1. **YouTube Spam Filtering**

   This Lab shows you how to build a simple spam filtering tool. Each data set has comment ID, author, date, and class (0=ham, 1=spam).

   (a) Download the YouTube Spam Collection data set from: `https://archive.ics.uci.edu/ml/datasets/YouTube+Spam+Collection`. We will mainly work with the Eminem file. Note that the chronological order of the comments were kept. (5 pts)

   (b) Select the first $\lfloor 0.8H \rfloor$ of H hams and the first $\lfloor 0.8S \rfloor$ of S spams as your training set and the rest as your test set. (10 pts)

   (c) Represent each comment using TF-IDF features. Consider each comment a document and the whole set of comments in the file (e.g. Eminem) as the corpus. Do NOT remove stop words.[1] (10 pts)

   (d) $\mathscr{L}_2$-penalized Logistic Regression

      i. Determine λ using five fold cross-validation on your training set. Consider $\log_{10}\lambda \in \{-5, -4, \ldots, 5\}$. (5 pts)

      ii. Train $\mathscr{L}_2$-penalized Logistic Regression using the λ you found in 1(e)i. Calculate the confusion matrix, accuracy, precision, recall, and F1 score as well as the ROC curve and AUC for your training set. (10 pts)

      iii. Test the algorithm on the test set and calculate the confusion matrix, accuracy, precision, recall, and F1 score as well as the ROC curve and AUC for the test set. Which one is more important in this application, precision or recall? Calculate F$_\beta$ score for $\beta \in \{0.1, 0.5, .9, 1, 5, 10\}$. (15 pts)

   (e) $\mathscr{L}_1$-penalized Logistic Regression

      i. Determine λ using five fold cross-validation on your training set. Consider $\log_{10}\lambda \in \{-5, -4, \ldots, 5\}$. (5 pts)

      ii. Train $\mathscr{L}_1$-penalized Logistic Regression using the λ you found in 1(e)i. Calculate the confusion matrix, accuracy, precision, recall, and F1 score as well as the ROC curve and AUC for your training set. (15 pts)

      iii. Test the algorithm on the test set and calculate the confusion matrix, accuracy, precision, recall, and F1 score as well as the ROC curve and AUC for the test set. (10 pts)

   (f) Binary Classification Using Naïve Bayes' Classifiers

      i. Solve the problem using a Naïve Bayes' classifier. Use Gaussian class conditional distributions. Report the confusion matrix, ROC, precision, recall, F1 score, and AUC for both the train and test data sets. (15 pts)

   (g) (Extra Credit, 5 points) Repeat 1(f)i using multinomial priors.[2]

[1] Some believe that removing stop words is not a good idea in spam filtering.

[2] We briefly covered them in the lecture without using the term multinomial. Research what they mean.

(h) (Extra Credit, 20 points) Create one table for each of the five data sets: Shakira, Eminem, LMFAO, KatyPerry, and Psy to compare accuracy, precision, recall, F1 score and AUC of $\mathscr{L}_1$-penalized and $\mathscr{L}_2$-penalized logistic regression, Naïve Bayes' with Gaussian and multinomial priors for the test sets. Use the first $\lfloor 0.8H \rfloor$ of H hams and the first $\lfloor 0.8S \rfloor$ of S spams as your training set and the rest as your test set. Show the best score in each column using boldface. A hypothetical table is shown below:

Psy

Instance	Accuracy	Precision	Recall	F1	AUC
NB-Gaussian	96%				
NB-Multinomial	**98.2**%				
$\mathscr{L}_2$	88%				
$\mathscr{L}_1$	85%				