# Introduction to Text Mining with R using packages

Venkateswarlu pynam [1], Kolli srikanth [2], Ashok Surgala [3], Aravind Bammidi [4]
*Assistant professors Department of Information Technology JNTUniversity KAKINADA*
*Department of Information Technology*
*University College of engineering, Vizianagaram, Andharapradesh, 525003*

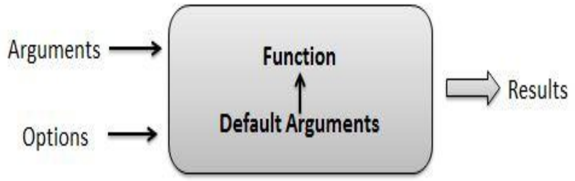
***ABSTRACT:*** *The fact that R is a language may deter some users who think "I can't program". This should not be the case for two reasons. First R is an interpreted language, not a compiled one meaning that all commands typed on the keyboard are directly executed without requiring tobuild a complete program like in most computer languages (C, FORTRAN, Pascal . . .). R's syntax is very simple and intuitive For instance a line a 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()).R can be considered as a different implementation of S and is much used in as an educational language and research tool. The main advantages of R are the fact that R is freeware and that there is a lot of help available online. It is quite similar to other programming packages such as Mat Lab, but more user-friendly than programming languages such as C++ or FORTRAN and python etc….*

**Keywords** — *R, S, programming packages***,** *text mining.*

## I. Introduction :

R is a language and environment for statistical [2] calculate ting and artwork. It is a GNU's not Unix linux proposal which is identical to the S language and situation which was developed at Bell Laboratories by John Chambers and colleagues. R can be treated as a distant operation of S [3]. There are some important differences but much code written for S runs unaltered under R. 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 and functions. The use of operators is relatively intuitive we will see the details later. An Rfunction may be sketched as



The arguments can be objects, some 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, or no argument has been defined in the function.

All the behaviors of R are done on substance stored in the active memory of the computer no short lived files are used. Fig. 1. The readings and writings of files are used for input and output of data and results. 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. Since the results are themselves objects they can be considered as data and analyzed as such Data can be read from the local disk or from a remote server through internet.

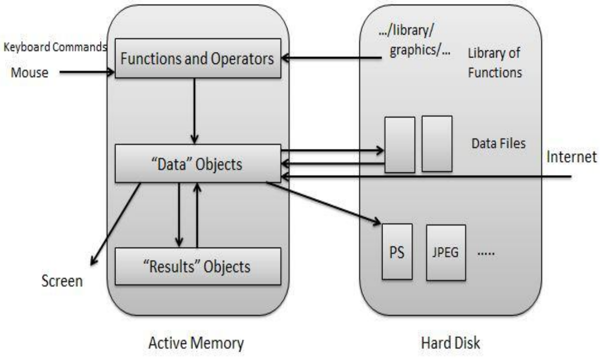
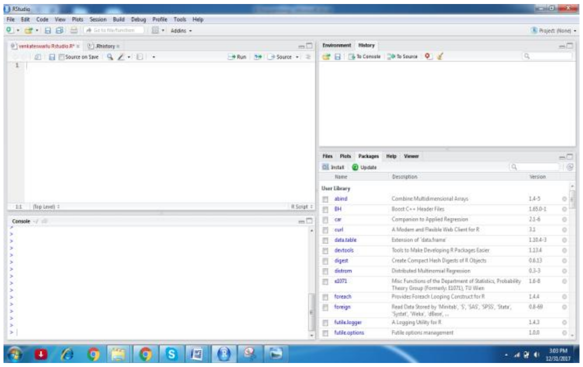


Figure 1: A schematic view of how R works.

The functions available to the user are stored in a library localized on the disk in a directory called R HOME/library. This directory contains packages [4] of functions which are themselves structured in directories. The package named paltry is in a way the core of R and consists of the basic functions of the language exceptionally for reading and manipulating data. Each package has a directory called R with a file named like the package R HOME /library/ base/ R/ base this file consists all the functions of the package.



