

**PROJECT REPORT**

**ON**

**SPAM MAIL ANALYSIS**

**USING CLASSIFIERS**

GUIDED BY: DEVELOPED BY:

MR. CHANDRA PRAKASH SANT SATYARTH BHAI PATEL

**CERTIFICATE**

I have the pleasure in certifying that Satyarth Bhai Patel, a bonafied student

of DIT UNIVERSITY has undergone summer training towards the partial

fulfilment of B.Tech Degree, from 22/05/19 to 21/06/19 at Tel Bhavan,

ONGC, Dehradun.

He has successfully completed his project work entitled under my supervision.

He has taken keen interest in their training and project work. I wish him a

bright future.

Project Guide

Mr.Chandra Prakash Sant

Manager(Programming)

ECPF Trust.

Date:

**ACKNOWLEDGEMENT**

This project report is a sincere attempt to carefully and systematically gather facts about email ham and spam data.

A number of individuals have contributed to the manuscript of this report. I would like to express my sincere thanks and deep gratitude to each of them.

I am thankful to Mr. C D Singh allowing me to work in ONGC and giving valuable information about the organisation and helping me to learn about professional world.

I express my deep gratitude to guide Mr. Chandra Prakash Sant for his kind initiative ,guidance, precious time and valuable suggestion without which the completion of this would not have been possible.

**ABOUT ONGC**

**THE ORGANISATION**

**Oil and Natural Gas Corporation Limited (ONGC)** is an Indian multinational oil and gas company headquatered in Dehradun, India. It is a Public Sector Undertaking (PSU) of the government of India, under the administrative control of the Ministry of Petroleum and Natural Gas. It is India’s largest oil and gas exploration and production company. It produces 69% of India’s crude oil (equivalent to around 30% of the country’s total demand ) and around 62% of its natural gas.

On 31 March 2013 ,its market capitalisation was INR 2.6 trillion(US$48.98 billion),making it India’s second largest publicly traded company. In government survey for FY 2011-2012, it was ranked 357th in Fortune Global 500 list of the world’s biggest corporation for year 2012. It is ranked 22nd among the Top 250 Global Energy Companies by Platts.

ONGC was founded on 14 August 1956 by Government of India, which currently

holds a 68.94% equity stake. It is involved in exploring for and exploiting hydrocarbons in 26 sedimentary basins of India, and owns and operates over 11,000 kilometers of pipelines in the country. Its international subsidiary ONGC

videsh currently has projects in 15 countries. ONGC has discovered 6 of the 7 commercially producing Indian basins. In terms of growth, there was a palpable shift in gears in the second-half of 2016 with buoyancy in financial markets and a cyclical recovery underway in manufacturing and trade which supported the global GDP growth rate at 3.1% (same as 2015). Further, as per the latest World Economic Outlook of IMF, growth in oil and gas industry is projected to rise from 3.1% in 2016 to 3.5% in 2017 and 3.6% in 2018.

**PROFILE**

**Global Ranking**

* ONGC received Dun & Bradstreet Award 2018 in the 'Oil and Gas Exploration' category
* ONGC received 4 PSE Excellence Awards from Indian Chamber of Commerce in 2016
* This **Top Energy Company in India**, ranked 11th globally as per Platts Top 250 Global Energy Rankings, 2017
* Ranked 464 in the Newsweek Green Rankings World's Greenest Companies 2016
* Ranked 14th among global Oil and Gas Operations industry in Forbes Global 2000 list, 2017 of the World's biggest companies for 2017; Ranked 443 in the overall list, 2017 - based on Sales (US$ 19.89 billion), 288 on Profits, 470 in Assets and 300 Market Value.
* Ranked 26 in 'Transparency in Corporate Reporting' among the world's 124 largest listed companies published by Transparency International, 2014(Up from 39 in 2012)

**Our Growth Story**

ONGC was set up under the visionary leadership of Pandit Jawahar Lal Nehru. Pandit Nehru reposed faith in Shri Keshav Dev Malviya who laid the foundation of ONGC in the form of Oil and Gas division, under Geological Survey of India, in 1955. A few months later, it was converted into an Oil and Natural Gas Directorate. The Directorate was converted into Commission and christened Oil & Natural Gas Commission on 14th August 1956

Not only had India..set up her own machinery for oil exploration and exploitation... an efficient oil commission had been built where a large number of bright young men and women had been trained and they were doing good work" said Pandit JawaharLal Nehru, India's first Prime Minister to Lord Mountbatten, on ONGC in 1959.

