

Assignment On

Implementing Naive Bayes Classification algorithm into PHP to classify given text as Sports, Finance, Religion or Politics. This application uses MySql as database.

[Course Teacher](https://www.facebook.com/sabrina.ria)

Syeda Sabrina Akter,

Lecturer *CSE,*

*Daffodil international university*

Paper By:

Md Shihab Ali,

ID: 151-15-4975

*11th Sem, CSE,*

Course Title: Data mining

Course Code: cse-450

Problem 2

**How to install the project-?**

* Create database in MySql- Or import naivebayes.sql database file
  + - mysql> create database naiveBayes;
    - mysql> use naiveBayes;
    - mysql> create table trainingSet (S\_NO integer primary key auto\_increment, document text, category varchar(255));
    - mysql> create table wordFrequency (S\_NO integer primary key auto\_increment, word varchar(255), count integer, category varchar(255));
* Open a terminal and move to project folder
* Edit database connection info in db\_connect.php file
* Execute main.php php main.php

**Task 1:** Taking input properly:

$classifier->train('Sport in Bangladesh is a popular form of entertainment as well as an essential part of Bangladeshi culture', $sports);  
$classifier->train('Sport can play a role in improving the lives of not only individuals but whole communities', $sports);  
  
$classifier->train('Financial risk protection and equity are major components of universal health coverage which is defined as ensuring access to health services for all citizens without any undue financial burden', $finance);  
  
$classifier->train('Secularism is established in Bangladesh and freedom of religion is guaranteed by constitution', $religion);  
$classifier->train('Bangladesh religious minorities have been facing attacks since the 2014 national election', $religion);  
  
$classifier->train('The politics regarding the bargaining of the students for their sports games interest with the university authority is called student politics', $politics);  
$classifier->train('A key risk related to the violence in Bangladesh from a rating perspective is that at some stage safety issues could deter foreigners from doing business there', $politics);

**Task 2:** Removing stopwords:

$stopWords = array('A', 'an', 'the', 'that', 'their', 'there', 'it',  
 'would', 'should', 'shall', 'will', 'into', 'unto', 'undo', 'in', 'of', 'to', 'from', 'for', 'by',  
 'but', 'not', 'is', 'are', 'have', 'has', 'as', 'at', 'and', 'can', 'could');  
  
*//removing all the characters which ar not letters, numbers or space*$sentence = preg\_replace("/[^a-zA-Z 0-9]+/", "", $sentence);  
  
*//converting to lowercase*$sentence = strtolower($sentence);  
  
*//an empty array*$keywordsArray = array();

**Task 3:** Implement Naïve Bayes Algorithm

*Naive Bayes Classifier Algorithm -* private function decide($keywordsArray)  
 {  
 $sports = Category::*$SPORTS*;  
 $finance = Category::*$FINANCE*;  
 $religion = Category::*$RELIGION*;  
 $politics = Category::*$POLITICS*;  
  
 *// by default assuming category to be sports* $category = $sports;  
  
 *// making connection to database* require 'db\_connect.php';  
  
 $sql = mysqli\_query($conn, "SELECT count(\*) as total FROM trainingSet WHERE category = '$sports' ");  
 $sportsCount = mysqli\_fetch\_assoc($sql);  
 $sportsCount = $sportsCount['total'];  
  
 $sql = mysqli\_query($conn, "SELECT count(\*) as total FROM trainingSet WHERE category = '$finance' ");  
 $financeCount = mysqli\_fetch\_assoc($sql);  
 $financeCount = $financeCount['total'];  
  
 $sql = mysqli\_query($conn, "SELECT count(\*) as total FROM trainingSet WHERE category = '$religion' ");  
 $religionCount = mysqli\_fetch\_assoc($sql);  
 $religionCount = $religionCount['total'];  
  
 $sql = mysqli\_query($conn, "SELECT count(\*) as total FROM trainingSet WHERE category = '$politics' ");  
 $politicsCount = mysqli\_fetch\_assoc($sql);  
 $politicsCount = $politicsCount['total'];  
  
 $sql = mysqli\_query($conn, "SELECT count(\*) as total FROM trainingSet ");  
 $totalCount = mysqli\_fetch\_assoc($sql);  
 $totalCount = $totalCount['total'];  
  
  
  
 *//p(Sports)* $pSports = $sportsCount / $totalCount; *// (no of documents classified as sports / total no of documents)  
  
 //p(Finance)* $pFinance = $financeCount / $totalCount; *// (no of documents classified as Finance / total no of documents)  
  
 //p(Religion)* $pReligion = $religionCount / $totalCount; *// (no of documents classified as Religion / total no of documents)  
  
 //p(Politics)* $pPolitics = $politicsCount / $totalCount; *// (no of documents classified as Politics / total no of documents)  
  
 //echo $pSports." "$pFinance." ".$pReligion." ".$pPolitics;  
  
