Edit Criteria**

User can exit the controls

**Activities: 2.3**

**Activity Name:** View Images

**Purpose**

To see how many users are what are the image shared

**Entry Criteria**

Valid admin name and password

**Edit Criteria**

User can exit the controls

**Activities: 2.4**

**Activity Name:** Logout

**Purpose**

The purpose is to completing their process to leave in this page.

**Module No: 03**

**Module Name:** User

**Overview:**

We can share, view, following friends and followers are saw

**Entry Criteria**

To give the valid email id and password.

**Input**

Email id

Password

**Output**

Now the user can use this tools

**Edit Criteria**

User exit the login of the moment he/she press the proceed or back button.

**Activities: 3.1**

**Activity Name:** User Home

**Purpose**

To view the home page of users

**Activities: 3.2**

**Activity Name:** Profile

**Purpose**

To see the user details

**Edit Criteria**

User can exit the controls

**Activities: 3.3**

**Activity Name:** Share

**Purpose**

To share any images

**Entry Criteria**

Valid details are given about the image

**Input**

To see the image details

**Output**

To see the image

**Edit Criteria**

User can exit the controls

**Activities:** 3.4

**Activity Name:** TimeLine

**Purpose**

To see the images

**Edit Criteria**

User can exit the controls

**Activities:** 3.5

**Activity Name:** Search Image

**Purpose**

To search the given images

**Entry Criteria**

We can search the specific images.

**Input**

To give the name of the images

**Output**

To view for name of the images

**Edit Criteria**

User can exit the controls

**Activities:** 3.6

**Activity Name:** Followers

**Purpose**

To see how many followers are seen in this site

**Input**

Search the images

**Output**

To see the images

**Edit Criteria**

User can exit the controls

**Activities:** 3.7

**Activity Name:** Followings

**Purpose**

To see how many users are followed yourself

**Input**

Search the images

**Output**

To see the images

**Edit Criteria**

User can exit the controls

**Activities:** 3.8

**Activity Name:** Logout

**Purpose**

The purpose is to completing their process to leave in this page.

**Exit Criteria:**

After completing this task, the user will be redirected at the home page for further process.

**Module No:** 04

**Module Name:** User Registration

**Purpose:**

If the new user means they have to give their details.

**Entry Criteria:**

User has login into the application.

**Input:**

To give the

6

1. First name
2. Last name
3. Email
4. Password
5. Date of birth
6. State
7. Gender
8. Country
9. Profile picture

**Output:**

To add the new user

**Edit Criteria**

User can exit the controls

**2.3 Vision Document****2.3.1 Introduction**

**1** **Joint Hypergraph Learning for Tag-based Image Retrieval** is a web application. It's helpful for social network users. The social network users can easily create their own account, and use this account. Through this web application, the users can share the images approach and Search for the image in efficient way using tag based image retrieval. While the user sharing the images, the images are dumped into search option and the same can be added into timeline news feed. This application not only limited to image sharing but also the users can able to follow or following other users in the friends list.

**2.3.2 Positioning****I Problem Statement**

The problem of	To maintain the image and user details.
Affects	The users.
The impact of which is	Time consuming to see the images easily.
The successful solution would be	It is a web application for developing this project to full fill the joint hyper learning for tag-based image retrieval Tagging requirements to provide into the successful solution.

## II Product Position Statement:

For	The Image retrieval Tagging
Who	Users.
The	Image retrieval Tagging
That	Maintenance of the image, user details.
Unlike	The existing system, this application approach to automatically use different visual features and tags for images.
Product	It is one of the important ways to find images contributed by social networking users.

## III STAKEHOLDER AND USER DESCRIPTION

### Stakeholder summary:

Name	Represent	Responsibilities
End user	The maintain details about Tag based image retrieval.	To monitor the details of admin to ensure their needs.

### User summary:

Name	Represent	Responsibilities
Admin	To maintain the records about users and image details.	Ensure whether the records have been maintained.

