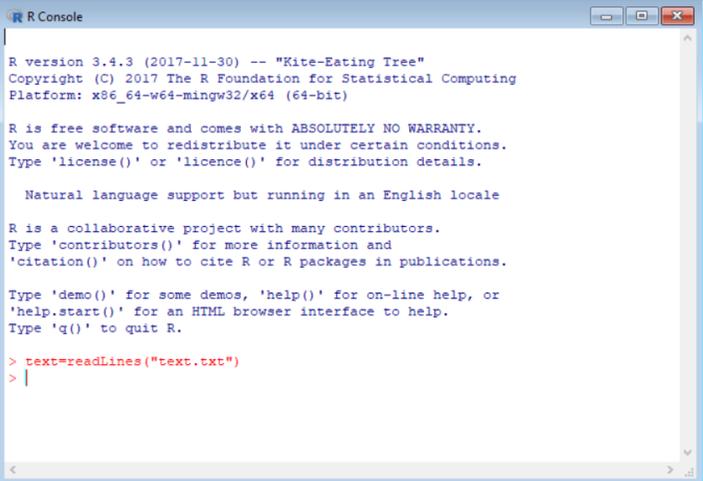
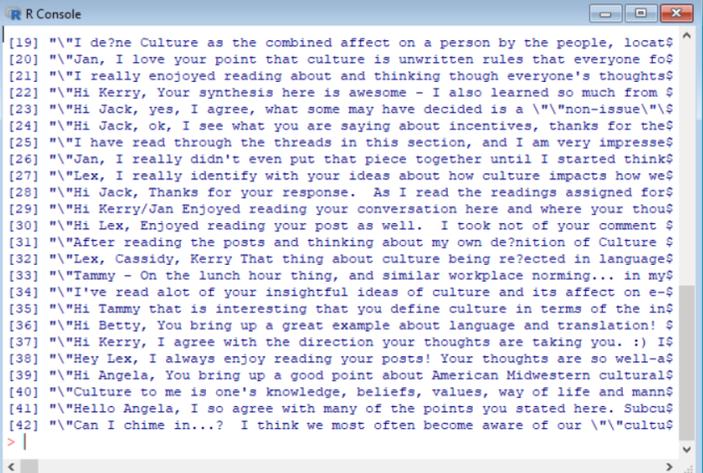


# R-Programming Fundamentals for Business Students — Cluster Analysis, Dendrograms, Word Cloud Clusters

Nick V. Flor, University of New Mexico (nickflor@unm.edu)

**Assumptions.** This tutorial assumes (1) that you had an Excel worksheet with a single column A, where each row represented a social media post (e.g., a tweet), and (2) that you at least removed carriage returns and/or line feeds from each post; and (3) that you saved that file as *text.txt* and, finally, (4) that you ran R and did a *File > Change dir...* to the folder containing *text.txt*

ACTION	REACTION
<b>IMPORTING A TEXT FILE INTO R</b>	
<ul style="list-style-type: none"> <li>Type <code>text=readLines("text.txt")</code> then press <b>Enter</b></li> </ul> <p><u>Explanation:</u> This command reads every post into a separate row in the variable <code>text</code>. If you do not see an error message, you have done this correctly! If you do get an error, retype the command (don't copy and paste)</p>	 <p>R Console</p> <pre>R version 3.4.3 (2017-11-30) -- "Kite-Eating Tree" Copyright (C) 2017 The R Foundation for Statistical Computing Platform: x86_64-w64-mingw32/x64 (64-bit)  R is free software and comes with ABSOLUTELY NO WARRANTY. You are welcome to redistribute it under certain conditions. Type 'license()' or 'licence()' for distribution details.  Natural language support but running in an English locale  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.  &gt; text=readLines("text.txt") &gt;  </pre>
<ul style="list-style-type: none"> <li>Type <code>text</code> (or whatever you named your variable) then press <b>Enter</b> to see your data</li> </ul> <p><u>Explanation:</u> Typing any variable allows you to see its content.</p> <p><u>Note:</u> The number of rows (e.g., 42) should agree with the number of rows in your Excel spreadsheet. If not, then you did not remove all Enters (carriage returns / line feeds) and you will have to go back and start over.</p>	 <p>R Console</p> <pre>[19] "\"I de?ne Culture as the combined affect on a person by the people, locat [20] "\"Jan, I love your point that culture is unwritten rules that everyone fo [21] "\"I really enjoyed reading about and thinking though everyone's thoughts [22] "\"Hi Kerry, Your synthesis here is awesome - I also learned so much from [23] "\"Hi Jack, yes, I agree, what some may have decided is a \"non-issue\" [24] "\"Hi Jack, ok, I see what you are saying about incentives, thanks for the [25] "\"I have read through the threads in this section, and I am very impress [26] "\"Jan, I really didn't even put that piece together until I started think [27] "\"Lex, I really identify with your ideas about how culture impacts how we [28] "\"Hi Jack, Thanks for your response. As I read the readings assigned for [29] "\"Hi Kerry/Jan Enjoyed reading your conversation here and where your thou [30] "\"Hi Lex, Enjoyed reading your post as well. I took not of your comment [31] "\"After reading the posts and thinking about my own de?nition of Culture [32] "\"Lex, Cassidy, Kerry That thing about culture being re?ected in language [33] "\"Tammy - On the lunch hour thing, and similar workplace norming... in my [34] "\"I've read alot of your insightful ideas of culture and its affect on e- [35] "\"Hi Tammy that is interesting that you define culture in terms of the in [36] "\"Hi Betty, You bring up a great example about language and translation! [37] "\"Hi Kerry, I agree with the direction your thoughts are taking you. :) [38] "\"Hey Lex, I always enjoy reading your posts! Your thoughts are so well-a [39] "\"Hi Angela, You bring up a good point about American Midwestern cultura [40] "\"Culture to me is one's knowledge, beliefs, values, way of life and mann [41] "\"Hello Angela, I so agree with many of the points you stated here. Subcu [42] "\"Can I chime in...? I think we most often become aware of our \"cultu &gt;  </pre>