| Name | Description | Version |
|---|---|---|
| User Library | | |
| abind | Combine Multidimensional Arrays | 1.4-5 |
| BH | Boost C++ Header Files | 1.65.0-1 |
| car | Companion to Applied Regression | 2.1-6 |
| curl | A Modern and Flexible Web Client for R | 3.1 |
| data.table | Extension of 'data.frame' | 1.10.4-3 |
| devtools | Tools to Make Developing R Packages Easier | 1.13.4 |
| digest | Create Compact Hash Digests of R Objects | 0.6.13 |
| distrom | Distributed Multinomial Regression | 0.3-3 |
| e1071 | Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien | 1.6-8 |
| foreach | Provides Foreach Looping Construct for R | 1.4.4 |
| foreign | Read Data Stored by 'Minitab', 'S', 'SAS', 'SPSS', 'Stata', 'Systat', 'Weka', 'dBase', ... | 0.8-69 |
| futile.logger | A Logging Utility for R | 1.4.3 |
| futile.options | Futile options management | 1.0.0 |

Figure 2: list of packages available in R

## II. Getting started

### II.I Install R

To install R on your computer go to the home website of R*:

- click download CRAN in the left bar
- choose a download site
- choose Windows as target operation system
- click base
- choose Download R 3.0.3 for Windows

It is also possible to run R and RStudio from a USB stick instead of installing them. This could be useful when you don't have administrator rights on your computer.

### II.II Install RStudio

After finishing this setup you should see an "R" icon on your desktop. Clicking on this would start up the standard interface. We recommend however to use the RStudio interface. To install RStudio go to: http://www.rstudio.org/and do the following (assuming you work on a windows computer):

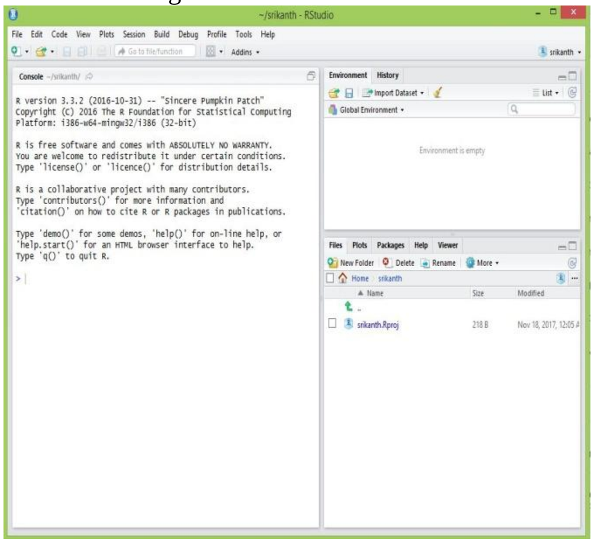
- click Download RStudio
- click Download RStudio Desktop
- click Recommended For Your System
- download the .exe file and run it

### II.III RStudio layout

The RStudio interface consists of several windows.

- Bottom left: console window (also called command window). Here you can type simple commands after the ">" prompt and R will then execute your command. This is the most important window, because this is where R actually does stuff.

- Start the R system, the main window (RGui) with a sub window (R Console) will appear as in figure 2.



```
R version 3.3.2 (2016-10-31) -- "Sincere Pumpkin Patch"
Copyright (C) 2016 The R Foundation for Statistical Computing
Platform: i386-w64-mingw32/i386 (32-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

>
```

Figure 3 : The R system on Windows

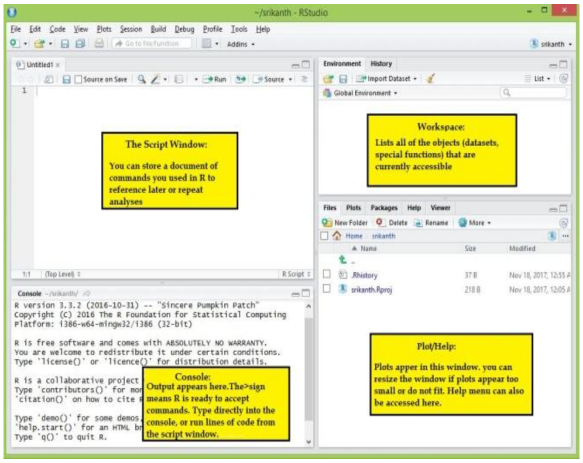
**3. RStudio you will see four windows:**

**Script:**
The script is a document to store a list of R commands. This window frame may not show up when you first accessible the RStudio. To create a new script, Click "File New → R Script." figure 3.

**Console:**
Output appears here. The > sign (also called the "prompt") means that R is ready to accept commands. You can enter the commands directly into the console is a good habit to instead enter into the script window and run commands from there. Nothing in the console can be saved. You can notwithstanding save your commands in a script file, and then rerun your analysis next. This is especially helpful if you are working on a big project or if you'd like to keep your code to refer back to later.

