

Image Spam Classifier

Minor Project

Submitted By:

Lakshay Hurria
Satyendar Kumar
Mujahid Omer

INTRODUCTION

In this modern era of technology, we are getting spammed from all kinds of places. Be it media, social network or even calls, spamming has been a very annoying yet dangerous activities.

In the present automated world, sharing information is important to be competitive and sustainable in business. Email is a rapid and inexpensive medium of communication. It is a popular medium of interaction between people and has become a part of life itself.

- Nowadays, spam classification is becoming a challenging area due to the complex nature of the spam.
- Complexity is defined as the modifications of content, such as tokenization and obfuscation, etc.. to change the information of features so as to create barriers in distinguishing spam from legitimate emails.



- Most e-mail readers spend a non-trivial amount of time regularly deleting junk e-mail (spam) messages, even as an expanding volume of such e-mail occupies server storage space and consumes network bandwidth.
- An ongoing challenge, therefore, rests within the development and refinement of automatic classifiers that can distinguish legitimate e-mail from spam.

Spam vs Ham

- Wikipedia describes Spam as “the use of electronic messaging systems to send unsolicited bulk messages, especially advertising, indiscriminately.”
- The key word here is **unsolicited**. This means that you did not ask for messages from this source.
- So if you didn't ask for the mail it must be spam, Right? That is true, however quite often people don't realize that they are signing up for mailers when they download free software, or sign up for a new service, or even when updating existing software.



- "Ham" is *e-mail that is not Spam*. In other words, "non-spam", or "good mail". It should be considered a shorter, snappier synonym for "non-spam".
- Its usage is particularly common among anti-spam software developers, and not widely known elsewhere; in general it is probably better to use the term "non-spam", instead.

