Solutional Journal of Computer Sciences and Engineering Open Access Research Paper Volume-4, Issue-9 E-ISSN: 2347-2693

Text Mining of Unstructured Data Using R

M. Siva Lakshmi¹ and MD. Arsha Sultana²

¹B.Tech Student, Computer Science and Engineering, NRI Institute of Technology, Pothavarappadu, India

² Assistant Professor, Computer Science and Engineering, NRI Institute of Technology, Pothavarappadu, India

Available online at: www.ijcseonline.org

Revised: 02/Sept/2016	Accepted: 20/Sept/2016	Published: 30/Sep/2016
e process of acquiring high-qualit	y information from text that is typical	lly borrowed through the
s such as statistical pattern learning	ng .It usually involves the process of	structuring the input text,
ructured data and finally evaluation	on and interpretation of the output. It ca	an help an organization to
usiness insights from text-based	content such as word documents, ema	ail and postings on social
k, Twitter and LinkedIn. Mining	g unstructured data with natural lang	guage processing (NLP),
ine learning techniques can be cha	llenging because natural language tex	t is often inconsistent. So,
ed data which is the most exhau	stive statistical analysis package and	it incorporates all of the
els and analyses for managing and	manipulating data.	
	Revised: 02/Sept/2016 e process of acquiring high-qualit s such as statistical pattern learni ructured data and finally evaluatio usiness insights from text-based c, Twitter and LinkedIn. Mining ine learning techniques can be cha ed data which is the most exhau- ils and analyses for managing and	Revised: 02/Sept/2016Accepted: 20/Sept/2016e process of acquiring high-quality information from text that is typicas such as statistical pattern learning .It usually involves the process ofructured data and finally evaluation and interpretation of the output. It causiness insights from text-based content such as word documents, emac, Twitter and LinkedIn. Mining unstructured data with natural langine learning techniques can be challenging because natural language texted data which is the most exhaustive statistical analysis package andls and analyses for managing and manipulating data.

Keywords- R,S,Text mining,Statistical Modeling

I. INTRODUCTION

Text Mining is one of the most convoluted scrutinizes in the industry of analytics because dealing with unstructured data is not certainly determine observation and variables (rows and columns). Thus for doing any kind of analytics, first transform this unstructured data into a structured dataset and then continue with normal modeling framework [3]. The additional step of transforming an unstructured data into a structured format is simplified by a Word dictionary especially crucial to do all kind of information eradication. To accomplish a sentiment scrutiny in dictionary is easily available on web world but for a few distinct scrutiny need to create a dictionary of our own.

The dictionary need in the firm obstacle looks like a very slot dictionary because we need to identify all names from the transaction free text. Probability of such a dictionary is reliably very low [1]. So, we need to devise a dictionary and then record integrated data set by using R to build an unstructured data model. R is very convenient to devise dictionary on smaller datasets.

Statistical modeling is a simplified, mathematicallydetermine way to approximate reality and make predictions from this approximation. It is collecting, encapsulating, analyzing and translating variable numerical data. Statistical methods can be contradicted with deterministic methods which are relevant where observations are precisely reproducible or assumed. Statistical methods are extensively used in life sciences, economics and agricultural science [7]. They also have an important role in the physical science in the study of evaluation errors of random phenomena such as radio activity or meteorological events and retrieving approximate results where deterministic solutions are hard to apply.

R is a programming language and environment developed for statistical scrutiny by practicing statisticians and researchers. It reflects well on a very competent community of computational statisticians [6]. R is free and open source software allowing anyone to use and to customize it. R is licensed under the GNU General Public License, with copyright held by The R Foundation for Statistical Computing. R has no license restrictions and so can run it anywhere and at any time, and even sell it under the conditions of the license.

II. OVERVIEW

Unstructured data is a universal design for characterize data especially not accommodated in a database or some other type of data structure textual or non-textual. Textual unstructured data is spawn in media like email messages, PowerPoint presentations, Word documents etc. Non-textual unstructured data is spawn in media like JPEG images, MP3 audio files and Flash video files. These consequences are in inconsistency and ambiguities that produce it difficult to interpret using traditional programs as compared to data stored in databases or annotated in documents.

In 1998, Merrill Lynch indicated a rule of thumb that around 80-90% of all probably usable business information may derive in unstructured form. This rule of thumb is not based on fundamental or significant research but however is accepted by some. IDC and EMC project data will become to

40 zettabytes by 2020 resulting in a 50-fold success from the beginning of 2010.Unstructured information might report more than 70%–80% of all data in organizations.

Simple techniques for structuring text generally contain manual tagging with metadata or part-of-speech tagging for more text mining. Unstructured Information Management Architecture (UIMA) implements a common structure for transforming information to extract content and devise structured data about the information.

Software that devises machine-process capable structure exploits the linguistic, auditory and visual structure built-in in all forms of human communication. Algorithms can interpret this built-in structure from text for instance by auditing word morphology, sentence syntax and other patterns. Unstructured information can be enhanced and identify to address ambiguities and relevancy-based techniques to simplify search and discovery [2]. Examples of "unstructured data" may consist of books, journals, documents, metadata, health records, audio, video, analog data, images, files, and unstructured text such as body of an e-mail message, Web page, or word-processor document. Main content comes packaged in files or documents that they have structure and are incorporate of structured and unstructured data, but generally this is still referred to as "unstructured data". For example, an HTML web page is identified, but HTML markup consistently serves simply for rendering. It does not capture the content of identified elements that support computerized processing content of the page. XHTML identified grant machine processing of elements, despite it consistently does not capture or transfer the semantic content of identified terms.

Unstructured data generally occurs in electronic documents. Document management system classifies entire documents which is often favored over data transfer and manipulating within the documents to transfer structure into document. Search engines turn into popular tools for indexing and searching data, especially text.

The generic S philosophy is important for users of S and R to understand because it arranged the moment for the design of the language itself which many programming expert find a bit different and complicated. It is important to understand that the S language had its origin in data scrutiny and do not arrive from a traditional programming language background. Its inventors were focused on computation out how to make data scrutiny easier, first for themselves and then finally for others. In the Evolution of S, John Chambers writes: "[W]e wanted users to be capable to begin in an interactive environment, where they did not consciously think of themselves as programming. Then as their needs became clearer and their sophistication increased, they should be able to slide gradually into programming, when the language and system aspects would become more important." The key part here was the transition from user to developer. They wanted to build a language that could easily service both "people". More technically, they needed to build language that would be suitable for interactive data analysis (more command-line based) as well as for writing longer programs (more traditional programming language-like).

In the early days, a key feature of R was that its syntax is very similar to S, making it easy for S-PLUS users to switch over. While the R's syntax is nearly identical to that of S's, R's semantics, while superficially similar to S, are quite different [4]. In fact, R is technically much closer to the Scheme language than it is to the original S language when it comes to how R works under the hood. Today R runs on almost any standard computing platform and operating system. Its open source nature means that anyone is free to adapt the software to whatever platform they choose. Indeed, R has been reported to be running on modern tablets, phones, PDAs, and game consoles.

One nice feature that R shares with many popular open source projects is frequent releases. These days there is a major annual release, typically in October, where major new features are incorporated and released to the public. Throughout the year, smaller-scale bug fix releases will be made as needed. The frequent releases and regular release cycle indicates active development of the software and ensures that bugs will be addressed in a timely manner [6]. Of course, while the core developers control the primary source tree for R, many people around the world make contributions in the form of new feature, bug fixes, or both. Another key advantage that R has over many other statistical packages (even today) is its sophisticated graphics capabilities. R's ability to create "publication quality" graphics has existed since the very beginning and has generally been better than competing packages.