**Workspace:**
This workspace frame lists the objects directly accessible to you the actions that are element of packages [4] will not show up here applicable functions that you write yourself or that are element of a previously saved workspace will show up here.

**Plot/Help:**
The final frame has a few tabs along with a help tab with a search feature. When you create plots they will show up in this frame which you can resize to get a improved view. The Files tab likewise shows you the files on your computer as one way to approach R Scripts you have already written. Be careful- deleting files in this frame to deletes them from your computer.


~/srikanth - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Addins
srikanth
Untitled1
Source on Save
Run
Source
Environment History
Import Dataset
List
Global Environment
Files Plots Packages Help Viewer
New Folder Delete Rename More
Home srikanth
Name Size Modified
.Rhistory 37 B Nov 18, 2017, 12:55
srikanth.Rproj 218 B Nov 18, 2017, 12:05
1:1 (Top Level) R Script
Console ~/srikanth/

```
R version 3.3.2 (2016-10-31) -- "Sincere Pumpkin Patch"
Copyright (C) 2016 The R Foundation for Statistical Computing
Platform: i386-w64-mingw32/i386 (32-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project
Type 'contributors()' for mor
'citation()' on how to cite R

Type 'demo()' for some demos,
'help.start()' for an HTML br
Type 'q()' to quit R.
```

The Script Window: You can store a document of commands you used in R to reference later or repeat analyses

Workspace: Lists all of the objects (datasets, special functions) that are currently accessible

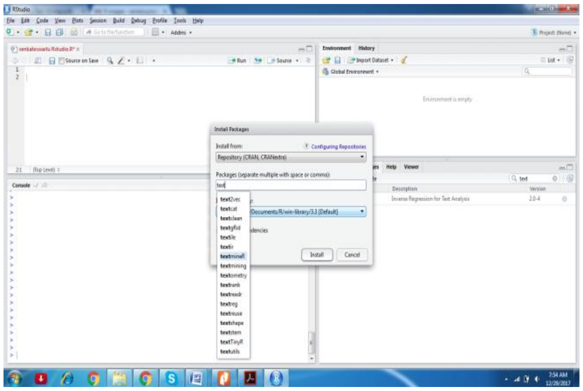
Console: Output appears here.The>sign means R is ready to accept commands. Type directly into the console, or run lines of code from the script window.

Plot/Help: Plots apper in this window. you can resize the window if plots appear too small or do not fit. Help menu can also be accessed here.

Figure 4 : The R system on Windows

## III. Introduction to text mining using R

Text mining has become a attractive access to analyzing and understanding massive datasets not responsive to traditional subjective research techniques. These appliances have been applied to a range of information complication, such as understanding argument in social media or facilitating information retrieval in unstructured data. Text mining [5] can be a deeply useful tool in the creation of research analysis, allowing the textual data to suggest argument and approach to the analyst during analysis figure 5.



Install Packages

Install from: Repository (CRAN, CRANextra)

Packages (separate multiple with space or comma): text

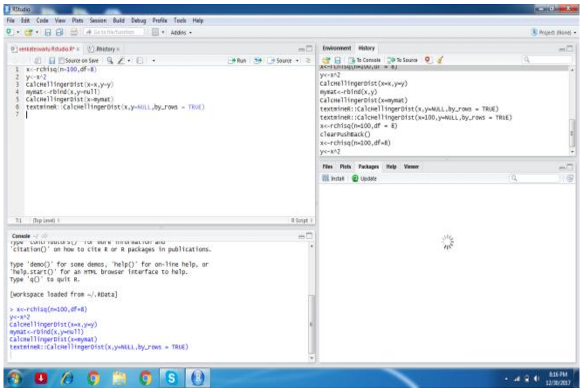
Documents/R/win-library/3.3 [Default]

text2vec
textcat
textclean
textgRid
textile
textir
textmineR
textmining
textometry
textrank
textreadr
textreg
textreuse
textshape
textstem
textTinyR
textutils

Figure 5 : The TextmineR package installation

Introduction to text mining in R utilizing the text mining framework provided by the textmineR package. We present design for knowledge import, corpus manipulation, metadata administration, and creation of term-document matrix[1]. Our spotlight is on the primary aspects of appropriate ting started with text mining in R an in-extent description of the text mining framework offered by textmineR.