In 1994, Oil and Natural Gas Commission was converted in to a Corporation, and in 1997 it was recognized as one of the Navratnas by the Government of India. Subsequently, it has been conferred with Maharatna status in the year 2010

In its 60 years of illustrious journey, ONGC has crossed many a milestone to realize the energy aspirations of India. The journey of ONGC, over these years, has been a tale of conviction, courage and commitment. ONGCs’ superlative efforts have resulted in converting earlier frontier areas into new hydrocarbon provinces. From a modest beginning, ONGC has grown to be one of the largest E&P companies in the world in terms of reserves and production.

**The Company’s Evolution can be summarized as under:**

1955 – Inception

1958 – First Oil in Cambay

1960 – Oil gas discovery in Gujarat

1963 – Oil in Assam

1965 – Concept of ONGC Videsh Operations

1970 – first Offshore well

1974 – Mumbai High discovered

1976 – Bassein Gas field of Mumbai High

1984 – GAIL formed out of ONGC

1993 – ONGC a limited company

1993 – Govt of India divest 2% share

1994 – 2% share to employees

1999 – Equity swap ONGC, IOC, GAIL

2003 – Acquired Mangalore Refineries Petrochemicals Ltd from Birla Group

2003 – Ist equity Oil & gas from Sudan / Vietnam

2004 – Govt of India divests 10%

2006 – Diversification – ONGC Petro additives Ltd and ONGC Mangalore Petro Ltd

2007 – ONGC Energy Centre formed

2010 – Coal Bed Methane Production

2013 – Oil at Kazakhstan/Mozambique

**Vision and Mission**

* To be global leader in integrated energy business through sustainable growth, knowledge excellence and exemplary governance practices.

World Class

* Dedicated to excellence by leveraging competitive advantages in R&D and technology with involved people.
* Imbibe high standards of business ethics and organizational values.
* Abiding commitment to safety, health and environment to enrich quality of community life.
* Foster a culture of trust, openness and mutual concern to make working a stimulating and challenging experience for our people.
* Strive for customer delight through quality products and services.
* Focus on domestic and international oil and gas exploration and production business opportunities.
* Provide value linkages in other sectors of energy business.
* Create growth opportunities and maximize shareholder value.
* Retain dominant position in Indian petroleum sector and enhance India's energy availability

**ABOUT ECPF TRUST**

The company has created a trust namely ,the ONGC Employees contributory provident Fund Trust for employees.

Employee’s Provident Fund Scheme takes care of following:

* Retirement
* Medical Care
* Housing
* Family Obligation
* Education Of Children
* Financing of insurance policies

All the employees(including casual, part time daily wage, contact etc.) other that c an excluded employee are required to be enrolled as members of the fund the day ,the Act comes into force in such establishment.

The employee’s provident Funds and Act.1952 Is enacted to provide a kind of social security to the workers. The security however differs from the security provided to them under the workmen’s compensation act or the employee’s state insurance Act .The employee’s provident funds and miscellaneous provision act mainly provide retirement or old age benefits, such as provident Fund ,pension, invalidation pension, family pension and deposit Linked insurance.

**PROJECT DESCRIPTION**

The notion of effective spam filtering has long since been a problem. The difficulty lies in the necessity of substantially small false rejection rates, as the misclassification of a valid email is rarely tolerable. Because senders of unsolicited email (“spammers”) very often disguise their messages to appear as valid correspondence, perfect filtering is impossible. Borrowing terminology from the field of biometrics, it would be possible to calculate the equal error rate of a given test set. From this value, we could produce a system that would be capable of balancing the false rejection rate with the false acceptance rate [1]. However, this approach will typically yield unacceptable results, as false rejections come as a much higher cost to the user - the result of missing an important email usually outweighs the improper acceptance of a particular spam message. Therefore, a virtually non-existent false rejection rate is necessary, so we instead focus on reducing the positive acceptance rate as much as possible while maintaining a minimal false rejection rate.