Today, with many more visualization packages available than before, that trend continues. R's base graphics system allows for very fine control over essentially every aspect of a plot or graph. Other newer graphics systems, like lattice and ggplot2 allow for complex and sophisticated visualizations of highdimensional data [2]. R has maintained the original S philosophy, which is that it provides a language that is both useful for interactive work, but contains a powerful programming language for developing new tools. This allows the user, who takes existing tools and applies them to data, to slowly but surely become a developer who is creating new tools.



Fig.1.Flow diagram

III. IMPLEMENTATION

R is an integrated suite of software facilities for data manipulation, calculation and graphical display. It includes 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 on-screen or on hardcopy, and a well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities [2]. The term "environment" is intended to characterize it as a fully planned and coherent system, rather than an incremental accumulation of very specific and inflexible tools, as is frequently the case with other data analysis software.

Many users think of R as a statistics system but prefer it to think of it of an environment within which statistical

techniques are implemented. R can be extended (easily) via packages. There are about eight packages supplied with the R distribution and many more are available through the CRAN family of Internet sites covering a very wide range of modern statistics [6]. R has its own LaTeX-like documentation format, which is used to supply comprehensive documentation, both on-line in a number of formats and in hardcopy.

R provides two packages for working with unstructured text – TM and Sentiment. TM can be installed in the usual way. Unfortunately, Sentiment has been archived in 2012, and is therefore more difficult to install. Once initially installed, each can be loaded later as library (name).

The next step is to load the data. This data was saved to a text file and loaded and processed

>setwd("G://r lng") >data ← read.csv ("reviews.csv",header=TRUE)

- Now, change the data csv to doc format >dtm DocumentTermMatrix(courpus) >courpus corpus(source) >courpus

- Next, remove nonessential characters such as punctuation, whitespaces, numbers etc from the text, before processing the actual words themselves.

>courpus tm_map(courpus.content_transformer(tolower))
>courpus tm_map(courpus,removePunctuation)
>courpus tm_map(courpus,stripwhitespace)

Denote as matrix and then Find Frequency of words, sort it and see head >dtm2←as.matrix(dtm) >frequency←colSums(dt2) >frequency←sort(frequency,decreasing=TRUE) >head(frequency)

Retrieve data from dataset >text←data\$text

Now, Finally represent in a graph >barplot(head(frequency))

IV. RESULTS

Studio Edit Code	v	iew Pl	ots Sessio	on Build Debug Tools Help			- a ×
1 🕲 - 1 6	1		I A Got	o fée/fundion			🕓 Project: (None)
Heart ×	da	ta ×			-0	Environment History	-0
2	YF	ilter			Q	🎯 🕞 📑 Import Dataset - 🥑	📃 Ust • 🧐
x	1	ating [±]	location	text a		🚳 Global Environment -	(Q.
1	1	5	US	SO ADDICTING DEFF DOWNLAOD ITS EPIC YOU CAT L	^	Data	
2	2	5	us	Great game Hove this game. Unlike other games they	10	• data 1000 obs. of 4 variables	1
3	3	5	nz	Sooo much FUN I would definitely recommend this g			
4	4	5	gb	AWESCIME Epic game so addictive Sstars <f0><u+00< td=""><td></td><td></td><td></td></u+00<></f0>			
5	5	5	us	Good game Great game			
6	6	1	nz	Meh Used to be good until World Cup upgrade. Now i			
7	7	3	nz	Not bad <u+6e38><u+620f><u+4e0d><u+9519><u< td=""><td></td><td>Files Plots Packages Help Viewer</td><td>-</td></u<></u+9519></u+4e0d></u+620f></u+6e38>		Files Plots Packages Help Viewer	-
8	8	5	nz	Clash of clans Clash is a awesome game		a a Place Jeport 9	
9	9	4	au	Cool Great game A+A+A			
10 1	0	4	nz	Candy crush review Addictive			
11 1	1	4	us	Great game Creat game			
12 1	2	4	US	Fun but I have a disagreement Whenever you try to p	v		
owing 1 to 12	of 1.	000 ent	ries				
nnsole C-/r/ho	a/ 2	3			-7		
is free su u are weld pe 'licen: is a coll: pe 'contri itation() pe 'demo() elp.start pe 'd()' setwd("G:	ofti com se() ibu ibu ibu ibu ibu ibu ibu	mare a e to r)' or rative tors() n how for so for a quit R lng")	ind comes edistrib 'licence project 'far mo to cite me demos in HTML b	<pre>with ABSOLUTELY NO WARRANTY. Jute it under certain conditions. {})' for distribution details. t with many contributors. sre information and R or R packages in publications. s, 'help()' for on-line help, or rowser interface to help.</pre>	~		

After installing required libraries and tm package, load the data in R

Fig.2. Input data

Change the data csv into doc format

RStudio	D							-	
Pile Edit Code View	Go to fil	e/function	 Addins 					🔋 Proj	ject: (None) 🔹
Console G:/r Ing/ 📣				Ð	Envi	ronment History			
Content: documer	ts: 1	documente reve	(Indexed).	×	- er	🕞 👘 Import Dataset	- 4		List • 🚱
> stopwords("engl	lish")					Clobal Environment -	-	0	
[1] "i"	"me"	"my"	"myself"	"we"		-		0	
[6] "our"	"ours"	"ourselves"	"you"	"your"	Dat	a			^
[11] yours	"bimsolf"	"sho"	"her"	"hers"	O da	ata	1000 obs. of 4 variables		
[21] "herself"	"it"	"its"	"itself"	"they"	val	ues			
[26] "them"	"their"	"theirs"	"themselves"	"what"	0 CC	ourpus	List of 1		
[31] "which"	"who"	"whom"	"this"	"that"	🚺 🔘 dt	tm	List of 6		
[36] "these"	"those"	"am"	"hoop"	"are"	Osc	ource	List of 1		
[41] was	"has"	"had"	"having"	"do"	L	av+	" CO ADDICTING DEED DOWNLADD THE EDITE VOL	CAT LOVE	рс. ыт. ^V
[51] "does"	"did"	"doing"	"would"	"should"	Files	Plots Packages H	lelp Viewer		
[56] COUID	ought "it's"	1 m "we're"	'they're	ne s	01	Install 💽 Lindate	Q.		R
[66] "you've"	"we've"	"they've"	"i'd"	"you'd"	-	Name	Description	Version	
[71] "ĥe'd"	"she'd"	"we'd"	"they'd"	"1'11"		Nume	Desciption	version	
[76] "you'll"	"he'11"	"she'11"	"we'11"	"they'll"	User	Library			~
[81] "isn't" [86] "haven't"	"hadn't"	"wasn't" "doesn't"	"weren t" "don't"	"didn't"		ISLR	Data for An Introduction to Statistical Learning with Application in R	s 1.0	0
[91] Worl t	"couldn't"	"mustn't"	"let's"	"that's"		NLP	Natural Language Processing Infrastructure	0.1-9	0
[101] "who's"	"what's"	"here's"	"there's"	"when's"		slam	Sparse Lightweight Arrays and Matrices	0.1-38	0
[106] "where's"	"why's"	"how's"	"a"	"an"		tm	Text Mining Package	0.6-2	0
[111] "the"	"and"	"but"	"if"	"or"			ick mining rackage	0.0 2	
[121] Decause	as "by"	"for"	"with"	"about"	Syste	em Library			
[126] "against"	"between"	"into"	"through"	"during"		boot	Bootstrap Functions (Originally by Angelo Canty for S)	1.3-18	0
[131] "before"	"after"	"above"	"below"	"to"		class	Functions for Classification	7.3-14	0
[136] "Trom" [141] "on" [146] "funther"	"off" "then"	"over"	"under"	"again" "there"		cluster	"Finding Groups in Data": Cluster Analysis Extended Rousseeuw et al.	2.0.4	0
[151] "when"	"where"	"why"	"how"	"all"		codetools	Code Analysis Tools for R	0.2-14	0
[156] "any"	"both"	"each"	"few"	"more"		compiler	The R Compiler Package	2.2.1	0
[161] "most"	"other"	"some"	"such"	"no"		complici	The R Compiler Package	2.2.4	
[166] "nor"	"not"	"only"	"own"	"same"		datasets	The K Datasets Package	5.5.1	8
> courpus <- tm_m	iap(courpus,re	moveWords,stop	words("englis	h"))		foreign	Read Data Stored by Minitab, S, SAS, SPSS, Stata, Systat, Weka, dBase,	0.8-66	8
> uim <- Document	renmatrix(co	ur pus j				graphics	The R Graphics Package	3.3.1	0
				*		arDevicer	The R Graphice Devices and Support for Colours and Fonts	221	<u> </u>

