A

Project report on

**EXTRACTION OF DATA**

Submitted in partial fulfillment of the requirements

For the award of the Degree of

**BACHELOR OF TECHNOLOGY**

**IN**

**COMPUTER SCIENCE AND ENGINEERING**

Submitted By

G.SATYA PRIYA CHOWDARY 14B21A0567

T.LAKSHMI NARAYANAMA 14B21A0563

R.SAI KUMAR 14B21A0564

P.RAM NIKHIL 14B21A05B3

V.PRAVEEN 15B25A0525

**Under the esteemed guidance of**

**Mr. K. VIJAY KUMAR, B.Tech, M.Tech, (PH.D).**

**Assistant Professor**



**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

**KAKINADA INSTITUTE OF ENGINEERING & TECHNOLOGY**

**(Approved By AICTE & Govt. Of AP And Affiliated to JNTUK)**

**Yanam Road, KORANGI-533461**

**2014-2018**

**KAKINADA INSTITUTE OF ENGINEERING & TECHNOLOGY**

**(Approved By AICTE & Govt. Of AP And Affiliated to JNTUK)**

**Yanam Road, KORANGI-533461**

**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

**CERTIFICATE**



This is certify that this project work entitled “**EXTRACTION OF DATA**” that is being submitted by

G.SATYA PRIYA CHOWDARY 14B21A0567

T.LAKSHMI NARAYANAMA 14B21A0563

R.SAI KUMAR 14B21A0564

P.RAM NIKHIL 14B21A05B3

V.PRAVEEN 15B25A0525

In the partial fulfillment of the requirements for the award of the degree of Bachelor Of technology in “**Computer Science & Engineering**” in **Kakinada Institute of Engineering And Technology** affiliated to **Jawaharalal Nehru Technological University, Kakinada** is Bonafied work carried out by him/her under our guidance and supervision.

The result embodied in this project have not been submitted to any other University or Institute for the award of any degree or diploma.

INTERNAL EXAMINER EXTERNAL EXAMINER

HEAD OF THE DEPARTMENT

**ACKNOWLEDGEMENT**

We would like to take the privilege of the opportunity to express our gratitude into Project work of **"EXTRACTION OF DATA"** enabled us to express our special thanks to our honorable Chairman of the institution **Sri P.V.VISWAM.**

We are thankful to our honorable **Director Sri JOHN UDAY KUMAR** and **Principal Prof. C.SUBASH** **CHANDRA**, M.Tech(Ph.D) who has shown keen interest in us and encouraged us by providing all the facilities to complete our project successfully.

We owe our gratitude to our beloved Head of the Department **Mr.M.CHINNA RAO,** M.Tech(Ph.D) for assisting us in completing our project work.

We express our sincere thanks to my supervisor **Mr. K.VIJAY KUMAR** who have been a source of inspiration for us throughout our project and for their valuable advice in making our project a success.

We wish to express our sincere thanks to all teaching and non-teaching staff of Computer Science & Engineering. We wish to express our special thanks to all the Faculty members of our college for their concern in subjects and their help throughout our course.

We are very thankful to all our friends who gave us good co-operation and suggestions throughout this project and helped us in successful completion.

G.SATYA PRIY CHOWDARY 14B21A0567

T.LAKSHMI NARAYANAMA 14B21A0563

R.SAI KUMAR 14B21A0564

P.RAM NIKHIL 14B21A05B3

V.PRAVEEN 15B25A0525

CONTENTS

ABSTRACT

EXISTING

DRAW BACK

OVER COMING IN PROPOSED IN SYSTEM

LITERATURE SURVIVAL

LANGUAGE USED DESCRIPTION

EXRACTION OF DATA

OFFLINE

ONLINE

UML DIAGRAMS

HARDWARE

SOFTWARE

TESTING ABOUT DATA EXTRACTION

VERFICATION AND VALDIATION

CODE

INPUT AND OUTPUT

**ABSTRACT**

This project is about “EXTRACTION OF DATA” from offline and online. This project gives Information about huge amount of data given by the users in the world, Therefore we started to collect the required data from some social networking sites like (twitter, facebook). This is mainly used in Major sources of abundant data, Business: Web, e-commerce, transactions, stocks, Science: Remote sensing, bioinformatics, scientific simulation, Society and everyone: news, digital cameras are used. For gathering Information about a particular topic and much more. It is one of the basic methodology in data extraction. We can also analysis what’s app chat data. The Data will be extracted only in English language. Now it converts into global languages. The previous version has more number of lines of code. It’s just gives the data it doesn’t gives the user name, time, date

In global languages according to the user requirement. From what’s app chat data we can make sentiment analysis. We can make frequently messaged word from the chat data. We can make total count on sentiment. Date and time can be obtained from what’s app chat data. The mean of message length and sender details can be obtained. Plotting of data from particular date and time can be done.

**1 INTRODUCTION**

1.1 What is Data Extraction 1

1.2Applications of Extraction 2

1.3Advantages of Extraction 3

**2. LITERATURE SURVEY** 4

**3 SYSTEM ANALYSIS**

3.1 Existing System 6

3.2 Disadvantages of Existing System 7

3.3 Proposed System 8

3.4Advantages Of Proposed System 8

**4 SYSTEM REQUIREMENTS**

4.1 Hardware Requirements 9

4.2 Software Requirements 9

**5. STSTEM STUDY**

**5.1** Feasibility study 10

**6. SOFTWARE DESCRIPTION** 11

7. **SYSTEM DESIGN**

7,1 System architecture 25

7.2 Dataflow diagrams 26

**8 IMPLEMENTATION**

8.1 Modules 32

8.2 Modules description 32

8.3 Sample code 33

9 **SYSTEM TESTING**

9.1 Unit testing 44

9.2 Integration testing 44

9.3 Acception testing 45

9.4 Validation testing 45

10 **SCREEN SHOTS** 46

11 **CONCLUSION** 52

12 **BIBLOGRAPHY** 53

13 **PUBLICATION** 55

**CHAPTER 1**

**INTRODUCTION:**

* 1. **What is Data Extraction?**

Data extract is the output of the data extraction process, a very important aspect of data warehouse implementation.

A data warehouse gathers data from several sources and utilizes these data to serve as vital information for the company. These data will be used to spot patterns and trends both in the business operations as well as in industry standards.

Since the data coming to the data warehouse may come from different source which commonly are of disparate systems resulting in different data formats, a data warehouse uses three processes to make use of the data. These processes are extraction, transformation and loading (ETL).

The data extraction process in general is performed within the source system itself. This is can be most appropriate if the extraction is added to a relational database. Some database professionals implement data extraction using extraction logic in the data warehouse staging area and query the source system for data using applications programming interface (API).

At a specific point in time, only the data that has changed since a well-defined event back in history will be extracted. This event may be the last time of extraction or a more complex business event like the last booking day of a fiscal period. To identify this delta change there must be a possibility to identify all the changed information since this specific time event. This information can be either provided by the source data itself like an application column, reflecting the last-changed timestamp or a change table where an appropriate additional mechanism keeps track of the changes besides the originating transactions. In most cases, using the latter method means adding extraction logic to the source system.

