# SENTIMENT ANALYSER of movie reviews USING NAÏVE BAYES

## AIM

We have build a Sentiment Analyser that will interprete the sentiment of the text and classify the movie reviews as Positive or Negative. We have achieved this by using Naïve Bayes algorithm for classification.

## Capturing Of Data

We have collected a text file containing 7086 reviews of books and movies from https://www.kaggle.com/c/si650winter11. The data is already labelled with 1 (positive sentiment) or 0 (negative sentiment). Further we divided this data into training and test data and checked the accuracy of model for both.

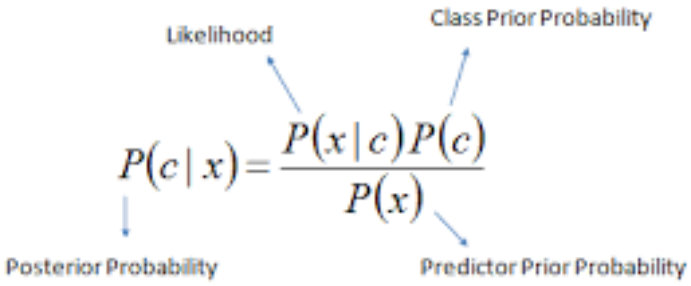
## Cleaning and Text Processing

The data was already pretty clean and arranged, then we classified the data into two columns named **'txt'** (contained the movie review) and **'liked'** (containing the comments 0 and 1 respectively for the reviews). This created a proper structured data for training the model. The csv file was read while doing the same.

## Use of TF-IDF Vectorization

- Firstly, we listed all the stop words in the language and then used TF-IDF Vectorization.
- TF-IDF is an abbreviation for Term Frequency-Inverse Document Frequency and is a very common algorithm to transform text into a meaningful representation of numbers. The technique is widely used to extract features across various NLP applications.
- Then the 'txt' section containing movie reviews is converted from text to features by using vectorization i.e. all the text is now represented by numbers
- We defined dependent variable 'y' that will be 0(didn't like the movie) and 1 (liked the movie) and independent variable as 'x' i.e. the movie review.
- Now our text file is converted to vector form containing two vectors x and y and finally ready for training our model.

**TRAINING THE NAÏVE BAYES CLASSIFIER**




$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

- We splitted our text file into training and test data
- We imported module named ‘naive_bayes’ from the library ‘sklearn’.
- Then finally trained our model by using the following commands-
- clf = naive_bayes.MultinomialNB()
- clf.fit(x_train,y_train)

## Testing The Accuracy

- we tested the accuracy of our model with training as well as test data which are following-

- Accuracy with Training data - **0.9998073132206913**
- Accuracy with Test Data - **0.9979292333245913**
- As we can see the accuracy of training and test data both are High, our model is a pretty good model.

## Conclusion

At the end, we ran our model on any random movie review and it correctly classified it as ‘1’ Or ‘0’ which respectively means ‘positive’ or ‘negative’.
Naive Bayes is a classification algorithm that is suitable for binary and multiclass classificati on. It is a supervised classification technique used to classify future objects by assigning class labels to instances/records using conditional probability. In supervised classification, training data are already labeled with a class. The above model could be made by using other machine learninig algorithms also like logistic regression, but the accuracy will be highest with Naïve Bayes. Some Other applications of Naïve Bayes are Weather Forecast and Fraud Detection.

## Packages Used

- Nltk, nltk.corpus,
- Sklearn, sklearn.metrics, sklearn.feature_extraction.text, sklearn.model_selection

## Refrences


- https://www.kaggle.com/c/si650winter11
- **https://youtu.be/oXZThwEF4r0**
- https://www.google.com/imgres?imgurl=https%3A%2F%2Fwww.saedsayad.com%2Fimages%2FBayes_rule.png&imgrefurl=https%3A%2F%2Fwww.saedsayad.com%2Fnaive_bayesian.htm&docid=skmzIxw8w4Lf1M&tbnid=kwLT20eBUyxVdM%3A&vet=10ahUKEwjE-dP3i5LmAhWDxTgGHVM8DbMQMwhiKAAwAA..i&w=461&h=264&bih=674&biw=1536&q=naive%20bayes%20formula&ved=0ahUKEwjE-dP3i5LmAhWDxTgGHVM8DbMQMwhiKAAwAA&iact=mrc&uact=8