Fig.3. Doc format

Delete whitespaces between rows

International Journal of Computer Sciences and Engineering

Vol.-4(9), Sep 2016, E-ISSN: 2347-2693

٥ ×

🔳 Project: (None) 🔻

<u>File Edit Code View Plots Session Build Debug Tools Help</u> 🍳 🗸 🤮 🖌 🔒 📄 🌈 Go to file/function 👘 🔀 🗸 Addins 🗸 6 Environment History 🞯 🔒 📑 Import Dataset 🕶 🔮 R version 3.3.1 (2016-06-21) -- "Bug in Your Hair" Copyright (C) 2016 The R Foundation for Statistical Computing Platform: x86_64-w64-mingw32/x64 (64-bit) 🎒 Global Environment 🕶 Data

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.

> setwd("G://r lng")
> data <- read.csv("reviews.csv",header = TRUE)
> text <- data\$text
> text <- paste(data\$text,collapse = " ")</pre>

RStudio

Console G:/r Ing/ 🔗

> text <- Unitation() text (= Data(dtaftext,collapse = " ") > text [1] " SO ADDICTING DEFF DOWNLAOD ITS EPIC YOU CAT LOVERS WILL FALL IN LOVE <3 Gr eat game I love this game, unlike other games they constantly give you woney to pl ay. They are always given you a bone. Keep up the good work. Sooo much FUN I woul d definitely recommend this game, it's fun for dress up and business. It's extreme ly entertaining, I'm hooked already. :) AMESOME Epic game so addictive Sstars <fo ><U+009F><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+00BD><U+00BD><U+00BD><U+00BD><U+00BD><U+00BD><U+00BD><U+00BD><U+00BD><U+00BD><U+00BD><U+00BD><U+00BD><U+00BD><U+00BD><U+00BD><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009B><U+009



Fig.4.Removing Whitespaces

Remove stopwords

RStudio								- 0) ×
<u>File Edit Code View</u>	Plots Session	<u>B</u> uild <u>D</u> ebug <u>T</u> o	ols <u>H</u> elp						
Q • 🕣 • 🔒 🔒	🚔 🛛 🏕 Go to file,	/function 📄 🔛	 Addins - 					🙁 Proje	ect: (None) 🔸
Console G:/r Ing/				6	Environme	ent History			
Motodata: corpus	specific: 0	document love	1 (indexed):	0	a	mport Dataset	- 4	= 1	list - @
Content: documen	ts: 1	document reve	(Indexed).	0		I Cardona and	1 x	0	
> stopwords("engl	ish")				Bata	i chvironinent •		(~	
[1] "i"	me	''my	"myself"	"we"	Data				
[6] Our	ours "vourself"	"vourselves"	'you "be"	"bim"	🔘 data		1000 obs. of 4 variables		
[16] "his"	"himself"	"she"	"her"	"hers"	Values				
[21] "herself"	"it"	"its"	"itself"	"they"	🔘 courp	ous	List of 1		
[26] "them"	"their"	"theirs"	"themselves"	"what"	o sourc	e	List of 1		
[31] "which"	"who"	"whom"	"this"	"that"	text		" SO ADDICTING DEEE DOWNLADD ITS EPIC YOU	CAT LOVER	S WT
[36] "these"	"those"	"ho"	"heen"	"are"	cente		bo hobiciting bein bonnends ins ente noo	entr corten	
[41] was	"has"	"had"	"having"	"do"	51				-
[51] "does"	"did"	"doing"	"would"	"should"	Files Pl	lots Packages He	eip viewer		
[56] "could"	"ought"	"i'm"	"you're"	"he's"	O, Install	I <u> Q</u> Update	Q.		6
[61] "she's"	"it's"	we're"	"they're"	"i've"	Nam	1e	Description	Version	
[66] you ve	we ve	they ve	1 d	you a	llser Libra	ary.			^
[76] "you'11"	"he'll"	"she'll"	"we'll"	"they'll"	USET LIDIA	ary			
[81] "isn't"	"aren't"	"wasn't"	"weren't"	"hasn't"	ISLR		Data for An Introduction to Statistical Learning with Application:	; 1.0	8
[86] "haven't"	"hadn't"	"doesn't"	"don't"	"didn't"	-		IN K	010	
[91] "won't"	"wouldn't"	"shan't"	"shouldn't"	"can't"			Natural Language Processing Infrastructure	0.1-9	0
[96] "cannot"	"couldn't"	"mustn't"	"let's"	"that's"	slam	n	Sparse Lightweight Arrays and Matrices	0.1-38	0
[101] "who's"	what s	here's	there's	when's	🖂 tm		Text Mining Package	0.6-2	0
[100] where S	"and"	"hut"	d "if"	"or"		-	, ,		
[116] "because"	"as"	"until"	"while"	"of"	System Li	brary			
[121] "at"	"by"	"for"	"with"	"about"	boot	t	Bootstrap Functions (Originally by Angelo Canty for S)	1.3-18	0
[126] "against"	"between"	"into"	"through"	"during"	Class	s	Functions for Classification	7.3-14	0
[131] "before"	"after"	"above"	"below"	"to"		ter	"Finding Groups in Data": Cluster Analysis Extended Pourseeuw	204	0
[136] "Trom"	"off"	"down"	"under"	"again"		cer	et al.	2.0.4	0
[146] "further"	"then"	"once"	"here"	"there"	C code	etools	Code Analysis Tools for R	0.2-14	0
[151] "when"	"where"	"why"	"how"	"all"			The D Council or Desidence	2.2.1	-
[156] "any"	"both"	"each"	"few"	"more"	com	npiler	The R Compiler Package	5.5.1	0
[161] "most"	"other"	"some"	"such"	"no"	data	isets	The R Datasets Package	3.3.1	0
[166] "nor" [171] "so"	"not" "than"	"only"	"own"	"same"	forei	ign	Read Data Stored by Minitab, S, SAS, SPSS, Stata, Systat, Weka,	0.8-66	0
> courpus <- tm m	ap(courpus.ren	novewords.stop	words ("enalis	ih"))			dBase,		
>	- Correspondences				grap	phics	The R Graphics Package	3.3.1	8
				~	- arDe	avicer	The R Graphics Devices and Support for Colours and Fonts	221	0 Y

Fig.5.Removing stopwords

Next, remove nonessential characters such as punctuation, whitespaces, numbers etc from the text, before processing the actual words themselves.