Many data warehouses do not use any change-capture techniques as part of the extraction process. Instead, entire tables from the source systems are extracted to the data warehouse or staging area, and these tables are compared with a previous extract from the source system to identify the changed data. This approach may not have significant impact on the source systems, but it clearly can place a considerable burden on the data warehouse processes, particularly if the data volumes are large.

1.2 **Applications of Extraction**

Text mining (news group, email, documents) and Web mining, Stream data extraction.Bioinformatics and bio-data analysis.

Data collection and data availability.

Business: Web, e-commerce, transactions, stocks,

Science: Remote sensing, bioinformatics, scientific simulation.

Society and everyone: news, digital cameras .Automated analysis of massive data.

## ****Data Extraction Applications in Sales/Marketing****

## Data extraction enables businesses to understand the hidden patterns inside historical purchasing transaction data, thus helping in planning and launching new marketing campaigns in a prompt and cost-effective way. The following illustrates several data extraction applications in sale and marketing.

## ****Data Extraction Applications in Banking****

## Distributed data extraction have been researched, modeled and developed to help credit card fraud detection. Data extraction is used to identify customers loyalty by analyzing the data of customer’s purchasing activities such as the data of frequency of purchase in a period of time, a total monetary value of all purchases and when was the last purchase. After analyzing those dimensions, the relative measure is generate for each customer.

## ****Data Extraction Applications in Health Care and Insurance****

The growth of the insurance industry entirely depends on the ability to convert data into the knowledge, information or intelligence about customers, competitors, and its markets. Data extraction is applied in insurance industry lately but brought tremendous competitive advantages to the companies who have implemented it successfully. The data extraction applications in the insurance industry are listed below.Data extraction enables to forecasts which customers will potentially purchase new policies. Data extraction allows insurance companies to detect risky customers’ behavior patterns. Data extraction helps detect fraudulent behavior.

## ****Data extraction Applications in Transportation****

Data extraction helps determine the distribution schedules among warehouses and outlets and analyze loading patterns.

**1.3 ADVANTAGES OF EXTRACTION**

Data extraction is an important part of [knowledge discovery process](http://www.zentut.com/data-mining/data-mining-processes/) that we can analyze an enormous set of data and get hidden and useful knowledge. Data extraction is applied effectively not only in the business environment but also in other fields such as weather forecast, medicine, transportation, healthcare, insurance, government…etc. Data extraction has a lot of advantages when using in a specific industry. Besides those advantages, data extraction also has its own disadvantages e.g., privacy, security, and misuse of information. We will examine those in different industries in a greater detail

**Marketing / Retail**

Data extraction helps marketing companies build models based on historical data to predict who will respond to the new marketing campaigns such as direct mail, online marketing campaign…etc. Through the results, marketers will have an appropriate approach to selling profitable products to targeted customers.

Data extraction brings a lot of benefits to retail companies in the same way as marketing. Through market basket analysis, a store can have an appropriate production arrangement in a way that customers can buy frequent buying products together with pleasant. In addition, it also helps the retail companies offer certain discounts for particular products that will attract more customers.

**Finance / Banking**

Data extraction gives financial institutions information about loan information and credit reporting. By building a model from historical customer’s data, the bank, and financial institution can determine good and bad loans. In addition, data extraction helps banks detect fraudulent credit card transactions to protect credit card’s owner.

**Manufacturing**

By applying unknown rdata extraction in operational engineering data, manufacturers can detect faulty equipment and determine optimal control parameters. For example, semiconductor manufacturers have a challenge that even the conditions of manufacturing environments at different wafer production plants are similar, the quality of wafer are a lot the same and some for easons even has defects. Data extraction has been applying to determine the ranges of control parameters that lead to the production of the golden wafer. Then those optimal control parameters are used to manufacture wafers with desired quality.

**Governments**

Data extraction helps government agency by digging and analyzing records of the financial transaction to build patterns that can detect money laundering or criminal activities.

**Extraction of Data**

Data extraction is where data is analyzed and crawled through to retrieve relevant information from data sources (like a database) in a specific pattern. Further data processing is done, which involves adding metadata and other data integration; another process in the data workflow.The majority of data extraction comes from unstructured data sources and different data formats. This unstructured data can be in any form, such as tables, indexes, and analytics.

Data in a warehouse may come from different sources, a data warehouse requires three different methods to utilize the incoming data. These processes are known as Extraction, Transformation, and Loading (ETL).The process of data extraction involves retrieval of data from dishevelled data sources. The data extracts are then loaded into the staging area of the relational database. Here extraction logic is used and source system is queried for data using application programming interfaces. Following this process, the data is now ready to go through the transformation phase of the ETL process.

**CHAPTER 2**

**LITERATURE SURVEY**

**R Language**

The fact that R is a language may deter some users who think “I can’t program”. This should not be the case for two reasons. First, R is an interpreted language, not a compiled one, meaning that all commands typed on the keyboard are directly executed without requiring to build a complete program like in most computer languages (C, Fortran, Pascal, .). Second, R’s syntax is very simple and intuitive. For instance, a linear regression can be done with the command lm(y ~ x) which means “fitting a linear model with y as response and x as predictor”. In R, in order to be executed, a function always needs to be written with parentheses, even if there is nothing within them (e.g., ls()). If one just types the name of a function without parentheses, R will display the content of the function. In this document, the names of the functions are generally written with parentheses in order to distinguish them from other objects, unless the text indicates clearly so. When R is running, variables, data, functions, results, etc, are stored in the active memory of the computer in the form of objects which have a name. The user can do actions on these objects with operators (arithmetic, logical, Comparison, .) and functions (which are themselves objects). The use of operators is relatively intuitive, we will see the details later (p. 25). An R function may be sketched as follows:

Function

↑

Default arguments

Arguments −→ function

↑

=⇒result

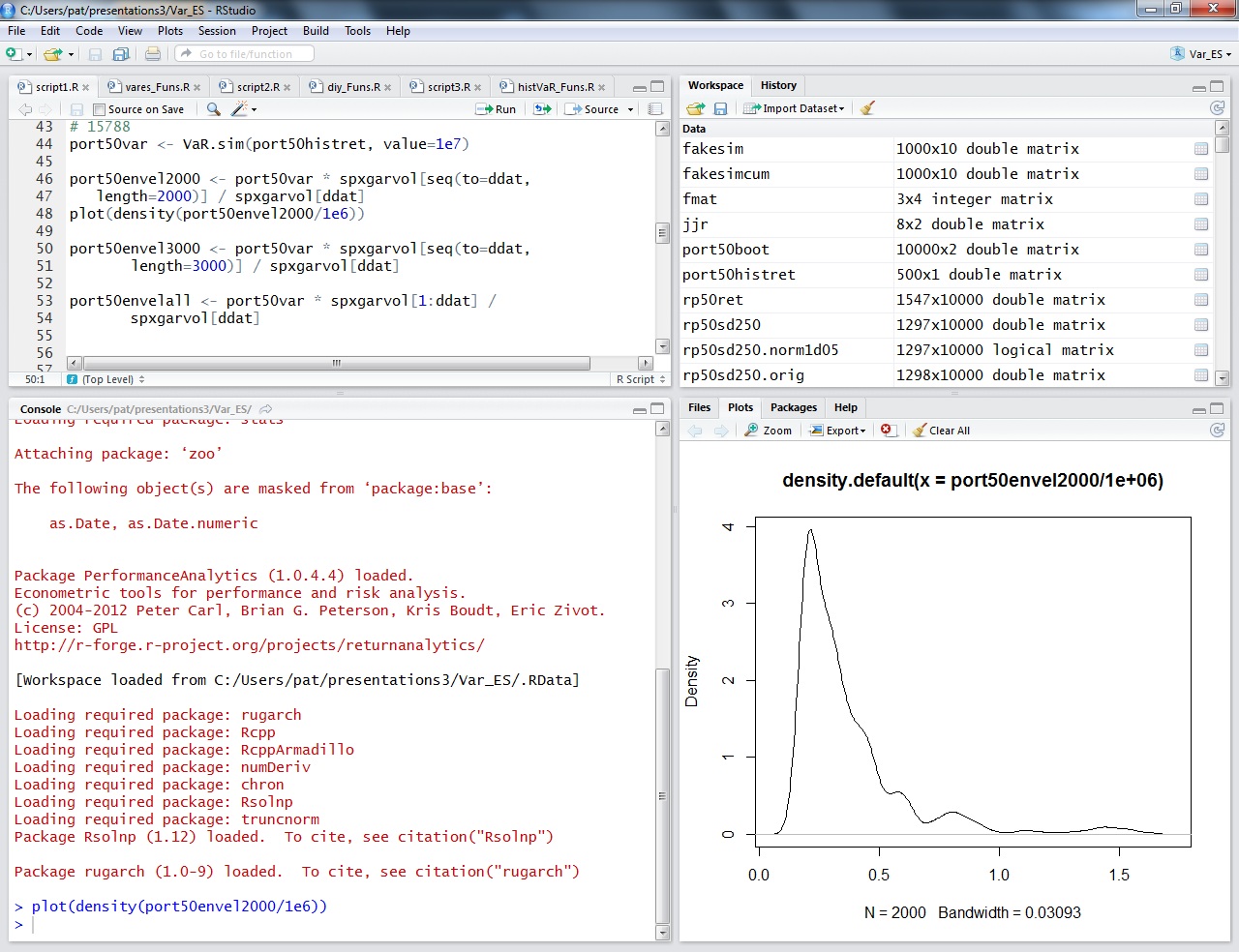
Options −→ default arguments

The arguments can be objects (“data”, formulae, expressions .) some 3 of which could be defined by default in the function; these default values may be modified by the user by specifying options. An R function may require no argument: either all arguments are defined by default(and their values can be modified with the options), or no argument has been defined in the function.

We will see later in more details how to use and build functions. The present description is sufficient for the moment to understand how R works. All the actions of Rare done on objects stored in the active memory of the computer: no temporary files are used. The readings and writings of files are used for input and output of data and results (graphics,). The user executes the functions via some commands. The results are displayed directly on the screen, stored in an object, or written on the disk (particularly for graphics). Since the results are themselves objects, they can be considered as data and analysed as such. Data files can be read from the local disk or from a remote server through internet.

**About R Studio**

R is a freely available environment for statistical computing. R works with a command-line interface, meaning you type in commands telling R what to do. R Studio is a convenient interface for using R, which can either be accessed online



The bottom left panel is the console. Here you can type code directly to be sent to R. The top left is called the R Script, and is basically a text editor that color codes for you and sends commands easily to R. Using a separate R script is nice because you can save only the code that works, making it easy to rerun and edit in the future, as opposed to the R console in which you would also have to save all your mistakes and all the output. We recommend always saving your R Scripts so you have the commands easily accessible and editable for future use. Code can be sent from the R Script to the console either by highlighting and clicking this icon: or else by typing CTRL+ENTER at the end of the line. Different R Scripts can be saved in different tabs. The top right is your Workspace and is where you will see objects (such as datasets and variables). Clicking on the name of a dataset in your workspace will bring up a spreadsheet of the data. The bottom right serves many purposes. It is where plots will be appear, where you manage your files (including importing files from your computer), where you install packages, and where the help information appears. Use the tabs to toggle back and forth between these screens as needed.

The R programming language is an [open source](http://whatis.techtarget.com/definition/open-source) [scripting language](http://searchwindevelopment.techtarget.com/definition/scripting-language) for [predictive analytics](http://searchbusinessanalytics.techtarget.com/definition/predictive-analytics) and data visualization.

The initial version of R was released in 1995 to allow academic statisticians and others with sophisticated programming skills to perform complex data [statistical analysis](http://whatis.techtarget.com/definition/statistical-analysis) and display the results in any of a multitude of visual graphics. The "R" name is derived from the first letter of the names of its two developers, Ross Ihaka and Robert Gentleman, who were associated with the University of Auckland at the time.

**Extraction of Data**

Data extraction is where data is analyzed and crawled through to retrieve relevant information from data sources (like a database) in a specific pattern. Further data processing is done, which involves adding metadata and other data integration; another process in the data workflow.

The majority of data extraction comes from unstructured data sources and different data formats. This unstructured data can be in any form, such as tables, indexes, and analytics.

Data in a warehouse may come from different sources, a data warehouse requires three different methods to utilize the incoming data. These processes are known as Extraction, Transformation, and Loading (ETL).

The process of data extraction involves retrieval of data from dishevelled data sources. The data extracts are then loaded into the staging area of the relational database. Here extraction logic is used and source system is queried for data using application programming interfaces. Following this process, the data is now ready to go through the transformation phase of the ETL process.

**CHAPTER 3**

**SYSTEM ANALYSIS**

**3.1 Existing System:**

Limited number of data will be extracted from social networking sites. The Data will be extracted only in English language. The previous version has more number of lines of code. Latterly this proposal was done, likely in Mid-winter season of 2015

**3.2 Disadvantages of existing system:**

Limited number of data can be extracted in previous versions

The extracted data will be in English language only.

It’s just gives the data it doesn’t gives the user name time,date.

**3.3 PROPOSED SYSTEM:**

In our project we have increased the size of extracting data nearly 1000.We are converting into global language. If we want particular data from any social networking sites it displays user name, time and data. Simple code we were use in this application.

**3.4 ADVANTAGES OF PROPOSED SYSTEM:**

In social media sentimental analysis like positive and negative expressions changes are going to be done.

Product ratings is also one of the best solution to data analysis.

Word count used to extract the common words from social networking sites.