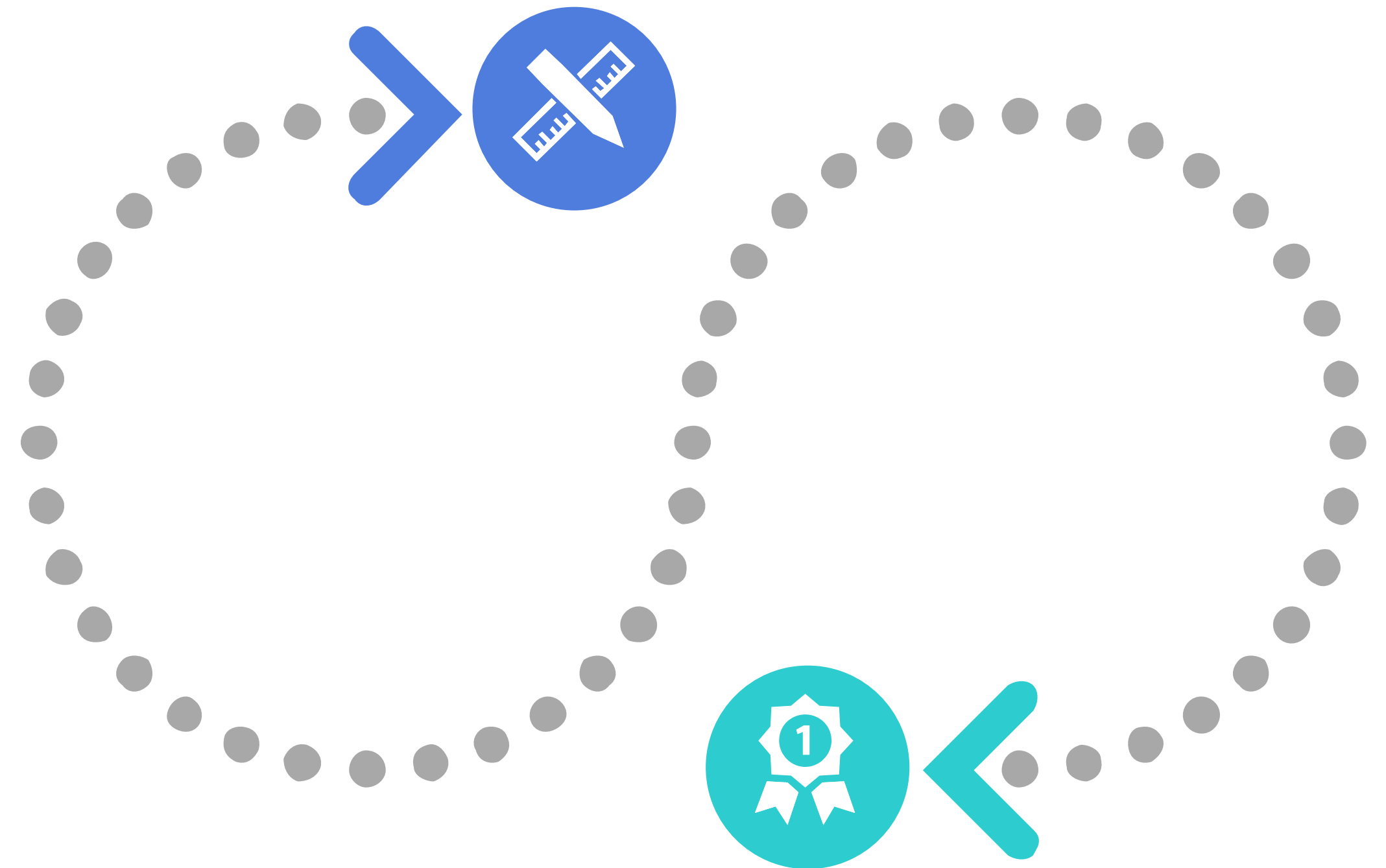
Our Vision

Make e-mails Spam Free

In this project, we aim to explore recurrent pattern detection system against image-based spam by using machine learning methods. We aim to find an intrinsic mechanism to match recurrent patterns across similar but not-identical images.

Contribute to the growing network of Spam Detections & Filtrations

A number of researches have been conducted in the area of classification. However, this research concentrates only on spam classification that has turned into a critical area of research in recent years. This study considers machine learning classifiers and possible ways of improving the learning capability of classifiers in a supervised learning environment.



Steps to Final Output



Collection of dataset which includes spam or ham emails.



Extract text from image using OCR (Optical Character Recognition).



Convert the words to lower case.



Converting the words to their root words.



A program maintaining two hash maps (string verses its number of occurrences in spam and ham).



Using Naive-Bayes Classifier, a Machine Learning Algorithm to predict possible outcomes of current input.