International Journal of Computer Sciences and Engineering

Vol.-4(9), Sep 2016, E-ISSN: 2347-2693

٥ ×

🖄 Project: (None) 🔻 _ List -Ć

_ _ C

 \odot

0

 \odot

 \odot

0

 \odot

 \odot

0

0

 \odot

-

<pre>Bite Eist East year years are provided by East Bedra Control and Device De</pre>	8 RStudio			- 6	1
Q. • Q. • Q. • Q. • Point Methander Point Methander Point Methander Conside Confind / D Conside Confind / D Conside Confind / D Conside Confind / D Verse guess of the point of the po	<u>File Edit Code View Plots Session Build Debug Tools Help</u>				
Conside Carl Ing/ Consider Carl Ing and the State of the	🔍 🗸 🥣 👻 🔒 🦾 🕼 Go to file/function 🔰 🔡 🗸 Addins 🗸			🙁 Proje	ct: (Nor
<pre>please red puyor into two redence dark burracks and years. I've gotten half my school in 2 years. I have been playing for 1 and a half years. I've gotten half my school in the but only the troop try would chock and with all go for git. I you can choose 20 of the same troop to move. Lvl 2 could be 25 troops move. Lvl 3 00 troops move to the tranget. Make it so you upgrade your spell actory to 10 3 costs 800000. Platter, Also new dark barracks into upgrade to 10 3 costs 800000 elixit. Also new dark barracks troop it can in stealth mode when on low health like e conduction to half the troop it can in stealth mode when on low health like e conduction the abstracks and adds 15 room to it max lvl is lvl 1. It costs per troop 300 dark elixit. Also new define, spotter short range unlocks at lvl 1 costs 4000000 gold and you can buy 4 and t's 222 room it does 15 damage per second and can affect only that dark barracks troop and pecka and gole m (firerease hit points of those 2 by 150) it is only single target but very good. T I kinda want my money back coz i've paid (on a bit of stars: 0. Kos oth egits) execute so no professional rewards! 2. I can't watch videos to get free stars or money any exc The nump courtes at vill dees to get free stars or money any exc. associate (i's simple source'' "source'' > courpus < corpus (sectret but when I reinstalled the game (due to glitches), everything was reset to: I kinda want my money back coz i've paid (on a bit of stars: 0. Kos the gits) reduction to Statistical Learning with Applications 1.0 m R MLP Natural Language Processing Infrastructure 0.1-9 m R Mutual Language Processing Infrastructure 0.1-9 m R Mut</pre>		Environment His	istory		_
<pre>Please read supercelling a new toeas dereince dark ourracks and spellic Congress on years, inve been playing for 1 and a haf years, inve gotten haff mischool in the play of the same troop to move, cut 2 could be 25 troops move to the target, wake it so you upgrade your spell factory to lv1 3 30 troops move to the target, wake it so you upgrade your spell factory to lv1 duta 1000 obs. of 4 variables elixtr. Also new dark barracks troop and pekk and got n hall max lv1 1 cost 4000000 gold and you can buy 4 and it's 2x2 room it dues 15 and play upgrade yours and it's 2x2 room it's attack damage is 100 per so the target in the target on your year year it's attack damage is 100 per so durate the play is not the same troop with a stack damage is 100 per so durate the play is not the same troop with a stack damage is 100 per so durate the play is not year year to it's a track damage is 100 per so durate the play is consistent woll is gotten to not wheal is 100 per so durate the play is consistent woll is gotten to not wheal is 2x2 room it due to it's 2x2 room it due to it's and the year gotten with a stack is play is a stack to play the same troop and pekk and got n hall max lv1 is cost 4000000 gold act a would have been synced with Game Center but when I reinstalled the game (due to glitches), everything was reset to is in experimenting are: is is cap't is all in-game center friend and they're dating me, so addictive" attr(, 'class') [11'vectorsource "simpleSource" "source" > courpus < tm.map(courpus, centerturation) > courpus < tm.map(courpus, centerturation) > courpus < tm.map(courpus, stripwhitespace) > courpus <</pre>	best game i ve prayed in a long enne. The developers and a certifie job mente.		port Dataret =		liet -
To the game and they are, Hocks to it. The new spell should be like a target choo Image: Comparison of the same troop to move. Lvl 2 could be 25 troops move. Lvl 1 you can choose 20 of the same troop to move. Lvl 2 could be 25 troops move. Lvl Odata 1000 obs. of 4 variables 3 00 troops move to the target. Make it so you upgrade you's pall factory to lvl of but you can't add another spell, so still 5 spells max. Lab cost for upgrade 8 ml Image: Courpus List of 1 0 data 1000 obs. of 4 variables 30 troops move. To the target. Make the you how health like it so you upgrade you's pall story to lvl of sots 8000000 It is to f 1 0 data 1000 obs. of 4 variables 30 troops move. To the same troop to move. Lvl 2 could be 25 troops move. Lvl 10 troops move to the target. Make the you how health like it to lvl 2 costs 4000000 gold and you can buy 4 and it's 12 sots 100 health and can't do it again. It costs 32500 It to to like the your can't so thow addition of the you do. It is pall the you do. It is attack and gole m (increase hit points of those 2 by 150) it is only single target but very good. Please please please relates are dithis supercell and constacts to orge free stars or momely arrise. The noly available contact is my game center friend and they're dating me, so addictive Is ma beckge heap Processing Infastructure 1.0 recorpus 1.1 can't call in-game center friend and they're dating me, so addictive Is ma beckge phocessing Infastructure 0.1-9 recorpus 1.1 can't call in-game centats or gent estars or momely arrise 1.1 can't call in-game contacts to coe for my pho	2 years, I have been playing for 1 and a half years. I've gotten half my school in		, .	=	LISE *
Ser but only the troop type you choose in your attack and they all go for it. L'J J you can choose 20 of the same troop to move. LVI 2 could be 25 troops move to the target. Make it so you upgrade your spell factory to lVI 3 0 troops move to the target. Make it so you upgrade your spell factory to lVI a 20 health. It has 500 health it takes up 15 room it's attack damage is 100 per second after it heals itself up to 100 health and can't do it again. It costs per troop 300 dark elixir. Also new defence, spotter short range unlocks at lVI 10 to 0 to upgrade dark barracks and adds 15 room to it is 2x2 room it's attack damage is 100 per second and can affect only that dark barracks troop and pekka and gole (increase please please read this supercell and consider it i hope you do. I really i damage per second and can affect only that dark barracks troop and pekka and gole so mo professional rewards 12. I can't lead to oll (itches), everything was reset I kinda want my money back coz I've paid for a bit of stars ok so the glitche s' me xepfiencing are: I. I can't autch videos to get free stars or momey any or e. Awsome It's awsome Ace Such a brilliant game, so addictive" attr(, class") [] Wetcorsource" "source" > courpus < tmmap(courpus, centerUnitation) > courpus < tmmap(courpus, serverVentuation) > courpus < tmmap(cou	to the game and they are, HOOKED to it. The new spell should be like a target choo	Global Environr	ment *	(G	
<pre>1 you can throose 20 of the same troop to move. UP 2 could be 25 troops move. to the target. Make tis by you upprade yours pail f actors pail f actors stated damage is 100 prosent to the same troop and pake tis by you upprade yours to this actors stated damage is 100 prosent to the same troop is and target. Makes the pake when on low health like 20 courpus List of 1 2 courpus dark barracks troop is can in stealth mode when on low health like a damage the second after is about the same functions to 11 as an it. It costs second on the same troop and pake health is also new dark barracks troop and pake have the on low health like a damage the second and the is also new dark barracks troop and pake have the same (locks at 11 low to statistical learning with Applications to 11 as and the same unlocks at 11 low to upprade dark barracks troop and pake have the same (locks at 11 low to upprade dark barracks troop and pake have the same (locks at 11 low to upprade dark barracks troop and pake have the same (locks at 11 low to upprade dark barracks troop and pake have the same (locks at 11 low to upprade dark barracks troop and pake have the same (locks at 11 low to upprade dark barracks troop and pake have the same (locks at 11 low to upprade dark barracks troop and pake have the same (locks at 11 low to upprade dark barracks troop and pake have the same (locks at 11 low to the same the same unlocks at 11 low to upprade dark barracks troop and pake have the same (locks at 11 low to the same the same the same unlocks at 11 low to the same the same the same data would have been synced with a same the same the same the same the same same data would have been synced with a same the same same data would have been synced with a same the same same data would have been synced with a same the same same data would have been synced with a same same data would have been synced with a same same data would have been synced with a same same data would have been synced with a same same data would have been synced with a same same</pre>	ser but only the troop type you choose in your attack and they all go for it. Lvl	Data			