(continued on next page)

## INSTALL PACKAGES

- If you haven't done so in a previous tutorial, install the following packages:  
`install.packages("tm")`  
`install.packages("SnowballC")`  
`install.packages("wordcloud")`

Note: This can take several minutes and freeze your keyboard. Just wait it out.

### Explanation:

"tm" is R's text mining package

"SnowballC" is a package to stem words (e.g., "turn", "turns", "turning", "turned", all become "turn")

"wordcloud" generates word clouds.

```
R Console
[2270] "\"You know something's up when your Dean offers to ply you with
> install.packages("tm")
Installing package into 'C:/Users/profe/Documents/R/win-library/3.4'
(as 'lib' is unspecified)
--- Please select a CRAN mirror for use in this session ---
trying URL 'https://cran.revolutionanalytics.com/bin/windows/contrib/3.4'
Content type 'application/zip' length 1272241 bytes (1.2 MB)
downloaded 1.2 MB

package 'tm' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
  C:/Users/profe/AppData/Local/Temp/RtmpuW5riL/downloaded_packages
> install.packages("SnowballC")
Installing package into 'C:/Users/profe/Documents/R/win-library/3.4'
(as 'lib' is unspecified)
trying URL 'https://cran.revolutionanalytics.com/bin/windows/contrib/3.4'
Content type 'application/zip' length 3082209 bytes (2.9 MB)
downloaded 2.9 MB

package 'SnowballC' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
  C:/Users/profe/AppData/Local/Temp/RtmpuW5riL/downloaded_packages
> install.packages("wordcloud")
Installing package into 'C:/Users/profe/Documents/R/win-library/3.4'
(as 'lib' is unspecified)
trying URL 'https://cran.revolutionanalytics.com/bin/windows/contrib/3.4'
Content type 'application/zip' length 566489 bytes (553 KB)
downloaded 553 KB

package 'wordcloud' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
  C:/Users/profe/AppData/Local/Temp/RtmpuW5riL/downloaded_packages
> |
<
```

- Type the following:  
`library("tm")`  
`library("SnowballC")`  
`library("wordcloud")`

Explanation: `install.packages` merely loads the packages onto your computer. The `library` command allows you to use the packages in your current R session.

```
R Console
trying URL 'https://cran.revolutionanalytics.com/bin/windows/contrib/3.4'
Content type 'application/zip' length 3082209 bytes (2.9 MB)
downloaded 2.9 MB

package 'SnowballC' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
  C:/Users/profe/AppData/Local/Temp/Rtmp8WYijQ/downloaded_packages
> install.packages("wordcloud")
Installing package into 'C:/Users/profe/Documents/R/win-library/3.4'
(as 'lib' is unspecified)
trying URL 'https://cran.revolutionanalytics.com/bin/windows/contrib/3.4'
Content type 'application/zip' length 566494 bytes (553 KB)
downloaded 553 KB

package 'wordcloud' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
  C:/Users/profe/AppData/Local/Temp/Rtmp8WYijQ/downloaded_packages
> library(tm)
Loading required package: NLP
> library(SnowballC)
> library(wordcloud)
Loading required package: RColorBrewer
> |
<
```

## CLEAN DATA (DATA CLEANING)

- Clean the data. One way to do so is to type:

```
text1=tolower(text)
text2=removeWords(text1, stopwords("english"))
text3=gsub("/", " ", text2)
text4=removePunctuation(text3)
text5=stemDocument(text4)
text6=stripWhitespace(text5)
```

