## Marketing Analytics Professor Swagato Chatterjee Vinod Gupta School of Management Indian Institute of Technology, Kharagpur Lecture 53 Text Mining and Sentiment Analytics (Contd.)

Hello everybody welcome to Marketing Analytics course, this is Dr. Swagato Chatterjee from VGSOM, IIT Kharagpur who is taking this course. We are in week 10 and this is session 3 we are discussing about Text Mining in the context of marketing.

(Refer Slide Time: 0:32)

I that				- 0 X
He Edit Code View Picts Senior Build Debug Pichle Cols Help 0 • 00 @* • 10 @ # • 11 @ # 0 catalant-senior 10 • Addes •				Thurst date
(F) (T), basic, 1 der, code A -	0=	Environment History Connections		-0
Concrete See A A      Concrete See A	-+ Run ++ -+ Source + - 2: A	🤓 🔒 🔐 Import Datant + 🔏		Q. (1)
they be "Approximation constitution incomptential superior before states to a set test - real loss (Hered) test, or a loss (Hered) test, or (Hered) effects (Lables(test, file - indicates (Hered)) effects (Lables(test, file - indicates (Hered)) effects (Lables(test, file - indicates (Hered))	HAR-Speech TYT"		Environment is energy	
13 tetr,1 ← read.cw('test_df.csv') 14 rest_lises 17 rest_lises 18 rest_lises 18 ded.cspps(weterborce(test_litest)) 19 ded.cspps(weterborce(test_litest))				
22 chapter the units of the doubt theory doubt theory doubt theory doubt theory doubt doubt theory of the doubt doubt theory of theory of the doubt theory doubt theory of the doubt theory doubt theory of the doubt theory of the doubt theory of the doubt theory of the doubt the doubt the doubt the doubt the doubt the doubt the doubt the doubt the doubt the doubt the doubt th	Riner r	Film Fully Facages Rep Vener		5a
Conside Imminut				
Console Revenue : C/Oners/RPTID Budie-B/Direkting/Birektil/Institut I/ -				
stol("Ci/unr/orts.tsofi-st/inistingendicionion")")				
🖬 Sand Newsland Window. 📕 🖨 🤌 👂 🗐 🗐 🗗 🦉	E # = 0 2	0		

So, if you see the file there is something called basic TDMcode.r, I have to thank my friend, Mr. Saptarsi Sarkar for sharing this this file with me, the actual file has been made by him but I am using this for the course and I think, he has given me the permission.

So the first thing in this particular thing, but in this particular class, what we are trying to do is that we will play with some text data and this text data is available online, freely available text data, and it is about Martin Luther King, Martin Luther King, I Have a Dream speech.

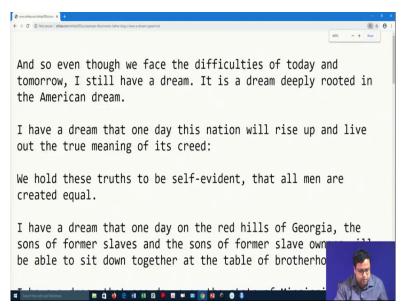
So that particular speech is what is there in the text and we will try to analyze that. So to do that, there are certain libraries that you require, you require a library called tm then Snowballc, wordcloud and RColorbrewer, we will discuss about them one by one. So first things, we have to check that whether these libraries are already installed or not. So I can see probably tm library, you can see here that tm library is not installed in my this thing. So I will try to call this library and it will say that error there is no package called tm.

So what I will I am going to do now in the next probably 2, 3 minutes, is to install all these packages so install .packages then tm , so we have to install quite a few package. So tm what else, Snowballc, so S capital, Snowballc, and then the next one is wordcloud, so wordcloud and the next one is RColorbrewer, so let us just install these files. So it might take some time depending on the internet connection.

So tm library is the classic library that we will be using. And they have, it has dependency called nlp and slam, so it has downloaded that, and the other ones have also been downloaded. So if you see that nlp, slam, tm, Snowballc, wordcloud and Colorbrewer all have been downloaded. So if I now call these four libraries together, I think they will work. So some of them saying that was built under 3.6.3, so tm built was 3.6.3, so you have to check whether your R version is compatible to this, but warning sign means it is okay we can go ahead.

So, I have till now what I did is, I install certain libraries, then what I will do? Then I will actually is not set working directory. So this I will delete and then I will set working directory to my source file location wherever this file is currently residing this port file and then I am putting this thing at file path is called to this particular link.

(Refer Slide Time: 4:06)



So if I just go to the link, what does it look like? Let us just check that. So if I just go to this link, so in this link, this whole text is there, I can probably make it bigger. And okay so, and so even though we face the difficulties of today and tomorrow I still have a dream it is a dream deeply rooted in the in the American dream and so on. So the great American Dream

Martin Luther King's speech, I have a dream that one day this nation will rise up and leave out the true meaning of its creed blah, blah, blah, is given that particular thing is written in this particular.

So we can read a text, it could have been some other file also. So there are different ways to read files from internet, text we can directly do that.

(Refer Slide Time: 4:52)

0       0	8 Code				- 0 X
<pre>Plot is an interval Plot is an interval Plot</pre>	File Edit Code View Plots Session Build Debug Profile Tools Help				
<pre>     () Concasts ( A + 1)</pre>	● + ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ●				📱 Pryett disnet -
A line of the second s	0 01, basic, 164, code 8 -	-0	Environment History Connections		-0
<pre>Visit intervision intervi</pre>		- Source + 2	💣 🔒 🕐 Import Dataset + 🧃		ii be - (ji
<pre>iter = the set of the set of</pre>		<u>^</u>	🐴 Global Environment +		Q
<pre>int is - not int int - not int int - not int int - not int int int int int int int int int in</pre>			values		
<pre>provide the set of the set o</pre>	4 19brary(Rcolortrever)				ole-files/martin-luther-king-i-have-a.
[2] Comparing the under of the Auduated (1) the Under Section (1) and Data (1), (1) and (1)	test - reaking(f)=rest) test, - reaking(f)=rest, (f) test, cortest, (f) = rest, (f) = re		test	ina (Tree)	
<pre>1                                       </pre>			First Parks Hall Mark		-0
<pre>1    def = texture interference (interference (interference)) 2    def = texture interference (interference) 2    def = texture interference 2    def = texture interference (interference) 2    def = texture interference 2    def = te</pre>	23 (nspect(docs)				HU.
if is a - backwidter, totage, "(i)";         if is a - backwidter, totag	24 toSpace <- content_transformer(function (x , pattern ) gsub(pattern, " ", x)) 35 dec to exclose toforma """				
Image: Section of the section of the section will rise and the section rest."	26 docs - tm_map(docs, toSpace, "8")				
Image: Section of the section of th	<pre>27 docs &lt;- tm_map(docs, toSpace, "\\\")</pre>				
11       A series values         12       The series values         12       The series values         13       The series values         14       The series values         15       The series values         15       The series values         15       The series values         16       The series values         17       The series values         18       The series values         18       The series values values         19       The series values values values values         10       The series values values values values values         10       The values values values values values values         10       The values values values values values values values values values         18       The values	29				
(1)     (1)     (1)       (2)     (1)     (1)       (2)     (1)     (1)       (1)     (1)     (1) <td><pre>10 docs &lt;- tm_map(docs, content_transformer(tolower))</pre></td> <td></td> <td></td> <td></td> <td></td>	<pre>10 docs &lt;- tm_map(docs, content_transformer(tolower))</pre>				
Control     Section       1     Interface     Section	17 dors - re-en-dors removationhers)	~			
Compared the interface interface in the integration of the integration	kta /http://www.it	R Sorget 2			
> Indef Charles Virtus Indef & Restands Week (Index Virtus Index ) > Indef Charles Virtus Indef Add Restands Week (Index Adop - Adope - Adop	Consule Terminal				
<ul> <li>I first is a "true", first, index calls that an an influence of the inter the inter-section -section -section -section -section -section -section -section -section - section - section</li></ul>					
the serior draw."  (1) " (2) " there a draw that one day this nation will rise up and The out the true maning of its orend:"  (3) " there a draw that one day this nation will rise up and The out the true maning of its orend:"	> filePath <- Thtp://www.sthda.com/sthda/HOpc/example-files/martin-luther-king-i-have-a-dream-speech.txt" > text <- read.ins(filePath) > text	Î			
[2] 'I have a drawn that one day this nation will rise and the act the true manning of its creed."	[2] "And so even though we face the difficulties of today and tomorrow, 1 still have a dream. It is a dream deep the American dream,"	bly rooted in			
	01 · · ·				-
	[4] "I have a dream that one day this mation will rise up and live out the true meaning of its creed:"				
(1)	(5)				
fill San hald share reache as he calif actions, shot all sans use researed each?	fill has hald share reather to be cald addapt show all one are reasted anoil "	Y			
🖬 Scretch the web and Mindows 🐘 🗿 🍪 🤮 📴 🔢 🖉 🦉 📓 🖬 🗰 🗰 🚳 👰 🖉 🌑 🛔	📽 Search the web and Windows 💼 🖨 🍪 🤮 🐏 👪 🖾 💆 💷 🗰 🚥	9 🛙	6 🔘 🚯		A Photo A

So I can I have to give file path. So file path is basically the link of this particular data. That is my file path and the text is read lines with lines fields line by line it will read, so read lines file path.

And that gives, if I now see the text, it will see the line by line has it has written. So, the first line was blank and then, so even though we face the difficulties of today, tomorrow, I shall have a dream that comes in the second line, third line was also blank, because if you remember if you if you remember how the text was there, you will see that the first line was blank.

And then second line onwards, there was some text is there here it was blank. There is some text then one blank then some text then again blank, then again text, so one single line by line it will just try to read. So whenever you press an enter in the text, when whoever created this text document whenever he has pressed and enter, another line got generated. So if that line was nothing written in that, then that line is blank, and then you go to the next line. So that is how we did this thing. So, I have some blanks and some text in this particular data.

Next work, next I will change this text data which is a character variable to a data frame. So as data frame text, text df = ask a different text. It is nothing but one single observation the same things with some blank values and some text values are there, why we do that? Because I can now save this.

