The Text Analysis R

Shivram Nanda[1], Er. Mohit Yadav[2]
[1]*M.TECH (CSE), Research Scholar,* [2]*Assistant Professor*
[12]*Department of Computer Science & Engineering*
OM Institute of Technology and Management Hisar, India

Abstract- Now a days computable text analysers in communication research is one of the most exciting research field with many applications .Acceptable point is, it is very difficult to apply because of it requires knowledge of many techniques and formulas and methods and most importantly is that the software which is required to perform most of these techniques is not commonly available in common software packages .After a lot of hard work and the study of different techniques this could be possible .With the help of teacher's corner method we address those barriers which provides an overview of different techniques ,steps and different operations in an computational text analysis projects which show us the path or we can say that give us an demo how each step is performed in the R statistical software for the text analysis .Basically R is an open source software or we can say that the platform in which we have an external user community that has the basic task to develop and also to maintain a wide range of text analysis packages of this platform .We will study some of these packages to perform advanced text analysis that will make it easier.

I. INTRODUCTION

The study of text analysis is start with the **Data Mining.** The basic meaning Of the data mining is to find the relevant information from the heap of the information. Ongoing through the process of the data mining we discover different patterns and the methods in database systems. The data mining is inter dependable subfield of the computer science to take the best methods form the set of the data and to transform it to the information's for the further use of it .The term data mining refers to discover knowledge from the databases it also include the data management, data preprocessing complexity consideration and post processing and the updating of the data. On large scale of data the data analysis and analytics are referred to get the actual methods and the machine learning are more appropriate then the data mining.

The actual data mining task is that to analysis large quantity of the data by an semi-automatic or automated process. The term data mining usually contains a database techniques called spatial indexing. This usually means that store the geometric patterns like polygons, lines, points in queuing data. The machine learning is an predictive language process. The data mining involves a process called KDD process known as knowledge discovery in database.

The basic process involves in the data mining for the KDD process are Selection, preprocessing, transforming and the data mining and last but not the least evaluation .Besides all these process data mining also involves six main classes of the tasks these tasks are:- problem definition and specific goals, Identifying text to be collected, Text organization,Feature, extraction, Analysis. Reach an insight recommendation for the output.

In text analysis is process in which we derive high quality information from the database in this we derive different patterns and trends which means statistical pattern learning.in the process of text analysis we come up with the new term called Text Analytics which has a set of machine learning techniques that model and structure the information .It also has an automated process in which to discover the present knowledge facts and relationships.

The text analysis process is rapidly growing and also used in in wide variety of government and business and research for record management and searching of the relevant documents.

II. RELATED WORKS

In today world with the increase of the database size to find the computational text analysis from the database many researchers face the most common challenges of learning how to implement an most advanced software for the text analysis .Currently one of the most popular and easily learnable environment for the computable text analysis in the field of data science is R statistical software. Perhaps for those researchers which don't have a well knowledge of the programming R can be a big challenge & performing a text analysis method in R can be a headache but with the help of teacher's corner method this fear can be overcome. A step by step information of each common techniques will help the researchers to get an familiar way to get the text analysis process with the R environment.

Now the Question is what is R environment and how it will help the researchers for the text analysis. R is a free under the GNU license and it is already installed in many operating systems like Linux and windows, open source programming environment besides most of the programming languages R is specially designed for the statistical analysis which make it most suitable for the data science applications. Learning the programming in R is not easy for the people which are not prior the programming experiences. The tools and the packages which are present in the R environment for the text analysis make it more powerful and accurate using only a few simple commands. R is an explosive growth has been most populated collections of the packages because these packages are supplied and maintained by the extensive user of the

community. Each packages functionality are extended from the core packages. Thus R is features a wide range of inter-compatibles packages which are maintained continuously by scholars and the researchers which can be installed further.

The most common and useful methods for the text analysis with the R environment that covers most of the steps to perform the text analysis is Teacher's corner method .In it from data preparation to text analysis it also provide easy way to replicate example code to perform each step.

Generally the interactivity with the R environment is very simple syntax. You can also make your work directory "$ mkdir" work for the use of that particular program or you can simply exit to that program after your work and you can use that directory for the further cases.

You can also give a special character for the "character string" like you can simply use

>help(solve)

An alternative is

> ?solve

However R is also a very case sensitive in some operating system like Unix. Perhaps R environment has also a specification of recall or correction of some commands in the system. You can also execute some commands by taking it from some other outsource or output file. All the variables functions, objects that are used in current R session are stored permanently in the current R session but you can also remove that particular one from current R session.