**CHAPTER 4**

**SYSTEM REQUIREMENTS**

* 1. **Hardware Requirements:**
* Hard ware : At least 32 MB of RAM, a mouse, and enough disk space
* Processor : An Intel –core 2 duo and above.
* Operating systems : Windows 2000/XP/2003/Vista/7/8/10 2012 Server/8.1/10.
* Network : A network connection for data recovering over network.

* 1. **Software Requirements:**
  + R studio
  + twitter (package)
  + Rcurl (package)
  + Twitter api Keys
  + R studio Supported OS(Windows, Linux, Macintosh, etc;)

**CHAPTER 5**

**SYSTEM STUDY**

**5.1 Feasibility study**

As the name implies, a feasibility study is used to determine the viability of an idea, such as ensuring a project is legally and technically feasible as well as economically justifiable. It tells us whether a project is worth the investment—in some cases, a project may not be doable. There can be many reasons for this, including requiring too many resources, which not only prevents those resources from performing other tasks but also may cost more than an organization would earn back by taking on a project that isn’t profitable.

**What is Extraction?**

Data extract is the output of the data extraction process, a very important aspect of warehouse implementation.

A data warehouse gathers data from several sources and utilizes these data to serve as vital information for the company. These data will be used to spot patterns and trends both in the business operations as well as in industry standards

**What is Twitter data?**

Twitter is on online news and social networking site where people communicate in short messages called tweets. Twitter is sending short messages to anyone who follows you on Twitter, with the hope that your messages are useful and interesting to someone in your audience. Another description of Twitter and tweeting might be microblogging. Some people also use Twitter to discover interesting people and companies online and to follow their tweets for as long as they are interesting.

**What is what’s app chat text?**

What’s app is day to day application it is individual or group text. Using what’s app chat text we can use sentimental analysis and text cleaning.

**CHAPTER 6**

**SOFTWARE DESCRIPTION**

**R Language**

R is an open source programming language and software environment, commonly used for statistical computing within data heavy roles such as data mining and statistics. R has had a resurgence in recent years with a growing number of programmers using its data generation and analysis capabilities within machine learning and other emerging data-dependant technologies. R is an integrated suite of software facilities for data manipulation, calculation and graphical display.

• An effective data handling and storage facility

• A suite of operators for calculations on arrays, in particular matrices,

• A large, coherent, integrated collection of intermediate tools for data analysis,

• graphical facilities for data analysis and display either directly at the computer or on hardcopy, and a well developed, simple and effective programming language (called ‘S’) which includes conditionals, loops, user defined recursive functions and input and output facilities. (Indeed most of the system supplied functions are themselves written in the S language.) The term “environment” is intended to characterize it as a fully planned and coherent system, rather than an incremental accretion of very specific and inflexible tools, as is frequently the case with other data analysis software. R is very much a vehicle for newly developing methods of interactive data analysis. It has developed rapidly, and has been extended by a large collection of packages. However, most programs written in R are essentially ephemeral, written for a single piece of data analysis

**About R packages**

**Rcurl**

The curl package provides bindings to the [libcurl](http://curl.haxx.se/libcurl/) C library for R. The package supports retrieving data in-memory, downloading to disk, or streaming using the R “connection” interface. Some knowledge of curl is recommended to use this package. For a more user-friendly HTTP client, have a look at the [httr](https://cran.r-project.org/package=httr/vignettes/quickstart.html) package which builds on curl with HTTP specific tools and logic.

**TwitteR**

twitteR is an R package which provides access to the Twitter API. Most functionality of the API is supported, with a bias towards API calls that are more useful in data analysis as opposed to daily interaction.

**ROAuth**

Provides an interface to the OAuth 1.0 specification allowing users to authenticate via OAuth to the server of their choice.

**About Whatsapp**

The means of communication has changed over time according to the situation and advancements in technology. The process of transferring data from one individual to another such as audio, video and images have grown beyond texting and evolved to enable the transmission of media not only between two individuals but also in a group where huge number of people can interact and have a talent to connect worldwide. What’s App is such an application which is used widely for transferring media, text, files as well as audio calling. This research paper predicts the level of addiction of an individual to the What’s App group as per the age group and gender with the help of R statistics software programme.

**Ggplot2**

ggplot2 is the most elegant and aesthetically pleasing graphics framework available in R. It has a nicely planned structure to it. This tutorial focusses on exposing this underlying structure you can use to make any ggplot. But, the way you make plots in ggplot2 is very different from base graphics making the learning curve steep. So leave what you know about base graphics behind and follow along. You are just 5 steps away from cracking the ggplot puzzle.

## Topics

1. The Setup
2. The Layers
3. The Labels
4. The Theme
5. The Facets
6. Commonly Used Features

**Lubri date**

Functions to work with date-times and time-spans: fast and user friendly parsing of date-time data, extraction and updating of components of a date-time (years, months, days, hours, minutes, and seconds), algebraic manipulation on date-time and time-span objects. The 'lubridate' package has a consistent and memorable syntax that makes working with dates easy and fun. Parts of the 'CCTZ' source code, released under the Apache 2.0 License, are included in this package. See <https://github.com/google/cctz> for more details.

**Scale**

The value of **scale** determines how column **scaling** is performed (after centering). If**scale** is (alike) a numeric vector with length equal to the number of columns of x , then each column of x is divided by the corresponding value from **scale** . ... If **scale**is FALSE , no **scaling** is done.

**Reshape2**

Reshape2 is a reboot of the reshape package. It's been over five years since the first release of reshape, and in that time I've learned a tremendous amount about R programming, and how to work with data in R. Reshape2 uses that knowledge to make a new package for reshaping data that is much more focused and much much faster.

This version improves speed at the cost of functionality, so I have renamed it to reshape2 to avoid causing problems for existing users. Based on user feedback I may reintroduce some of these features.

* considerably faster and more memory efficient thanks to a much better underlying algorithm that uses the power and speed of subsetting to the fullest extent, in most cases only making a single copy of the data.
* cast is replaced by two functions depending on the output type: dcast produces data frames, and acast produces matrices/arrays.
* multidimensional margins are now possible: grand\_row and grand\_col have been dropped: now the name of the margin refers to the variable that has its value set to (all).
* some features have been removed such as the | cast operator, and the ability to return multiple values from an aggregation function. I'm reasonably sure both these operations are better performed by plyr.
* a new cast syntax which allows you to reshape based on functions of variables (based on the same underlying syntax as plyr):
* better development practices like namespaces and tests.
* the function melt now names the columns of its returned data frame Var1, Var2, ...,VarN instead of X1, X2, ..., XN.
* the argument variable.name of melt replaces the old argument variable\_name.

Initial benchmarking has shown melt to be up to 10x faster, pure reshaping cast up to 100x faster, and aggregating cast() up to 10x faste

**Tm**

This vignette gives a short introduction to text mining in R utilizing the text mining framework provided by the tm package. We present methods for data import, corpus handling, preprocessing, metadata management, and creation of term-document matrices. Our focus is on the main aspects of getting started with text mining in R—an in-depth description of the text mining infrastructure offered by tm was published in the Journal of Statistical Software (,Feinereretal., 2008). An introductory article on text mining in R was published in R News (Feinerer,2008).