So I will save this in my file wherever the files are stored. I can save this so that the next time by chance the internet is off, and I want to do this thing once more, I do not have to read the file again from internet once more because if internet is not there, then there is a problem. So this is how the data is now saved here, text\_df.

(Refer Slide Time: 7:16)

	LI'LE	NERT	add i av	NIT 10	DATE AT	8451	and an	VIDA								Activation Fail													-
		-our																											
	* 1		fi																										
A	1	c	0	1		1.6	5	N 1	1		κ.	0	м	N	0		Q		5	1	U	V.	W	х	×.	2	AA	AB	M
	text																												
	1																												
	2 And so	even thoug	h we fac	e the diffic	ulties of to	day and	tomorrow	r, i still have a	a dream. P	t is a dre	am dee	ply rootes	I in the An	verican de	eatt.														
	Albert	draam that	eee da	this outline	will rise i	n and by	a out the	true meanin	a of 21.000	-																			
	5	Cream tra		COLO TRADUC			in cost con		Tourse																				
	6 We hole	these trut	ts to be	self-evider	t, that all s	ten are	created e	oual.																					
	7																												
	8 I have a	dream that	one da	on the red	hills of Ge	orgia, th	e sons of	former slave	s and the	sons of t	ormer s	lave own	es will be	able to sit	t down tog	pether at th	table of I	brotherho	od,										
	9																												
		dream that	one da	even the r	nate of Mi	aissippi,	a state p	weltering with	th the heat	t of injust	tice, tw	eltering	with the he	at of opp	ression, w	ill be transf	cented into	o an casis-	of freedor	n and justs	ce.								
	11																												
		dream that	my fou	little child	ren will or	e day liv	e in a nat	ion where the	ey will not	t be judy	red by th	the color o	their skin	but by th	e content	of their cha	racter.												
	13																												
	14 I have a	dream tod	N1																										
								sts, with its g			-					1.1.1.1				and the second						and stands			-
	10 Theve a	cream that	000 04	, down in A	caberna, m	Ch ICS VI	oous no	so, won is p	overnor h	word us	spi an	bbed in	n the word	IS OF HOME	posicion a	ne number	201, 014 0	of olliging	ere in Ala	Dertia littor	black boy	s and brac	e pris wa	De able to	ion hands	with sittle	write boys	and white	feor
	17 18 i have a	dealers had																											
	10	(****** 100	· ·																										
	20 I have a	dream that	one da	every sall	ey shall be	exalted	and ever	y hill and mo	untain she	l be m	de lou	the roug	a places with	il be mad	e clain, an	d the crook	ed places	will be me	de straigh	t and the	elory of the	Lord sha	be revea	led and all	fiesh shall	see it tope	ther.		
	21																												
	22 This is o	ur hope, an	d this is	the faith th	at I go bac	to the !	South wit	h.																					
	23																												
	24 With th	is faith, we	will be a	ble to hew	out of the	mounta	in of desp	air a stone of	fhope. W	th this f	ath, we	will be a	He to trans	form the	jangling d	isconds of a	or nation i	into a beau	tiful symp	hony of be	otherhood	L with th	s faith, we	will be ab!	le to work t	together, to	o pray toge	ther, to str	uggie
	25																												
		will be the	day, th	is will be th	e day whe	all of G	ied s child	iren will be a	ble to sing	with ne	w meat	sing																	
	27																												
	28 My cou						¥.																						
	29 Land set					pride,																							
	30 From et 31 And if A																												
	32 And so							chies																					
	33 Let free							seee.																					
	34 Let free																												
	35 Let free																												
	36 Let free																												
	37																												
	38 But not																												
	29 Let free																												
	40 Let free																												
	41 Let free					Assusip	spi.																				1	16	
	42 From et	ery mount	enside,	et treedor	ing				1							-											21		
	text																												
		_							-														-			100		1	VA.
- C																													

Okay, so if you check it carefully here, there is another column that has come up, which is this 1, 2, 3, 4, 5 thing things, these things have come up. So these are basically row names and then when there is no row names, the serial numbers of the rows comes up.

## (Refer Slide Time: 7:32)

1 min			- a x
He Lát Cole Vev Pott Sector Build Debug Pothe Cole Help D • Og get • St Cole States -			
			T Project (blane) +
P. R. Law, Mar, and A	Environment History Connection		-0
Control Courses Control See C	🐨 🖯 🖓 import Dataset + 🦸		11 List + 👔
2 library(snoballc)	🕼 Gobal Environment •		٩,
3 1fbrary(wordcloud) 4 1fbrary(kcolontrawer)	Data Otext_df	46 obs. of 1 variable	
5	values	46 005. OF 1 Variable	
6 ffl@math <= "http://www.sthda.cov.sthda/Rooc/axample_ffles/martin-luther_king-1-have-a-dream-speech.txt" 7 text <= read.ines(ffl@math) 8	fileFath	"http://www.sthda.com/sthda/RDoc/example-files/mart1 chr [1:46] ""	in-luther-king-1-have-a.
9 10 test_df-as.data.frame(text) 11 write.csv(text_df.test_df.text_df.tev')			
12 11 ##rite.table(text, file = "webdata.txt", sep = "".[Weinamet = FACIA]			
15 tent_1 <- read.csv("tent_df.csv") 16 #tot_1Steat 17			
18 Fichal the data as a corpus 19 does - Corpus(vectorSource(test_litext)) 20			
21 22 #Inspect the content of the document	Files Plots Fackages Help 1	lever	-0
23 (segac(doc)) 24 (segac(	(w. e) / Inn ( Bibpot -	0 2	
28 30 dex - trupp(dex, content_transformer(tshowr)) 12 # dexner maker: 30 def _ transformer(tshowr)			
Console Terminal = m			
> tac_fac.da.fn vietac.da > +fac.acifac.df > f			
객 Sant Strandson 📲 이 이 은 내 13 55 7 🕅 위 📾 🧿 13	0 0 3		

If you do not want that, if you do not want that, you have to write . CSV text df then CSV and then row names row.names = false you can write this that , row.names = false. It will just append the previous thing.

(Refer Slide Time: 7:57)

* 1 X J & test																					
A 8 C D 8 F 0	н		1	x		M N	0		0		5	T	u	v	w	x	Y	2	44	AR	ĸ
text																					
And so even though we face the difficulties of today and tomorrow,	I still have a c	fream. It is a d	ream deepl	y rooted in	the Ame	rican dream.															
I have a dream that one day this nation will rise up and live out the t	nie meaning	d its cened																			
I name a arream shar one day this nation will not up and not out the t	of stand	a na orea.																			
We hold these truths to be self-evident, that all men are created eq	ual.																				
I have a dream that one day on the red hills of Georgia, the sons of h	ormer slaves i	and the sons o	t former sla	ve owners	will be a	ole to sit down	together at	the table of	brotherho	и.											
I have a dream that one day even the state of Mississippi, a state sw	eltering with	the heat of inj	ustice, swel	period mg	h the hea	t of oppression	, will be tran	isformed int	to an casis	of freedor	n and justi	ce.									
I have a dream that my four little children will one day live in a natio		and and he is		color of the	also data la	the the cost	and of the local	hander													
Index a dream that my your inter children will one day live in a natio	in white they	ann noc oe ju	office of the	color or or	oper paint o	ut by the cons	and the funder of	naracter.													
I have a dream today!																					
I have a dream that one day, down in Alabama, with its vicious racist	s, with its gov	ernor having	his lips dripp	ping with t	he words	of interpositio	n and nullifi	cation, one-	day right th	ere in Ala	barna little	black boy	s and black	girls will b	be able to)	join hands	with little	white boys	s and white	e girls as sit	ters and
have a dream today!																					
I have a dream that one day every valley shall be exalted, and every	hill and mour	stain shall be i	made low, t	he rough p	faces will	be made plain	, and the cro	oked places	will be ma	de straigh	t; and the	giory of th	e Lord shall	be reveal	led and all	flesh shall	see it tog	ether.			
This is our hope, and this is the faith that I go back to the South with.																					
tins is our rope, and this is one rath that i go cack to the south with.																					
with this faith, we will be able to hew out of the mountain of despa	ir a stone of h	cos. With the	faith, we w	vil be able	to transf	om the langlis	e discords ci	our nation	into a beau	ailul suma	hony of b	otherhood	. with the	faith, we	will be abl	le to work	together.	to pray tope	ther, to str	nuggle tope	ther, to
																	-				
And this will be the day, this will be the day when all of God s childre	en will be abl	e to sing with	new meaning	w.																	
My country tis of thee, sweet land of liberty, of thee I sing.																					
Land where my fathers died, land of the Pilgrim s pride,																					
From every mountainside, let freedom ring! And if America is to be a great nation, this must become true.																					
And so let freedom ring from the prodigious hilltops of New Hamps	ire.																				
Let freedom ring from the mighty mountains of New York.																					
Let freedom ring from the heightening Alleghenies of Pennsylvania.																					
Let freedom ring from the snow-capped Rockies of Colorado.																					
Let freedom ring from the curvaceous slopes of California.																	1.00				
																			-		
But not only that:																			-	1	
at freedom ring from Stone Mountain of Georgia.																			1	T	
Let freedom ring from Lookout Mountain of Tennessee.																			at	1	
Let freedom ring from every hill and molehill of Mississippi. From every mountainside, let freedom ring.																			10/19		
		-																	1 -	WAR.	
* text df (+)																			and the	100	

Now see there is no row names. So 1, 2, 3, 4, 5, 6, 7, 8 has gone. So, this is something that you can try to do.

