SPL-1 Project Report 2018

SPAM SMS DETECTION

SE 305: SPL-1

Submitted by

ISHRAT JAHAN EMU

BSSE Roll No. : 0927

BSSE Session: 2016-17

Supervised by

Dr. Mohammad Shoyaib

Designation: Professor

Institute of Information Technology

Institute of Information Technology

University of Dhaka

30-05-2018

Table of Contents

1. Introduction 3

1.1. Background Study 3

1.2. Challenges 4

2. Project Overview 4

3. User Manual 5

4. Conclusion 7

References 7

1. Introduction

There are two kinds of SMS.

1. Spam SMS
2. Ham SMS

Spam SMS is any kind of junk message delivered to a mobile phone as text messaging through the SMS. Spam messages include advertisements, free services, promotions, awards etc. It refers to unwanted commercial text. Spam SMS detection is harder than spam e-mail detection. The purpose of my project is to detect spam SMS from SMS dataset.

1.1. Background Study

The domain of my project is natural language processing. Natural language processing (NLP) is a technology created from the need for machines to understand and communicate with humans in human language. It is a way of analyzing texts by computerized means. NLP involves gathering of knowledge on how human beings understand and use language.

To complete my project, I've to know about tokenization, stop word removal, stemming, frequency count and classifier algorithm.

I've learned about 2 classifier algorithm.

I. K-Nearest Neighbor Algorithm : K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions).

II. Naïve Bayes Algorithm : The Naive Bayes Classifier technique is based on the so-called Bayesian theorem and is particularly suited when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods.

1.2. Challenges

I've to face challenges –

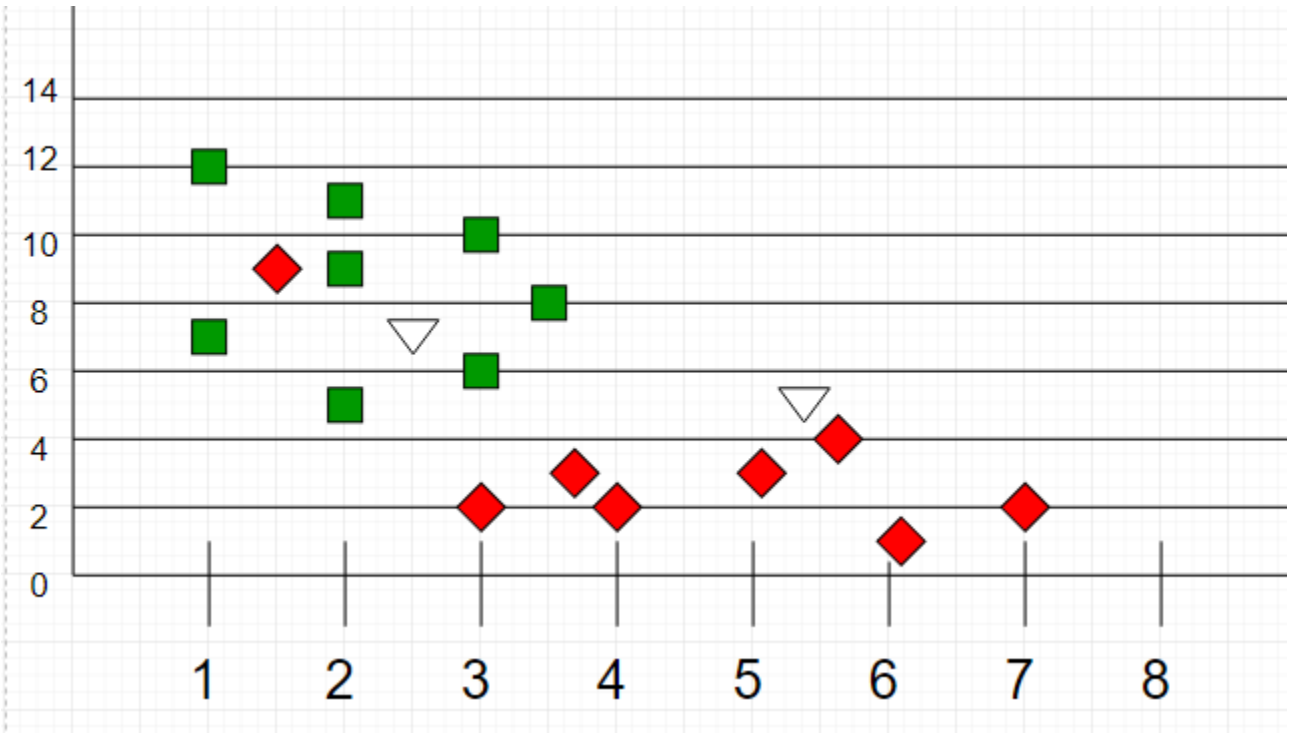
- Frequency Count
- To select testing and training dataset (As I'm new in machine learning, so it is difficult to me select testing and training data. Then after studying about it, I can overcome my problem)