but you can't add another spell, so still 5 spells max. Lab cost for upgrade spell, so still 5 spells max. Lab cost for upgrade spell Values Ulion. To upgrade spell, so still 5 spells max. Lab cost for upgrade spell Output takes up 15 costs \$000000 ell xir. Also new dark barracks troop it can in stealth mode when on low health it is 100 per spectro dark barracks and adds 15 room it max takk damage is 100 per spectro 300 dark ell xir. Also new defence, spotter short range unlocks at 1V1 lows to the add this to ont to than X hui is 1V1 1. It costs 200000 gold and you can buy 4 and the 's 2x2 room it does 15 damage per second and can affect only that dark barracks troop and pekka and gold. Please please read this supercell and consider it 1 hops you do. I really 1 doet 10 and the y teen source 10 and the y teen to the source 10 and the y teen the plus the source 10 and the y teen source 10 and the y teen the plus the source 10 and the y teen the source 10	3 30 troops move to the target. Make it so you upgrade your spell factory to 1v1 6	0 data	1000 obs. of 4 variables		
<pre>llion. To upgrade spell to 1v1 2 costs 5000000. To upgrade to 1v1 3 costs 8000010 elixir. Also new dark barracks troop it can in stealt mode when on low health lit e 20 health. It has 500 health it takes up 15 room it's attack damage is 100 prevent and can't do it again. It costs 32500 00 to upgrade dark barracks and adds 15 room to it max 1v1 is 1v1. It costs 32500 00 to upgrade dark barracks and adds 15 room to it max 1v1 is 1v1. It costs 32500 00 to upgrade dark barracks troop and page unlocks at 1v1 10 tors n hall max 1v1 L costs 4000000 gold and you can buy 4 and it's 2x2 room it does 1 damage persecond and can affect only that dark barracks troop and pake ka and gold m (Increase hit points of those 2 by 10) it is only simple target but very good nove playing this game! Help I thought the game data would have been synced with came center but when I reinstalled the game data would have been synced with came center but when I reinstalled the game data would have been synced with came center only available contact is my game center friend and they're dating me, so o no professional rewards! 2. I can't watch videos to get free stars or money any etc. Alkonem it's awsome it's awsome</pre>	but you can't add another spell, so still 5 spells max. Lab cost for upgrade 8 mi	Values			
<pre>elistic Also hew dark barracks (roop it can in steatch angue with on low health it is takes up is a track dange is 100 per second after it heals itself up to 100 health and can't do it again. It costs per troop 300 dark elistic. Also new defence, spotter short range unlocks at Iv1 is torist is to orbit max iv1 is torist so those 20 kg is 100 and you can buy 4 and its 2x2 room it does is 1 dange per second and can affect only that dark barracks troop and pekka and gole m (increase hit points of those 2 by 150) it is only is not it sax viou to 'up and it's 2x2 room it does is 1 dange per second and can affect only that dark barracks troop and pekka and gole m (increase hit points of those 2 by 150) it is only is not is avaint is out very good. Please please read hit supercell and consider it I do per you do. ir eall have bescription were the provide to take the troop and pekka and gole m is the provide dark barracks to come for my photo sho ots. The only available contact is my game center friend and they're dating me, so addictive" attr(, 'class'') [1] 'vectorsource'' "simpleSource'' source'' source'' source'' source'' source'' source'' source'' source'''s courpus <- compus (courpus, removePunctuation) <- courpus <- tm.map(courpus, remove</pre>	llion. To upgrade spell to lvl 2 costs 5000000. To upgrade to lvl 3 costs 8000000	O courpus	List of 1		
<pre>courpus < Corpus Specific: 0, document level (indexed): 0 courpus < tmmap(courpus, removePunctuation) </pre> text "SO ADDICTING DEFF DOWNLAOD ITS EPIC YOU CAT LOVERS WILL text "SO ADDICTING DEFF DOWNLAOD ITS EPIC YOU CAT LOVERS WILL text "SO ADDICTING DEFF DOWNLAOD ITS EPIC YOU CAT LOVERS WILL text "SO ADDICTING DEFF DOWNLAOD ITS EPIC YOU CAT LOVERS WILL text "SO ADDICTING DEFF DOWNLAOD ITS EPIC YOU CAT LOVERS WILL text "SO ADDICTING DEFF DOWNLAOD ITS EPIC YOU CAT LOVERS WILL text "SO ADDICTING DEFF DOWNLAOD ITS EPIC YOU CAT LOVERS WILL <pre> text "SO ADDICTING DEFF DOWNLAOD ITS EPIC YOU CAT LOVERS WILL </pre> <pre> text "SO ADDICTING DEFF DOWNLAOD ITS EPIC YOU CAT LOVERS WILL </pre> <pre> text "SO ADDICTING DEFF DOWNLAOD ITS EPIC YOU CAT LOVERS WILL </pre> <pre> text "SO ADDICTING DEFF DOWNLAOD ITS EPIC YOU CAT LOVERS WILL </pre> <pre> text "SO ADDICTING DEFF DOWNLAOD ITS EPIC YOU CAT LOVERS WILL </pre> <pre> text "SO ADDICTING DEFF DOWNLAOD ITS EPIC YOU CAT LOVERS WILL </pre> <pre> text "SO ADDICTING DEFF DOWNLAOD ITS EPIC YOU CAT LOVERS WILL </pre> <pre> text "SO ADDICTING DEFF DOWNLAOD ITS EPIC YOU CAT LOVERS WILL </pre> <pre> text "SO ADDICTING DEFF DOWNLAOD ITS EPIC YOU CAT LOVERS WILL </pre> <pre> text "SO ADDICTING DEFF DOWNLAOD ITS EPIC YOU CAT LOVERS WILL </pre> <pre> text "SO ADDICTING DEFF DOWNLAOD ITS EPIC YOU CAT LOVERS WILL </pre> <pre> text "SO ADDICTING DEFF DOWNLAOD ITS EPIC YOU CAT LOVERS WILL </pre> <pre> text "SO ADDICTING DEFF DOWNLAOD ITS EPIC YOU CAT LOVERS WILL </pre> <pre> text "SO ADDICTING DEFF DOWNLAOD ITS EPIC YOU CAT LOVERS WILL </pre> <pre> text "SO ADDICTING DEFF DOWNLAOD ITS EPIC YOU CAT LOVERS WILL </pre> <pre> text "SO ADDICTING DEFF DOWNLAOD ITS EPIC YOU CAT LOVERS WILL </pre> <pre> text "SO ADDICTING DEFF DOWNLAOD ITS EPIC YOU CAT LOVERS WILL </pre> <pre> text "So ADDICTING DEFF DOWNLAOD ITS EPIC YOU </pre> <pre> text "So ADDICTING DEFF DOWNLAOD</pre>	e 10 health. It has 500 health it takes up 15 room it's attack damane is 100 per s	0 source	List of 1		
00 to upgrade dark barracks and adds 15 room to it max lvl is lvl 1. It costs per troop 300 dark elixit. Also new defence, spotter short range unlocks at lvl 10 to toots per second and can affect only that dark barracks troop and pekka and gots are second and can affect only that dark barracks troop and pekka and gots are second and can affect only that dark barracks troop and pekka and gots are second and can affect only that dark barracks troop and pekka and gots are second in super cell and consider it i hope you do. I really it does 15 Please please please read this super cell and consider it i hope you do. I really it ove playing this game! Help! I thought the game data would have been synced with Game center but when I reinstalled the game (due to glitches), everything was reset. I. Kinda want my money back coz I've paid for a bit of stars ok so the glitch is a server that its awsome accessional rewards! 2. I can't watch videos to get free stars or money anym ore. Awsome It's awsome Acce Such a brilliant game, so addictive" Istal Were Ill "vectorsource" "SimpleSource" "Source" Spate library > courpus <- Corpus (source)	econd after it heals itself up to 100 health and can't do it again. It costs 32500	text	" SO ADDICTING DEFF DOWNLAOD ITS EPIC YO	U CAT LOVER	S WI
Tho a Jum Arrive Links A new defence, sporter short range unlocks at VI 10 tots In hall max IV1 1 costs 400000 gold and you can buy 4 and it's 2x2 room it does is damage per second and can affect only that dark barracks troop and pekka and gole m (increase hit points of those 2 by 150) it is only single target but very good. Please please read this supercell and consider it I hope you do. I really it ove playing this game! Help I thought the game (date would have been synced with can be been synced with a want my money back to zo I've paid for a bit of stars Ok so the glicter is a wome new arest it. I kind awant my money back to zo I've paid for a bit of stars Ok so the glicter is a wome Ace Such a brilliant game, so addictive? start (, class') ISLR Data for An Introduction to Statistical Learning with Applications 1.0 with the class - a conset is a wasome Ace Such a brilliant game, so addictive? NLP Natural Language Processing Infrastructure 0.1-9 if "vectorsource" "simpleSource" "source" source" 0 boot Boottrap Functions (Originally by Angelo Canty for S) 1.3-18 courpus <- th_map(courpus, stripwhitespace)	00 to upgrade dark barracks and adds 15 room to it max lvl is lvl 1. It costs per				