## (Refer Slide Time: 8:05)

Robbie He Ede Code View Picts Sesson Build Debug Profile Tools Help				- a x
· • • • • • • • • • • • • • • • • • • •				E Project (None) -
0 (1) save time costs #* -	-0	Environment History Connector		-0
O D H Cloure in See Q. F + 1		er 🔒 (Primport Datavet + 🖌	1	E 04 • 10
1. Thrany(th)	A	Gisbal Environment -		9
2 Tibrary(Snowballic)		Cata		1.0
3 Tithrary (wordc1oud) 4 Tithrary (RosTortzewer)		0.6653	List of 45	
5		Other 1	45 obs. of 1 variable	<u>.</u>
6 filePath <= "http://www.sthda.com/sthda/Rtoc/example-files/martin-luther-king-i-have-a-dream-speech.txt"		Otest of	45 obs. of 1 variable	
7 text readLines(filerath)			40 005. OF 1 Variable	
9		values	These of the set of the set in the inter in	xample-files/martin-luther-king-i-have-a.
10 text_df-as.data.frame(text) 11 write.csv(text_df,'text_df',csv',row.names = FALSE) 12		filerath text	<pre>"http://www.sthda.com/sthda/#boc/e chr [1:46] ""</pre>	xample-files/martin-luther-king-1-have-a.
12 11 Perite.table(test, file = "webdata.tst", sep = "",row.names = FALSE) 14				
15 text_1 ← read.csv("text_df.csv") 16 ftext_litext 17				
18 Pcsed the data as a corpus 19 docs Corpus(vectorSource(text_litext))	-			
20 21				
22 #Enspect the content of the document 23 Inspect (docs)		Files Plots Fachages Help		=0
24 tospace <- content_transformer(function (x , pattern ) gsub(pattern, " ", x))		Jopot -		
25 docs <- tm_map(docs, toSpace, "/")				
<pre>26 docs tm_map(docs, toSpace, "0")</pre>				
27 docs - tm_map(docs, toSpace, "\\\") 28				
22				
<pre>10 docs &lt;&lt; tm_map(docs, content_transformer(tolower))</pre>				
31 # Remove numbers 32 dors 2. Im man/dors removationhars)	~			
101 Diplove 7	R Sout 1			
Losade Teminal :	-0			
C/Owe/WPTB Dudie-H/Desking/WestBl/Session 3/	-			
Road the data as a corpus				
docs <- Corpus(vectorSource(text_lStext))				
				100 March 1
				CONTRACTOR OF A STATE
🗉 Search the web and Windows 🛛 📑 🖨 🎽 😫 🔁 🖬 🗱 📕 💷	0 12			
		and the second s		

Now, what I will be doing is that my text one I will, I am reading the text once more. So text one = read.CSV, text\_df .CSV I am reading the text once more, why? I am reading the text once more, because now I can play with this text. So, this is text document that I have bought. So, if I just want to see what are the texts that are there, that is basically text 1 dollar text is the column name. So text one not from, sorry, text\_one text\_1 dollar text and that is this text.

This is what I will be playing with. Now once I have a column of data set, there is something called corpus corpus is a special form like if you remember factor was a special form of screen variable while you can analyze the screen data with numerical representations. Similarly here, this text\_df is one kind of data set whose text I can convert to something called corpus.

Corpus is a special form which can be worked with tm library tm library can easily analyze this corpus form of that data. And this is generally takes data and corpus can be formed from text files from CSV files, from PDF files from folders. If you put lots of folders inside lots of text documents inside a folder, and then if you read a folder and then change that folders content into corpus that can also create lots of text.

So, it is basically a bunch of text put together in its very raw form is called corpus. Why it is required? Because it reduces size and the moment it reduces size, you can actually analyze the data with faster speed and lots of accuracy. So, I will fast change his data to its corpus corresponding.

So, if it is vector source, because my column with there is a column of data, when for other things this vector source things will change, but now, because they are in vector in law most of the situation in marketing, you will get customer reviews let say and customer reviews will be in a column or our blog reports will be in a column. So, that can be a vector and you can use vector source as your function to convert the vector while screen data to a corpus data.

(Refer Slide Time: 11:00)

47 · · · · · · · · · · · · · · · · · · ·				- 8
de Code View Plots Session Build Debug Profile Book Help				
🔍 📽 • 🔛 🚱 🖄 🖞 A Contraction 👘 🖄 • Addes •				Project (No
(Jaok) dim, code Africa	-0	Environment History Connections		-
F B Coursenser Q F + 1 -+ Ran Ne -+ Sour	8.41.2	🐨 🔒 🖙 Import Datavet + 🧃		List +
10rary(te)	A.	Gisbai Environment +		9
1 (Br ary (Sroaba11c) 1 (Br ary (solid c) (solid c)	- 1	Data		
10prary(RColorBrewer)		0.4003	List of 45	
filePath <- "http://www.sthda.com/sthda/RDOC/example-files/martin-luther-king-1-have-a-dream-speech.tat"		Otest_1	45 obs. of 1 variable	
text < read, ines (filerath)		O text_df	46 obs. of 1 variable	
		values		
text_df-as.data.frame(text)	- 1	filePath		umple-files/martin-luther-king-i-have-a.
write.csv(text_df, 'text_df.csv',row.names = FALSE)	- 1	text	chr [1:46] **	
Awrite.table(text, file = "webdata.txt", sep = "",row.names = FALSE)				
text_1 <- read.csv("text_df.csv") Prant_lStext				
	- 5			
<pre>PLoad the data as a corpus docs &lt;- Corpus(VectorSource(text_Litext))</pre>				
docs (- corpus(vector source(text_streat))				
		film Path Pashapen Help Ver		
VENSPECT the content of the document (respect(docs)				-
tospace content_transformer(function (x , pattern ) esub(pattern, " ", x))		Jibpot - 0		
docs ← tm_map(docs, tofpace, "/") docs ← tm_map(docs, tofpace, "#")				
<pre>docs &lt;- tm_map(docs, toSpace, "\\\")</pre>				
docs <- tm_map(docs, content_transformer(tolower))				
# Kenove numbers				
ders 2. en man/ders ranciakunhars) Diplosij :	R Sout 1			
ar Innia				
an VIII Budio - McDedito / Beek D/Besian V - P	-			
set freedom ring from every hill and molehill of Hississippi.				
From every mountainside, let freedom ring.				
and when this happens, when we allow freedom ring, when we let it ring from every village and every hamlet, from e and every city, we will be able to speed up that day when all of God's children, black men and white men. Jews and	very s			
and every city, we will be able to speed up that day when all or God's children, black men and white men, sews and Protestants and Catholics, will be able to join hands and sing in the words of the old wegro spiritual:	Genti			
Free at last! Free at last!				
				10000
				and the second sec
thank God Aleighty, we are free at last!	- 1			
	- 1			and a second
	1.0			
en la constantia fan	¥			ANNU PROFESSION
Search the web and Windows 💼 💼 💼 🍁 🤮 🖷 🕫 🗱 🕺 📕 🗰 💼 🏟	12	and the second se		

So, this is what we are doing the first thing I am creating a docs, so, you check are lots of con things are there in docs. So if I just inspect docs, this is how it looks like. And along this thing, it will be very difficult to inspect the whole thing. So doc is a metadata which is a corpus specific = 1 document level 0. So it is not indexed. And the documents is 45 there are 45 documents. And if I inspect let us first 10 generally, that is what we do because the corpus in real life situation will be very long.

So then I say let us have 1 to 10, I get the, sorry, not this one. So if I just write within docs, yeah. So I get the first 10 documents in this thing. So that is how we create the corpus. Then there are lots of things that we have to do, you have to understand that there are some changes that we have to do.

So we are doing content transformation. So content transformer we are transforming some content. How? So it will be doing this thing that particular function is called two space, what it will do? It will be a basically a function of x and pattern where it will change a pattern with a blank. Whenever you give a source a pattern under blank it will change that in that in the source called x.

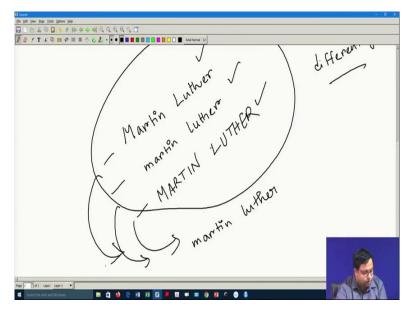
It will find out the pattern and corresponding pattern it will replace it with a blank space for example, is a function just a function that we have written for example, let us say doc tm\_map and then doc's 2 space and then this. So whenever this particular sign comes up, it will use this to space function here. Okay and that is how it will convert it. So, that means it will convert this whenever this particular backslash is coming up, it will convert it to space.

(Refer Slide Time: 13:25)

Care G G G A Contestantes   E + Addes +				B Project dia
esc tán cola #* :	-0	Environment History Connections		-
C B Cloure in See Q P + 11	Aun Na - Source + 2	2 8 Pinport Datavet - 1		ii ber -
r ne w y Generale (s.) 1 the ary (wordc loud)	A	Global Environment +		9
library(RColorBrever)		Data		
fileFath <= "http://www.sthda.com/sthda/Rtoc/example-files/martin-luther-king-i-have-a-dream-speed	h tur'	0 docs	List of 45	
text - readLines(fileFath)	0.80	0 text_1	45 obs. of 1 variable	
		Otext_df	46 obs. of 1 variable	
text_df-as.data.frame(text)		values		
write.csv(text_df,"text_df.csv",row.mames = FALSE)		filePath	"http://www.sthda.com/sthda/RDoc/examp chr [1:46] ""	ole-files/martin-luther-king-i-have-a.
Awrite.table(text, file = "webdata.txt", sep = "",row.names = FALSE)		text tospace	function (x,)	
text_1 ← read.csv("text_df.csv")		Copere	function (A, 111)	
text_A <- read.csv(text_dr.csv) ftext_litext				
PLoad the data as a corpus docs <- Corpus(VectorSource(text_Sitext))				
#Inspect the content of the document				
inspect(docs)		film Plots Packages Help Ver		
tospace <- content_transformer(function (x , pattern ) gsub(pattern, " ", x)) docs <- tm.map(docs, tospace, "/")		Jim Jippet -		
docs «- tm_map(docs, toSpace, "0")				
docs <- tm_map(docs, toSpace, "\\\")				
<pre>docs tm_map(docs, content_transformer(tolower))</pre>				
# Eenove numbers docs <- tm_map(docs, removenumbers)				
# Remove english comon scopwords	Ψ.			
Pop Level 2	A Solut 1			
Terminal -				
/WPTEL Munito-PL/Techlog/Week30/Session 3/ 17				
pect the content of the document ect(docs)	<u>^</u>			
1ecorpus				
ta: corpus specific: 1, document level (indexed): 0				
t: documents: 45				
nd so even though we face the difficulties of today and tomorrow, I still have a dream. It is a dr	eam deeply rooted in			
erican dream.				
Y.				
1				
have a dream that one day this nation will rise up and live out the true meaning of its creed:				
The second				
e hold these truths to be self-evident, that all men are created equal.				
e nore trepe trains to be perreproduct, that are sen are treated equal.				

