The Text Analysis R

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

Abstract **-** R environment is an open-source data analysis environment and programming language. The method of converting data into knowledge, insight and understanding is Data analysis, which is a significant part of statistics. For the successful processing and analysis of big data, it allows user to carry out a number of tasks that are important. R consists of several ready-to-use statistical modelling algorithms and machine learning which allow users to make reproducible research and develop data products. Even though big data processing may be proficient with other tools as well, it is when one step on to the data analysis that R really stands only one of its kind, remaining to the huge amount of built- in statistical formulae and third-party algorithms available.

I. INTRODUCTION

R has three things: a project, a language, and a software environment. As a project, R is a part of GNU free software project (www.gnu.org), an worldwide attempt to allocate software on a free basis, without license limitations. As a result, R does not cost the user everything to use. The development and licensing of R are done under the values that software should be free and not proprietary. This is good for the user, even if there are some disadvantages. Mainly, that *"R environment is free software and comes with ABSOLUTELY NO WARRANTY."* This testimonial comes up on the screen whenever we start R. There is no quality control team of a software company adaptable R as a product.

The R project is largely an intellectual venture, and most of the contributors are statisticians. The R project was on track in 1995 by a group of statisticians at University of Auckland and has persistent to grow ever since. Since statistics is a cross-disciplinary science, the use of R has appealed to intellectual researchers in different fields of applied statistics. There are a lot of niches in conditions of R users, as well as: *environmental Statistics, econometrics, medical and public health applications and bioinformatics, along with others.*

As a language R is a idiom of the S language , an object-oriented statistical programming language developed in the late 1980's by AT&T's Bell labs.

Related software and documentation - There are lots of books which give us the details about how to get use of the R data analysis & documentation & statistics by keeping in mind the difference between R and S implementation.

The given introduction to the R environment did not mean that Statistics, perhaps many people use R environment as it is a Statistics system. R is as of an environment within which many classical and modern statistical techniques have been implemented. Those few which are supplied as **packages** are built into the base R environment**.** There are about 25 standard packages that are supplied with R(which are also known as "standard" and "recommended" one) & lot more that can available through the Internet sites and elsewhere.

The important difference between the other statistical systems and R.

In S intermediate results being stored in objects as a statistical analysis is normally done as a series of steps. This will give copious output from a regression or discriminant analysis whereas SAS and SPSS, R will give the results in a fit object for subsequent interrogation minimal output and store by further R functions.

The most convenient way to use R is running a windowing system is at a graphics workstation. In particularly we will occasionally refer to the use of R on an X window system.

Most users can directly interact with the OS from the even spend of the time they can also find its necessary. The mainly interaction with the operating system on UNIX machines.

When we use the R program it issues a prompt when it expects input commands. The system default prompt is '>', which is on UNIX system might be the same one as the shell prompt ,& so it may be also appear that nothing is happening . However, as we can see, if you wish it is easy to change to a different R prompt. We use the UNIX shell prompt is '$' as assumption .

Technically R has a very simple syntax is an expression language . It is case sensitive A and a are different symbols and would refer to different variables as are most UNIX based packages, so. The R names depends and country within which R is being run (technically on the **locale** in use) on the operating system.

Elementary commands have their either of assignments or expressions . If an which has the expression that is given in an a command, which is evaluated, and also printed (unless that is specifically made invisible), which also has the value is lost.

The entities manipulates are known as objects that R has creates. These may be variables, arrays of members, character strings, functions, or more general structures built from such components.

All objects stored permanently created during an R session in a file for use in future R sessions. At the end of each R session we are given the opportunity to save all the currently available objects. If we indicate that we want to do this, the objects are written to a file called **.RData** in the current directory, used in the session are saved to a file the command lines called **.Rhistory**.

When the R is started at later time it reloads the workspace from this file from the same directory. And at that same time the associated command history is also get reloaded.

Benefits of using R - Packages Ecosystem-One of R's strongest qualities is the vastness of package ecosystem. There's a lot of functionality that's built in and that's built for statisticians.

R is extensible- R provides rich functionality for developers to build their own tools and methods for analyzing data. Lots of people aroused to it from other fields such as biosciences and even humanities. People can extend it without a need to ask permission.

Free software-At the time when R first came out, the biggest advantage of it was that it was free software. It is available to look at every single thing and source code.

R's graphics and charting capabilities-For data manipulation and plotting the dplyr and ggplot2 packages, respectively have literally improved quality of life.

R's strong ties to academia-Any new research has an associated R package to go in the field probably. So R stays progressive. The caret package also offers a pretty smart way of doing machine learning in R through a relatively API. A lot of popular machine learning algorithms are implanted in R.

II. OBJECTIVES

R in Growth - In 2015, IEEE had listed R at 6th position in the top 10 languages of 2015. In addition to this, as the amount of intensive data work increases, Market demand for tools which such as R which is particularly used for data-mining, visualization & processing will also increase.