## V Assumptions and dependence



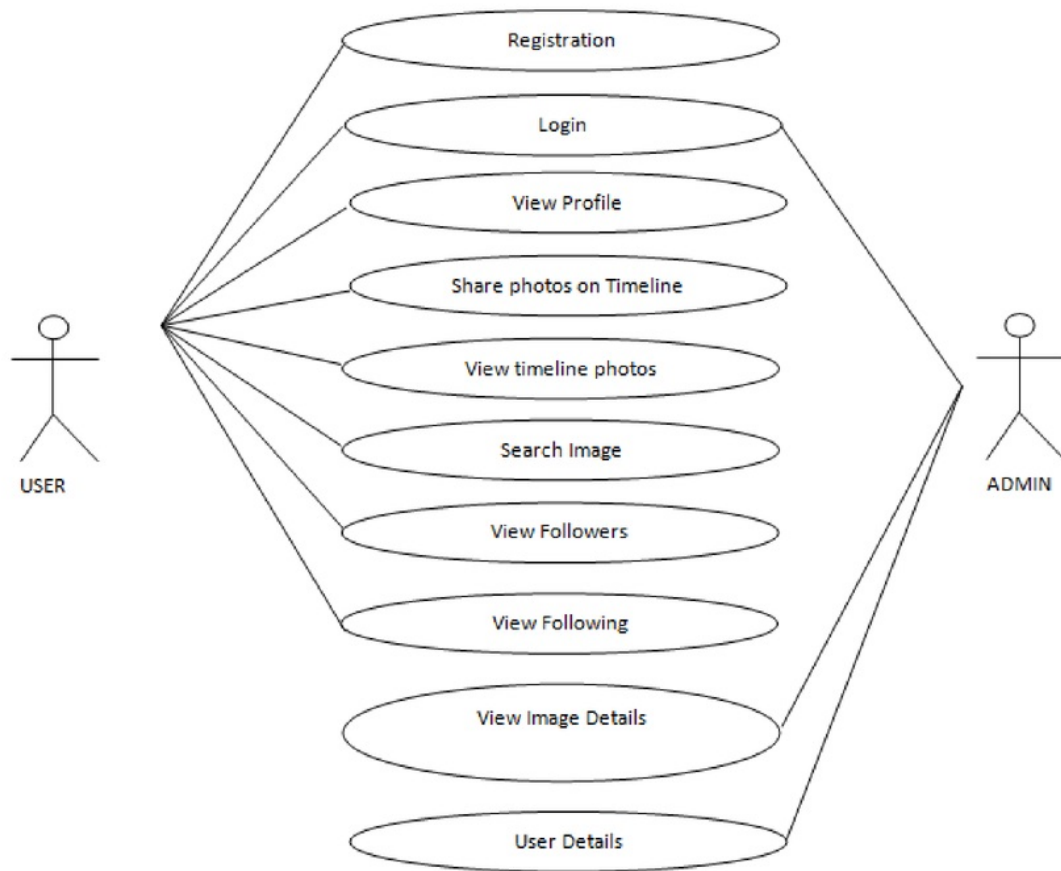
**Hardware:**

Processor : Intel i3  
Ram : 2 GB more  
Storage : 50 GB  
Input\Output : Printer, pen drive, key board, mouse.  
Operating system : Windows XP/7

**Software:**

Front Tier : JAVA/J2EE  
Data Tier : MYSQL  
Software Tools : NetBeans IDE 7.2.1

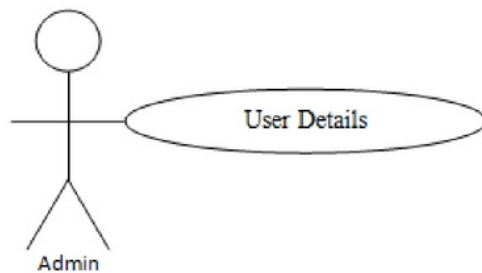
**2.4 USE CASE DIAGRAM**



**1. Use Case Name: User Details**

**1.1. Brief Description:**

This Use Case describes about view the all the user details.



### 1.1.2. Actors:

Admin

### 1.1.3. Precondition:

The user has separate login to add new users.

### 1.2. Flow of Events:

The use Case begins when the user move to see add the new user details.

## 2. Use Case Name: View Image Details

### 2.1. Brief Description:

This Use Case describes about view the all the images.



### 2.1.2. Actors:

Admin

### 2.1.3. Precondition:

The user has separate login to add their images.

### 2.2. Flow of Events:

The use Case begins when the user move to see the image details.

