**Naive Bayes and Binary Naive Bayes**

**Introduction:-**

Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes’ theorem with the “naive” assumption of independence between every pair of features. Naive Bayes learners and classifiers can be extremely fast compared to more sophisticated methods. In spite of their apparently over-simplified assumptions, naive Bayes classifiers have worked quite well in many real-world situations, famously document classification and spam filtering.

Sentiment Analysis is the process of determining whether a piece of writing is positive, negative or neutral. It’s also known as opinion mining, deriving the opinion or attitude of a speaker. A common use case for this technology is to discover how people feel about a particular topic.

**Methodology:-**

Data structures, important variables:-

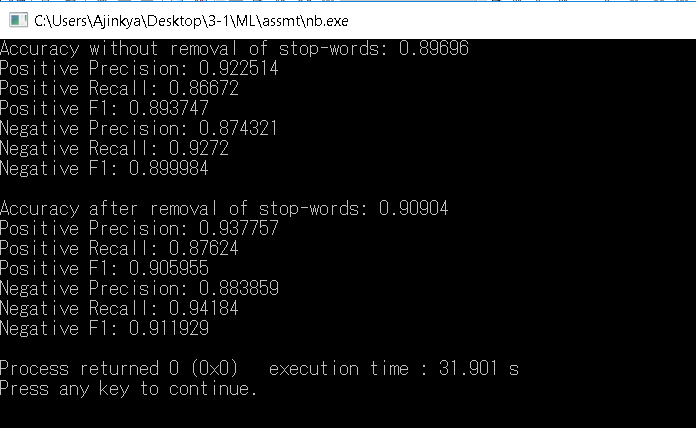
* Total\_words = total words in the document.
* Word\_freq = 2d array representing the frequencies of words in positive and negative documents.
* Correct = no. of documents correctly predicted.
* Count\_pos = no. of words in positive documents.
* Count\_neg = no. of words in negative documents.
* Word index = a map which maps words against their corresponding integers.

Important Functions:-

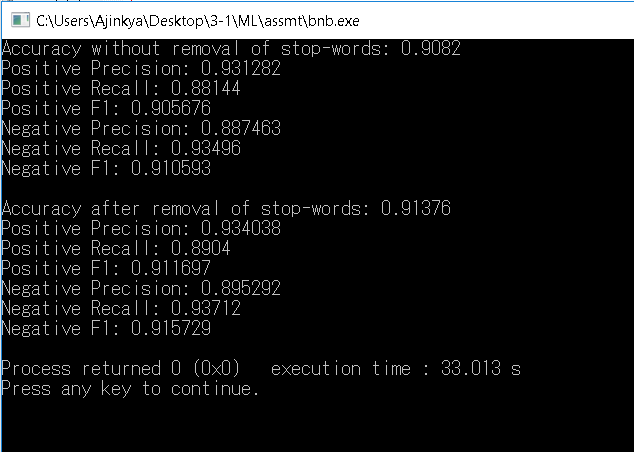
* Create\_word\_freq = function to read and store the word-frequencies in a 2d matrix
* Calc\_porb = function to calculate probability for given word
* Pred\_Acc = function to calculate probabilities of all documents
* Create\_index = create an index that maps words to their corresponding numbers
* Remove\_stop\_words = function to remove stop words from training and testing data

**Results:-**

Following are the results that we observed after applying naïve bayes algorithm to the training data –



Following are the results that we observed for binary naïve bayes algorithm –



**Conclusion:-**

* We observe that binary naïve bayes and naïve bayes algorithms are reasonably good algorithms for predicting the sentiment of users based on their reviews.
* We also observe that the accuracy of our predictions increased after removing stop words. This shows that stop words don’t play a role in measuring the sentiment of users.
* Finally we observed that binary naïve bayes performs better than the traditional naïve bayes algorithm. This shows that the existence/non-existence of a word plays a more important role in predicting the sentiment of users than its frequency.