damage per second and can affect only that dark barracks troop and pekka and gole m (increase hit points of those 2 by 150) it is only single target but very good. Drease please please read this supercell and consider it I hope you do. I really 1 ove playing this game! Help I thought the game data would have been synced with Game Center but when I reinstalled the game (due to glitches), everything was rese t. I kinda want my money back coz I've paid for a bit of stars. ok so the glitch s I'm experiencing are: 1. I can't call in-game contacts to come for my photo sho ots The only available contact is my game center friend and they're dating me, so on op rofessional reverds! 2. I can't watch videos to get free stars or money anym ore. Awsome It's awsome Ace Such a brilliant game, so addictive" attr(, "class") [1] "VectorSource" "SimpleSource" "Source" > courpus <- Corpus (source) > courpus <- tm_map(courpus, content_transformer(tolower)) > courpus <- tm_map(courpus, stripwhitespace) > courpus <- tm_map(courpus, stripwhitespace)	n hall max lvl 1 costs 4000000 gold and vou can buy 4 and it's 2x2 room it does 15	Files Plots Pa	ackages Help Viewer		-
m (increase hit points of those 2 by 150) it is only single target but very good. Name Description Version Please pl	damage per second and can affect only that dark barracks troop and pekka and gole	🧾 Install 🗕 🖸 U	Jpdate	2	
Please please flease flease flease flease flease this supercell and consider it I hope you too. I really if ove playing this game? Help I thought the game data would have been synced with Game Center but when I reinstalled the game (due to glitches), everything was rese to response to the synchrony available contact is my game center friend and they're dating me, so no professional rewards! 2. I can't watch videos to get free stars or money anym ore. Awsome It's awsome Ace Such a brilliant game, so addictive" User library user library IsIR Data for An Introduction to Statistical Learning with Applications 1.0 in R on oprofessional rewards! 2. I can't watch videos to get free stars or money anym ore. Awsome It's awsome Ace Such a brilliant game, so addictive" NLP Natural Language Processing Infrastructure 0.1-9 attr(, "class") [1] "vectorsource" "simplesource" "source" 0.6-2 System library > courpus <- corpus (source)	m (increase hit points of those 2 by 150) it is only single target but very good.	Name	Description	Version	
Game Center but when I reinstalled the game due to glitches), everything was reset L. I kinda want my money back coz i ve paid for a bit of stars ok so the glitch s I'm experiencing are: 1. I can't call in-game contacts to come for my photo sho ots The only available contact is my game center friend and they're dating me, so on op professional rewards! 2. I can't wardth videos to get free stars or money anym ore. Awsome It's awsome Ace Such a brilliant game, so addictive" attr(,"class") [1] "vector Source" "Source" > courpus <- Corpus (source) > courpus secific: 0, document level (indexed): 0 contents: 1 > courpus <- tm_map(courpus, content_transformer(tolower)) > courpus <- tm_map(courpus, stripwhitespace) > courpus <- tm_map(courpus, stripwhitespace) > courpus secific: 0, document level (indexed): 0 courpus <- tm_map(courpus, stripwhitespace) > courpus secific: 0, document level (indexed): 0 content: documents: 1 > courpus secific: 0, document level (indexed): 0 courpus <- tm_map(courpus, stripwhitespace) > courpus secific: 0, document level (indexed): 0 content: documents: 1 > courpus secific: 0, document level (indexed): 0 content: documents: 1 > courpus secific: 0, document level (indexed): 0 content: documents: 1 > courpus secific: 0, document level (indexed): 0 content: documents: 1 > courpus secific: 0, document level (indexed): 0 content: documents: 1 > courpus secific: 0, document level (indexed): 0 content: documents: 1 > courpus secific: 0, document level (indexed): 0 content: documents: 1 > courpus secific: 0, document level (indexed): 0 content: documents: 1 > courpus secific: 0, document level (indexed): 0 content: documents: 1 > courpus secific: 0, document level (indexed): 0 content: documents: 1 > courpus secific: 0, document level (indexed): 0 content: documents: 1 > courpus secific: 0, document level (indexed): 0 content: documents: 1 > courpus secific: 0, document level (indexed): 0 content: documents: 1 > courpus secific: 0, document level	ove playing this game! Help! I thought the game data would have been synced with	User Library			
estime experiencing affeit 1. I can't call in-game contacts to come for my photo shore NLP Natural Language Processing Infrastructure 0.1-9 ots The only available contact is my game center friend and they're dating me, so on professional rewards! 2. I can't watch videos to get free stars or money anym isam Sparse Lightweight Arrays and Matrices 0.1-38 attr(, "class") [1] "vectorsource" "simplesource" "source" 0.6-2 0 > courpus <- corpus (source)	Game Center but when I reinstalled the game (due to glitches), everything was rese t. I kinda want my money back coz I've paid for a bit of stars. Ok so the glitch	ISLR	Data for An Introduction to Statistical Learning with Applicati in R	ons 1.0	8
<pre>o no professional rewards! 2. I can't watch videos to get free stars or money anym ore. Awsome It's awsome Ace Such a brilliant game, so addictive" attr(, 'class'') (]1 "vectorSource" "simpleSource" "source" > courpus <- corpus (source) > courpus <- corpus specific: 0, document level (indexed): 0 Content: documents: 1 > courpus <- tm_map(Courpus, content_transformer(tolower)) > courpus <- tm_map(Courpus, stripwhitespace) > cou</pre>	ots The only available contact is my game center friend and they're dating me. s	NLP	Natural Language Processing Infrastructure	0.1-9	8
ore. Awsome It's awsome Ace Such a brilliant game, so addictive"	o no professional rewards! 2. I can't watch videos to get free stars or money anym	slam	Sparse Lightweight Arrays and Matrices	0.1-38	0
attr("class") System Library [1] "vector Source" "SimpleSource" "Source" boot Bootstrap Functions (Originally by Angelo Canty for S) 1.3-18 > courpus courpus courpus courpus 7.3-14 Courpus > courpus courpus courpus 7.3-14 Courpus > courpus courpus courpus 7.3-14 Courpus > courpus courpus courtent: 1 courpus courpus 7.3-14 Courpus > courpus courpus content: documents: 1 courpus	ore. Awsome It's awsome Ace Such a brilliant game, so addictive"	🗹 tm	Text Mining Package	0.6-2	8
[1] VectorSource Simplesource Simplesource Source Source Source Simplesource Simplesource Simplesource Simplesource Source Source 1.3-18 > courpus Courpus (Source) 1.3-18 > courpus Courpus (Source) 1.3-18 > courpus Cass Functions for Classification 7.3-14 > courpus specific: 0, document level (indexed): 0 Courpus (Courpus, content_transformer(tolower)) 2.0.4 1 > courpus <- tm_map(courpus, content_transformer(tolower))	attr(,"class")	System Library			
<pre>> courpus - courpus specific: 0, document level (indexed): 0</pre>	<pre>> courpus <- Corpus(source)</pre>	🔲 boot	Bootstrap Functions (Originally by Angelo Canty for S)	1.3-18	ø
<pre><vvcorpus>> </vvcorpus></pre> <pre>(corpus systerific: 0, document level (indexed): 0 </pre> <pre>Content: documents: 1 </pre> <pre> courpus <- tm_map(courpus,content_transformer(tolower)) > courpus <- tm_map(courpus,setripwhitespace) > courpus <- tm_map(courpus,setripwhitespace) > courpus </pre> <pre> content: documents: 1 </pre> <pre> codetols codetols codetols codetols codetols codetols compiler The R Compiler Package</pre>	> courpus	class	Functions for Classification	7.3-14	8
<pre>> courpus <- tm_map(courpus,content_transformer(tolower)) > courpus <- tm_map(courpus,removePunctuation) > courpus <- tm_map(courpus,stripwhitespace) > courpus <- tm_map(courpus,stripwhitespace) > courpus>> Metadata: corpus specific: 0, document level (indexed): 0 Content: documents: 1 </pre>	<vvcorpus>> Metadata: corpus specific: 0, document level (indexed): 0 Content: documents: 1</vvcorpus>	cluster	"Finding Groups in Data": Cluster Analysis Extended Rousseeu et al.	w 2.0.4	8
<pre>> courpus <- tm_map(courpus,removePunctuation) > courpus <- tm_map(courpus,stripwhitespace) > courpus < curpus < <vcorpus>> Wetadata: corpus specific: 0, document level (indexed): 0 Content: documents: 1 > </vcorpus></pre>	<pre>> courpus <- tm_map(courpus,content_transformer(tolower))</pre>	codetools	Code Analysis Tools for R	0.2-14	8
<pre>> courpus <- tim_map(courpus,stripwritespace) > courpus < curpus <- tim_map(courpus,stripwritespace) > courpus < cvcorpus>> < cvcorpus>> < cvcorpus specific: 0, document level (indexed): 0 Content: documents: 1 > </pre>	<pre>> courpus <- tm_map(courpus,removePunctuation)</pre>	compiler	The R Compiler Package	3.3.1	8
<pre><</pre> <pre></pre> <pre> </pre> <pre> <pre> <pre> <pre> <pre> <pre> <pre> <pre> <pre> <pre> <pre> <pre> <pre> <pre> <pre> <pre> <pre> <pre> <pre> <pre> <pre> <pre> <pre> <pre> <pre> <pre> <pre> <pre> <pre> <pre> <pre> <pre> <pre> <pre> <pre> <pre> <pre> <pre> <pre> <pre> <pre> <pre> <pre> <pre> <pre> <pre> <pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre></pre>	> courpus <- cm_map(courpus,scripwrrcespace) > courpus	datasets	The R Datasets Package	3.3.1	8
Content: accuments: 1 Image: second secon	< <vcorpus>> Metadata: corpus specific: 0, document level (indexed): 0</vcorpus>	foreign	Read Data Stored by Minitab, S, SAS, SPSS, Stata, Systat, Weka dBase,	, 0.8-66	8
TarDepicer The R Granhier Depicer and Support For Colours and Easter 331	Content: documents: 1	graphics	The R Graphics Package	3.3.1	8
	×	- orDevicer	The R Graphics Devices and Support for Colours and Fonts	331	