Vectors and Assignment-The R has different types of the objects vector is one of the simplest form in which a single entity contains of the order of the number. Basically in the vector arithmetic each element are performed one by one. Vector occurs in an expression need not to be of same length. If they are not the value of smaller expression is recycled for the largest one. Besides this vectors also has the logical values Like "TRUE & FALSE &NA(not available)".In any case if the vector value is not present the a with the vector value of it is reserved by assigning a simple value NA.We know that the vectors are one of the most important type of object in R environment but there are many more objects than vector like matrices , factors , list, data frames.

Reading Data from the Files- R also give us an important feature that is to read a file from an external source either entering the whole values. Input of the data from an external source is very simple but also inflexible one can easily modified the input files from some other tools. As we strongly suggest that most of the variables are held in the data frames, then they can directly read from the function **read.table()** we can also use the **scan()** function for direct access.

The**read.table()** function this will read all the data frame directly.In this first line has the name of the each data frame & all the additional lines first has the row label then the details of the each values.

Input File From Rows and Labels:

	Price	Floor	Areas	Rooms
01	53.45	189.00	750	5
02	64.70	220.60	600	4
03	78.00	280.00	800	5

The **scan()** function these will explain with an example let us assume there are the data vectors of equal length which is to be read in parallel in which first mode is the character mode and remaining mode is the numeric mode. The scan() function is used to read all the three vectors in a list.
Now the next thing is accessing the built in datasets there are approx. 100 datasets are supplied with the R environment and some are also present in the packages. One can easily check all the datasets present in R with a simple command **data()**
To access data from a particular package, use the package ,for example **data(package="rpart")**
If one of the package which has been linked by library, its datasets which are present in it are automatically get included in the search results. Editing data is also a simple process because when we invoked a data frame R environment brings a spate sheet like environment for editing. This is useful for making small changes once the data is set to be read.
Probability Distribution-R environment is one of the most convenient use is to provide a comprehensive set of statistical tables. Different functions are used to evaluate the cumulative distribution function. Given set of the data can be examined in large numbers of ways. And the simplest one to examine that numbers displayed the numbers by Stem(a 'stem and leaf' plot).A stem-and-leaf plot is like a histogram, and R Environment has a function hist() to plot histograms.
Grouping & Looping-One of the most important feature of R is grouped Expression R is an expression language in the sense that it's only command type is a function or expression which returns a result even the result may be a value. Commands may be grouped together in braces,{erp_1,….exp_m}. There are also some controlled statements like if statements like > if (expr_1) expr_2 else expr_3.There is also repetitive execution for loops, repeat and while.we can also construct the loop in the form > for (name in expr_1) expr_2 where the name of the loop can variable.
Write your own function-As we know R tool give us a power of creating our own objects mode function. These R functions that are stored in a special internal form and may be used in further expressions and so on. In the process, the language gain power, learning to write useful functions is one of the main important ways to make the use of R environment comfortable. Those symbols that occur in the body of function can be divided inti three classes ; formal parameters, local variables and free variables these are the main 3 classes The formal parameters of a function are that one functions which occurs in the list argument . Local variables are those whose values are determined by the evaluation of expressions in the

body of the functions. Those Variables which are no formal parameters and they are also not local variables these type of variables are free variables. One can customize their R environment in several different ways. There is a way that is site initialization. Special functions that are used is**.First**& the **.Last** .The location for the site initialization file is taken from the value of the **R_PROFILE** environment variable.

Statistical models in R-The reader has some familiarity with statistical methodology, Such as With the regression analysis and the analysis of variance. The requirements for the statistical models are make it possible to construct general tools that can be applied in a broad spectrum of problems in R environment environment provides an inter locking facilities that make fitting of the statistical models simple.

III. GRAPHICAL FACILITIES

R environment are an important and extremely versatile component. In R environment one can use the facilities to display a wide variety of statistical graphs and also to build completely new types of graphs designs. It can be used in both interactive and batch modes, in most of the cases, interactive used. Because its use is also easy because at startup time. Basically R environment opens a new graphics window for the display of interactive graphics which is initiates by a graphics device driver . Although this is done automatically, with the command used is X11() under UNIX, windows() under Windows and quartz() under macOS. A new device always can be opened by dev.new().When the device driver is running, R plotting commands that are used to to create entirely new kinds of display produce by a variety of graphical displays .

