EE 580 Lab 3 Part3
Summer 2017 Nazarian **Score:___/100**

Student ID: _____ **Name: _____**
Assigned: Wednesday May 31st
Due: Wednesday June 7th at 11:59PM
Late submissions will be accepted only in the first two days after deadline with a maximum penalty of 15% per day: For each day, submissions between 12 and 1am: 2%, 1 and 2am: 4%, 2 and 3am: 8% and after 3am: 15%.

Notes:

- **All assignments including this lab are based on individual work. No collaborations (including no discussions) are allowed.**
- **We may pick some students in random to demonstrate their design and simulations. Please watch the first lecture of this course regarding the academic integrity policies and also refer to the syllabus for a summary of AI policies (including the penalties for any violation).**
- **If you have any concerns or doubts about what is or is not allowed or prohibited in this course, please contact the instructor.**
- **ATTENTIONS: Start early otherwise you cannot finish this lab on time.**

Naive Bayes Classifier Design

Overview

The goal of this assignment is to get some experience with text classification using the basic machine learning technique. You will be working with an email dataset and perform binary classification: SPAM or HAM (not spam).

We are providing two sets of data. One is labeled, and one is specifically for testing and we are not providing the labels. You will use the labeled data to train your model and submit your classification result from the test data.

Data Set

On Blackboard, we will post two sets of data: labeled and unlabeled (test) as well as enron.vocab. All data has already been cleaned up (leaving only the text parts of the subjects and bodies) and tokenized with tokens separated by spaces.

All email data are stored as zipped files and you'll need to unzip the files. The archives contain a large number of individual files (one per email) divided into subfolders "ham" and "spam". You will need to write a C++ program to convert all these individual files into a single file in the project data format.

Naive Bayes Classifier in C++

You need to write a C++ program nblearn.cpp that will generate a model (nb.model) from training data set (labeled data set), and nbclassify.cpp will use a model to classify the unlabeled data.

The format for nb.model is up to you but should contain sufficient information for nbclassify.cpp to process the unlabeled data and for each line print to result.txt the name of the more probable class (one per line).

SPAM
SPAM
HAM
SPAM
……

Smoothing, and common, rare and unknown tokens

For the Naive Bayes classifier, you need to consider these issues. The reference solution written by the TAs will be using add-one smoothing on the labeled data. For tokens unique to the unlabeled (testing) data, the reference solution will simply ignore these tokens (i.e., pretend they did not occur).

Self Check

You can check how powerful your classifier is by dividing the labelled data into two subsets. Use 80% of the data for training and use 20% of the data for testing (assume they are unlabeled). Compare classification result with the original data label.

Submission Checklist (submit one lab3_part3_firstname_lastname.zip file containing the following files)

nblearn.cpp
nbclassify.cpp
nb.model (can be in any format)
result.txt
any additional supporting .cpp file (i.e., change the data format)
readme.txt (tell the TA how to run your code)

Important Notice: Do not contain any data file. Put all the training data and the test data in the original folder.