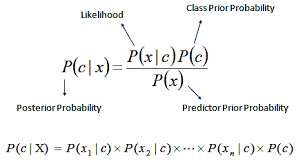
This project is developed using R language and Machine Learning Algorithm.

**METHODOLOGY**

* Collecting data.
* Exploring and preparing the data.
* Training a model on the data.
* Evaluating model performance.
* Improving model performance.

**Naïve Bayes**

It is a probabilistic method based on classification. Classification here refers to problem of identifying to which of categories a new observation belongs. It is based Bayes’ Theorem finds the probability of an event occurring given the probability of another event that has already occurred. Bayes’ theorem is stated mathematically as the following equation



Naïve Bayes Formula

**SOFTWARE PLATFORM USED**

**R-STUDIO** is a family of powerful, cost-effective disk recovery software. Originally developed by R-Tools Technology, Inc. for experienced data recovery professionals, R-Studio has been redesigned as a scalable, user-friendly all-in-one data recovery tool. By coupling our most advanced file recovery and disk repair technology with an intuitive user-interface, R-Studio provides enterprise and professional-level data recovery specialists the tools they need without hindering the experience of entry-level users

R-Studio is a cross-platform disk recovery suite with stable, actively supported releases for Windows, Mac, and Linux. R-Studio for Windows, R-Studio for Mac, and R-Studio for Linux each deliver the same powerful disk recovery tools and user-friendly interface on their respective platforms. Regardless of the host operating system, R-Studio can read, write, and recover drives, partitions, and files created by Windows, Mac, or Linux. For example, R-Studio for Windows can recover data from a Unix hard drive (be it an ext4 of Linux or UFS of BSD), R-Studio for Mac can recover data from a Windows hard drive (FAT/NTFS), R-Studio for Linux from a Mac hard drive (HFS+), and so on.

To see how R-Studio can recover data in a specific case, you may download R-Studio free, install it on your computer, and run it in the Demo mode (for a nonbootable machine, an R-Studio Emergency CD/DVD can be created). If you do not have hard drive recovery experience, we advise you to download and read our Data Recovery Manual before you start. You'll find step-by-step instructions and recommendations prepared by our data recovery specialists. For any additional questions you may always contact our technical support team even if you haven't purchased the software license yet. When lost files are found you may recover files with a size of less that 256KB each. Other files, supported by the pre-viewer built in the software, can be previewed to estimate chances for successful file recovery. If you are satisfied with the result, you may purchase an R-Studio license online right away. Upon receiving a registration key, you may register R-Studio on the fly without even closing the program. As soon as R-Studio is registered, you may continue recovering files.

Moreover, using R-Studio in the Demo mode you may create images of your logical disks or entire hard drives. Then you may perform all data recovery actions with those images to keep the source disks safe from accidental data corruption. These images are especially important when you are working with hard drives that start showing signs of dying, to prevent losing data from the drives for good.

**SOFTWARE CYCLE USED**

* Feasibility Study
* System Requirement Specifications
* Design
* Coding
* Testing
* Maintenance

**FEASIBILITY STUDY**

Feasibility studies aim to objectively and rationally uncover the strengths and weakness of an existing business or proposed venture, opportunities and threats present in environment, the resources required to carry through and ultimately the prospects for success. In simple terms the likelihood that the system will be useful to the organisation and it is an important outcome of the preliminary investigation. It is divided into following categories.

* **Economical Feasibility**
* **Technical Feasibility**
* **Operational Feasibility**

**ECONOMICAL FEASIBILITY:** The economical feasibility step of business development is that period during which break-even financial model of the business venture is developed based on all costs associated with taking the product from idea to market and achieving sales sufficient to satisfy debt or investment requirements. The purpose of the economic feasibility assessment is to determine the positive economic benefits to the organisation that the proposed System will provide. It includes qualification and identification of all the benefits expected. This assessment typically involves a cost/benefits analysis.

**Technical Feasibility**: It is focused on gaining an understanding of the present technical resources of the organization and their applicability to the expected needs of proposed system. It is based on the outline design of system requirements in terms of input, process , outputs, fields, programs, procedures. This can be quantified in terms of volumes of data, trends and frequencies of updating etc. In order to estimate whether the new system will perform adequately or not. It is an evaluation of the hardware and software and how it meets the need of the proposed system.

