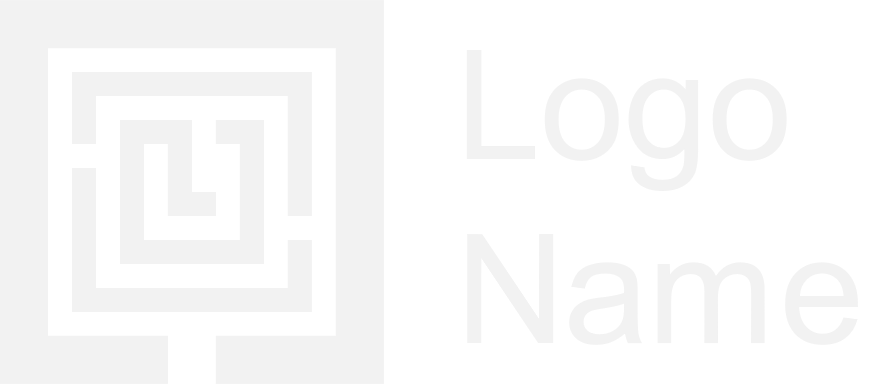


|  |
| --- |
| Email Spam Detection |
|  |
| JULY 15  Sir Mazhar Javed Awan  Authored by: Irtaza malik(F2019332006) BS-CDS |



Exploratory Data Analysis (EDA)

Exploratory Data Analysis is a very important process of data science. It helps the data scientist to understand the data at hand and relates it with the business context.

The open source tools that I will be using in visualizing and analyzing my data is Word Cloud.

Word Cloud is a data visualization tool used for representing text data. The size of the texts in the image represent the frequency or importance of the words in the training data.

Steps to take in this section:

1. Get the email data
2. Explore and analyze the data
3. Visualize the training data with Word Cloud & Bar Chart

Get the spam data

Data is the essential ingredients before we can develop any meaningful algorithm. Knowing where to get your data can be a very handy tool especially when you are just a beginner.

Below are a few of the famous repositories where you can easily get thousand kind of data set for free :

1. [UC Irvine Machine Learning Repository](https://archive.ics.uci.edu/ml/index.php)
2. [Kaggle datasets](https://www.kaggle.com/datasets)
3. [AWS datasets](https://registry.opendata.aws/)

For this email spamming data set, it is distributed by Spam Assassin, you can click this [link](https://spamassassin.apache.org/old/publiccorpus/) to go to the data set. There are a few categories of the data, you can read the *readme.html*to get more background information on the data.

In short, there is two types of data present in this repository, which is ham (non-spam) and spam data. Furthermore, in the ham data, there are easy and hard, which mean there is some non-spam data that has a very high similarity with spam data. This might pose a difficulty for our system to make a decision.

## Ham

This looks like a normal email reply to another person

## Spam

One of the spam training data does look like one of those spam advertisement email in our junk folder

# ****Train Test Split****

It is important to split your data set to**training set** and **test set**, so that you can evaluate the performance of your model using the test set before deploying it in a production environment.

One important thing to note when doing the train test split is to make sure the distribution of the data between the training set and testing set are similar.

What it means in this context is that the percentage of spam email in the training set and test set should be similar.

# Data Preprocessing

**Text Cleaning**

Text Cleaning is a very important step in machine learning because your data may contains a lot of noise and unwanted character such as punctuation, white space, numbers, hyperlink and etc.

Some standard procedures that people generally use are:

* convert all letters to lower/upper case
* removing numbers
* removing punctuation
* removing white spaces
* removing hyperlink
* removing stop words such as a, about, above, down, doing and the list goes on…
* **Word Stemming**
* **Word lemmatization**