So this is what I am writing, so, convert all this funk punctuations this one and then this one and probably this one also all of them to space admits that they will be blank spaces. Now if I try to inspect, if I just try to inspect you will see that those kind of signs are probably removed from the data and there are specific things that are still there.

So you have to check that when whenever this particular thing was there and now it has got removed. So, that is how we are slowly cleaning the data so that the data can be used for further analysis. (Refer Slide Time: 14:20)



What next? Now, there can be so many different kinds of transformation, why this transformation is needed? Let us say somebody says that there is Martin Luther and some other guys have written martin Luther and then some other guys have written, M. Luther. So ideally these 3 tests will be different or M. Luther might be should be different actually M, I am not saying, let us say this one. So that M. Luther will be different in any way. But let us say MARTIN and LUTHER.

Now, how do you deal with this kind of a situation? You think that see all these 3 are different and if you ask the guy to analyze it, it will analyze it differently. And you have a problem that you do not know that whether you should analyze them differently or not. There are situations which, where you want to analyze them differently.

There are situations where you do not want. If you do not want, then if you want to analyze them as a same thing, then what do you have to do basic thing you have to change all of them to their lower form. So let us say I write it, all of them becomes Martin and then Luther. Something like that. So, to do that, while I have to do is convert them to their lower form. So, that is something that I have to do as that fast.

## (Refer Slide Time: 16:12)

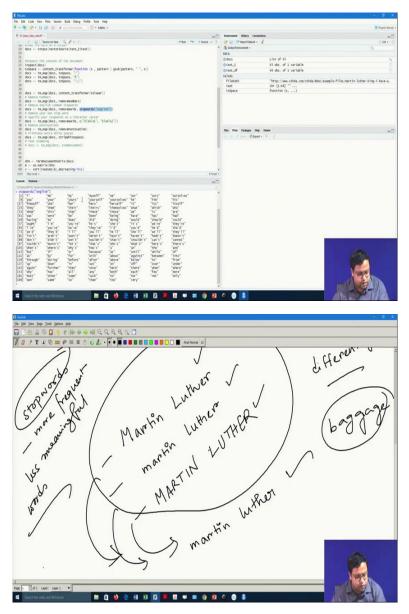
3 Khude				- 0 X
File Lás Code View Piots Session Build Debug Profile Tools Help 0 • Old Off • La Data A Constitutional 10 • Addes •				Thurst None -
♥ 01_basis_tilde_code.d* >	-0-	Environment History Conne		-0
	+Run 🍋 🕂 Source + 🗦	😅 🔒 🖙 Import Dataiet +	1	iii lat + 👘
<pre>19 docs</pre>	^	🐴 Global Environment •		Q, 1
20 21		Data		104-0
22 Finspect the content of the document		0 docs	List of 45	٩.
23 inspect(docs) 24 totpace <- content_transformer(function (x , pattern ) osub(pattern, " ", x))		0 text_1	45 obs. of 1 variable	0.1
24 toppace - content_transformer(function (x , pattern ) gsub(pattern, , x)) 25 docs - tm_map(docs, tofpace, "/")		O text_df	46 obs. of 1 variable	
<pre>26 docs += tm_map(docs, toSpace, "0")</pre>		values		
<pre>27 docs &lt;- tm_map(docs, toSpace, "\\ ") 28</pre>		filePath	<pre>chr [1:46] **</pre>	xample-files/martin-luther-king-i-have-a.
29		tospace	function (x,)	
<pre>10 docs &lt;- tm_map(docs, content_transformer(tolower)) 31 # Remove number;</pre>		coopere	Survey of the second	
12 docs - tn_map(docs, removements)				
31 # Remove english comon stopwords				
34 docs <-> tm_map(docs, removements, stopwords("english") 35 # Remove your own stop word				
16 # specify your stopwords as a character vector				
17 docs tm_map(docs, removementds, c("blabla1", "blabla2")) 38 # memory punctuations				
10 docs tm.map(docs, removePunctuation)		Files Plots Packages Help	Viewer	-0
40 # «Tintrate extra white spaces		J bpot		
41 docs <- tm_map(docs, stripshitespace) 42 # Text stemping		A Deper		
43 # docs ↔ tm.map(docs, stembocument)				
44 45				
46				
47 dtm TermDocument#atrix(docs)				
48 m ← as.matrix(dtm) 49 v ← sort(rowsums(m),decreasing=TRUE)				
3455 Dip Level 2	R Souge 1			
Consule Seminul -	-0			
(_/bes/MP()) State H/Deskto/Street//Sesters.)/ =	MLI.			
> docs <- tm_map(docs, removementers)				
warning message:				
In tm_map.simplecorpus(docs, removenumbers) : transformation drops documents				
>				
				1074523
				and the second se
📽 Search the web and Windows 📰 🛍 🍪 🤮 🖬 🗱 🖾 🦉 🔚		a 🕛 👔		AN A CONTRACT OF

So I am here if you check the text, I am converting them to their lower form, docks = tm map docks content transformer to tool lower. So this is something that I am doing for the first thing done, if I now inspect my content, so let us say dock, if I inspect docks, everything is in lowercase. Everything all the text is in lowercase there is nobody which is in uppercase, then what then there are certain things which are numbers.

So I will just check whether numbers are there or not. In this particular text numbers is not there but number is probably a very standard cleaning process here that we have to follow because sometimes we have numbers, I would say pretty often times, we have numbers here we do not have numbers. But still, if you have to remove numbers, the code is like this doc, tm map docs is comma removed number.

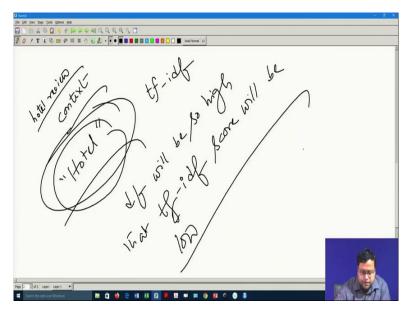
So you pick up docs, that particular corpus, find out the numbers, remove them, and then again, save it in docs. So that is what I am doing. Then what then there is something called stopwords and we are focusing on English language. So stopwords\_English suffers within bracket English, so I will just copy that and paste it here.

(Refer Slide Time: 17:42)



So these are the stopwords, what a stopwords? Stopwords are basically unnecessary words, words which have no contribution or more frequent less meaningful words are basically stopwords. So, this we have to remove to reduce the baggage. So, if my data set has lots of baggage, baggage means unnecessary words, then what I feel is that I feel heavy I do not cannot analyze and sometimes the insights generated are not so useful. So, we have to remove stopwords.

(Refer Slide Time: 18:40)



Now, you can remove stopwords by just mentioning that this some of the cases there can be certain words, which itself are stopwords but for example, if you remember I told in a hotel review context in a hotel review context, one stopword itself is hotel. This itself is a stopword.

So, if this itself is a stopword that that particular thing has to be removed because hotel is very common. If we do tf idf the guy's df will be so high that tf idf score will be much low. So in such kind of a situation, we have to remove them. So this is something that I am doing now.

(Refer Slide Time: 19:29)

e Edit Code View Pots Sesion Build Debug Profile Tools Help → QB (at + 12) (2) (2) (4) A Catallationaux (2) + Addes +			E Proved Bloc
Disector column	Environment History Connector	-	
- CHOSecretary G. F. I	en 🔒 🖙 Import Dataset - 🖌	(	i ut -
	🐴 Global Environment +		Q,
20 21	Cuta		
22 Finspect the content of the document	0 docs	List of 45	
23 inspect(docs) 24 toSpace << content_transformer(function (x , pattern ) qsub(pattern, "", x))	Otext_1	45 obs. of 1 variable	
25 docs - tm.map(docs, tofpace, "/")	0 text_df values	46 obs. of 1 variable	
26 docs tm_map(docs, tofpace, "0") 27 docs tm_map(docs, tofpace, "\\\")	fileFath	"http://www.sthda.com/sthda/#Doc/ex	xample-files/martin-luther-king-1-have-a
28	text	chr [1:46] ***	
19 10 docs tm_map(docs, content_transformer(tolower))	toSpace	function (x,)	
1 / Finese routies:     2 / 6454 - thoughdes, removembers)     7 / Finese regulation counts (foregradies)     4 Finese regulation (counts foregradies)     4 Finese your time top and			
27 dod - transforder, research, c("blabla", "blabla") 28 dod - transforder, transforder, research, c("blabla", "blabla") 29 dod - transforder, research, component,	files Pada Packages Help 20 Euport -		
2 2 First iteming 42 First iteming 44 45 47 47 47 48 49 49 49 40 40 40 41 41 41 41 41 41 41 41 41 41			
48 v ↔ sort(rowsums(m),decreasing=110,c) v 54.01 (top leve) 1 R longs 1			
anale benind : mil			
Alses/WTE: Dudie It/Sedates/WesktlyTession 17-1			
Disput         Open         Open <thopen< th="">         Open         Open         <t< td=""><td></td><td></td><td></td></t<></thopen<>			
Diff         Control         Control <thcontrol< th=""> <thcontrol< th=""> <thcont< td=""><td></td><td></td><td></td></thcont<></thcontrol<></thcontrol<>			

These are my stopwords, and if you want that, okay, there is one more stopword, which is hotel let us say. So you said they say, so remove stopword, tm then docs is your x y from where you will remove the function is remove words and what will you remove stopwords and then this English, so if I just run this in the whole docs, this particular words have been removed. So thank you Almighty free last, free last free last every mountainside let freedom ring.

So, there are certain words you can understand now let, freedom ring every heel. So, let freedom ring from every hill was there to another to heal Mississippi or something like that that kind of to form the all the pronouns all the prepositions conjunctions has been removed. Let us say in this particular thing I also want to remove this let this let is coming very common and let has no meaning for my case. So, let us remove let so, how will I remove let I will just create docs and here instead of stopwords I would write let let has got removed.