**Snowball**

An R interface to the C libstemmer library that implements Porter's word stemming algorithm for collapsing words to a common root to aid comparison of vocabulary. Currently supported languages are Danish, Dutch, English, Finnish, French, German, Hungarian, Italian, Norwegian, Portuguese, Romanian, Russian, Spanish, Swedish and Turkish.

**Word cloud**

The procedure of creating word clouds is very simple in **R** if you know the different steps to execute. The text mining package (tm) and the **word cloud** generator package (**wordcloud**) are available in **R** for helping us to analyze texts and to quickly visualize the keywords as a **word cloud**.

**Rcolorbrewer**

Provides color schemes for maps (and other graphics) designed by Cynthia Brewer

**Stringr**

There are four main families of functions in stringr:

1. Character manipulation: these functions allow you to manipulate individual characters within the strings in character vectors.
2. Whitespace tools to add, remove, and manipulate whitespace.
3. Locale sensitive operations whose operations will vary from locale to locale.
4. Pattern matching functions. These recognise four engines of pattern description. The most common is regular expressions, but there are three other tools.

**Syuzhet**

This vignette demonstrates use of the basic functions of the Syuzhet package. The

package comes with four sentiment dictionaries and provides a method for accessing

robust, but computationally expensive, sentiment extraction tool developed in the NLP

group at Stanford. Use of this later method requires that you have already installed the

coreNLP package

The goal of this vignette is to introduce the main functions in the package so that you

can quickly extract plot and sentiment data from your own text files.This document will use a short example passage to demonstrate the functions and the various ways that

the extracted data can be returned and or visualized

**dyplr**

When working with data you must:

* Figure out what you want to do.
* Describe those tasks in the form of a computer program.
* Execute the program.

The dplyr package makes these steps fast and easy:

* By constraining your options, it helps you think about your data manipulation challenges.
* It provides simple “verbs”, functions that correspond to the most common data manipulation tasks, to help you translate your thoughts into code.
* It uses efficient backends, so you spend less time waiting for the computer.

This document introduces you to dplyr’s basic set of tools, and shows you how to apply them to data frames. dplyr also supports databases via the dbplyr package, once you’ve installed, read vignette("dbplyr") to learn more.

**CHAPTER 7**

**System Design**

**7.1 System architecture**

**7.2 Data flow diagram:**

**Uml diagrams**

UML stands for Unified Modeling Language. UML is a standardized general-purpose modeling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object Management Group. The goal is for UML to become a common language for creating models of object oriented computer software. In its current form UML is comprised of two major components: a Meta-model and a notation. In the future, some form of method or process may also be added to; or associated with, UML. The Unified Modeling Language is a standard language for specifying, Visualization, Constructing and documenting the artifacts of software system, as well as for business modeling and other non-software systems



. The UML represents a collection of best engineering practices that have proven successful in the modeling of large and complex systems. The UML is a very important part of developing objects oriented software and the software development process. The UML uses mostly graphical notations to express the design of software projects.



**CHAPTER 8**

**Implementation**

**8.1 Modules:**

**Online**

Here the data is extracted directly from the Source for processing in the staging area, that’s why it’s called online extraction. During Extraction we connect directly to the source system and then access the source tables. There is no need of any external staging.

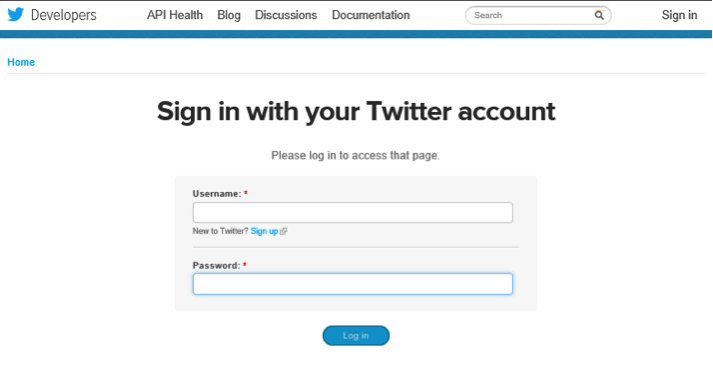


sign in with your twitter account.

1. Enter your **full name**, **phone number** or**email address**,**date of birth**, and a **password**.
2. Click **Sign up for Twitter**.

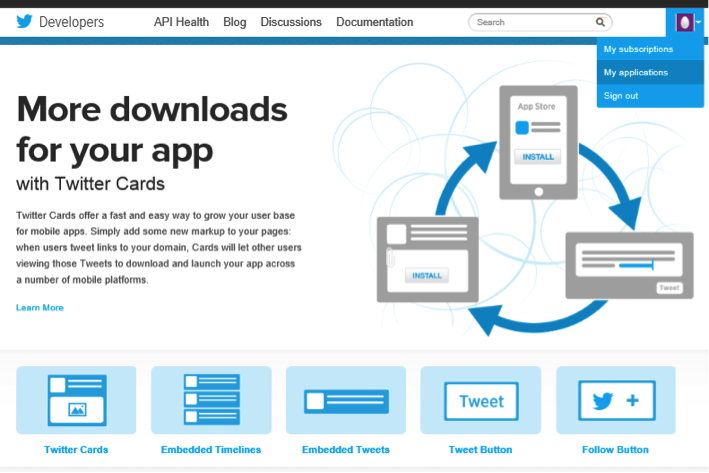
In order to verify your phone number, we will send you an SMS text message with a code. You may also request a voice call to verify your phone number. Enter the verification code in the box provided. Learn more about having a phone number is associated.

1. Once you've clicked **Sign up for Twitter,** you can select a **username** (usernames are unique identifiers on Twitter) — type your own or choose one we've suggested. We'll tell you if the username you want is available.
2. **Double-check** your name, phone number, password, and username.
3. Click **Create my account**. You may be asked to complete a Captcha to let us know that you're human.



