Naive Bayes Classifier Used For Text Classification.

**Problem Statement:** Using a Naive Bayes classifier to state whether a sentence is positive or negative.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name | Roll No. | Section | Semester | Branch |
| Jenit | 9 | D | 3 | CSE |
| Somiya | 43 | D | 3 | CSE |
| Ankur | 63 | D | 3 | CSE |
| Rishab | 42 | D | 3 | CSE |
| Ekam | 26 | D | 3 | CSE |

**Problem Statement:** To classify an input statement into 2 groups, positive and negative, based on the information provided by the developer during the training of the classifier.

**Algorithm:**

1. The entire file containing the dataset is read.
2. A set is made which contains all possible different words in the dataset.
3. The file is read again, sentence by sentence, and for each sentence a vector is made (number of elements in the vector=number of words in the set). There are just 2 values possible in the vector : 0,1. If the element is 0 then the corresponding word in the set doesn’t exist in the sentence read from the file and if it does exist then the element is 1.
4. 2 1-D arrays are created – p1num, p0num – for which all the elements are set to 1 and 2 variables are createrd – p1denom, p0denom – which are set to 2.
5. The vector form of the sentences that are pre-classified as positive are added to the p1num array and the vector form of the sentence that are pre-classified as negative are added to the p0num array.
6. The sum of the vector form of the sentences that are pre-classified as positive are added to the p1denom and the sum vector form of the sentence that are pre-classified as negative are added to the p0denom.
7. 2 probability vector are formed by taking the log of the division of p1num and p0num by p1denom and p0denom respectively. These vectors show the probabilty of a word in the set being negative or positive.
8. A sentence is then taken as an input using the applet viewer and then sent to the classify method. Over here the vector is multiplied to both the probabilty vectors elementwise and the product for all the elements is summed up. After finding the 2 sums (p0,p1), they are compared to check whether the sentence is more probable to be a positive sentence or a negative sentence and a boolean value is returned accordingly and based on the returned value and output is displayed on the appletviewer.

**Concepts:**

* Use of Static function and arrays.
* Basic File I/O.
* Set Interface.
* Exceptions.
* Swings.
* Applets.
* Event Handling.