Now in my introduction Let Freedom Ring blah blah blah blah this has also gone way. Now let us say, I want to remove 2 words instead of 1 word, what will I so will let us say will, will be free, will be the day, will be able to, so we will has nothing to do. And then there is another word let us say, which is let us say another word can be will is 1 word it is coming up pretty common, Four will and four this is another one that I will be removing, let us just say that.

(Refer Slide Time: 21:42)

1 cade				- 0 X
He Ede Code View Plots Session Build Debug Profile Tools Help				
P + 9 2 + H B   B   A tristeriore   1 E + Adex +				📱 Project (None) +
• It, Basin, John, code. R	-0	Environment History Connec	Rons	-0
(a) (a) (a) Chourse on Saw ( (3, 2 + 1))	-+Run ++ -Source + 2	😅 🔒 🛛 👘 Import Dataset +	1	III UH + 1 ()
19 docs - Corpus(VectorSource(text_litext))	^	🚳 Global Environment +		9
20 21		Cuta		
22 #Inspect the contant of the document		0 docs	List of 45	٩.
2) inspect(dots) 24 toSpace <- content_transformer(function (x , pattern ) gsub(pattern, " ", x))		0 text_1	45 obs. of 1 variable	
24 toppace <- content_transformer(function (x , pattern ) glub(pattern, , x)) 25 docs <- tm.map(docs, tospace, "/")		Otext_df	46 obs. of 1 variable	
<pre>26 docs tm_map(docs, toSpace, '0")</pre>		values	These of the state of the later (and	ample-files/martin-luther-king-i-have-a.
<pre>27 docs &lt;- tm_map(docs, toSpace, "\\\") 28</pre>		text	chr [1:46] ""	apte-rites/martin-luther-king-t-nave-a.
29		tofpace	function (x,)	
<pre>30 docs &lt;- tm_map(docs, content_transformer(tolower)) 31 r Remove numbers</pre>				
12 docs tm_map(docs, removements)				
33 # Bemove english common stopwords 34 docs + tm_map(docs, removewords, stopwords("english"))				
15 # Remove your own stop word				
36 # specify your stopwords as a character vector 37 docs <- tm_map(docs, removements, c("blabla1", "blabla2"))				
38 # Remove punctuations				
19 docs < tm_map(docs, removePunctuation) 40 r Eliminate extra white spaces		Files Flots Fackages Help	Vewer	=0
41 docs tm_map(docs, stripshitespace)		John Street	*10 I X	
42 # Text stemming 43 # docs ← tm_map(docs, stmmDocument)				
44				
45				
47 dtm <- TermbocumentHatrix(docs)				
48 m ← as.matrix(dtm) 49 v ← sort(rowdums(m),decreasing=TBLX)				
147 V - Set (resses (e), secressing-rest)	A Sout 2			
Conside Terminal -				
C2/verv/WFTE Dudia-H/Dealop/Wesh1/Sealon 3/ -9				
[29] every mountainside, freedom ring:				
[30] america great nation, must become true.				
[31] freedom ring prodigious hilltops new hampshire.				
[32] freedom ring mighty mountains new york.				
[33] freedom ring heightening alleghenies pennsylvania.				
[34] freedom ring snow-capped rockies colorado.				
[15] freedom ring curvaceous slopes california.				
[36]				
📽 Search the web and Windows 🔚 🖬 🏟 🤮 🖬 🕅 🕅 🕅	1 📫 🖬 🌖 👔	a 🕛 👔		

So then what I will do is, here I will write C will and then four now, if I inspect the docs, the will has gone away and the four has gone away. So all of these things are so, that is how you clean unnecessary words. Now after cleaning necessary words, so see here you see that blah

blah blah and blah blah2, these words are not there, but this is remove it some example has been given here.

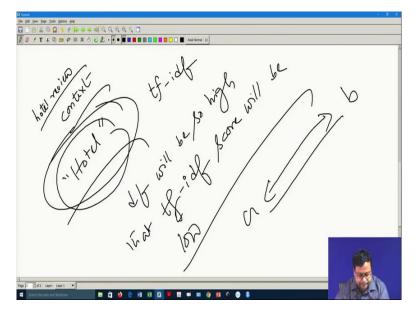
(Refer Slide Time: 22:18)

<pre></pre>	- đ				
<pre>//pacestructure// actions// act</pre>	E hued do				
10     Figure 1     ****     ****       20     Ref = Constrainty (***)     ****       21     State Amount     ****       22     State Amount     ****       23     State Amount     ****       24     State Amount     ****       25     State Amount     ****       26     State Amount     ****       26     State Amount     ****       26     State Amount     ****       26     State Amount     ****       27     State Amount     ****       28     State Amount     ****       29     State Amount     ****       20     State Amount     *****       21     State Amount     **********       21     State Amount     ************************************					
<pre>b Vote: Display of an interplay b Vote: Display of an interplay of an interplay b Vote: Display of an interplay of an interplay b Vote: Display of an interplay of an</pre>					
<pre>13 ddd = (organisettificate(text_1)text)) 25 ddd = (organisettificate(text_1)text_1)text)) 25 ddd = (organisettificate(text_1)text_1)text)) 25 ddd = (organisettificate(text_1)text_1)text)) 25 ddd = (organisettificate(text_1)text</pre>	List +	1			C G Source on Sale 1 9, 7 + 1
<pre>3</pre>	9			^	<pre>ics &lt;= Corpus(vectorSource(text_l(text)))</pre>
<pre>20 critique fut formation of the denset 20 critique fut formation of the denset 30 critique fut formation of the denset 31 critique fut formation of the denset 3</pre>					
<pre>i tobic: c.tragefore (results) of points of points</pre>					inspect the content of the document
<pre>i doi = may finds, respective, respec</pre>					spect(docs)
<pre>i dot =</pre>		46 obs. of 1 variable			
dis.     text     or (1):40 *       dis.     text     text       dis.     text     or (1):40 *       dis.     text     text       dis.     text     or (1):40 *       dis.     text     text       dis.     text<					<pre>ics &lt;- tm_map(docs, toSpace, "0")</pre>
Pois - Recurré danse:     Topase     Foreitier 0,)       Pois - Recurré danse:     Foreitier 0,)	r-king-1-have-a.				<pre>ics &lt;- tm_map(docs, toSpace, "\\(")</pre>
and to the specific constraints (Salar) (Salar					
dar - traggidar, reservedar; traggidar, reservedar; reserved, reserved, re		Function (x,)	tospace		
<pre>rtess: eight beine tragent specify word there is a there is an intervent specify word there is a there is a there is a there is a specify word there is a there is a there is a specify word there is a there is a there is a specify word there is a specify the is a specify the specify the is a specify the is a specify the specify the is a specify the is a specify the specify the is a specify the is a specify the specify the is a specify the is a specify the specify the is a specify the is a specify the is a specify the specify the is a specify the is a specify the is a specify the specify the is a specify the is a specify the is a specify the specify the is a specify the is a specify the is a specify the is a specify the specify the is a specify the is a specify the is a specify the specify the is a specify the specify the is a specify the specify the is a specify the is a specify the is a speci</pre>					Remove numbers
test - tradigidor, remover/in, statusticity       test - tradigidor, remover/in, (Sublidir, "Initial"))       test - tradigidor, remover/initial")       test - tradigidor, remover/initial") </td <td></td> <td></td> <td></td> <td></td> <td></td>					
<pre>right(r) root therefore a target restor root to the root to the root to the root restor root to the root to the root to the root root root to the root root root root root root root roo</pre>					<pre>ics &lt;- tm_map(docs, removewords, stopwords("english"))</pre>
doc - tracycletc, resourching, (Schaltz, "Listatz"))     Tax. Tracycletc, resourching, (Schaltz, "Listatz"))       doc - tracycletc, resourching, (Schaltz, "Listatz"))     Tax. Tracycletc, resourching, (Schaltz, "Listatz"))       doc - tracycletc, resourching, (Schaltz, "Listatz"))     Tax. Tracycletc, resourching, (Schaltz, "Listatz"))       doc - tracycletc, resourching, (Schaltz, "Listatz"))     Tax. Tracycletc, resourching, (Schaltz, "Listatz"))       doc - tracycletc, resourching, (Schaltz, "Listatz"))     Tax. Tracycletc, resourching, (Schaltz, "Listatz"))       doc - tracycletc, resourching, (Schaltz, "Listatz"))     Tax. Tracycletc, resourching, (Schaltz, "Listatz"))       doc - tracycletc, resourching, (Schaltz, "Listatz"))     Tax. Tracycletc, resourching, (Schaltz, "Listatz"))       doc - tracycletc, resourching, (Schaltz, "Listatz"))     Tax. Tracycletc, resourching, (Schaltz, "Listatz"))       doc - tracycletc, resourching, (Schaltz, "Listatz"))     Tax. Tracycletc, Tax. Tracycle					Remove your own stop word
Form and instantial i					spectry your stopports as a character vector ics = te man(docs, removamords, c("blablat", "blabla2"))
A (linear output affaire space af all the set of a affaire space affaire a standing affaire a standing affaire affaire a standing affaire aff					Remove punctuations
des - tragologies, strafettesser des - tragologies, strafettesser des - tradocentent des - tradocentent		Vever	Files Flots Fackages Help Vie		<pre>ics &lt;= tm_map(docs, removePunctuation)</pre>
r Fore India, Freedom (I. Secondar) de - Transformer (I. Secondar) freedom ring I solated association torressee Freedom ring I solated association torressee			m m / han Ji bpot - 0		cs <- te map(docs, striph(tespace)
fire - Introducementari (das) v - Introduced V, derrasting-fisch mentari mentari fireden rig: Toolact sources: fireden rig: Toolact sources: fireden rig: Toolact sources:					Text steming
de « tredecouerraris/sicio) « » autorization « per la construction » « per la construction » en la constru					docs << tm_map(docs, stemDocument)
2 as backfordsomerserification)					
I i i i i i i i i i i i i i i i i i i i					
V = or (Variansis), development) ( and the second of the					
Andref Forder     Andref					sort (routing (n), derreas (n), third (n)
weeters daaks entoesen weeters ook ook ook ook ook ook ook ook ook oo				A Solet 1	
andréfi ladar érbana hannarkanna (z. d				-7	Irrend :
; freedow ring wavy Mill wellenill adjustisalignt				mu.	
freedow ring weary hill webstill afastistight				A	
					reedom ring lookout mountain tennessee
					reedom ring every hill molehill mississippi
every mountainside freedom ring					very mountainside freedom ring
happens allow freedom ring ring every village every hanlet every state every city able speed day g				able speed day o	hammens allow freedom ring ring every village every hamlet every state every ci
s children black men white men jews gentiles protestants catholics able join hands sing words old regro spiritu		the second se		ords old negro spiritu	ildren black men white men jews gentiles protestants catholics able join hands sing
] free last free last					a last free last
	<b>.</b>				
	131				
i) thank god almighty free last					ink god alwighty free last T
	Rev Vinn	model		2	
		1.0005			
Search the web and Windows: 📰 🛍 🍁 😌 🖷 🛤 🖾 🐺 🖾 🗰 📾 🚳 🔞 🖉 🖉 🚳 📲			0 0 1		art the set and Windows 💦 📑 👘 🐴 斗 🖓 👘 🖬 🕅 🦉

So then the remove punctuations, so we will punctuation. So, you include just trying that, and then if you inspect the document, all the punctuation in this document has been removed. So, that is what we are doing.