## 2.5 Conclusion

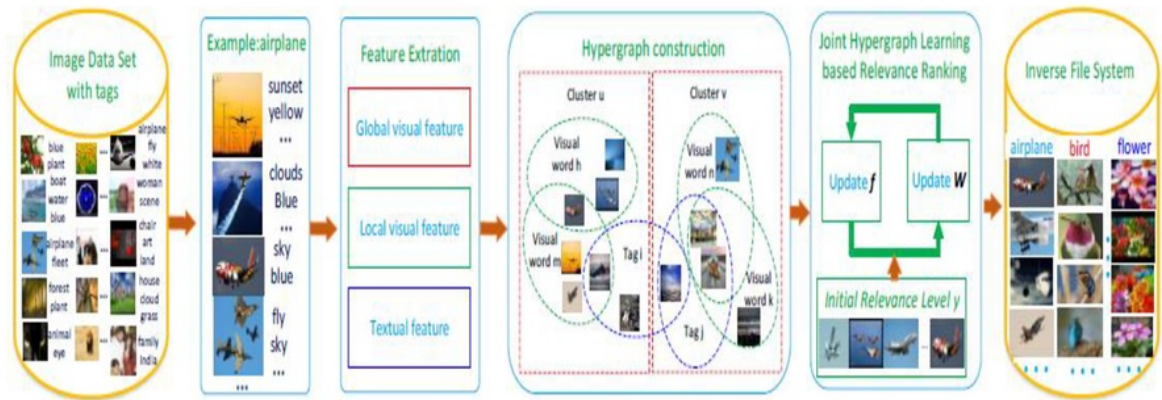
This web application has been developed for Joint Hypergraph Learning for Tag-based Image Retrieval. It has been developed by the JAVA and the database is MYSQL server. It can be helpful for the public social users.

### 3. DESIGN DOCUMENT

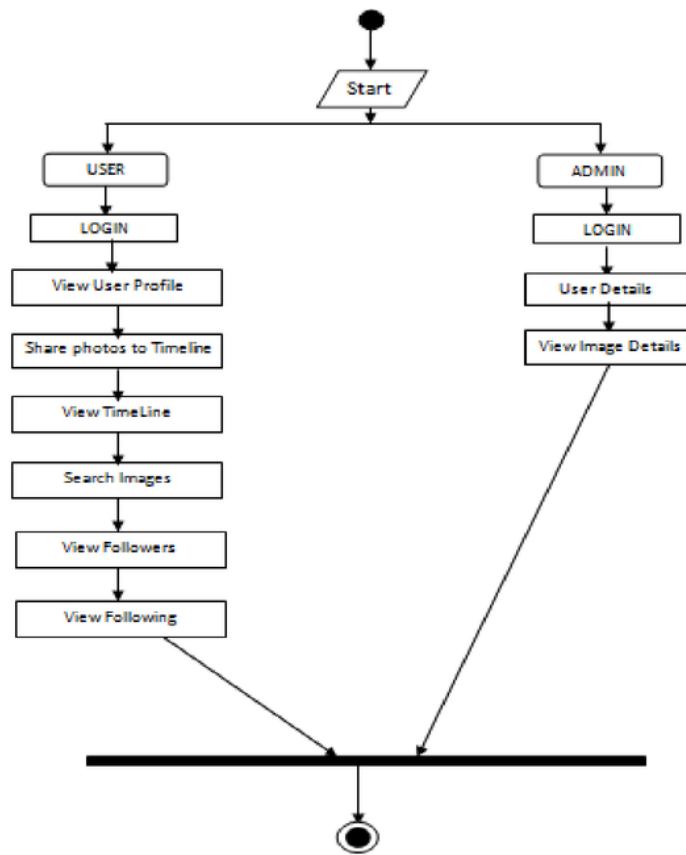
#### 3.1 Introduction

This chapter tells the design diagram. It gives brief introduction about activity diagram, architecture design, and database design, interface design and procedural design at last the test case.