**OPERATIONAL FEASIBILITY**: it is a measure of how well a proposed system solves the problems and takes advantage of opportunities identified during scope definition and how it satisfies the requirement identified in the requirements analysis phase of system development. The operational feasibility assessment focuses on the degree to which the proposed development projects fits in with the existing business environment and objectives with regard to development schedule ,delivery date , corporate culture and existing business processes.

**SYSTEM REQUIREMENTS**

**SPECIFICATIONS**

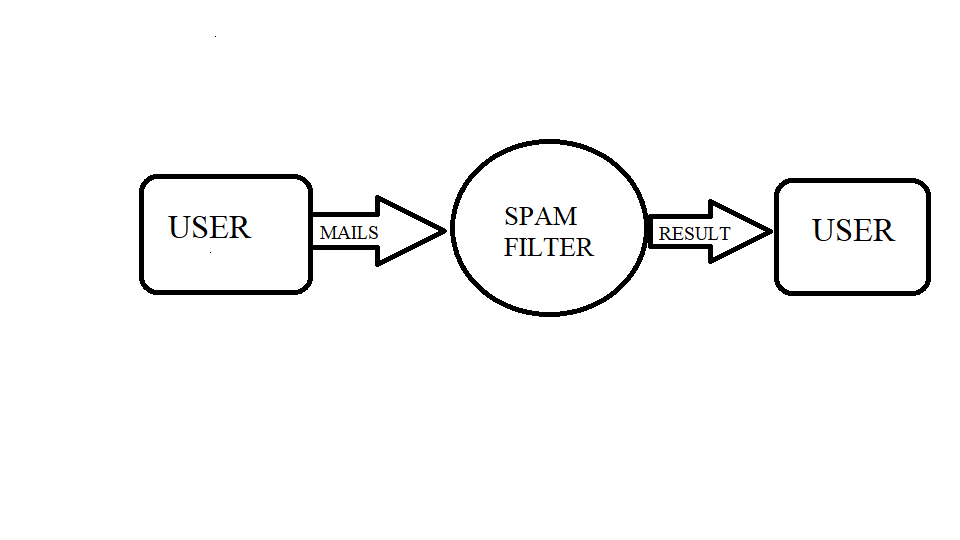
**Software Requirements:**

* Platform: cross platform
* Programming Language Used: R
* Software Used :R-Studio