And then there are white spaces also blank spaces. So, tm map docs pre white space, so that will actually reduce remove all the white spaces, not the enters but if you have given tabs in between, let us say a tab b, so, a tab b will create a situation like this.

(Refer Slide Time: 22:50)



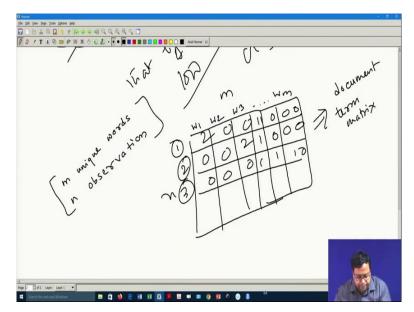
You write a and then a blank and then B so this blank will be removed by whitespace remove whitespace.

(Refer Slide Time: 23:04)

Rade			- 0
ie Liefe Code View Parts Sension Build Debug Profile Tools Help • One GP + GP GP → GP GP → Grammatananan III + Addes +			Thurst day
Pittanucion, on chi Di Sanzin Sa	Environment History Con		-
an econo con una so a company	🐨 🔒 🌁 Import Datavet	• 6	ii List -
19 docs - Corpus(vectorSource(text_litext))	Gobal Environment •		Q.
20 21	Cuta		
22 #Inspect the content of the document	0 docs	List of 45 45 obs. of 1 variable	0
33 [inspect(docs) 34 tospace <- content_transformer(function (x , pattern ) glub(pattern, " ", x))	Otext_1 Otext_df	45 obs. of 1 variable	
25 docs in th_hap(docs, tofpace, "/")	values	we opp, or a variable	
20 docs te_map(docs, toSpace, "0") 27 docs te_map(docs, toSpace, "\\")	filePath	"hanna / hone and da ann/anda/anna	/example-files/martin-luther-king-i-have-a.
28 Sector	text	chr [1:46] ***	example integral construction scriptionare a.
8	tospace	function (x,)	
10 docs <- tm_map(docs, content_transformer(tolower)) 11 Filtenzye number:			
12 docs +- tm_map(docs, removementers)			
13 # Remove english common stopwords 14 docs tm_map(docs, removemends, stopwords("english"))			
15 # Remove your own stop word			
17 docs - th_map(docs, removements, c("blablat", "blablat")) 18 F menove purctuations			
10 docs tm.map/docs, removePunctuation)	Film Picts Packages Help Vener		-
40 # sliminate extra unite spaces 41 docs ← tm.map(docs, stripunitespace)			
42 0x3 <- th_Nep(ox3, stripenitespace) 42 / Text steming			
43 # dots th suppliers, stateouvert)			
44			
45			
47 dtm <> TermbocumentHatrix(docs)			
48 m ← as.matrix(dtm) 49 v ← sort(rowSums(m),decreasing=TNUE) v ↓			
11 Dig Look 1 Andrew 1 (1990) Harring Hills (1990)			
main immini - mili			
Alan APTI) their It/Testing/Net/U/Jestin () 1			
(Elificitar stort, allite gines misse Taussii). (traggi TigleGroups(Bits), striphtespeck) : rangeTigleGroups(Bits), striphtespeck) : randfracture drop document.			
E Sana than and an fordama	0 0 1		

So, after this cleaning this part of to this part is called pre processing of the data after this pre processing, what we try to do is the first thing that I will do is we will create a document term matrix.

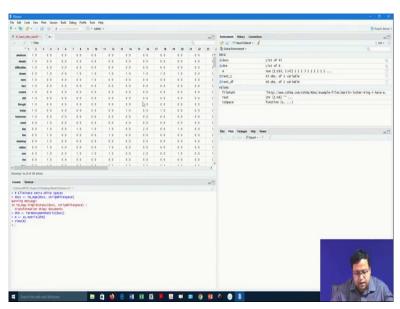
(Refer Slide Time: 23:20)



What is document term matrix? Document term matrix is something which looks like this, let us say I have in my in the whole corpus, I have m unique words. And in observations, that means n number of variables, m number of unique words, and n number of observations. So what we do in this case is a m by n matrix where here each word is written word 1, word 2, word 3, up to word m. So each word is given here and correspondingly, we just see that in observation number 1, whether some of the words has occurred or not or how many times it has occurred.

So, this has occurred 2 times let us say 00100 something 0 and then in a second entry 0021000, the third entry it is 0001110 and the fourth entry and so on these are basically frequencies. So, if this particular matrix is called document term matrix generally we do not convert this matrix to visible matrix form, but we do is a special form of matrix which is document term matrix and I will tell you what that particular thing means.

(Refer Slide Time: 24:52)

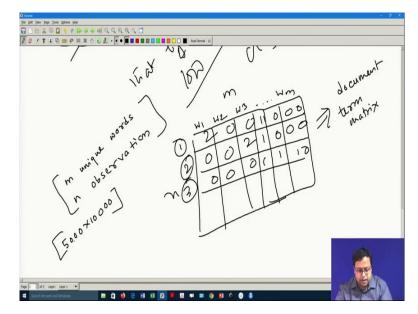


So, this is something and the purpose is also something that I will tell you so, DTM converts is to D document of matrix and DTM is this thing if I by chance convert DTM to its matrix form m, then the M will look like this, see there are I, so if you check here 163 unique text is there unique words are there and this 1 to 45 is done number of observations, so, sorry in the other way it has done in the other way.

So, 163 unique rows are unique words are there, which are rows and 45 columns are there and this particular column is particularly saying that in the first row, which words are occurring, so, 111311111, So, which words are occurring, so if I do the row sums of this particular matrix form that you are seeing, then what you will get is the whether this particular were is the total frequency of that particular word, you will get that, but actually in

document of matrix if you see, what is DTM? DTM gives me the i, the j and the v. See remember, if it is a if there were a number of words with 45 observations, I got 163 unique words.

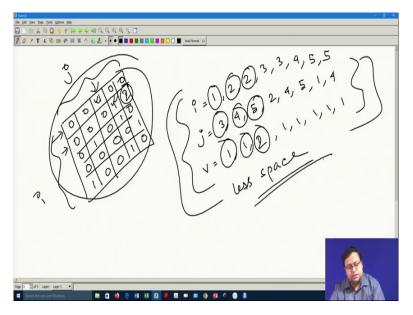
(Refer Slide Time: 26:26)



So now, if I have let us say around our data set of 5000 observations, and how many unique what will I have? I will have probably 10,000 15,000 unique part. So, then  $5000 \times 10,000$  matrix will be so, big matrix that that analyzing that will be a difficult thing forget about analyzing, reading that in your art will be a very hectic thing.

So, then also you have to remember that there will be lots of 0 you see that the it will see this particular document or matrix most of the words not occurring in a particular observation in a particular column probably at max 100 words will occur, but even 5000 words. So in a particular observation, maximums will be 0. So there are lots of 0 and some of the positions where these numbers are positive, then why do not why do this?

(Refer Slide Time: 27:20)



So instead of, let us say creating let us I have a five by five matrix 12345 12345, so this is 0100, this is 11000,01010,00001,10010. So this can maximum 0 if you want.

Instead of writing this thing can I write that these are my i's, these are my j's and these are values. So I = 1, then I = to 2, 2, then 3, 3, then 4, and then 5 5 and j = for the first one, it is 3, then 4, then 5, then two then 4, then 5, then 1 then 4, and the V, which is value is on 1, or let us this is 2, so 1,1,2,1,1,1,1,1 but how do I read this these 3 vectors, how do I read this?

So, the first entry is i = 1 j = 3 value =1. So, first row, third column the value will be 1, the second is second row, fourth column value will be 1, the second row, fourth column value is 1, second row, fifth column values 2, second rows fifth column there is 2 and that is how you create the matrix and exactly.

Now, this formation this formation of matrix takes less space, you have equally whatever information you have here, you have the same information you have here, but it takes less space.