#### 3.2 Architecture Design

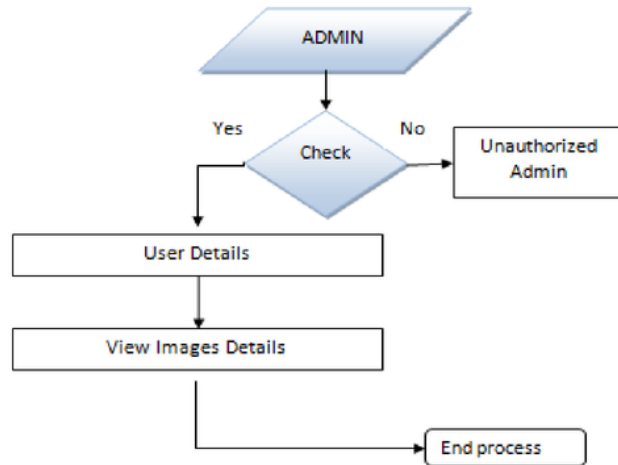


### 3.3. Activity design

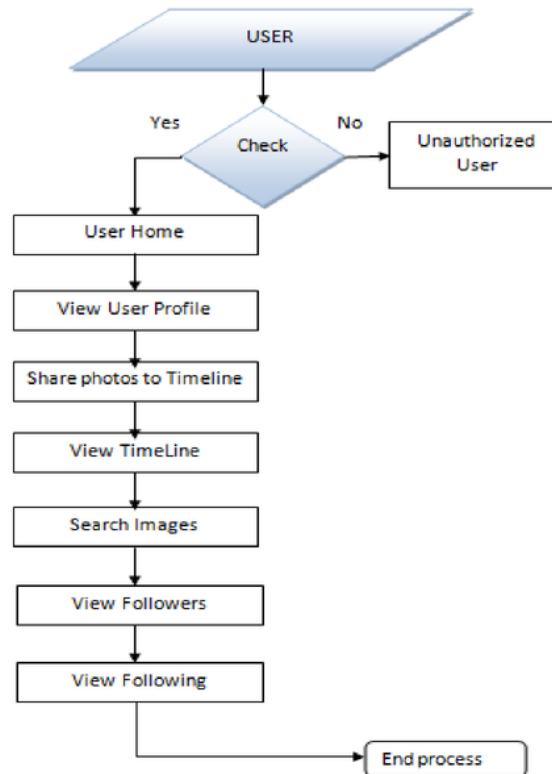


### 3.4 DATA FLOW DIAGRAM

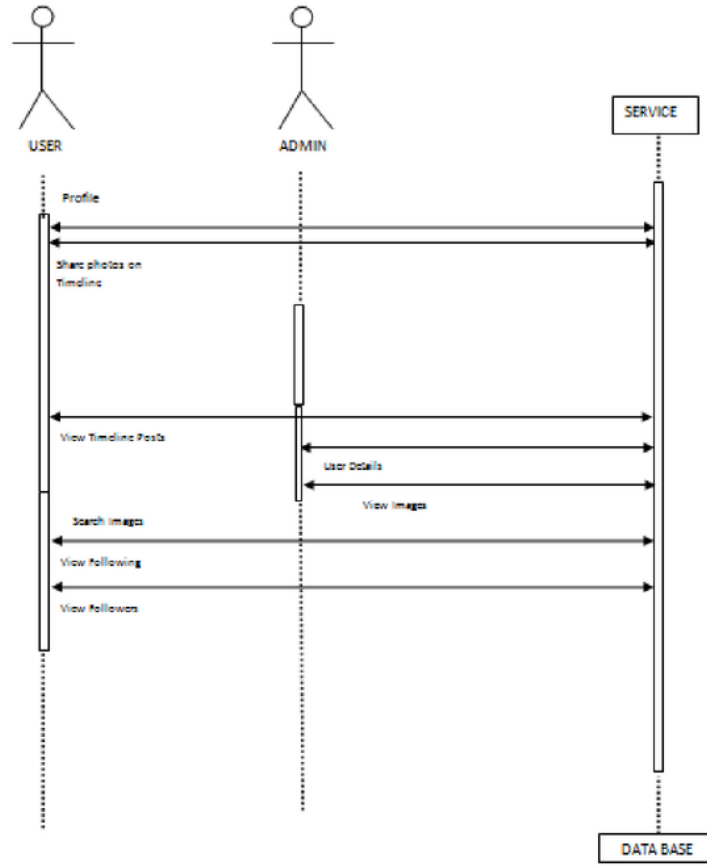
#### Admin



#### User



### 3.5 Sequence Diagram



### 3.7 Test case Design

#### Login:

TSC_ID	TSC_1
Objective	To Admin Login
Test Design	ST1: Go on admin page first login .in that login without giving any data.
	EO: No error should occur and proper message should be handled
	ST2: Click on cancel without giving any data
	EO: No error should occur and window should be closed
	ST3: Enter user name and password
	EO: User name should be displayed as it is given and password should be displayed as special character
	ST4: Enter valid user name and password and click on login
	EO: Admin page should be opened
	ST5: Enter valid user name and valid password and click on cancel
	EO: No error message should be displayed, text boxes should be cleared
	ST6: Enter valid user name and password and click on login
	EO: No error should occur and proper message should be displayed to get correct data
	ST7: Enter valid user name and password and click on login
	EO: No error should occur and proper message should be displayed to get correct data
Expected output	The admin is logged in the application.