- Implementing Naïve Bayes (Naive Bayes algorithm was new to me. But after studying about it ,I can overcome)

2. Project Overview

In my project, I've to implement-

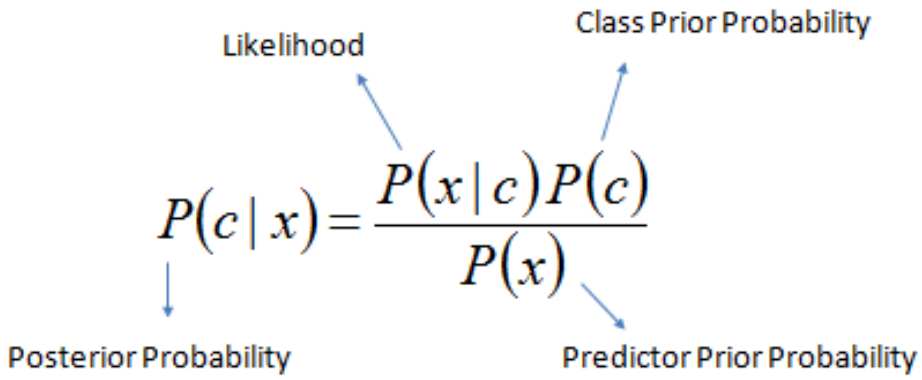
- Tokenization
- Stop word removing
- Root word frequency
- Classifier (Using K-nearest neighbor algorithm)



I take 90% Training data and 10% Testing data for K-NN algorithm implementation. I test the whole dataset 10 times. Then finally show the average accuracy, which is 72.6316%.

- Classifier (Using Naïve-Bayes algorithm)

I also take here 90% Training data and 10% Testing data. Then for 10% Testing data, I show the average accuracy which is 88.2353%.