**Hardware Requirements:**

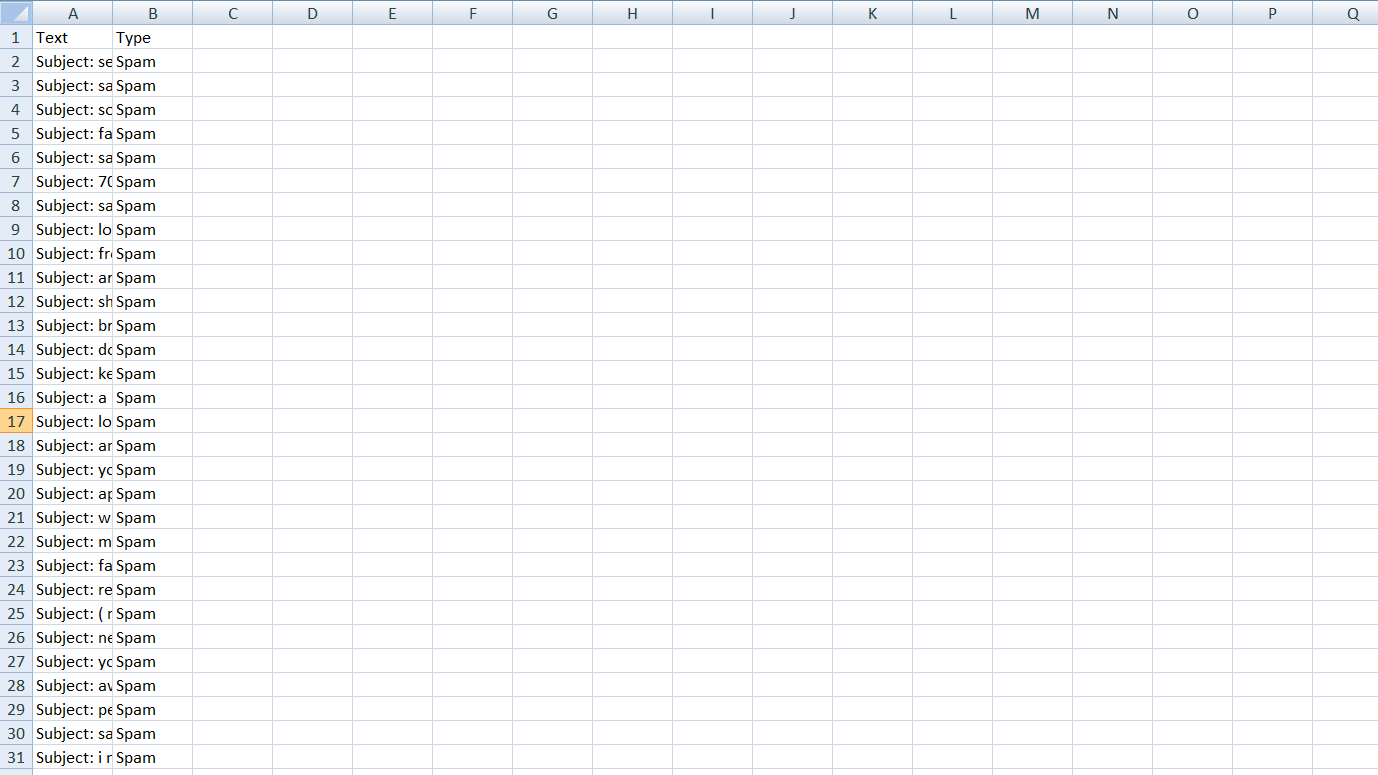
* Processor: Intel Core i5-6200u CPU @ 2.30GHz
* Installed RAM : 8.00 GB
* Graphics Video Card: 2GB

