CSC 395 : Information Retrieval
Fall 2019
Assignment 02
Spam Detection & Machine Learning
Full Credit: 100

Assigned Date: September 28 2019
Mid-Progress Presentation: October 03 2019
(In class, 5 min presentation/demonstration)
Final Submission Due: October 10 1019 (10:30 PM)

Problem Description:

As a student of Information Retrieval class, we all know that our goal is to find the latent features from the so-called "unstructured" and abnormally "large" collection of data. In this assignment, we aim to solve (or at least try our best) the "email-spam detection" problem.

The **spam filtering problem** is one kind of text categorization problem with the categories being spam and ham. The structure of email is richer than that of flat text, with meta-level features such as the fields found in MIME compliant messages. Researchers have recently acknowledged this, setting the problem in a semi-structured document classification framework. Several solutions have been proposed to overcome the spam problem. Among the proposed methods, much interest has focused on the machine learning techniques in spam filtering. They include rule learning, **Naive Bayes, decision trees, support vector machines**, or **combinations of different learners**. The basic and common concept of these inductive approaches is that using a classifier to filter out spam and the classifier is learned from training data rather than constructed by hand.

Your job is to choose **two possible machine learning approaches** to solve the problem and analyze their performance. Then you will report possible mechanisms to improve the current models you implemented.

Corpus/Dataset:

You will use TREC 2007 Spam Corpus (https://trec.nist.gov/data/spam.html) as the standard dataset for this assignment.

Grading Rubric:

We are free to choose any platform (programming language, tools, etc.). At the end, we will submit the followings:

[60 points] A directory with fully documented source code
[10 points] A README
[30 points] A 3-page (at least) technical report

You may find the following sources very helpful:

https://trec.nist.gov/data/spam.html

https://scikit-learn.org/stable/

https://hackernoon.com/how-to-build-a-simple-spam-detecting-machine-learning-classifier-4471fe6b816e

https://en.wikipedia.org/wiki/Naive_Bayes_spam_filtering

https://medium.com/analytics-vidhya/building-a-spam-filter-from-scratch-using-machine-learning-fc58b178ea56

https://trec.nist.gov/pubs/trec15/papers/hit.spam.final.final.pdf