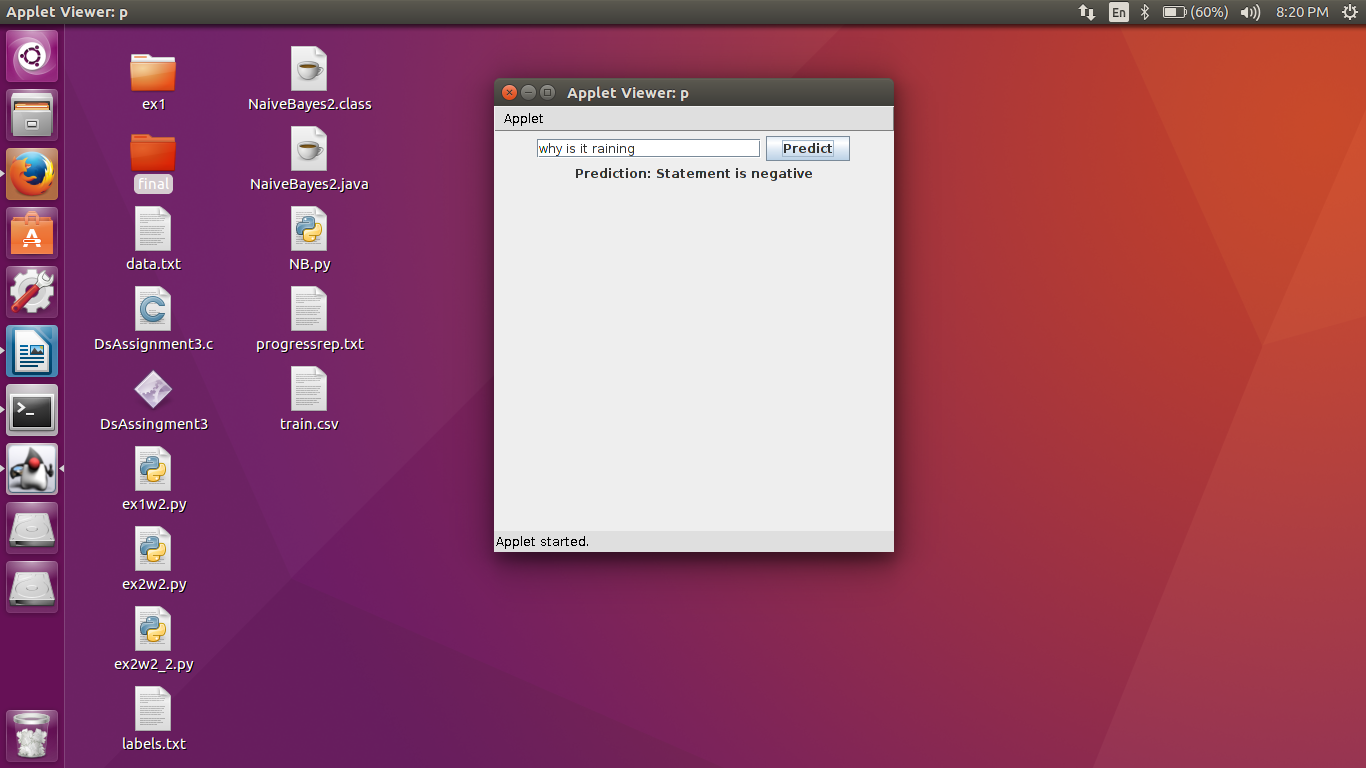
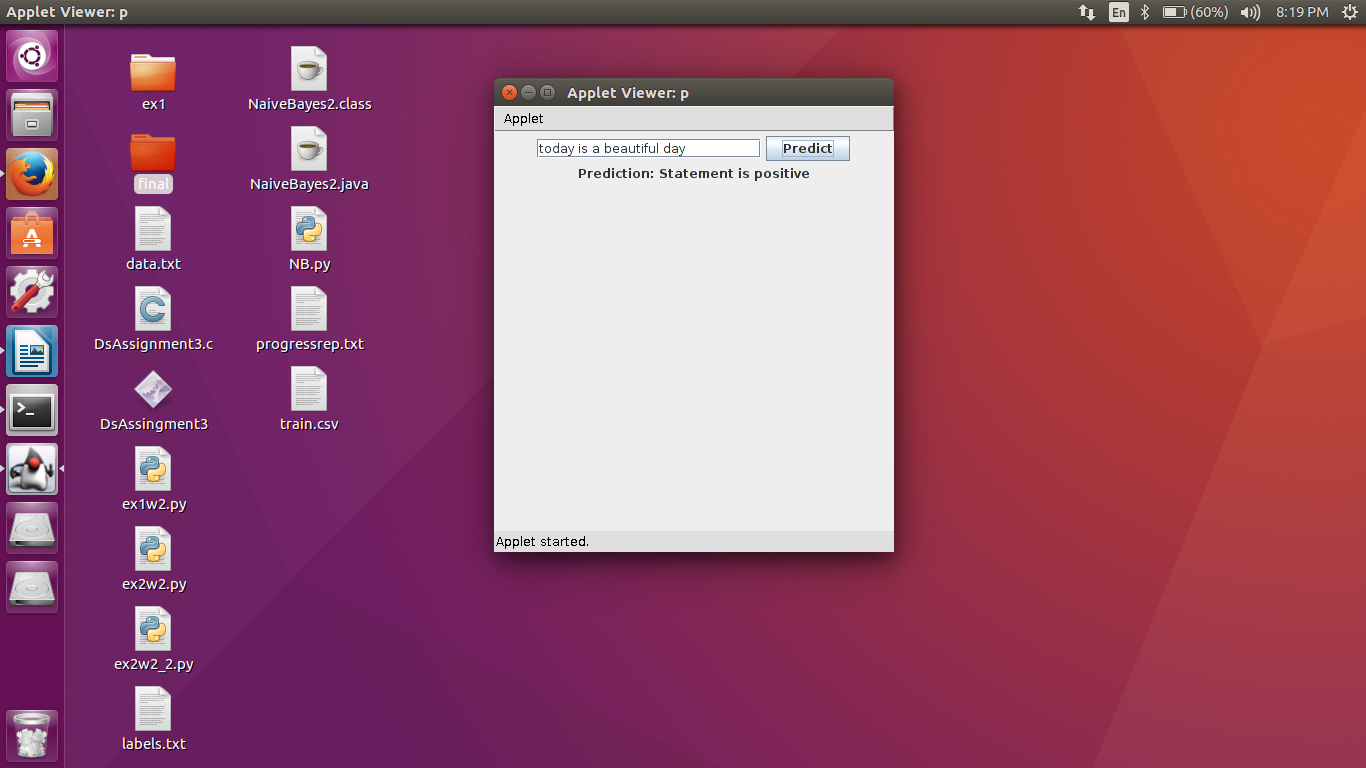
**User Defined Function:**