Fig.6. Removing nonessential characters

Denote as matrix and then Find Frequency of words, sort it and see head

RStudio File Edit Code Vi	au Plots Service	Ruild Dahug	fools Help					- ć	J ×
Q ∰ -	🔋 📇 🚺 🍌 Go to file	/function	Addins +					🔳 Proje	ect: (None) 🔹
Console G:/r Ing/	>			Ð	Envir	onment History			_
[21] "herself"	"it"	"its"	"itself"	"they"		Import Datas	et - 💰	=	List - 📿
[26] "them"	"their"	"theirs"	"themselves"	"what"		lobal Environment -	-	0	
[31] "which"	"who"	"whom"	"this"	"that"	Dur				
[36] "these"	"those"	"am"	"is"	"are"	🚺 🔘 da	ita	1000 obs. of 4 variables		^
[41] Was [46] "have"	"has"	"bad"	"baving"	"do"	dt	:m2	num [1, 1:2860] 1 1 1 2 1 1 11 4 1 1		
[51] "does"	"did"	"doina"	"would"	"should"	valu	ues			
[56] "could"	"ought"	"i'm"	"you're"	"he's"	0.00	urnus	List of 1		
[61] "she's"	"it's"	"we're"	"they're"	"i've"	Odt		List of 6		_
[66] 'you've'	"we've"	"tney ve	"they'd"	you d	Gut				100
[76] "vou'11"	"he'11"	"she'11"	"we'll"	"they'11"		equency	Named num [1:2860] 917 249 241 236 222 149	130 130	128 v
[81] "isn't"	"aren't"	"wasn't"	"weren't"	"hasn't"	Files	Plots Packages	Help Viewer		
[86] "haven't'	"hadn't"	"doesn't"	"don't"	"didn't"	- Inc.	The Tuckages			
[91] "won't"	"wouldn't"	"shan't	"shouldn't"	"can't"		nstall 🔮 Update	<u> </u>		
[90] Cannot	"what's"	"here's"	"there's"	"when's"		Name	Description	Version	
[106] "where's'	' "why's"	"how's"	"a"	"an"	User	Library			^
[111] "the"	"and"	"but"	"if"	"or"		ISLR	Data for An Introduction to Statistical Learning with Application	x 10	
[116] "because'	"as"	"until"	"while"	"of"		IDEN.	in R	3 1.0	
[121] "at" [126] "aesiset"	"by"	"for"	"With"	"about"		NLP	Natural Language Processing Infrastructure	0.1-9	8
[131] "before"	"after"	"above"	"below"	"to"		clam	Sparce Lightweight Arrays and Matrices	0.1-38	0
[136] "from"	"up"	"down"	"in"	"out"			T INC I D I	0.1 50	
[141] "on"	"off"	"over"	"under"	"again"		tm	lext Mining Package	0.6-2	8
[146] "further'	'"then"	"once"	"here"	"there"	Syste	m Library			
[151] "when" [156] "ppv"	"where"	"why" "oach"	"how"	"all"	Ċ.	boot	Bootstrap Functions (Originally by Angelo Canty for S)	1.3-18	0
[161] "most"	"other"	"some"	"such"	"no"			Eventions for Classification	7.2.14	
[166] "nor"	"not"	"only"	"own"	"same"		ciass	Functions for classification	7.5-14	
[171] "so"	"than"	"too"	"very"			cluster	"Finding Groups in Data": Cluster Analysis Extended Rousseeuw	2.0.4	8
> courpus <- tr	n_map(courpus,re	moveWords,sto	pwords("englis	h"))		and shared a	et al. Code Application Teolo fee D	0.2.14	
> dtm2 <- as ma	enciernmatrix(com atrix(dtm)	urpus)				codecools	Code Analysis loois for R	0.2-14	
> frequency <-	colSums(dtm2)					compiler	The R Compiler Package	3.3.1	8
> frequency <-	sort(frequency,	decreasing =	TRUE)			datasets	The R Datasets Package	3.3.1	8
> head(frequence game great f	y) un good love	get				foreign	Read Data Stored by Minitab, S, SAS, SPSS, Stata, Systat, Weka, dBase,	0.8-66	8
91/ 249 2	241 236 222	149				graphics	The R Graphics Package	3.3.1	8
1				~		orDevicer	The R Granhics Devices and Sunnort for Colours and Fonts	221	~ v