**DESIGN**



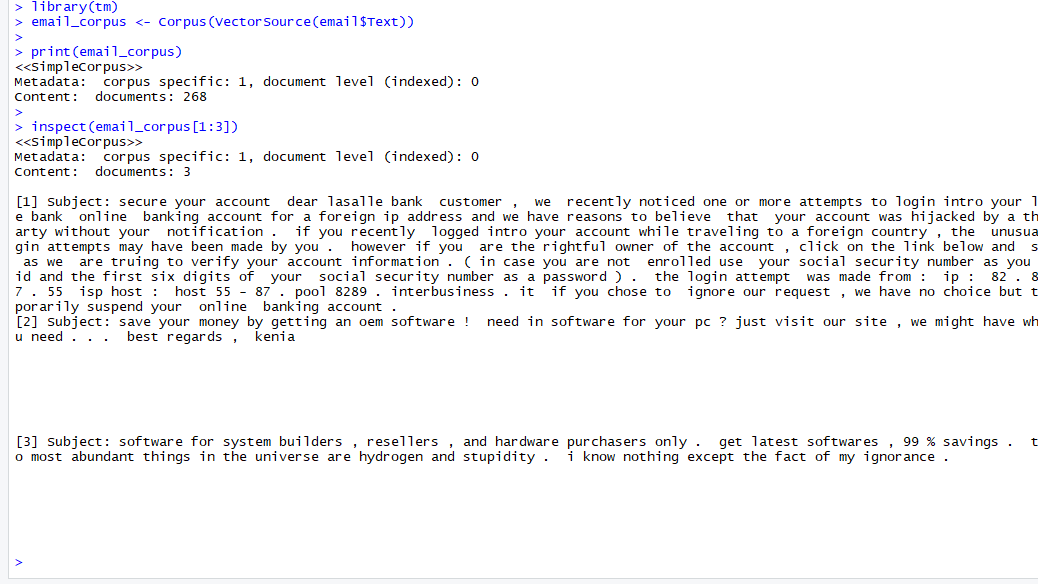
**DATA SAMPLE**

* Collecting the data

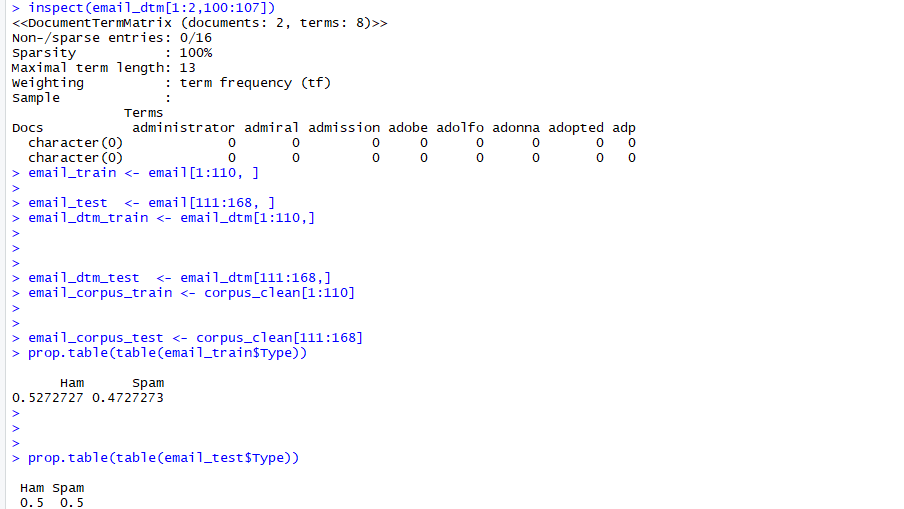


**READING DATA**

* Reading first three rows of data

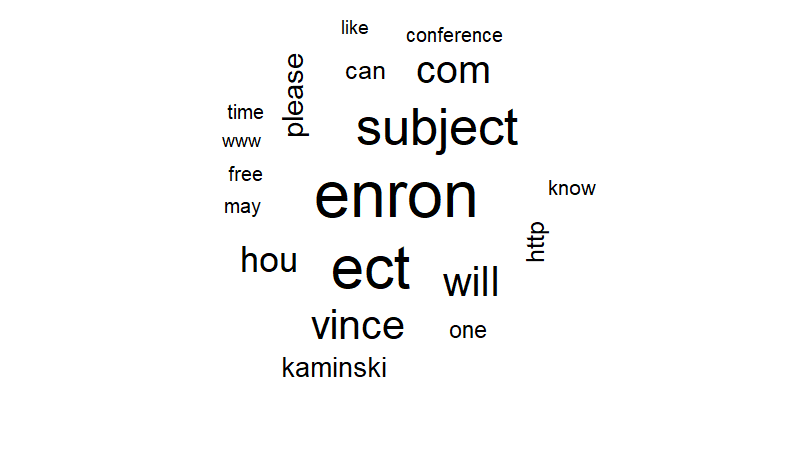


**SPARSE MATRIX**

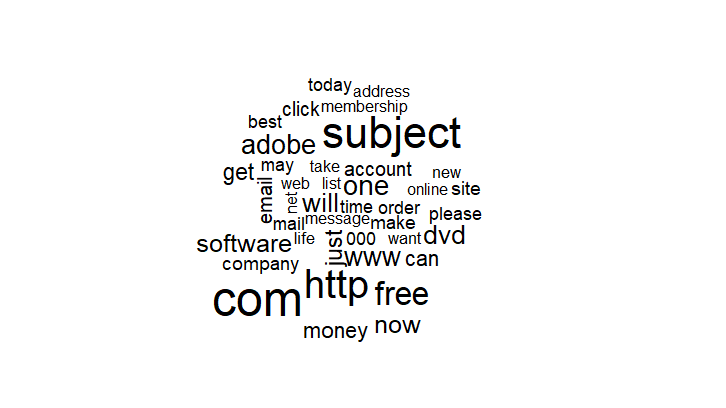


**REPEATED WORDS**

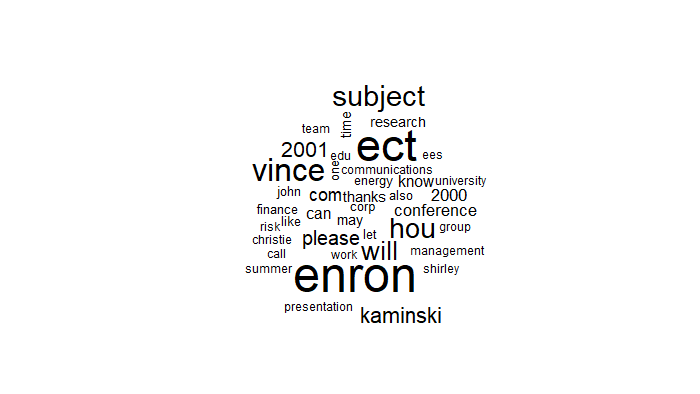
* Words having frequency 40 and order is random



