

Spam Mail Detection Project

Submitted by:

Pooja Pratik Rajput

**ACKNOWLEDGMENT**

On this great occasion of the successful completion of my computer project, I would like to thank Javatpoint website and YouTube to give the knowledge about NLP. Also thanks to Stack Overflow site to solve my error.

**INTRODUCTION**

* Business Problem Framing

Spam Detector is used to detect unwanted, malicious and virus infected texts and helps to separate them from the nonspam texts. It uses a binary type of classification containing the labels such as ‘**ham’** (nonspam) and **spam**. Application of this can be seen in Google Mail (GMAIL) where it segregates the spam emails in order to prevent them from getting into the user’s inbox.

* Conceptual Background of the Domain Problem

Describe the domain related concepts that you think will be useful for better understanding of the project.

* Review of Literature

**Natural Language Processing (NLP)** is the ability of a machine to read, write, understand and derive meaning from a human language.

**Steps in NLP**

* Tokenization
* Stemming
* Lemmatization
* Part-of-speech (POS) tagging
* Named entity recognition
* Chunking

1. **Tokenization**: We break down the text into **tokens**. Check the example below to see how this is done.

**Text:** The cat sat on the bed. **Tokens:** The, cat, sat, on, the, bed

1. **Stemming**: We remove the prefixes and suffixes to obtain the root word. Check the example below to see how it’s done.

**List of words**: Affection, Affects, Affecting, Affected, Affecting  
**Root word**: Affect

1. **Lemmatization**: We group together different inflected forms of a word into a base word called **lemma**. Check the example below how it’s done.

**List of words**: going, gone, went  
**Lemma**: go

1. **POS tagging**: We identify the parts of speech for different tokens. Check the example below to see how it’s done.

**Sentence:** The dog killed the bat.  
**Parts of speech**: Definite article, noun, verb, definite article, noun.

1. **Named entity recognition**: We classify named entities mentioned in the text into categories such as “People,” “Locations,” “Organizations,” and so on. Check the example below to see how it’s done.

**Text**: Google CEO Sundar Pichai resides in New York.  
**Named entity recognition**:  
Google — Organization  
Sundar Pichai — Person  
New York — Location

1. **Chunking**: We pick up individual pieces of information and group them into bigger pieces.

* Motivation for the Problem Undertaken

Implementing spam filtering is extremely important for any organization. Not only does spam filtering help keep garbage out of email inboxes, it helps with the quality of life of business emails because they run smoothly and are only used for their desired purpose. Spam filtering is essentially an anti-malware tool, as many attacks through email are trying to trick users to click on a malicious attachment, asking them to supply their credentials, and much more.

**Analytical Problem Framing**

* Data Sources and their formats

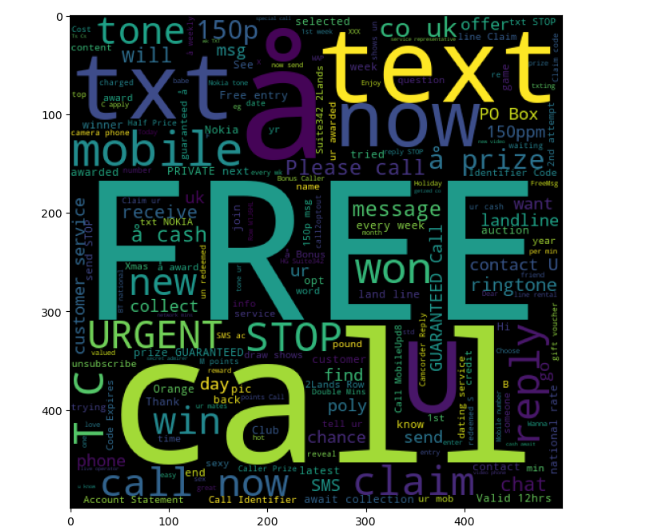
The files contain one message per line. Each line is composed by two columns: v1 contains the label (ham or spam) and v2 contains the raw text.

This corpus has been collected from free or free for research sources at the Internet:

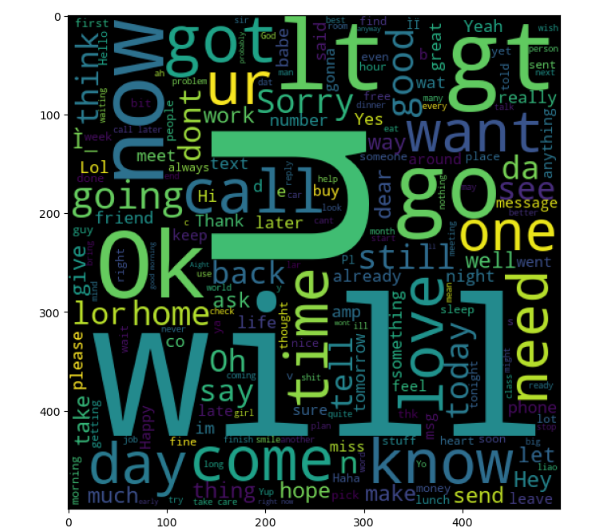
-> A collection of 5573 rows SMS spam messages was manually extracted from the Grumbletext Web site. This is a UK forum in which cell phone users make public claims about SMS spam messages, most of them without reporting the very spam message received. The identification of the text of spam messages in the claims is a very hard and time-consuming task, and it involved carefully scanning hundreds of web pages.

-> A subset of 3,375 SMS randomly chosen ham messages of the NUS SMS Corpus (NSC), which is a dataset of about 10,000 legitimate messages collected for research at the Department of Computer Science at the National University of Singapore. The messages largely originate from Singaporeans and mostly from students attending the University. These messages were collected from volunteers who were made aware that their contributions were going to be made publicly available.

* Data Preprocessing Done
  + In this project for data cleaning first we have to remove the unwanted column and then rename the column name.
  + We rename v1 to labels and v2 to messages
  + Now we visualize the spam words with the help of wordcloud Library



* + Now we visualize the ham words with the help of wordcloud Library



* + Now encoding our labels means if message is span it display 1 or if message is ham then it will display 0 in dataset
  + Now we apply NLP on data set
  + After NLP is completed we split our data into train and test model.
  + So, here we can see that X\_train have 4179 data and X\_test have 1393 data which means 75% data is for training data and 25% data for testing purpose
  + Model will not understand the word so we need to convert it into a vector means in number

# Training and testing data is ready ,Now we train our model using Naive Bayes algoritham

* + After training our model we check our model accuracy. so here our model is **85%** accurate.