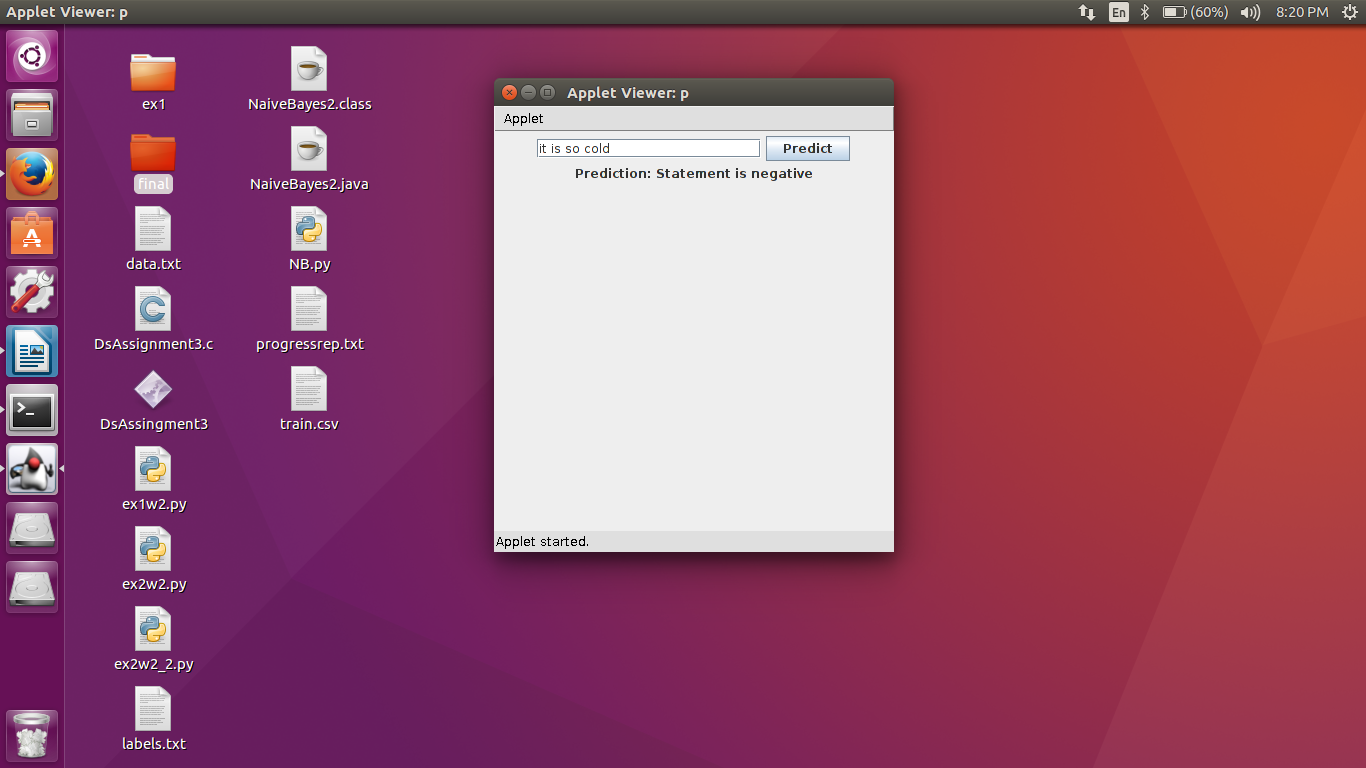
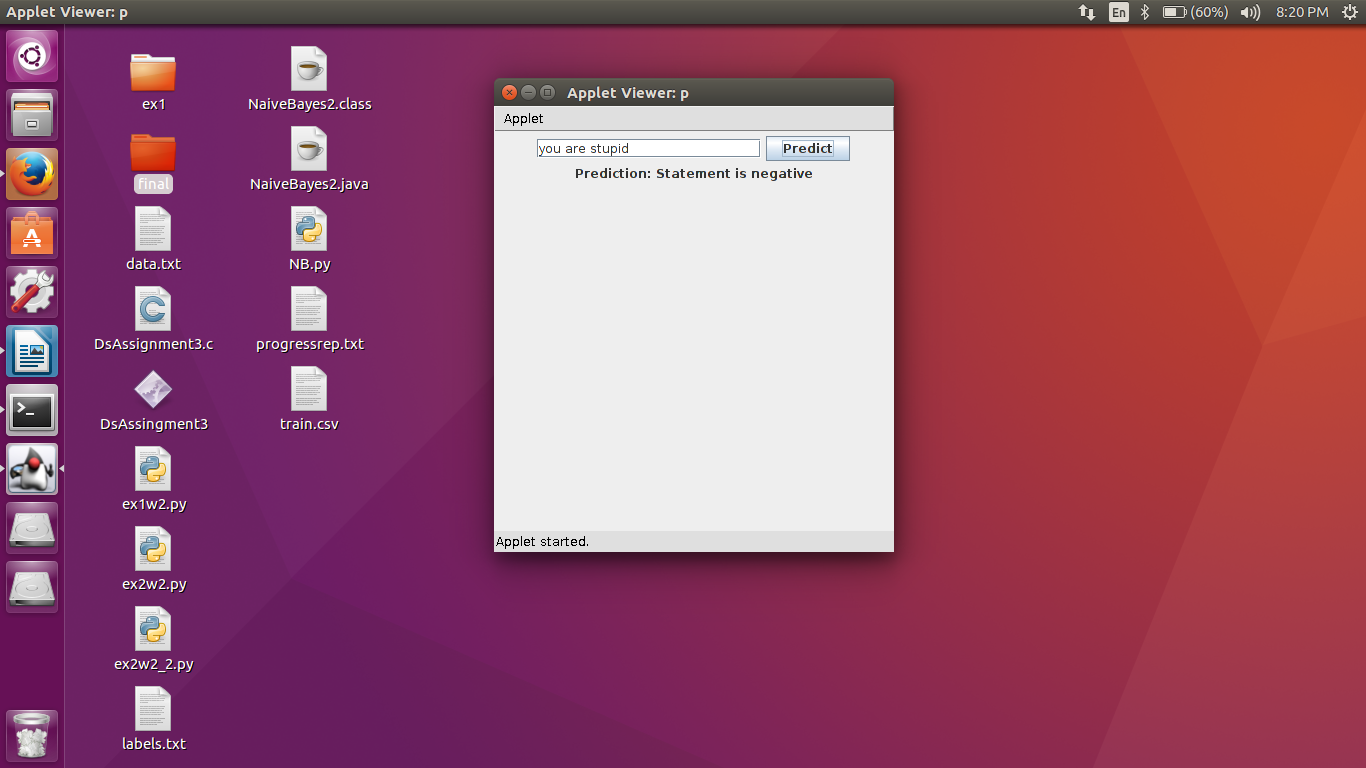
* CreateSet() : To create the set
* initiate() :To call the relevant methods and finally display and output
* setOfWords2Vec() :To convert sentences to their vector forms
* trainNB() :To train the classifier
* classify() :To classify the sentence
* init() :Calls the driver method
* driver() :Adds elements to the displayed window and

calls the initiate method

**I/O Specifications:** The input and ouput are taking place through the appletviewer.

**Results:**





**Conclusion:** The Naive Bayes Classifier works is a good application of the Bayes decision theory and Conditional probability and it gives accurate results as seen above even if the number of training examples available are few.

**Reference:** Machine Learning In Action, Peter Harrington.