RELATED WORK



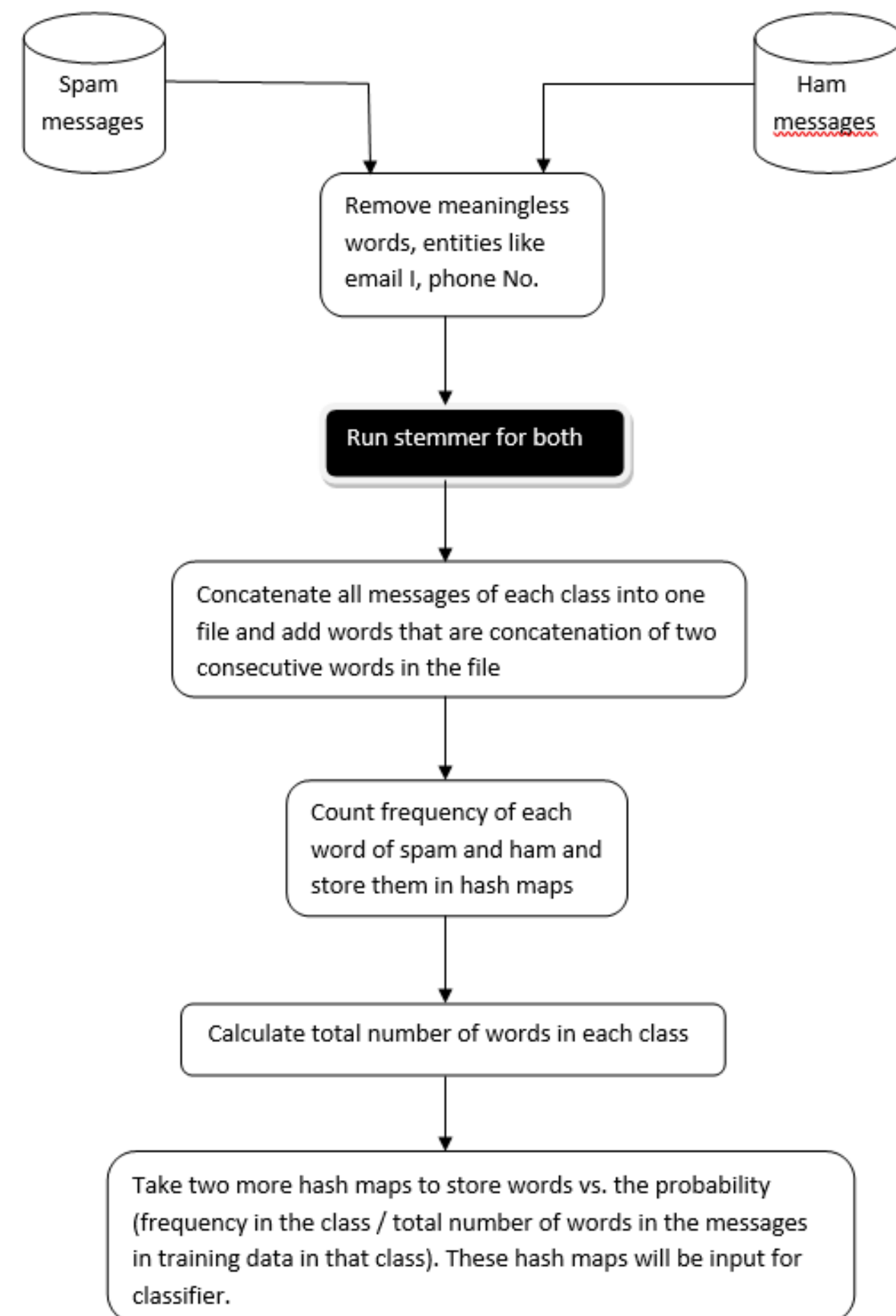
A number of researches have been conducted in the area of classification. However, this research concentrates only on spam classification that has turned into a critical area of research in recent years. This study considers machine learning classifiers and possible ways of improving the learning capability of classifiers in a supervised learning environment. A variety of research has been reported in literature where different classifiers have been tested on different email datasets.

The **Bayesian classifier** was introduced by Lewis (1998). This method involves traditional text mining approaches based on content and domain knowledge about the documents.

A common problem of misclassification seen in machine learning classifiers happens due to poor sampling of the dataset during training stage. Boosting algorithms attempt to solve this problem with the help of a voting mechanism and have therefore found a prominent place in the literature.

SVM has been identified to be a strong classifier and has a respectable place in classification literature. Drucker et al. (1999) have done a comparative study where the performance of SVM was compared with various machine learning classifiers. The result of this study was in favor of SVM and boosted decision trees in terms of accuracy and speed. However, the training time of boosted decision tree was longer than that of SVM.

PROPOSED METHADODOLOGY



Flowchart for trainer of proposed method

Idea behind Proposed Algorithm

The classifier tends to extract meaning from the message and judge whether the email message is legitimate or not.

The existing Naive-Bayes Classifier does not follow the context of the message, that is, the order of words doesn't change the meaning that classifier derives from the message, when it should.

So the idea is to add more contexts to the meaning, to influence the probabilities that each word carries with it for each class (ham and spam) should depend on the set of words before it.