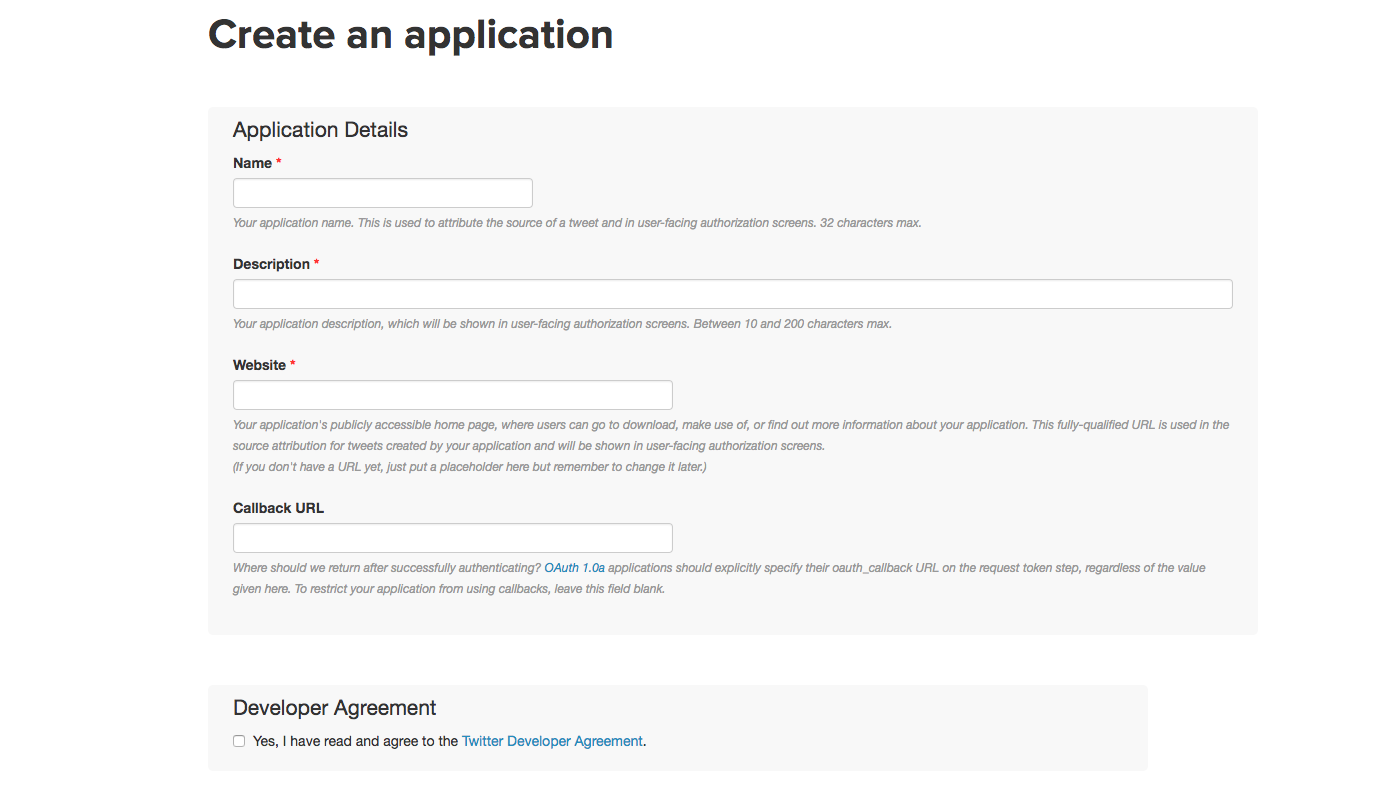
**Downloads for your app with twitter cards**

Now you can see your Profile picture in the upper right corner and a drop-down menu. In this menu you can find “**My Applications**”. Navigate to “**My Applications**” in the upper right hand corner.



**Create an application**:

Give your application a name. Give a description for application in few words, provide your website’s URL or your blog address . Leave the Callback URL blank for now. Complete other formalities and create your twitter application.

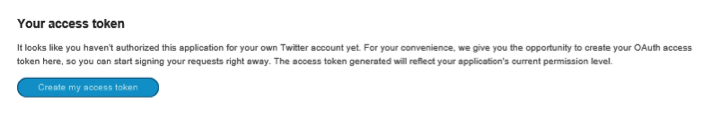


**Scroll down and click on “Create my access token” button:**

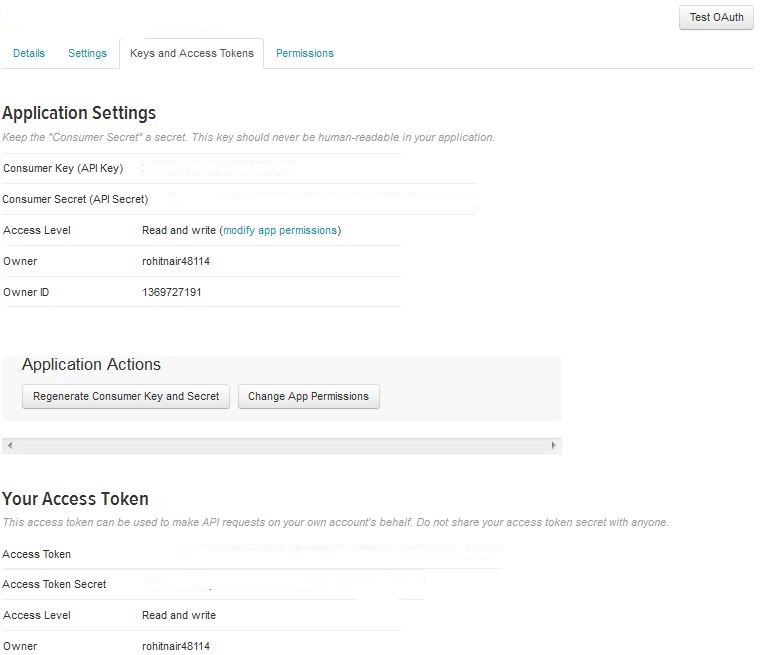
I have removed the Access Token and Consumer keys from the above image which were I provided by twitter for Security purpose of my app..

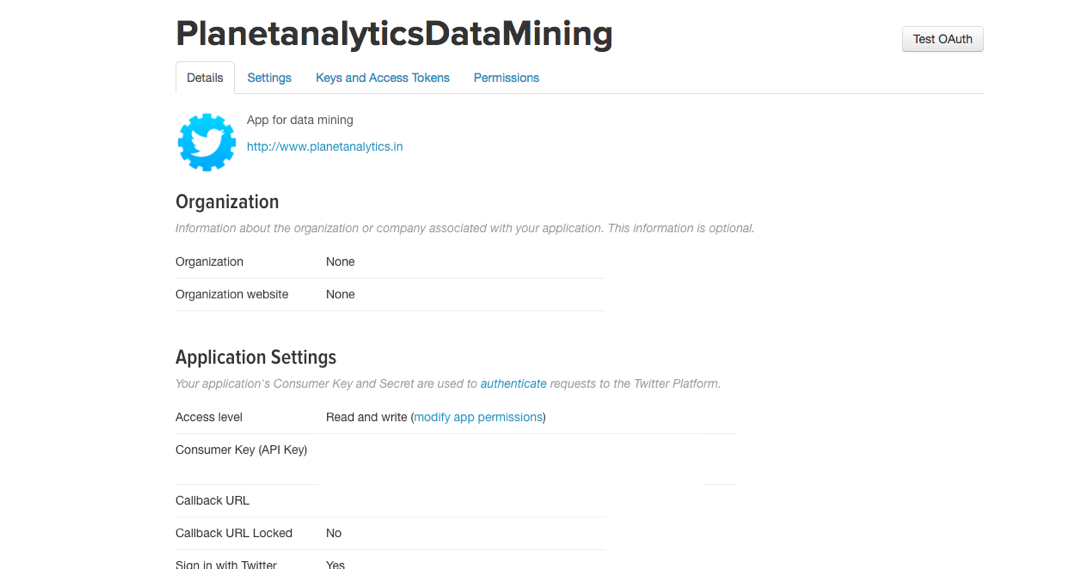
· Please note the Consumer key , Consumer Secret, Access Token and Access Token Secret numbers as they will be used in R later.