Plotting commands in R environment are divided into three basic groups:

- High-level plotting functions create a new plot in the graphics device, with axes,labels, titles and so on.
- Low-level plotting functions are used to add more information to an existing one, such as extra points, lines and labels.
- Interactive graphics functions allow users to add interactive information to, or extract information

From an existing plot, using a mouse or some other pointing device.

In addition, R also has some different functions to customize your plots.

Packages-All the R functions data sets are stored in the form of packages. When we load a package all the contents present in the packages are get available.>**library()** is used to see which packages are installed.

Users can also update and load the packages while they are connected to the internet. All the package have namespaces, which do three things: they allow the package writer to hide function and datasets that are meant only for the internal use, they prevent functions from breaking down when a user or some other package writer picks a name that clashes with one in the package, and they provide a useful way to refer to an new object within a particular package.

IV. OPERATING SYSTEM FACILITY

R has an advanced feature to access the OS facility under which it is running this allows that to be used as a scripting language and that ability is much used by R itself. Otherwise lot of time will be spending to make it feasible. R environment has many functions to manipulate files and directories.

To create an new (empty) file or new directory, use "file.create" or "dir.create"

Files can be removed by "file.remove" or unlink:

For directory listings use "list.files" or "list.dirs".

Many types of information on a filepath can be found by file.info.

These are functions "file.exist", "file.access" and "file_test"

There is some support for links in the filesystem: see functions "file.link" & "sys.readlink".

Functions like **system** and **system2** are used to invoke a system command.**system2** is a little more general it has the main advantage is that it is easier to write cross-platform code using it.**system** command have different behave on Windows from other OS . Recent versions of R environment have facility to read and write compressed files. Reading of files in R is to a very large extent done by connections, and the file function which is used to open a connection to a file (or a URL) and is able to identify the compression of file. The type of compression which has been supported for longest is gzip compression, and that remains a good general compromise. Files compressed by the earlier Unix compress utility can also be read, but these are becoming rare with the pass of time. File archives are single files that contain a collection of files. By these ways one can read and write the file and also can compression of the files.

V. CONCLUSION

R is one of the most efficient tool for the database analysis but it's real evolution or we can say that its real use come after the evolution of the Java programming languages. In starting all the developed packages are enough for the data analysis but later on as per the requirement the data packages are modified as per the need .Moreover the R environment is free & it's not as much difficult to learn so we can say that the scholar's and researchers can easily become familiar to R environment . Basically the selection of a package can be done by the various techniques. All the packages present in the teacher's Corner method provides a good starting point also there are many other some great packages. Firstly "we shape our tools" then later on "our tools shape us".

VI. REFERENCES

[1]. http://www.r-statistics.com/tag/hadley-wickham/
[2]. http://www.infoworld.com/article/2940864/application-development/r-programming-language-statisticaldata-analysis.html
[3]. http://spectrum.ieee.org/computing/software/the2015-top-ten-programming-languages
[4]. http://www.analytics-tools.com/2012/04/r-basicsintroduction-to-r-analytics.html
[5]. http://blog.revolutionanalytics.com/

[6]. http://www.r-bloggers.com/handling-large-datasetsin-r/
[7]. http://www.analytics-tools.com/2012/04/r-basicsintroduction-to-r-analytics.html
[8]. http://data.vanderbilt.edu/~hornerj/brew/useR2007.r html
[9]. http://blog.ukdataservice.ac.uk/the-power-of-rmethods-for-processing-big-data/
[10]. http://bigdatauniversity.com/moodle/course/view.ph p?id=522
[11]. http://aisel.aisnet.org/cgi/viewcontent.cgi?article=378 5&context=cais
[12]. http://www.revolutionanalytics.com/what-r
[13]. http://blog.revolutionanalytics.com/2013/12/tips-oncomputing-with-big-data-in-r.html
[14]. http://blog.ukdataservice.ac.uk/the-power-of-rmethods-for-processing-big-data/

AUTHOR'S BIOGRAPHIES

Shivram Nandais a M.TECH. student in department of Computer Science & Engineering from OM Institute of Technology and Management, Hisar (Haryana). He received B.TECH degree in Computer Science & Engineering from Ch. Devilal State Institute of Engineering & Technology, Sirsa (Haryana). His research include the some important features of R text analysis.