Machine Learning (BITS F464) - Assignment 3
Naïve Bayes Classifier

Maximum Marks:20
Submission Deadline: 13/11/2017, 9AM

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.

Your task is to classify whether a given review has a positive or negative tone using naive Bayes classifier.

Download the dataset from - http://ai.stanford.edu/~amaas/data/sentiment/

Dataset has 12,500 positive and 12,500 negative reviews of movies for training and testing separately. Each review is in separate file. Convert the data into a format that is comfortable for training and testing.

Apply the naive Bayes classifier along with smoothing techniques for the probabilities.

Extend Naive Bayes classifier using following techniques:

1. **Removing Stop Words**: Stop words are the most common words which occur in text. For stop words consider 'Default English stopwords list' from this website - https://www.ranks.nl/stopwords . To read about stop words refer to the textbook attached below.

2. **Binary Naive Bayes**: This is a variant of Naive Bayes which you can read in the textbook referred below under the section '6.4 Optimizing for Sentiment Analysis' and the topic 'binary NB'.

For each of the above extensions see how the results change from the base version of naive Bayes.

Languages allowed: C, C++, Java. It goes without saying that you are not allowed to use any packages which implement the aforementioned algorithms or copy code.

Report:
1. Precision, Recall and F1 measure for positive and negative sentiment in a table for the naive Bayes classifier and extensions.

2. Possible reasoning for the change in the results (increase or decrease) from the basic version.

Evaluation:
1. Results
2. Understanding of results.
3. Ability to reason the derived results.
4. Code documentation.
5. Final demo.

Submission should be through **CMS** only.

Refer to this book- Speech and Language Processing by Daniel Jurafsky and James H.Martin.
PDF for Naïve Bayes and Sentiment Analysis chapter -
https://web.stanford.edu/~jurafsky/slp3/6.pdf

Contact the following Teaching Assistants for any clarification on this assignment.
SriHarshitha Velivelli f20130847@hyderabad.bits-pilani.ac.in
Rajitha p2015409@hyderabad.bits-pilani.ac.in