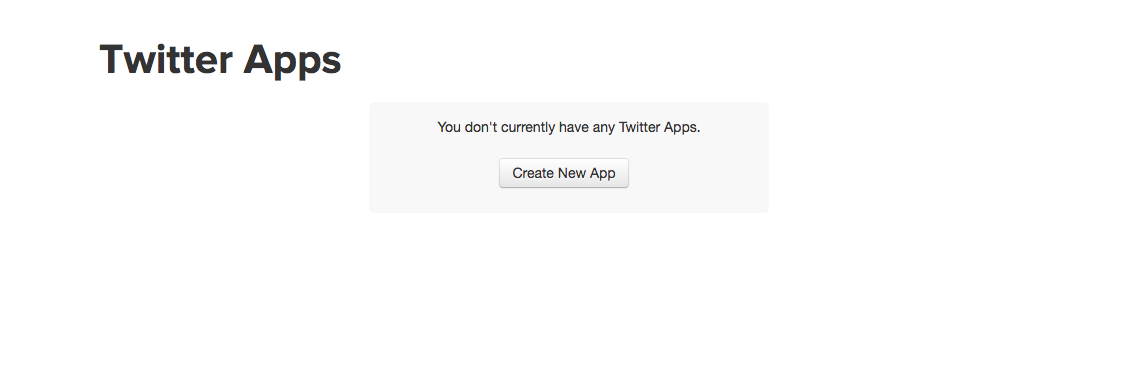
Once the Twitter Application is ready we can now move forward towards programming in R to extract data from Twitter.



Once, all the steps are done, the created application will show as below.







Twitter is a free social networking microblogging service that allows registered members to broadcast short posts called tweets. ... But unlike IMs that disappear when the user closes the application, tweets are also posted on the Twitter website.

**Offline:**

Here the data is not extracted directly from the source, but instead it’s taken from another external area which keeps the copy of source. The external area can be Flat files, or some dump files in a specific format. So when we need to process the datawe can fetch the records from the external source instead of the actual source

**8.2 Modules description**

.

**8.3 Sample code for twitter**

library(twitteR)

library(ROAuth)

library(RCurl)

shinyApp(

ui = fluidPage(

fluidRow(

column(4, textInput("searchkw", label = "search:", value = "#")),

column(4, textInput("num", label = "number of tweets:", value = 10)),

column(12, tableOutput('table'))

)

),

server = function(input, output) {

# OAuth authentication

consumer\_key <- "ezLejWfArzFgV6446tLlj5FUp"

consumer\_secret <- "jFV3R40FR7y7d0Z7ZqNH14jFbQN7Pn3ux5p9Fp17c4qEsMgRti"

access\_token <- "908658144699572225-10BHbiNw0yYWeSeNp6Qz7vbvjkxq4r7"

access\_secret <- "Ej35Q5KjtlOCzjWIeP9N5y1v037xgLUmOu0yBMPkgmFxy"

options(httr\_oauth\_cache = TRUE) # enable using a local file to cache OAuth access credentials between R sessions

setup\_twitter\_oauth(consumer\_key, consumer\_secret, access\_token, access\_secret)

# Issue search query to Twitter

dataInput <- reactive({

tweets <- twListToDF(searchTwitter(input$searchkw, input$num , lang = "en" ))

tweets$created <- as.character(tweets$created)

tweets <- tweets

})

# Create a reactive table

output$table <- renderTable(

dataInput()[, c("text", "screenName", "created")]

)

}

)

**Sample code for what’s app**

#Load required packages

library(ggplot2)

library(lubridate)

library(Scale)

library(reshape2)

library(tm)

library(SnowballC)

library(wordcloud)

library(RColorBrewer)

library(stringr)

library(syuzhet)

library(dplyr )

#get the data from whatsapp chat

text <- readLines("xyz.txt")

#let us create the corpus

docs <- Corpus(VectorSource(text))

#clean our chat data

trans <- content\_transformer(function (x , pattern ) gsub(pattern, " ", x))

docs <- tm\_map(docs, trans, "/")

docs <- tm\_map(docs, trans, "@")

docs <- tm\_map(docs, trans, "\\|")

docs <- tm\_map(docs, content\_transformer(tolower))

docs <- tm\_map(docs, removeNumbers)

docs <- tm\_map(docs, removeWords, stopwords("english"))

docs <- tm\_map(docs, removePunctuation)

docs <- tm\_map(docs, stripWhitespace)

docs <- tm\_map(docs, stemDocument)

#create the document term matrix

dtm <- TermDocumentMatrix(docs)

mat <- as.matrix(dtm)

v <- sort(rowSums(mat),decreasing=TRUE)

#Data frame

data <- data.frame(word = names(v),freq=v)

head(data, 10)

#generate the wordcloud

set.seed(1056)

wordcloud(words = d$word, freq = d$freq, min.freq = 1,

max.words=200, random.order=FALSE, rot.per=0.35,

colors=brewer.pal(8, "Dark2"))

#fetch sentiment words from texts

#count the sentiment words by categorySentiment <- get\_nrc\_sentiment(text)

head(Sentiment)

text <- cbind(text,Sentiment)

TotalSentiment <- data.frame(colSums(text[,c(2:11)]))

names(TotalSentiment) <- "count"

TotalSentiment <- cbind("sentiment" = rownames(TotalSentiment), TotalSentiment)