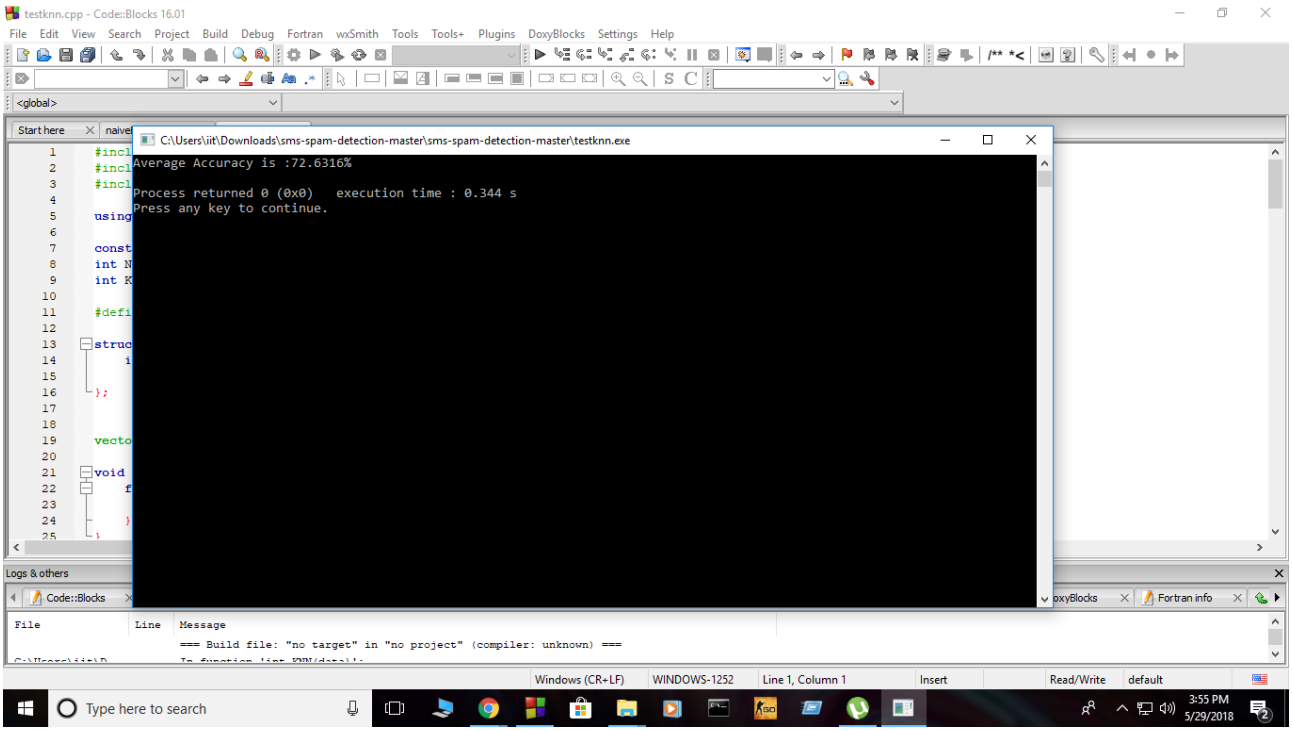


$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

3. User Manual

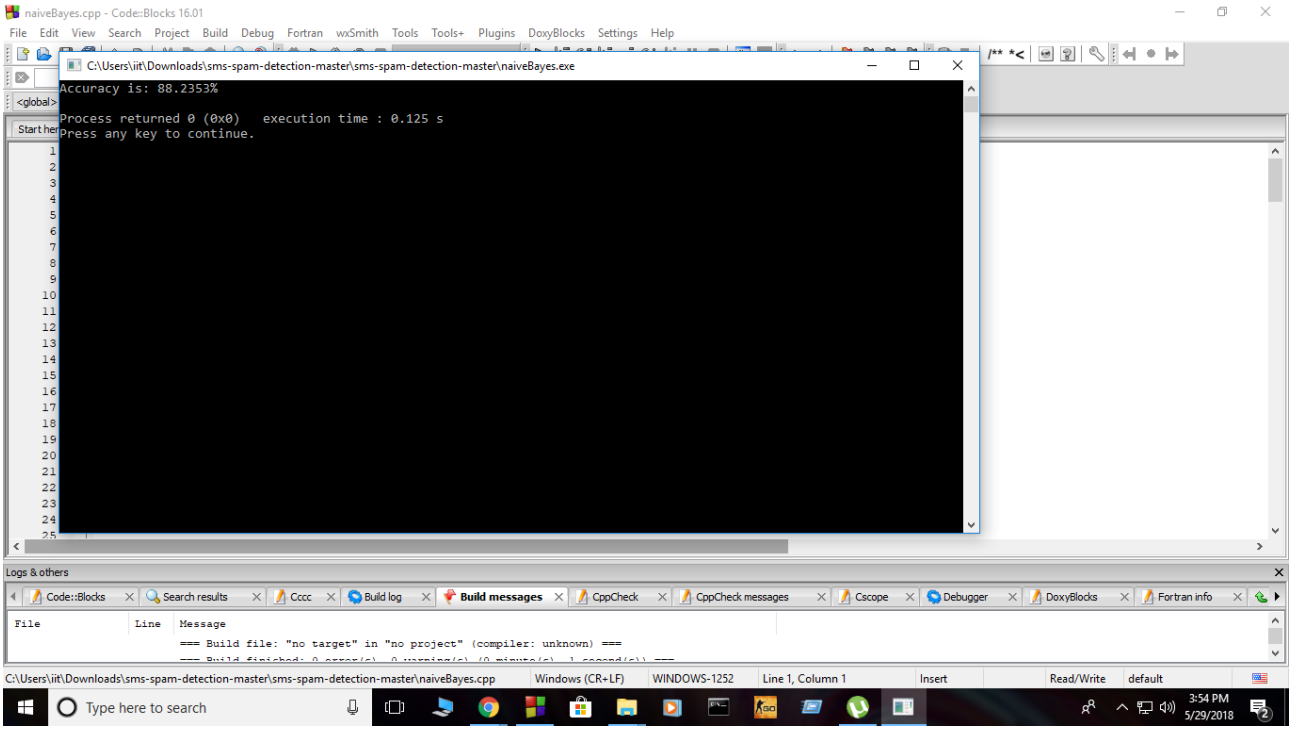
Actually, I use two classifier algorithms for detecting spam SMS.

For K-NN algorithm, my sample output is



Here ,my input file is out2.txt. And here, average accuracy is 72 .6316% .

For Naïve Bayes algorithm, my sample output is



Here my input file is smsDataset.txt. And here, accuracy is 88.2353%.

4. Conclusion

The Spam SMS problem is increasing nowadays with the increase in the use of text messaging. Spam SMS filtering is the big challenge these days. I use two classifier algorithm to detect spam SMS. Naive Bayes algorithm shows more accuracy than K-NN.

References

[1] My github link: https://github.com/ishrat98/sms-spam-detection

[2] Dataset link : https://archive.ics.uci.edu/ml/datasets/sms+spam+collection