(Refer Slide Time: 29:17)

1 Paula										- a x
File Edit Code V O - 😋 😅 -	lev Plots Session	Build Debug F	hofile Tools Helj							📕 Project, Dianej +
0 01, basic, tdm, cod								Environment History Connections		-0
	Source on Save	97.0					-+Run Ne -Source -	🗧 🚰 🔒 🜁 Import Dalaxet + 🦸		E DR + 10
28	calimetricocs, co	opers, 1117						<ul> <li>Global Environment +</li> </ul>		9.10000
	tm_map(docs, co		mer(tolower))					Data O docs	List of 45	0.
11 # Renove 12 docs <- 1	numbers tm_map(docs, re	novenumbers)						Odta	List of 6	A.
34 docs	tm_map(docs, re	stopwords, sto	pwords ("englis	h"))				n Otest_1	rum [1:163, 1:45] 1 1 1 3 1 1 1 1 1 1 45 obs. of 1 variable	
	your own stop y your stopword tm_map(docs, re							Otext_df	46 obs. of 1 variable	0
				bla2"))				values		
	tn_map(docs, re							fileFath text	<pre>"http://www.sthda.com/sthda/RDoc/example-files/s chr [1:46] **</pre>	nartin-luther-king-i-have-a.
41 docs	tn_map(docs, st	ripublitespace)						tospace	function (x,) Named num [1:163] 13 12 11 11 9 8 8 7 4 4	
41. docs ← 42 # Text s 43 # docs ← 45 46 47. dts ← T 48 m ← as, 49 v ← ss, 50 v ← dat 51. head(d,	tenting - ta_map(docs, st	stenDocument)						v	Named rum [1:161] 13 12 11 11 9 8 8 7 4 4	
44 45										
46 47 dtm - 7	ernDocumentHatr	1x(docs)								
48 m <- 45. 49 v <- 500	matrix(dtm) t(rowSums(m),de	creasing-TRUE)						files Plots Fackages Help Vi	iner .	-0
50 d dat. 51 head(d.	ermbocumentHatr matrix(dtm) t(rowSums(m),de a.frame(word = 10)	names(v), freq-	v)					the first of these of the end		
52										
53 54 barplot(	d[1:10,]ifreq, col ="lightblue ylab = "word fr	las - 2, names	arg = d(1:10.	]Sword,						
55 56 57	ylab = "word fr	equencies")	Trequest word							
58								v		
501 (top Level)							R Se			
Console Terminal										
> # Elisinate	extra white spa	ICes								
> docs <- true warning messag	ap(docs, stripa	hitespace)						1		
In tn_map, Simp transformati	leCorpus(docs, on drops docume	stripshitespace	(e) 1							
Contemporter that > # Eliminate > docs <- tm_m warning messag tr ansformati > dtm <- TermD > mice as.matr > view(m)	ocumentMatrix(d	locs)								
> v - sort(re	uSums (n) , decr ea	sting-TRUE)								
fre	ring	01926	iliy .	entry.		وانفت	topither			
nation	muntain	shall	faith	free	today	nen	state			
4 children	little	black	4 white	4 nade	god	3 new	3 sing			
1 Tand	last	) even	3 Tive	meaning	true	brotherhood	former			1.4
georgia	sons	2 heat	mississippi	sweltering	alabama	2 boys	girls 2			
2 hards	2 join	2 words	2 h111	places	2 hope	2 stone	2 thee			March 1
nountainside	american		difficulties	face	rooted	still	2 though			
NOUTCATIO 100	anerican	Deepity	difficulties	TACE	rooted	still	though	v	756	
Search the			- C C	1 📦 🤤	<b>H N</b>	X 📕 🛛	1 📫 🖬 🌖 1	1 4 0 1		
Rouse Ne Edit Code 1	lev Picts Section	Build Debug P	hotile Tools Hel							- σ ×
Koude     Ne Eds Code V     O • Og @* •     O • Ottawictom.cod     wood     medican these     Mag mag     detam ditam	Alti d+ Martini d+ Martini 10 10		frafile Tools Helj				ų	Invisonment History Connections Constant Property Connections Constant Invisonment - Constant Odd Odd Odd	163 obs. of 2 variables List of 45	- 0 X
Ne Lee Code 1 0 • Og er • 0 01,1aoc,1de,ceo 1 medium Tech Needum Tech Needum Tech Needum Tech Needum Tech	alto Car Nar Mar 15		Poplie Socis Helj				q	Cobe Encounter + Cobe Encounter + Cobe Encounter + Cod Code Co	163 obs. of 2 variables List of 45 List of 6	-0
Ne Let Cole 1 0 - Op er . 0 (1).best, tim, tes needen tero neg mg dram dram dram dram dram gram	alt of a		Profile Tools Helj				q	Cobe freesest -      Cobe freesest -     Cobe freesest -     Cobe freesest -     Cobe freesest -     Cobe freesest -     n	161 obs. of 2 variables List of 45 List of 45 Fram [Li64, 1043] 1 1 1 1 1 1 1 1	-0
Ne Lee Cole 1 O - Or	at a de la composición de la encomposición de la composición de l		hrafile Tools Helj				q		163 obs. of 2 variables List of 45 List of 6 num (1:169, 1:45) 1 1 1 1 1 1 1 1	-0
Ne Let Cole 1 0 - Op er . 0 (1).best, tim, tes needen tero neg mg dram dram dram dram dram gram	n 1 0 0 1 n 1 0 0 n 1		hyblie Toolo Hel				ų.	Provide the set of the set o	363 dos. of 2 voriables List of 43 List of 6 mp [1:16] 113 3 11 1 1 1 1 45 dos. of 1 voriable 46 dos. of 1 voriable 70ttp://www.sthe.	-0 
File Edit Code 1 0 - 0 - 0 - 0 - 0 - 0 - 0 - 0 - 0 - 0 -	at A         at -           at A         at -           read         at -		holie Tooh Hey				q		363 dos. of 2 voriables List of 43 List of 6 mp [1:16] 113 3 11 1 1 1 1 45 dos. of 1 voriable 46 dos. of 1 voriable 70ttp://www.sthe.	-0 
File Let Code 1 C - Og C - Og	at A         at -           at A         at -           read         at -		halie Tooh Hey				q	Provide the set of the set o	363 dos. of 2 voriables List of 43 List of 6 mp [1:16] 113 3 11 1 1 1 1 45 dos. of 1 voriable 46 dos. of 1 voriable 70ttp://www.sthe.	-0 
For Lot Code 1 0 - Qe Qe + 0 (1) Anny Line, code 1 month memory means degr corp desens dess degr corp dess dess degr corp anno code solution code mesonition code mesonition code mesonition code mesonition code mesonition code	at A         at -           at A         at -           read         at -		hydie Sock Helj				Q.		161 dec. of 3 variables Liter of 45 Liter of 5 Press [1149, 144] 1 1 3 1 1 1 1 1 1 43 dec. of 1 variable	-0 
No Lite Colo 1 O O O O O O O O O O O O O O O O O O O	at A         at -           at A         at -           read         at -		hutlin Sock Help				q	ar Ed 20 mayo basi + d & Gadrimovant + ★ Data Odd Odd Odds Odds Otes.i Otes.d Otes.d Otes.t Otes.t Otes.t Otes.t Otes.t Otes Ot	$ \begin{array}{l} \text{(1) dis, of 2 variables} \\ \text{(1) if of 6} \\ \text{(1) if of 6} \\ \text{(2) if of 6} \\ \text{(3) if of 6} \\ \text{(3) if of 6} \\ \text{(3) variable} $	are tio-luther-king-l-have-a.
For Lot Code 1 0 - Qe Qe + 0 (1) Anny Line, code 1 month memory means degr corp desens dess degr corp dess dess degr corp anno code solution code mesonition code mesonition code mesonition code mesonition code mesonition code	at A         at -           at A         at -           read         at -		hrefile Dock Hel				q	tar Ed (2 mayor basis + € & Conta mount + & Conta Odds	131 ans. of 3 unrelates 1410 of 4 1410 of 4 1410 of 4 1410 of 4 1410 of 1411 1411 1 1410 as. of 1 unrelate 1410 of 1411 1411 1411 1 1410 of 1411 1411 1411 1411 1411 1411 1411 1	-0 
No Ede Code V C - Qe Code V C - Code V C - Classic Line, close Hendens Hend Hendens Hend Hendens Hend Hendens Hend Hendens Hend Hendens Hende Hendens Hen	at A         at -           at A         at -           read         at -		hrefile Doth Hel				q	ar Ed 20 mayo basi + d & Gadrimovant + ★ Data Odd Odd Odds Odds Otes.i Otes.d Otes.d Otes.t Otes.t Otes.t Otes.t Otes.t Otes Ot	131 ans. of 3 unrelates 1410 of 4 1410 of 4 1410 of 4 1410 of 4 1410 of 1411 1411 1 1410 as. of 1 unrelate 1410 of 1411 1411 1411 1 1410 of 1411 1411 1411 1411 1411 1411 1411 1	are tio-luther-king-l-have-a.
No for Con 1 Con Con 1 Con Con 1 Con Con 1 C	Image         Image           Image <td></td> <td>holie Sook Hej</td> <td></td> <td></td> <td></td> <td>Q</td> <td>tar Ed (2 mayor basis + € &amp; Conta mount + &amp; Conta Odds</td> <td>131 ans. of 3 unrelates 1410 of 4 1410 of 4 1410 of 4 1410 of 4 1410 of 1411 1411 1 1410 as. of 1 unrelate 1410 of 1411 1411 1411 1 1410 of 1411 1411 1411 1411 1411 1411 1411 1</td> <td>are tio-luther-king-l-have-a.</td>		holie Sook Hej				Q	tar Ed (2 mayor basis + € & Conta mount + & Conta Odds	131 ans. of 3 unrelates 1410 of 4 1410 of 4 1410 of 4 1410 of 4 1410 of 1411 1411 1 1410 as. of 1 unrelate 1410 of 1411 1411 1411 1 1410 of 1411 1411 1411 1411 1411 1411 1411 1	are tio-luther-king-l-have-a.
No Ede Cook V C C C C C C C C C C C C C C C C C C	Image         Image           Image <td></td> <td>hadila Toola Hoji</td> <td></td> <td></td> <td></td> <td>q*</td> <td>tar Ed (2 mayor basis + € &amp; Conta mount + &amp; Conta Odds</td> <td>131 ans. of 3 unrelates 1410 of 4 1410 of 4 1410 of 4 1410 of 4 1410 of 1411 1411 1 1410 as. of 1 unrelate 1410 of 1411 1411 1411 1 1410 of 1411 1411 1411 1411 1411 1411 1411 1</td> <td>are tio-luther-king-l-have-a.</td>		hadila Toola Hoji				q*	tar Ed (2 mayor basis + € & Conta mount + & Conta Odds	131 ans. of 3 unrelates 1410 of 4 1410 of 4 1410 of 4 1410 of 4 1410 of 1411 1411 1 1410 as. of 1 unrelate 1410 of 1411 1411 1411 1 1410 of 1411 1411 1411 1411 1411 1411 1411 1	are tio-luther-king-l-have-a.
No for Colo 1 Color Color 1 Color	Image         Image           Image <td></td> <td>hulia Tuck Hol</td> <td>•</td> <td></td> <td></td> <td>ą</td> <td>tar Ed (2 mayor basis + € &amp; Conta mount + &amp; Conta Odds</td> <td>131 ans. of 3 unrelates 1410 of 4 1410 of 4 1410 of 4 1410 of 4 1410 of 1411 1411 1 1410 as. of 1 unrelate 1410 of 1411 1411 1411 1 1410 of 1411 1411 1411 1411 1411 1411 1411 1</td> <td>are tio-luther-king-l-have-a.</td>		hulia Tuck Hol	•			ą	tar Ed (2 mayor basis + € & Conta mount + & Conta Odds	131 ans. of 3 unrelates 1410 of 4 1410 of 4 1410 of 4 1410 of 4 1410 of 1411 1411 1 1410 as. of 1 unrelate 1410 of 1411 1411 1411 1 1410 of 1411 1411 1411 1411 1411 1411 1411 1	are tio-luther-king-l-have-a.
No for Code 1 C C Code 1 C C C C C C C C C C C C C C C C C C C	Image         Image         Image           me         Image         Image           me         10         Image           Image         10         Image		hrdie Took Hel				Q	tar [1] (Preparbater + € S Carls formations + € S Carls (Carls + € Carls (Carls + € Carls + €	131 ans. of 3 unrelates 1410 of 4 1410 of 4 1410 of 4 1410 of 4 1410 of 1411 1411 1 1410 as. of 1 unrelate 1410 of 1411 1411 1411 1 1410 of 1411 1411 1411 1411 1411 1411 1411 1	are tio-luther-king-l-have-a.
No for Code 1 Control Code 1 Code 1	Image         Image         Image           me         Image         Image           me         10         Image           Image         10         Image		hvilie Sout Hel				9	Bartonese	131 ans. of 3 unrelates 141 of 4 141 of 4 141 of 4 141 of 4 141 of 141 141 of 141 of	are tio-luther-king-l-have-a.
No for Code 1 C C Code 1 C C C C C C C C C C C C C C C C C C C	Image         Image         Image           me         Image         Image           me         10         Image           Image         10         Image		hvilis Toot Hel	:			9	tar [1] (Preparbater + € S Carls formations + € S Carls (Carls + € Carls (Carls + € Carls + €	131 ans. of 3 unrelates 141 of 4 141 of 4 141 of 4 141 of 4 141 of 141 141 of 141 of	are tio-luther-king-l-have-a.
No Leo Colo X Colored Colored	MA         Gr           MA         Gr           Max         Gr           Max         To           Max         To <t< td=""><td>Gete Hepkender</td><td>l + Addres</td><td></td><td>flat</td><td>alary</td><td>a</td><td>Bartonese     Bartonese     Bartonese</td><td>131 ans. of 3 unrelates 141 of 4 141 of 4 141 of 4 141 of 4 141 of 141 141 of 141 of</td><td>are tio-luther-king-l-have-a.</td></t<>	Gete Hepkender	l + Addres		flat	alary	a	Bartonese	131 ans. of 3 unrelates 141 of 4 141 of 4 141 of 4 141 of 4 141 of 141 141 of 141 of	are tio-luther-king-l-have-a.
No Leo Colo X Colored Colored	At 1         A -           At 2         A -           At 3         A           At 4         A           At 5         A           At 5         A           At 5         A           At 5         A           At 6         A           At 7         A	Gran Medicadore	roled	eulte	field	glay j	Q 	Bartonese	131 ans. of 3 unrelates 141 of 4 141 of 4 141 of 4 141 of 4 141 of 141 141 of 141 of	are tio-luther-king-l-have-a.
No Leo Colo X Colored Colored	At - 4 - 4 - 4 - 4 - 4 - 4 - 4 - 4 - 4 -	Ga to the functions	crooked	exalted 1 isee	straight	valley	a ta ta ta	Bartonese	131 ans. of 3 unrelates 141 of 4 141 of 4 141 of 4 141 of 4 141 of 141 141 of 141 of	are tio-luther-king-l-have-a.
File         Eine Code         %           Ø         • (3) anne (2) des (2	A L 4 4 A A	Ga to Machaeldee Streams 17 -01 victors 1 despate 1 despate	Crokers	eultei i i i i i i i i	straight jail	valley jangling	a ba bat bang	Bartonese	131 ans. of 3 unrelates 141 of 4 141 of 4 141 of 4 141 of 4 141 of 141 141 of 141 of	are tio-luther-king-l-have-a.
No Leo Con Y P (2) Jane (An Con P (2) Jane (	At a control of a	Gran Herberten Victors revealed despite	croata iranga discosta isymptory	e alter aster i ss i s i s i s i s i s i s i s i s i	straight jail work	valley jangling country	G Ind Ind Ind Ind Ind Ind Ind Ind Ind Ind	Bartonese	131 ans. of 3 unrelates 141 of 4 141 of 4 141 of 4 141 of 4 141 of 141 141 of 141 of	are tio-luther-king-l-have-a.
No Leo Con Y C 20 Jane Jan Con Y C 20 Jane Jane Jane Jane Jane Jane Jane Jane	Att         Att           Att	Grandender Victor result degage stronge	crosket rook tisorio juptoin juptoin	eulted it it it it it it it it it it it it it	straight jail work pride	valley jangling country america	a Indi	Bartonese	131 ans. of 3 unrelates 141 of 4 141 of 4 141 of 4 141 of 4 141 of 141 141 of 141 of	are tio-luther-king-l-have-a.
10         EM         EM<	At a for a second secon	Gronologie Vicinia Vicinia Grand Strong	r ader r radi discritication symbols hilling	alla alla alla alla alla alla alla all	1 straight jail york pride tj nighty	valley jangling country america sountains	A bed bad bad bad bad bad bad bad bad bay bay bay bay bay bay bay bay bay bay	Bartonese	131 ans. of 3 unrelates 141 of 4 141 of 4 141 of 4 141 of 4 141 of 141 141 of 141 of	are tio-luther-king-l-have-a.
No Link Const () See 10 () S	At A A A A A A A A A A A A A A A A A A	Growenser (* 1995) Victoria Victoria Growenser Bangatoria Bangatoria	orski i radi isodo	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	1 straight jail 1 work pride aighty 1 snowcapped	valley jangling country america mountains california	A here been a hore a ho	Bartonese	131 ans. of 3 unrelates 141 of 4 141 of 4 141 of 4 141 of 4 141 of 141 141 of 141 of	are tio-luther-king-l-have-a.
The last case of the second se	termine and a second se	Gronologie Vicinia resulta depart strongle defa defa defa defa defa defa defa def	crosted received received spectrost	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	1 straight jail york pride tj nighty	valley jangling country america sountains	A bed bad bad bad bad bad bad bad bad bay bay bay bay bay bay bay bay bay bay	Bartonese	131 ans. of 3 unrelates 141 of 4 141 of 4 141 of 4 141 of 4 141 of 141 141 of 141 of	are tio-luther-king-l-have-a.
No Link Const () See 10 () S	ALL CALL CALL CALL CALL CALL CALL CALL	Growense V / victors resultations dida bagatori ba	crosted received received spectrost	alla alla alla alla alla alla alla all	1 straight jail 1 work pride aighty 1 snowcapped	valley jangling country america mountains california	A here been a hore a ho	Bartonese	131 ans. of 3 unrelates 141 of 4 141 of 4 141 of 4 141 of 4 141 of 141 141 of 141 of	are tio-luther-king-l-have-a.
The last case of the second se	ALL CALL CALL CALL CALL CALL CALL CALL	Gronologie Vicinia resulta depart strongle defa defa defa defa defa defa defa def	orski i radi isodo	ealted state profession attantes attantes state	1 straight jall jall pride nighty snowcapped catholics	valley jangling country america nountains california city	4 1 1 1 1 1 1 1 1 1 1 1 1 1	Bartonese	131 ans. of 3 unrelates 141 of 4 141 of 4 141 of 4 141 of 4 141 of 141 141 of 141 of	are tio-luther-king-l-have-a.
The last Case of the second se	termine and a second se	Growholde Vicing Growholde Growhold Gro	crosted received received spectrost	ealted state profession attantes attantes state	1 straight jall jall pride nighty snowcapped catholics	valley jangling country america nountains california city	4 1 1 1 1 1 1 1 1 1 1 1 1 1	Bartonese	131 ans. of 3 unrelates 141 of 4 141 of 4 141 of 4 141 of 4 141 of 141 141 of 141 of	are tio-luther-king-l-have-a.
In the last case of the second	And a second sec	Growholde Vicing Growholde Growhold Gro	crosted received received spectrost	ealted state profession attantes attantes state	1 straight jall jall pride nighty snowcapped catholics	valley jangling country america nountains california city	4 1 1 1 1 1 1 1 1 1 1 1 1 1	Bartonese	131 ans. of 3 unrelates 141 of 4 141 of 4 141 of 4 141 of 4 141 of 141 141 of 141 of	are tio-luther-king-l-have-a.