R in Business - R was originated as an open-source version of the S programming language in the 90's. R has gained a lot of support for most of the companies since that time, mostly R Studio and Revolution Analytics which are used to create different packages, and various types of services that are related to the language. Basically R has help & support system the form of large companies that power to the some of the largest relational databases in the world.

R in Higher Education - R is also originated in academia. Ross Ihaka and Robert Gentleman in New Zealand at the University of Auckland created it, and it's also been widely adopted in graduate programs that include intensive study of statistics. Massive open online course such as the Coursera Data Science Program also makes use of R.

R has a diverse community - The R community is diverse, along with many individuals coming from unique professional backgrounds. This list includes statisticians, business analytics, academics, scientists and professional programmers. The comprehensive R Archive Network (CRAN), different community members maintains packages that are been created by them .All the Packages that exist in R in such an order to create different maps and also to perform stock market analysis.

R is fun - R is FUN! R has an ability to generate charts and plots in very few lines of code. Tasks which are done with the multiple lines in the different languages are done by few lines in the R. While it's been considered strange when you compare it with many popular languages, it also includes the powerful features towards data analysis.

R's Challenges - R has its share of shortcomings which are become our objectives as well as challenges which are really important to improve. These as follows:-

Memory
Management
Speed
Efficiency

These are probably the biggest challenges R faces. Also, people coming to R from other languages might also consider R odd.

When working with very large data sets the design of the language can sometime lead to problems. Data has to be stored in physical memory. But this can become a minor issue, as nowadays computers have plenty of memory.

R must have the abilities such as security that are not to be built. We can't use it for Web-like or Internet-like apps. It was primarily next to impossible to use R as back-end server to perform calculations due to lack of security over the Web. For a long time, there was not a lot of interactivity in the language. Languages such as JavaScript still have to enter in to fill this gap. Although the analysis that can be done in R, the end results might be accomplished in different types of language like JavaScript.

III. SOME STRATEGIES OF BIG DATA IN R

Big Data can be tackling with R, using four different strategies as follows:

Sampling - If data is too large in size to be analyzed completely, the size of it can be reduced by the means of the method sampling. Eventually, the question stands up whether sampling decreases the performance of a model or not. Much data is always better than little data of course. If in any case sampling that is needed to be avoided it is also recommendable that to use the another Big Data Strategy. But it for some reason sampling is necessary, it still can lead to various types of the satisfying models, especially those which when the sample is a kind of the big in total numbers, not much small in proportion to the full data set and not biased as well.

Bigger Hardware - Keeping all objects in memory, but this is a problem when the data gets too large in size. One of the easiest ways to deal with Big Data in R is to simply increase the machine's memory. Today, R it runs on 64-bit machines and it can address to 8 TB of RAM. In many situations this is a sufficient improvement compared to about 2 GB addressable RAM on 32-bit machines.

Store objects on hard disc and to analyses it chunk wise

As an alternative, there are various packages available to avoid storing data into the memory. Instead, all the objects are analyzed chunk wise instead of it is stored on hard disc. As a side effect, the chunking also leads to parallelization naturally, if the algorithms allow parallel analysis of the chunks. A negative side of this strategy is, only those

algorithms and R functions can be executed that are designed explicit to deal with datatypes that are hard disc specific.

IV. RESULTS AND DISCUSSION

To build a powerful and reliable statistical model, data transformation, estimate of several models options, and imagining the results are vital. This is the motive why the R language has proven so popular: its interactive language boost up exploration, amplification and presentation. Revolution R Enterprise provides the big-data support and speed to permit the data scientist to replicate through this procedure quickly.

For statistical data analysis, R environment is an open source software platform. Mostly because of its basis nature, R is quickly accepted by statistics departments in universities around the world, concerned by its extensible nature as a platform for academic research. Free in cost surely played a part as well. And it wasn't extensive before researchers in data science, statistics and machine learning in progress to publish papers in academic journals along with R code applying their new techniques. R constructs this process very easily and anyone can create an R package to CRAN that means for Comprehensive R Archive Network it accessible to everyone.

An excellent open-source interactive expansion environment has been generated by R Studio for the R language, further improving the efficiency of R users all over the place.

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".

V. 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 http://www.analytics-tools.com/2012/04/r-basicsintroduction-to-r-analytics.html
[4]. http://blog.revolutionanalytics.com/
[5]. http://www.r-bloggers.com/handling-large-datasetsin-r/
[6]. http://www.analytics-tools.com/2012/04/r-basicsintroduction-to-r-analytics.html
[7]. http://data.vanderbilt.edu/~hornerj/brew/useR2007.r html
[8]. http://blog.ukdataservice.ac.uk/the-power-of-rmethods-for-processing-big-data/
[9]. http://bigdatauniversity.com/moodle/course/view.ph p?id=522
[10]. http://aisel.aisnet.org/cgi/viewcontent.cgi?article=3785&context=cais
[11]. http://www.revolutionanalytics.com/what-r
[12]. http://blog.revolutionanalytics.com/2013/12/tips oncomputing-with-big-data-in-r.html

Author's Biographies - Shivram Nanda is 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 includes the some important features of R text analysis.