#### New user registration:



TSC_ID	TSC_2
Objective	New User Register
Test design	ST1: Click on register without giving any data.
	EO: No error should occur and proper message should be handled
	ST2: Enter mandatory fields and click on add
	EO: The user details will be add in the database

## Conclusion

This web application has been developed for **Joint Hypergraph Learning for Tag-based Image Retrieval**. It has been developed by the JAVA and the database is MYSQL server. It can be helpful for the public social users.

## 4. IMPLEMENTATION

### 4.1 Overview of the Implementation

**Joint Hypergraph Learning for Tag-based Image Retrieval** application is used to see the image details. The purpose of this document to collect, analyze and define the basic requirements, functional units of public user's. This system mainly focuses on see the image retrieval through offline.

- User Details
- View Images

#### Steps (In bootstrap web application)

- Create a Application and edit NetBeans IDE
- Create a Java file and run it in local host.

#### Steps (MYSQL)

- First need to install the software MYSQL server
- Next open the MYSQL Server and connect the MYSQL Server
- The need to create and insert the table fields.

### Steps (Microsoft Visio-2003)

Need to install the software Microsoft Visio for the diagram purpose.

It only for the diagrams to draw

DFD, Use case, Sequence, architecture diagram has been drawn by this software.

File→New→Software→UML Diagram.

## 5. TESTING DOCUMENT

### 5.1 UNIT TESTING

Unit testing focuses verification efforts on the smallest unit of software design of the modules. This is also referred as “Module testing”.

ADMIN LOGIN

Login	
Admin ID	<input type="text"/>
Password	<input type="text"/>
<input type="button" value="Login"/>	

USER LOGIN:

Login	
Email	<input type="text"/>
Password	<input type="text"/>
<input type="button" value="Login"/>	

USER REGISTRATION

User Registration			
FIRST NAME	<input type="text"/>	LAST NAME	<input type="text"/>
EMAIL	<input type="text"/>	PASSWORD	<input type="text"/>
DATE OF BIRTH	<input type="text"/>	STATE	<input type="text"/>
GENDER	<input type="text"/>		
COUNTRY	<input type="text"/>	PROFILE PICTURE	<input type="text"/>
<input type="button" value="INSERT"/>			

SHARE IMAGE:

**SHARE IMAGES AT TIMELINE**

IMAGE NAME

IMAGE TAGS

IMAGE DESCRIPTION

SELECT CATEGORY

SELECT IMAGE

## 5.2 Integration Testing

In this combination testing there are two kinds of testing exists that is top-down coordination and base up joining.

## 6. CONCLUSION

The project <sup>1</sup> **JOINT HYPERGRAPH LEARNING FOR TAG-BASED IMAGE RETRIEVAL** is completed and satisfying this application. It has been developed by the web application and the database is MYSQL server.

# Image Retrieval

---

## ORIGINALITY REPORT

---

<b>11</b> %	<b>5</b> %	<b>4</b> %	<b>6</b> %
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

---

## PRIMARY SOURCES

---

<b>1</b>	<b>Yaxiong Wang, Li Zhu, Xueming Qian, Junwei Han. "Joint Hypergraph Learning for Tag-Based Image Retrieval", IEEE Transactions on Image Processing, 2018</b>	<b>4</b> %
	Publication	
<b>2</b>	<b>Submitted to Sacred Heart College</b>	<b>3</b> %
	Student Paper	
<b>3</b>	<b>admissionsinindia.com</b>	<b>1</b> %
	Internet Source	
<b>4</b>	<b>Submitted to University of Greenwich</b>	<b>1</b> %
	Student Paper	
<b>5</b>	<b>www.nitc.ac.in</b>	<b>1</b> %
	Internet Source	
<b>6</b>	<b>www.behavior.net</b>	<b>1</b> %
	Internet Source	
<b>7</b>	<b>Submitted to College of Engineering Trivandrum</b>	<b>1</b> %
	Student Paper	

---

---

Exclude quotes      Off  
Exclude bibliography      Off

Exclude matches      Off