 // no of distinct words (used for laplace smoothing)* $sql = mysqli\_query($conn, "SELECT count(\*) as total FROM wordFrequency ");  
 $distinctWords = mysqli\_fetch\_assoc($sql);  
 $distinctWords = $distinctWords['total'];  
  
 $bodyTextIsSports = log($pSports);  
 foreach ($keywordsArray as $word) {  
 $sql = mysqli\_query($conn, "SELECT count as total FROM wordFrequency where word = '$word' and category = '$sports' ");  
 $wordCount = mysqli\_fetch\_assoc($sql);  
 $wordCount = $wordCount['total'];  
 $bodyTextIsSports += log(($wordCount + 1) / ($sportsCount + $distinctWords));  
 }  
  
 $bodyTextIsFinance = log($pFinance);  
 foreach ($keywordsArray as $word) {  
 $sql = mysqli\_query($conn, "SELECT count as total FROM wordFrequency where word = '$word' and category = '$finance' ");  
 $wordCount = mysqli\_fetch\_assoc($sql);  
 $wordCount = $wordCount['total'];  
 $bodyTextIsFinance += log(($wordCount + 1) / ($financeCount + $distinctWords));  
 }  
  
  
 $bodyTextIsReligion = log($pReligion);  
 foreach ($keywordsArray as $word) {  
 $sql = mysqli\_query($conn, "SELECT count as total FROM wordFrequency where word = '$word' and category = '$religion' ");  
 $wordCount = mysqli\_fetch\_assoc($sql);  
 $wordCount = $wordCount['total'];  
 $bodyTextIsReligion += log(($wordCount + 1) / ($religionCount + $distinctWords));  
 }  
  
  
 $bodyTextIsPolitics = log($pPolitics);  
 foreach ($keywordsArray as $word) {  
 $sql = mysqli\_query($conn, "SELECT count as total FROM wordFrequency where word = '$word' and category = '$politics' ");  
 $wordCount = mysqli\_fetch\_assoc($sql);  
 $wordCount = $wordCount['total'];  
 $bodyTextIsPolitics += log(($wordCount + 1) / ($politicsCount + $distinctWords));  
 }  
  
 if ($bodyTextIsSports >= $bodyTextIsFinance && $bodyTextIsSports >= $bodyTextIsReligion && $bodyTextIsSports >= $bodyTextIsPolitics) {  
 $category = $sports;  
 }  
 elseif ($bodyTextIsFinance >= $bodyTextIsSports && $bodyTextIsFinance >= $bodyTextIsReligion && $bodyTextIsFinance >= $bodyTextIsPolitics) {  
 $category = $finance;  
 }  
 elseif ($bodyTextIsReligion >= $bodyTextIsSports && $bodyTextIsReligion >= $bodyTextIsFinance && $bodyTextIsReligion >= $bodyTextIsPolitics){  
 $category = $religion;  
 }  
  
 else {  
 $category = $politics;  
 }  
  
 $conn->close();  
  
 return $category;  
 }  
}

**Task 4:** Case handling

From [Wikipedia](https://en.wikipedia.org/wiki/Bayes%27_theorem):

**P(A | B)** = **P(B | A) \* P(A) / P(B)** where A and B are events

and **P(B) != 0**

**P(A | B)** is a conditional probability: the likelihood of event A occurring given that B is true.

**P(B | A)** is also a conditional probability: the likelihood of event B occurring given that A is true.

**P(A)** and **P(B)** are the probabilities of observing A and B independently of each other, this is known as the marginal probability*.*

**Task 5:** Finding the category

$category = $classifier->classify('The issues of religion politics became interconnected in Bangladesh ');  
echo $category;  
echo " ";  
  
$category = $classifier->classify('This article argues that the interconnection of religion and politics in the context of Bangladesh is linked with the modes of governance ');  
echo $category;  
echo " ";  
  
$category = $classifier->classify('Religious minorities are subject of threats in bangladesh in several cases of political issues');  
echo $category;  
echo " ";  
  
$category = $classifier->classify('In other words it is a problem of politics not religion ');  
echo $category;  
echo " ";  
  
$category = $classifier->classify('The people of Bengal want freedom of religion they do not want any interference in religious matters ');  
echo $category;  
echo " ";  
  
$category = $classifier->classify('Religion can not be used for political ends');  
echo $category;

**Result:**

* Politics
* Politics
* Religion
* Politics
* Religion
* Religion

*Reference:*

* [*http://stackoverflow.com/questions/9996327/using-a-naive-bayes-classifier-to-classify-tweets-some-problems*](http://stackoverflow.com/questions/9996327/using-a-naive-bayes-classifier-to-classify-tweets-some-problems)
* [*https://github.com/ttezel/bayes/blob/master/lib/naive\_bayes.js*](https://github.com/ttezel/bayes/blob/master/lib/naive_bayes.js)