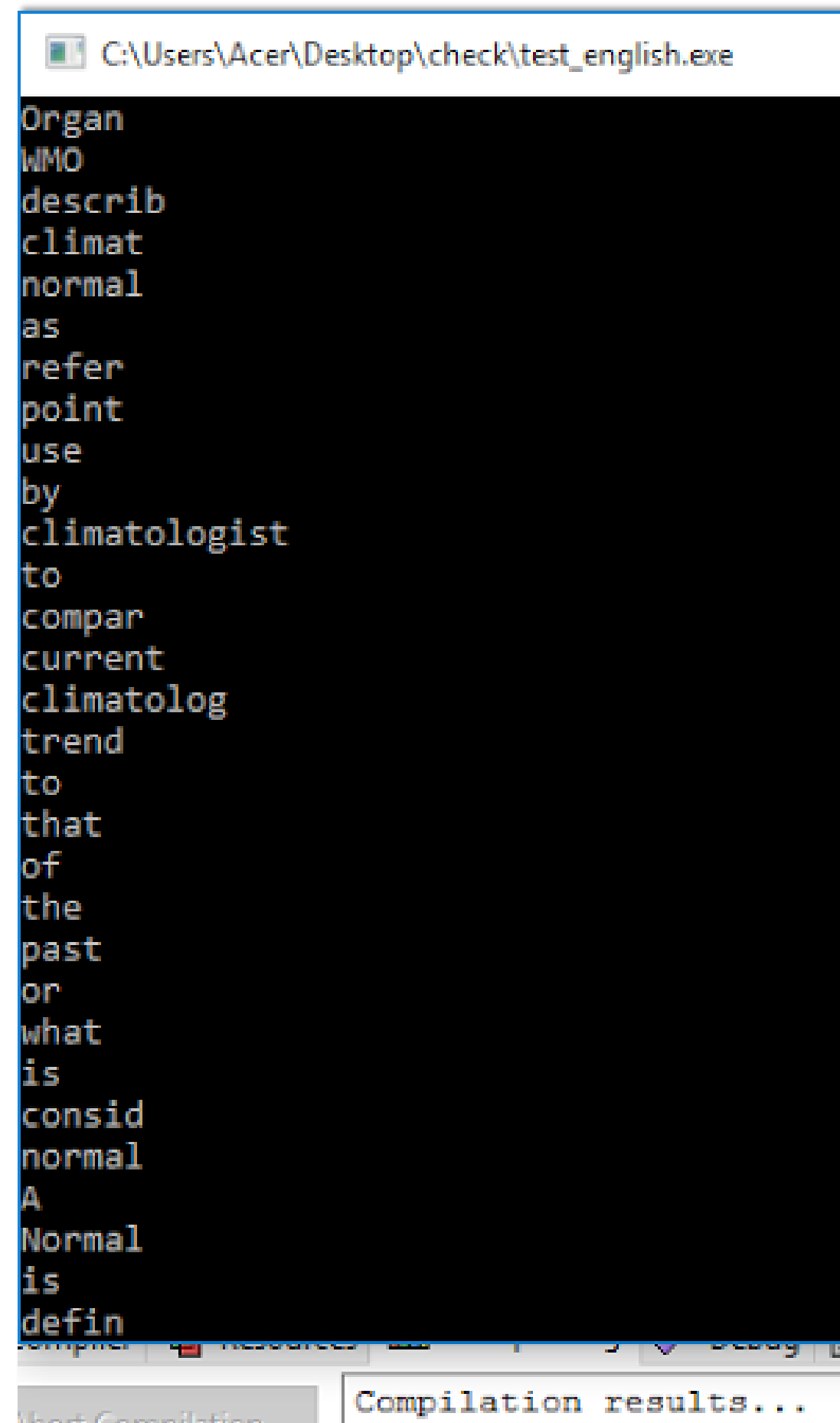
So we calculate the probabilities in linear combinations of size 'two' in the test message also. That is, we take two consecutive words at a time from the test message and calculate the probability that they contribute in the message being spam or ham.