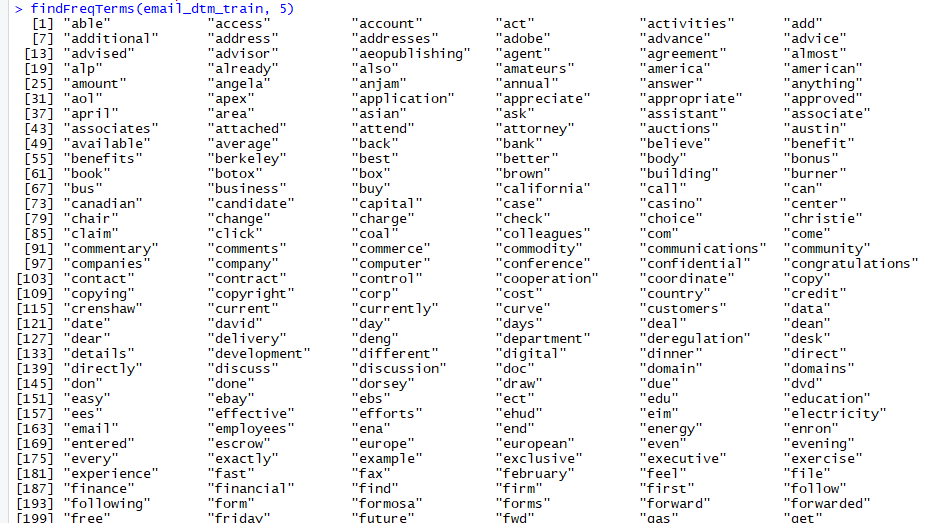
REPEATED WORDS IN SPAM



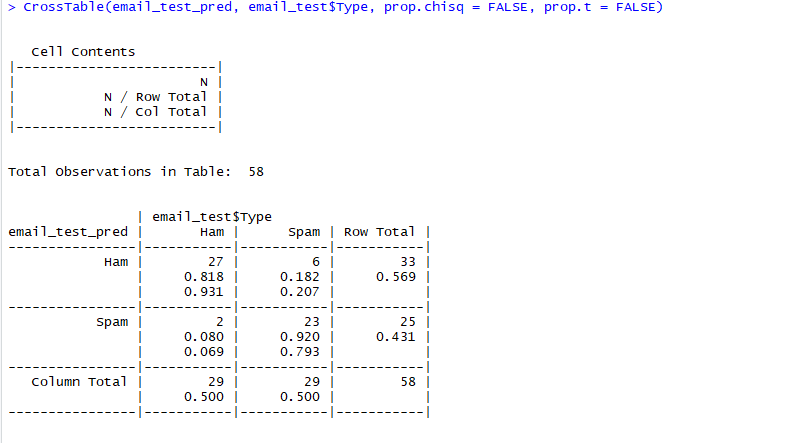
**REPEATED WORD IN HAM**



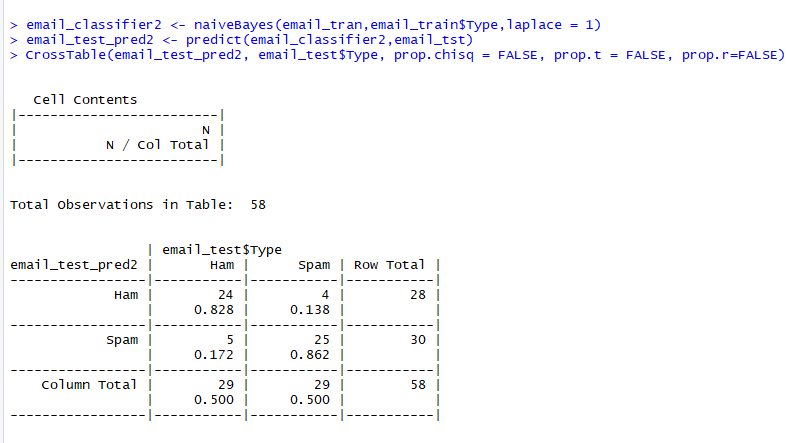
**FREQUENT TERMS**



**EVALUATING MODEL PERFORMANCE**



**IMPROVING MODEL PERFORMANCE**



**TESTING**

Testing is a phase where the errors remaining from all the previous phases must be detected.

Test techniques include the process of executing a program or application with the intent of finding software bugs (errors or other defects). As the number of possible tests for even simple software components is practically infinite, all software testing uses some strategy to select tests that are feasible for the available time and resources, as a result, software testing typically (but not exclusively) attempts to execute a program or application with the intent of finding software bugs ( errors or other defects). The job of testing is an iterative process as when one bug is fixed. It can illuminate other, deeper bugs, or can even create new ones.

Software testing can provide objective, independent information about the quality of software and risks of its failure to users or sponsors.

**Testing Levels**

There are generally four recognised level of tests:

* Unit Testing
* Integration Testing
* Component Interface Testing
* System Testing

**Unit Testing**

It is a level of software testing where individual units/ components of a software are tested. The purpose is to validate that each unit of the software performs as designed. A unit is the smallest testable part of any software. It usually has one or a few inputs and usually a single output. In procedural programming, a unit may be an individual program, function, procedure, etc. In object-oriented programming, the smallest unit is a method, which may belong to a base/ super class, abstract class or derived/ child class. (Some treat a module of an application as a unit. This is to be discouraged as there will probably be many individual units within that module.) Unit testing frameworks, drivers, stubs, and mock/ fake objects are used to assist in unit testing.

**Integration testing**

It is a level of software testing where individual units are combined and tested as a group. The purpose of this level of testing is to expose faults in the interaction between integrated units. Test drivers and test stubs are used to assist in Integration Testing.