Fig.7. Freq of Words

International Journal of Computer Sciences and Engineering

Vol.-4(9), Sep 2016, E-ISSN: 2347-2693

2 - 🚭	•		Go to	file/function Addins -							🖄 P	roject: (Nor	1e) -
Heart	× 🔲	data ×				Enviro	onment	History				-	
	2 8	Filter			(Q	1		Import Dataset 🗸 🛛 🎻				List -	C
	x	rating $\hat{}$	location $^{\circ}$	text \diamond		🐴 GI	obal Envir	ronment -			Q		
1	1	5	us	SO ADDICTING DEFF DOWNLAOD ITS EPIC YOU CAT L	^	Data							
2	2	5	us	Great game I love this game. Unlike other games they		0 dat	ta	1000 c	bs. of 4 varial	les			1
3	3	5	nz	Sooo much FUN I would definitely recommend this g		Valu	es						
4	4	5	gb	AWESOME Epic game so addictive 5stars <f0><u+00< td=""><td></td><td>te</td><td>xt</td><td>Factor</td><td>w/ 986 levels</td><td>" !!!!! It's abo</td><td>out time I ca</td><td>in log</td><td></td></u+00<></f0>		te	xt	Factor	w/ 986 levels	" !!!!! It's abo	out time I ca	in log	
5	5	5	us	Good game Great game									
6	6	1	nz	Meh Used to be good until World Cup upgrade. Now i									
7	7	3	nz	Not bad <u+6e38><u+620f><u+4e0d><u+9519><u< td=""><td></td><td>Files</td><td>Plots</td><td>Packages Help Viewe</td><td>r –</td><td></td><td></td><td>_</td><td></td></u<></u+9519></u+4e0d></u+620f></u+6e38>		Files	Plots	Packages Help Viewe	r –			_	
8	8	5	nz	Clash of clans Clash Is a awesome game				Zoom Zoom Export - 0					C
9	9	4	au	Cool Great game A+A+A									
10	10	4	nz	Candy crush review Addictive									
11	11	4	us	Great game Great game									
12	12	4	us	Fun but I have a disagreement Whenever you try to p	~								
showing 1	to 12 of	1,000 ent	ries										
Console	G:/r Ing/	~ 			-0								
You are Type 'l R is a	e welco license collab	ome to r e()' or porative	edistrib 'licence project	ute it under certain conditions. ()' for distribution details. with many contributors.	Ŷ								
Type 'c 'citati	ion()'	on how	to cite	re information and R or R packages in publications.	- 1								
Type o 'help.s Type o	tart()	for so for a quit F	ome demos an HTML b R.	rowser interface to help.									
> setwo > data > text >	i("G:// <- rea <- dat	r lng"; d.csv(a\$text) 'reviews.	csv",header = TRUE)									

Retrieve data from dataset





RStudio		– 0 ×
<u>File Edit Code View Plots Session Build Debug Tools Help</u>		
🔍 🔹 🚰 👻 🔜 📄 🧼 Go to file/function 🛛 🔡 👻 Addins 🗸		🔳 Project: (None) 🚽
Console G:/r Ing/	Environment History	
[26] "thom" "thoir" "thoirs" "thomsolves" "who	" A 🔿 🕞 Planport Data	set v 🖌 📃 list v 📿
[31] "which" "who" "whom" "this" "that		
[36] "these" "those" "am" "is" "are		
[41] "was" "were" "be" "been" "bei [46] "bayo" "bac" "bad" "bayon" "de"	g" 🔘 data	1000 obs. of 4 variables
[40] have has had having do	1d" dtm2	num [1, 1:2860] 1 1 1 2 1 1 11 4 1 1
[56] "could" "ought" "i'm" "you're" "he'	" Values	
[61] "she's" "it's" "we're" "they're" "i'v	O courpus	List of 1
[66] 'you've "we've they've 1'd you [71] "be'd" "sbe'd" "we'd" "they'd" "i'l	d dtm	List of 6
[76] "you'll" "he'll" "she'll" "we'll" "the	'11" frequency	Named num [1:2860] 917 249 241 236 222 149 136 136 128
[81] "isn't" "aren't" "wasn't" "weren't" "has	't"	
[86] "haven't" "hadn't" "doesn't" "don't" "did	t" Files Plots Packages	Help Viewer
[91] Won't Wouldn't Shan't Shouldn't Can [96] "cappot" "couldn't" "mustn't" "let's" "tha	's" 200m 2	Evport y 🗿 🖌 🙃 Publish 🥥
[101] "who's" "what's" "here's" "there's" "whe	's"	
[106] "where's" "why's" "how's" "a" "an"		
[111] "the" "and" "but" "if" "or"		
[116] "because" "as" "until" "while" "of"		
[121] "at" "by" "for" "with" "abo	t"	
[126] against between into through dur	ng	
[131] Defore after above below to		
[141] "on" "off" "over" "under" "aga	n" 00	
[146] "further" "then" "once" "here" "the		
[151] "when" "where" "why" "how" "all		
[156] "anv" "both" "each" "few" "mor		
[161] "most" "other" "some" "such" "no"	9	
[166] "nor" "not" "only" "own" "sam		
[171] "so" "than" "too" "very"		
<pre>> courpus <- tm_map(courpus,removewords,stopwords("english"))</pre>		
<pre>> dtm <- DocumentTermMatrix(courpus)</pre>	N N	
<pre>> dtm2 <- as.matrix(dtm)</pre>		
> frequency <- colsums(dtm2)		
<pre>> frequency <- sort(frequency,decreasing = TRUE)</pre>		
> nead(Trequency)	game	e great fun good love get
017 240 241 236 222 140	3	5
> harplot (head(frequency))		
>		
>		

Fig.9. Graph representation

V.CONCLUSION

R is used to mine unstructured data which is the most exhaustive statistical analysis package and it incorporates all of the standard statistical tests, models and analyses for managing and manipulating data. By using R only useful information can be gathered by removing unnecessary nonessential characters from unstructured data. So, we can easily predict the status of firm retrieving header frequency from unstructured data.

REFERENCES

- Mr. Rahul Patel, Mr. Gaurav Sharma,"A survey on text mining techniques", Int. Journal of Engineering and Computer Science, Volume-03, Issue 5, Page No. (5621-5625), May 2014.
- [2] Minakshi R. Shinde1, Parmeet C. Gill, "Pattern Discovery Techniques for the Text Mining and its Applications", Int. Journal of Science and Research (IJSR), Volume-03 Issue 5, May 2014.
- [3] S.S. Patil and V.M. Gaikwad, "Developing New Software Metric Pattern Discovery for Text Mining", International Journal of Computer Sciences and Engineering, Volume-02, Issue-04, Page No (119-125), Apr -2014
- [4] Abhilasha Singh Rathor, Dr. Pankaj Garg," Analysis on Text Mining Techniques", Int. Journal of Advanced Research in Computer Science and Software Engineering, Volume -06, Issue 2, Page No (132-137), February 2016.
- [5] Vishakha D. Bhope and Sachin N. Deshmukh, "Comparative Study on Information Retrieval Approaches for Text Mining", Int. Journal of Computer Sciences and Engineering, Volume-03, Issue-3, Page No (102-106), Mar 2015.
- [6] Vishakha D. Bhope and Sachin N. Deshmukh, "Comparative Study on Information Retrieval Approaches for Text Mining", International Journal of Computer Sciences and Engineering, Volume-03, Issue-03, Page No (102-106), Mar -2015
- [7] StatisticalModeling,https://en.wikipedia.org/wiki/Statistical_m odel, June 2016.

AUTHORS PROFILE

M. Siva Lakshmi is presently B.Tech Student, Dept. of Computer Science & Engineering, NRI Institute of Technology, Pothavarappadu, India.

MD. Arsha Sultana is presently Assistant Professor, Dept. of Computer Science & Engineering, NRI Institute of Technology, Pothavarappadu, India.