Converting

```
\quantities over a period ranging from months to thousands or millions of years the classical period is  years as defined by the world meteorological organization wmo these quantities are most 2
\often surface variables such as temperature precipitation and wind climate in a wider sense is the state including a statistical description of the climate system the world meteorological 2
\organization wmo describes climate normals as reference points used by climatologists to compare current climatological trends to that of the past or what is considered normal a normal is 2
\defined as the arithmetic average of a climate element eg temperature over a year period a  year period is used as it is long enough to filter out any interannual variation or anomalies but also 2
\short enough to be able to show longer climatic trends the wmo originated from the international meteorological organization which set up a technical commission for climatology in  at its 2
\wiesbaden meeting the technical commission designated the thirtyyear period from  to  as the reference time frame for climatological standard normals in  the wmo agreed to update climate normals 2
\and these were subsequently completed on the basis of climate data from january  to december  the difference between climate and weather is usefully summarized by the popular phrase climate is 2
\what you expect weather is what you get over historical time spans there are a number of nearly constant variables that determine climate including latitude altitude proportion of land to water 2
\and proximity to oceans and mountains these change only over periods of millions of years due to processes such as plate tectonics other climate determinants are more dynamic the thermohaline 2
\circulation of the ocean leads to a  c  f warming of the northern atlantic ocean compared to other ocean basins other ocean currents redistribute heat between land and water on a more regional 2
\scale the density and type of vegetation coverage affects solar heat absorption water retention and rainfall on a regional level alterations in the quantity of atmospheric greenhouse gases 2
\determines the amount of solar energy retained by the planet leading to global warming or global cooling the variables which determine climate are numerous and the interactions complex but there 2
\is general agreement that the broad outlines are understood at least insofar as the determinants of historical climate change are concerned two levels of abstraction are employed in the 2
\definition of latitude and longitude in the first step the physical surface is modeled by the geoid a surface which approximates the mean sea level over the oceans and its continuation under the 2
```

- The convertor.cpp converts all the words to lower case.
- It takes both the Spam & Ham text files as input & converts them into lower case files upon output.

Find Root of Word



```
C:\Users\Acer\Desktop\check\test_english.exe
Organ
WMO
describ
climat
normal
as
refer
point
use
by
climatologist
to
compar
current
climatolog
trend
to
that
of
the
past
or
what
is
consid
normal
A
Normal
is
defin
Compilation results...
```

Output of test_english.cpp

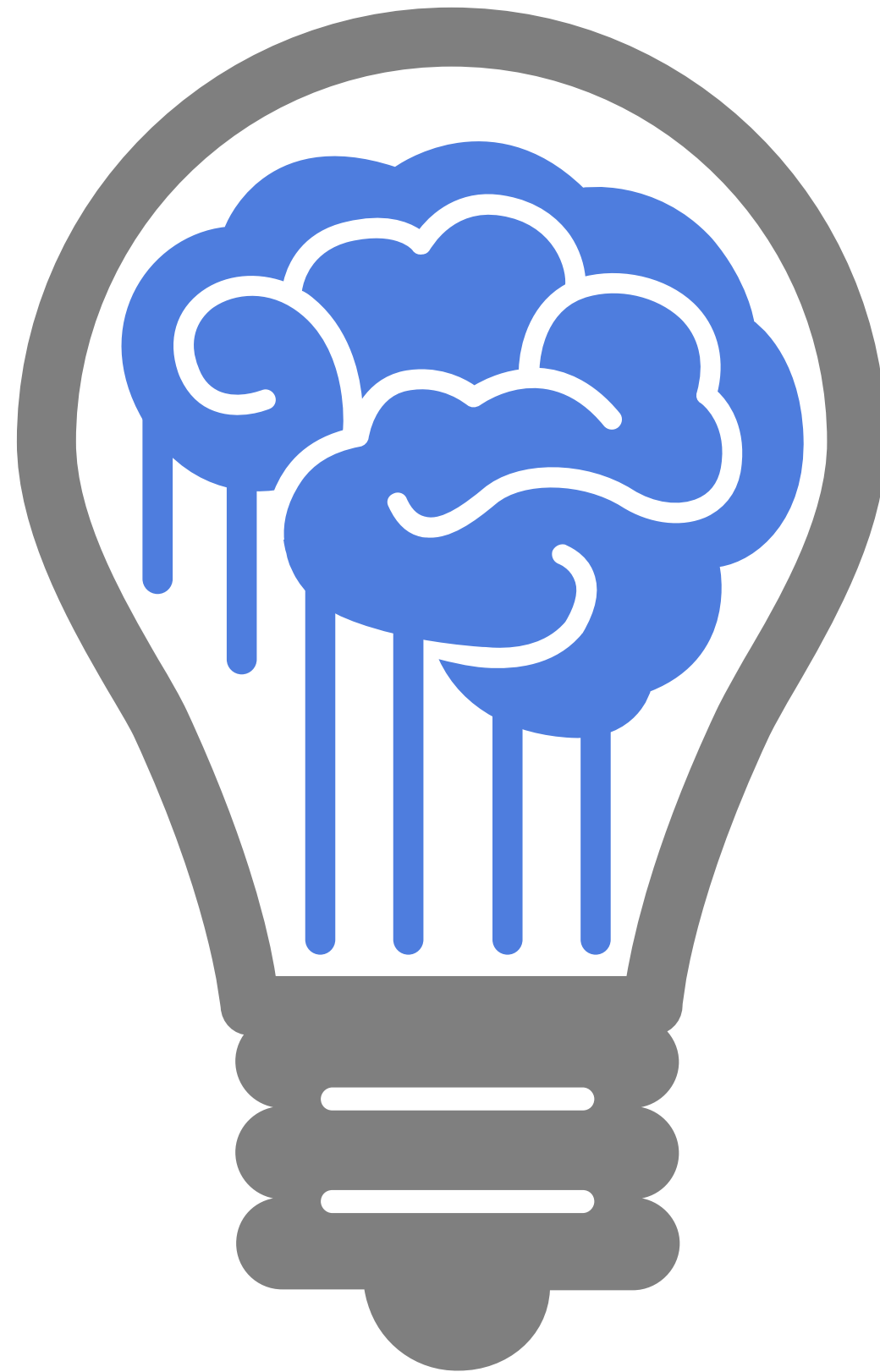
- We have written a code *test_english.cpp* where-in we transform the given word to their respective root word.
- The code includes a header file including "*english_stem.h*" which is one of the few languages that has the algorithm to convert the words to their root form. I.e, a word "*normals*" is automatically converted to "*normal*" and so on.
- The code takes the input from a pre-made file (soon to be replaced by OCR technology), and outputs them to "*ham_root.txt*" respectively.