**Component Interface Testing**

It is defined as a software testing type which verifies whether the communication between two different software systems is done correctly. A connection that integrates two components is called interface. This interface in a computer world could be anything like API's, web services, etc. Testing of these connecting services or interface is referred to as Interface Testing. An interface is actually software that consists of sets of commands, messages, and other attributes that enable communication between a device and a user.

**System Testing**

It is a level of software testing where a complete and integrated software is tested. The purpose of this test is to evaluate the system’s compliance with the specified requirements. The process of testing an integrated system to verify that it meets specified requirements.

**MAINTENANCE**

Once the software is delivered and deployed, it enters the maintenance phase. Software needs to be maintained not because of some of the components wear out but because there are often some errors remaining in the system that must be removed as they are discovered.

Software Maintaining Processes

This section describes software maintenance processes as:

1. Corrective maintenance:

Corrective maintenance of a software product may be essential either to rectify some bugs observed while the system is in use, or to enhance the performance of the system.

1. Adaptive maintenance:

This includes modifications and updations when the customers need the product to run on new platforms, on new operating systems, or when they need the product to interface with new hardware and software.

1. Perfective maintenance:

A software product needs maintenance to support the new features that the users want or to change different types of functionalities of the system according to the customer demands.

1. Preventive maintenance:

This type of maintenance includes modifications and updations to prevent future problems of the software. Its goal to attend problems, which are not significant at this moment but may cause serious issues in future.

**BIBLOIGRAPHY**

* About Company: https://www.ongcindia.com/wps/wcm/connect/en/about-ongc/ongc-at-a-glance/corporate-profile/
* Dataset: https://www.kaggle.com/uciml/sms-spam-collection-dataset
* Vision and Mission https://www.ongcindia.com/wps/wcm/connect/en/about-ongc/vision-and-mission/
* R-studio: https://www.r-studio.com/data-recovery-software/
* Testing: http://softwaretestingfundamentals.com/system-testing/
* Maintenance:https://www.geeksforgeeks.org/software-engineering-software-maintenance/