TwitMiner

Team Name: Bazinga

Team Members:
- Rahul Huilgol, B.Tech 2nd year, Dept of CSE, IIT Guwahati
- Simrat Singh Chhabra, B.Tech 2nd year, Dept of CSE, IIT Guwahati

This machine learning program to classify tweets is coded in Python 2.7.3.
The program uses a self-developed implementation of Naïve Bayes classifier to classify the tweets.
The implementation is as follows:

- We define a list of 100 common words which we call the 'stop' words. These words are not used during the implementation of the algorithm.
- First we normalize the case of all words. We remove punctuation marks.
- Then we create 2 dictionaries: sports and politics. (a dictionary corresponding to a class)
- The sports dictionary contains all the words present in sports labelled tweets (except the stop words) found out during training and the politics dictionary contains all the words present in politics labelled tweets (except the stop words) found out during training. We also keep a count of the number of occurrences of each word.
- Using these we find the probability of a word occurring given the class.
- When we run the program, for each word in the tweet, we find out the probability of the class being sport and the probability of the class being politics using Bayes Theorem.

- If the product of the probabilities of all the words is greater for the class sport, then we classify it as sport otherwise as politics.

.