**CS6140: Machine Learning**

**Fall 2018**

**Professor: Craig Martell**

**Email: c.martell@northeastern.edu**

**Homework 1: Naïve-Bayes Spam Classifier**

In this assignment you will build a naïve-Bayes classifier to determine whether an email is *spam* or not (*ham*). You will be provided three email sets: Spam, Ham and Test. Your delivery will be code that produces the *output table* below (in whatever data structure you want). You will also deliver a report describing naïve Bayes in your own words, as well as what your code does. It is important that you don’t simply use a naïve-Bayes API from a library. You must write your own classifier.

**Output Table**:

|  |  |  |
| --- | --- | --- |
| **Email ID** | **Classifier Output** | **Truth** |
| *The id of the email being tested* | *Label your classifier produces* | *Truth - provided by instructor* |
| *Example:* 105 | spam | ham |

**Input (provided by instructor):** Zip file containing three directories: Ham, Spam and Test. They contain the training data for Ham and Spam and the test files to test your classifier on.

**To be delivered:**

1. A program that implements a naïve-Bayes spam classifier using the input sets. It should produce a data structure containing the elements in the *Output Table* above. Bring this program to class ready to run it. It does not need to be submitted.
2. A report that
   1. Describes a naïve-Bayes classifier in your own words; and
   2. Describes what each section of your code does and how it implements a naïve-Bayes classifier.

**Due date:** The report needs to be turned in electronically before the start of class on 9/14/2018. The program needs to be complete and available in class on 9/14/2018.