So, to do that, this kind of a document of matrix, you see i j and v that values ith, so that values are given and then you can calculate. So, that is what we have created. Now, I have created the matrix formation of this particular observation also in real life situation this matrix will be so large and you will be not able to see like this, so, the suggestion is do not do this conversion.

Now, the row sums will be the basically the, as I told the row sums will be the value so, if I see that the most common this thing is freedom Martin Luther King, he will talk about freedom, right? When ring the dream, then day one day will come that day, we will do these that day we will do that.

So, day was, so, every morning will be very good and blah, blah blah. So, these kind of words are coming at the top. And then if I convert these two a data frame where frequency is called v and the data is called the data, the data from D looks like this, the words and their corresponding frequency. So, this is called the document term frequency data set.

So, we can create this data set the same thing in a much smaller code also I will show you at a later video that how to do it in a smaller code.



(Refer Slide Time: 30:57)

And now, with this D, but I will do is I will ask you to plot a simple bar plot which looks like this, which is basically talking about which words are common and which words are not common. So, here what I have done till this point is actually we have read the data we cleaned the data pre processing of the data has been done and we created a bar plot to show that which word is most common which one is less common, how to find out the term frequency of the words we have found out.

In the next video, we will try to see that how this kind of concepts can be done and be used for analyzing the data in a better way. Thank you for being with me. I will see you in the next video.