**Title:** Data Analysis Report

**Date:** 21 November 2017

**Student ID Number:** 17350942-1

**Name:** Tittaya MAIRITTHA

**# Purpose**

Filtering text messages as Spam or Ham using the supervised machine learning algorithm a naïve Bayes classifier and logistic regression.

**# Dataset**

I used all the datasets from [SMS Spam Collection Data Set from UCI Machine Learning repository](https://archive.ics.uci.edu/ml/datasets/sms+spam+collection). The SMS Spam Collection is a public set of SMSs labeled messages that have been collected for mobile phone spam research. This includes 5,572 text messages; ham 4825 and spam 747.

**Example dataset**

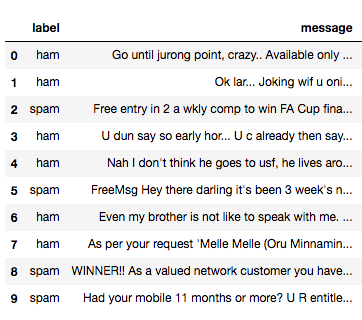
****

Figure 1 An example of datasets

**# Data Entry and Forming Method**

1. Converting raw text file into data frame
2. Convert label to a numerical variable
3. Define X and Y (from the data frame)
4. Split X and Y into training 75% and testing sets 25%
5. Representing text as numerical data using CountVectorizer (use it to tokenize and count the word occurrences of a minimalistic corpus of text documents:)
6. Tuning the vectorizer
   1. Remove English stop words
   2. Include 1-grams and 2-grams

**# Methods**

1. Naïve Bayes classifiers are a class of simple linear classifiers which are conditional probability models based on Bayes Theorem. In this case, used the multinomial Naïve Bayes classifier which implements the naive Bayes algorithm for multinomially distributed data, and is suitable for classification with discrete features (e.g., word counts for text classification).
2. Logistic Regression is another way to determine a class label, depending on the features. Logistic regression takes features that can be continuous (for example, the count of words in SMS texts) and translate them to discrete values (spam or not spam).

**# Results**

**Compute accuracy, precision, recall, F-measure**

1. The precision is the ratio tp / (tp + fp) where tp is the number of true positives and fp the number of false positives.
2. The recall is the ratio tp / (tp + fn) where tp is the number of true positives and fn the number of false negatives. The recall is intuitively the ability of the classifier to find all the positive samples.
3. The F- measure score can be interpreted as a weighted harmonic mean of the precision and recall, where an F- measure score reaches its best value at 1 and worst score at 0.

**Results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F-measure** |
| **Naive Bayes** | 0.99 | 0.99 | 0.97 | 0.98 |
| **Logistic Regression** | 0.98 | 0.99 | 0.93 | 0.95 |

**Confusion matrix**

|  |  |
| --- | --- |
| Naive%20Bayes%20CNF.png | Logistic%20Regression%20CNF.png |

**Examples**

|  |  |  |
| --- | --- | --- |
| **Example** | **Multinomial Naive Bayes** | **Logistic Regression** |
| **The false positives (ham incorrectly classified as spam** | Hey...Great deal...Farm tour 9am to 5pm $95/pa... | - |
| **the false negatives (spam incorrectly classified as ham)** | Xmas & New Years Eve tickets are now on sale f... | 449050000301 You have won a å£2,000 price! To ... |

**# Summary**

**Naïve Bayes classifiers**

A total of only 2 + 10 = 12 of the 1,393 SMS messages were incorrectly classified (0.86%).

Among the errors were 10 out of 1,221 ham messages that were misidentified as spam, and 2 of the 172 spam messages were incorrectly labeled as ham.

**Logistics Regression**

A total of only 24 of the 1,393 SMS messages were incorrectly classified (1.72%).

Among the errors were 24 out of 1,237 ham messages that were misidentified as spam, and 0 of the 156 spam messages were incorrectly labeled as ham.

The true messages that were incorrectly classified as spam could cause significant problems for the deployment of our filtering algorithm, because the filter could cause a person to miss an important text message.