**III.I Cosine Similarity** - This measure helps to find similar documents. It's one of the frequently used orbit metric used in text analysis. For a given 2 vectors A and B of length n each cosine similarity can be determined as a dot product of two unit vectors: figure 6.



```
x<-rchisq(n=100,df=8)
y<-x^2
calcHellingerDist(x=x,y=y)
mymat<-rbind(x,y=null)
calcHellingerDist(x=mymat)
textmineR::CalcHellingerDist(x,y=NULL,by_rows = TRUE)
```

```
y<-x^2
calcHellingerDist(x=x,y=y)
mymat<-rbind(x,y)
calcHellingerDist(x=mymat)
textmineR::CalcHellingerDist(x,y=NULL,by_rows = TRUE)
textmineR::CalcHellingerDist(x=100,y=NULL,by_rows = TRUE)
x<-rchisq(n=100,df = 8)
clearPushBack()
x<-rchisq(n=100,df=8)
y<-x^2
```

```
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Workspace loaded from ~/.RData]

> x<-rchisq(n=100,df=8)
y<-x^2
calcHellingerDist(x=x,y=y)
mymat<-rbind(x,y=null)
calcHellingerDist(x=mymat)
textmineR::CalcHellingerDist(x,y=NULL,by_rows = TRUE)
```

Figure 6 : The cosine similarity on TextmineR package

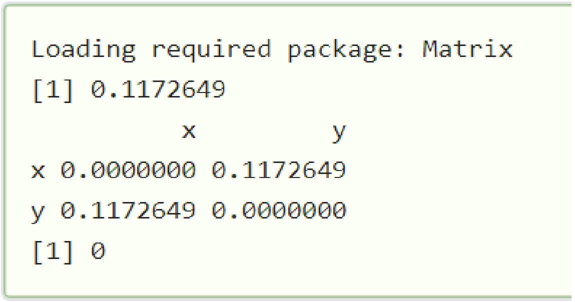
The functionality offered by the text mining package in R, assuming little knowledge of the R language and text mining generally. Although there are numerous and resources in this area an end-to-end exploration of the common functionalities must still be cobbled together across instructional sources, which can be difficult for novice users. The intention is to cover the primary tools and process, such that readers may dive in with their own text collections immediately and produce necessary data for the users.

### III.II Calculate Hellinger Distance

If x is a matrix returns a square and symmetric matrix. The i, j access compare to the Hellinger Distance between the rows of x (or the columns of x if by_rows = FALSE).

If x and y are angle entry to a numerical scalar whose expense is the Hellinger Distance between x and y. figure 7.

```
Library (textmineR)
x <- rchisq (n= 50, df= 20)
y <- x 2
CalcHellingerDist(x= x, y= y)
mymat <- rbind(x, y)
CalcHellingerDist (x= mymat)
textmineR:: CalcHellingerDist(x=3, y=2, by_rows=false)
```



```
Loading required package: Matrix
[1] 0.1172649
          x         y
x 0.0000000 0.1172649
y 0.1172649 0.0000000
[1] 0
```

Figure 7 : Calculate Hellinger Distance

## IV. Conclusion

Using text mining models the massive volume of text being generated by companies, social media and etc.. There is going to be a surge in demand for people who are well versed with text mining & natural language processing. There is a huge amount of documentation, materials dedicated towards teach [6] both novice and advanced tools in R. For those new to the area of text mining in R, this can be overwhelming and require cherry-picking instructions from numerous resources. Brings together the most common analysis techniques that can be applied to text data, providing an introduction to the powerful text mining functionality available with the textmineR package. Based on this paper we can extend the paper with the help of node package Manager to execute the text mining as a web based application.

## REFERENCES

1. Feinerer, I and Hornik, K. (2015). tm: Text MiningPackage,Rpackage version0.6-2. Retrieved from http://CRAN.R-project.org/ package=tm

2. Meyer, D. and Hornik, K. and Feinerer, I. (2008). Text Mining Infrastructure in R. Journal of Statistical Software, 25 (5). pp. 1-54.
3. R Core Team. (2016). R: A language and environment for statistical computing. Retrieved from https://www.r-project.org/about.html
4. https://cran.r-project.org/web/packages/text-mineR /vignettes/tm.pdf
5. https://www.google.co.in/search?q=text+mining+in+r&safe=active&ei=Y6xIWsCAJYmAvQSr8prQAw&start=10&sa=N&biw=1093&bih=510
6. https://www.hackerearth.com/practice/machine-learning/advanced-techniques/text-mining-feature-engineering-r/tutorial/