Training Data

```
1 a 0.00956522
2 aarti 0.000434783
3 aartiand 0.000217391
4 aarticannot 0.000217391
5 ababy 0.000217391
6 able 0.000217391
7 ableto 0.000217391
8 about 0.00130435
9 aboutbusiness 0.000217391
10 abouther 0.000434783
11 abouthis 0.000434783
12 aboutthe 0.000217391
13 above 0.000217391
14 abovethe 0.000217391
15 absorption 0.000217391
16 absorptionwater 0.000217391
17 abstraction 0.000217391
18 abstractionare 0.000217391
19 abundance 0.000217391
20 abundanceof 0.000217391
21 ac 0.000217391
22 accepts 0.000434783
23 acceptshim 0.000217391
24 acceptstanya 0.000217391
25 accident 0.000434783
26 accidentand 0.000217391
27 accidentwhen 0.000217391
28 accomplice 0.000217391
29 accompliceof 0.000217391
30 according 0.000217391
31 accordingto 0.000217391
32 accuracy 0.000217391
33 accuracyis 0.000217391
34 accurate 0.000217391
35 accurateapplications 0.000217391
36 accurately 0.000217391
37 accuratelymodeled 0.000217391
38 achemical 0.000217391
39 aclimate 0.000217391
40 acoma 0.000217391
41 acompound 0.000217391
42 acontinuously 0.000217391
43 acquire 0.000217391
44 acquirethe 0.000217391
45 acrosssection 0.000217391
```

- A program maintaining two hash maps (string verses its number of occurrences in spam and ham).
- We take – in the file “*ham_converted_to_small.txt*” (from the output of the previous step) & calculate the trained probability of the spam and ham respectively.
- The output is saved to “*train_prob_spam*” & “*train_prob_ham*” respectively. The image shown in the output of trained probability of spam.

What's left to do ...



- ✓ **Classifier**

Implement the Naïve Bayes Classifier which is the next step in line.

- ✓ **Integrate OCR**

We have to integrate OCR in such a way that it automatically processes the image(s) downloaded from E-mail and detects whether it is a spam or ham.

- ✓ **Use more Algorithms**

If time permits, try to implement Multilayer Perceptron (MLP) or C4.5 Decision Tree Classifier which will make the whole project more efficient & reliable.



Thank You