rownames(TotalSentiment) <- NULL

#total sentiment score of all texts

ggplot(data = TotalSentiment, aes(x = sentiment, y = count)) +

geom\_bar(aes(fill = sentiment), stat = "identity") +

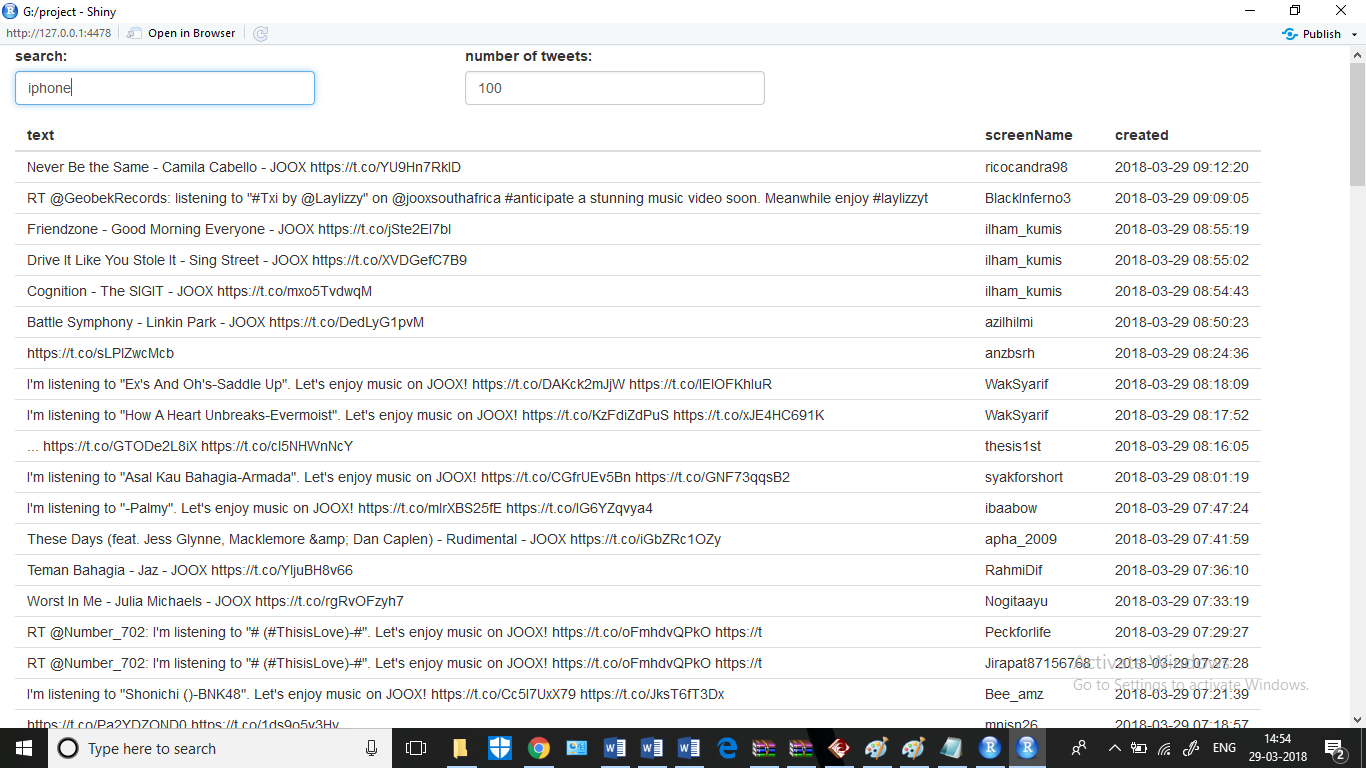
theme(legend.position = "none") +

xlab("Sentiment") + ylab("Total Count") + ggtitle("Total Sentiment Score")

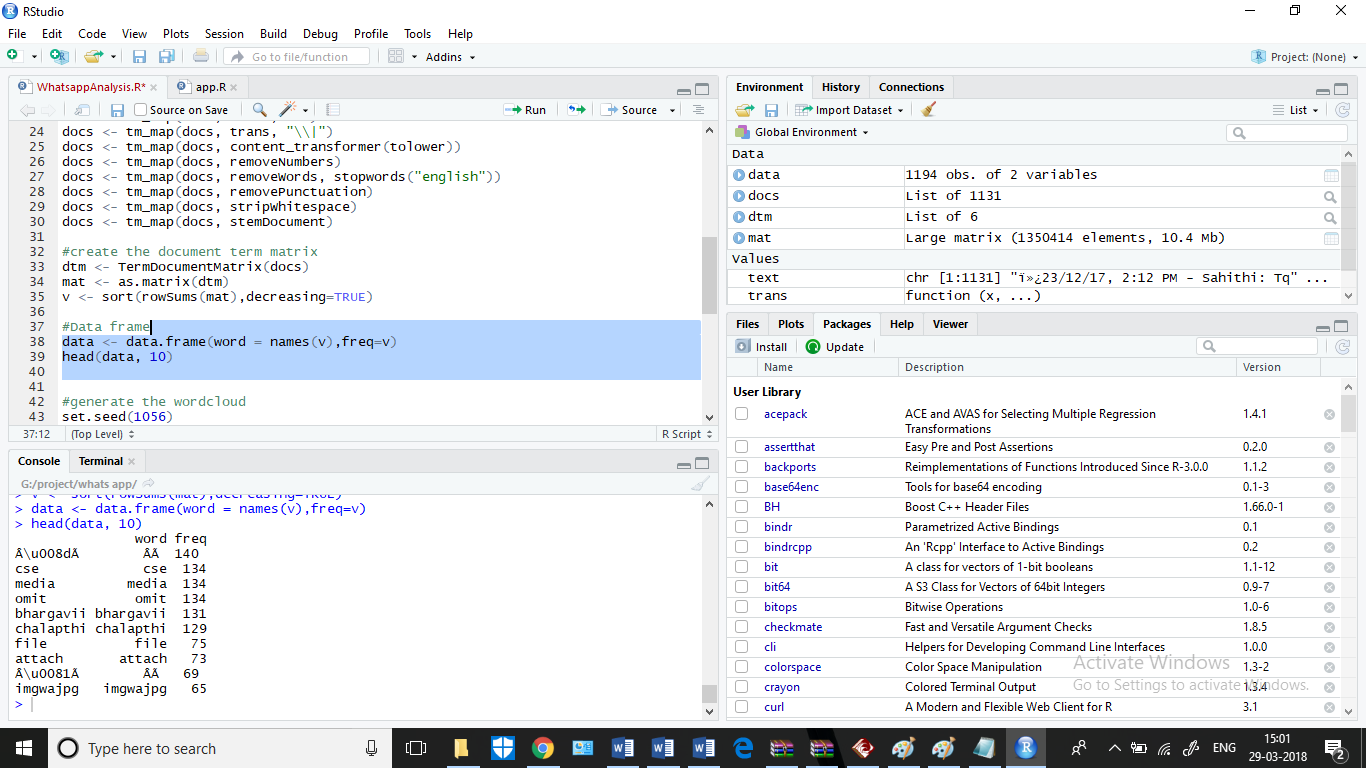
**CHAPTER 9  
 System Testing**

**CHAPTER 10**

**Screen Shorts**



Twitter output.



What’s app most frequently massege output.

 smost messaged during a day  sentiment analysis of group chat

 Number of messages date wise.

**CHAPTER 11**

**Conclusuion**

**CHAPTER 12**

**Biblography**

https://en.wikipedia.org/wiki/R\_(programming\_language)

About R programming language

https://en.wikipedia.org/wiki/Twitter

About twitter

https://en.wikipedia.org/wiki/WhatsApp

About WhatsApp

R for Data Science

Hands-on Programming with R

by Grolemund (Author), Garrett (Author)

Applied Predictive Modeling

applied-predictive-modelingThis book is written by Max Kuhn and Kjell Johnson.