Explanation: Converts the original text to lower case, while removing stopwords, converting slashes to spaces, removing punctuation, stemming the document, and stripping whitespace.

```
R Console
> install.packages("wordcloud")
Installing package into 'C:/Users/profe/Documents/R/win-library/3.4'
(as 'lib' is unspecified)
trying URL 'https://cran.revolutionanalytics.com/bin/windows/contrib/3.4'
Content type 'application/zip' length 566489 bytes (553 KB)
downloaded 553 KB

package 'wordcloud' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
  C:/Users/profe/AppData/Local/Temp/RtmpuW5riL/downloaded_packages
> library("tm")
Loading required package: NLP
Warning message:
package 'tm' was built under R version 3.4.3
> library("SnowballC")
> library("wordcloud")
Loading required package: RColorBrewer
Warning message:
package 'wordcloud' was built under R version 3.4.3
> text1=Corpus(VectorSource(text))
> |
<
```

## CREATE HIERARCHICAL CLUSTER & PLOT DENDROGRAM

- Type `corp=Corpus(VectorSource(text6))`

Explanation: a **Corpus** is a special data structure which we need to create first before we can create a **DocumentTermMatrix**. We store the corpus in the variable `corp`.

```
R Console
[20] "jan love point cultur unwritten rule everyon follow make think dif
[21] "realli enojoy read think though everyon thought cultur elearn deci
[22] "hi kerri synthesi awesom also learn much discuss denit differ stat
[23] "hi jack yes agre may decid nonissu can mean differ someone day kerr
[24] "hi jack ok see say incen thank claric rememb rst began olit progr
[25] "read thread section impress level thought extent teach experi embe
[26] "jan realli even put piec togeth start think other say contribut fa
[27] "lex realli identifi idea cultur impact can translat inform rememb
[28] "hi jack thank respons read read assign week also includ iron silk
[29] "hi kerri jan enjoy read convers thought lead question probabl subj
[30] "hi lex enjoy read post well took comment emili comment languag cul
[31] "read post think denit cultur elearn decid approach point view hush
[32] "lex cassidi kerri thing cultur reect languag realli struck spent f
[33] "tammi lunch hour thing similar workplac norm opinion incumb manag
[34] "read alot insight idea cultur affect elearn cultur valu belief tra
[35] "hi tammi interest defin cultur term individu never look perspect a
[36] "hi betti bring great exampl languag translat mani languag mean can
[37] "hi kerri agre direct thought take think general awar affect cultur
[38] "hey lex alway enjoy read post thought wellarticul great bring ques
[39] "hi angela bring good point american midwestern cultur vast differ
[40] "cultur one knowledg belief valu way life manner general form assoc
[41] "hello angela agre mani point state subcultur can formul around us
[42] "can chime think often becom awar cultur program collid other diffe
> corp=Corpus(VectorSource(text6))
> |
<
```

- Type `dtm=DocumentTermMatrix(corp)`

Explanation: The **Corpus** object (`corp` in our example) is converted to a **DocumentTermMatrix**, which we store in the variable `dtm`. The rows of this matrix correspond to the individual text lines. The columns, however, for each row, correspond to every single word across all text lines.

Typing `dtm` summarizes the matrix, but does not show you the contents of the matrix.

```
R Console
[21] "realli enojoy read think though everyon thought cultur elearn deci
[22] "hi kerri synthesi awesom also learn much discuss denit differ stat
[23] "hi jack yes agre may decid nonissu can mean differ someone day kerr
[24] "hi jack ok see say incen thank claric rememb rst began olit progr
[25] "read thread section impress level thought extent teach experi embe
[26] "jan realli even put piec togeth start think other say contribut fa
[27] "lex realli identifi idea cultur impact can translat inform rememb
[28] "hi jack thank respons read read assign week also includ iron silk
[29] "hi kerri jan enjoy read convers thought lead question probabl subj
[30] "hi lex enjoy read post well took comment emili comment languag cul
[31] "read post think denit cultur elearn decid approach point view hush
[32] "lex cassidi kerri thing cultur reect languag realli struck spent f
[33] "tammi lunch hour thing similar workplac norm opinion incumb manag
[34] "read alot insight idea cultur affect elearn cultur valu belief tra
[35] "hi tammi interest defin cultur term individu never look perspect a
[36] "hi betti bring great exampl languag translat mani languag mean can
[37] "hi kerri agre direct thought take think general awar affect cultur
[38] "hey lex alway enjoy read post thought wellarticul great bring ques
[39] "hi angela bring good point american midwestern cultur vast differ
[40] "cultur one knowledg belief valu way life manner general form assoc
[41] "hello angela agre mani point state subcultur can formul around us
[42] "can chime think often becom awar cultur program collid other diffe
> corp=Corpus(VectorSource(text6))
> dtm=DocumentTermMatrix(corp)
> |
<
```

- Type  
`mat=as.matrix(dtm)`  
`mat`

Explanation: `as.matrix` converts the **DocumentTermMatrix** into a matrix, which we store in `mat`, that you can examine directly by typing the variable name. However, the matrix is so big that R doesn't display it properly.

Note: If you want to examine the matrix, you can type `write.csv(mat, "mat.csv")` and import the file into Excel (not shown).

```
R Console
19 0 0 0 0 0 0 0 0 0 0 0 0 0 0
20 0 0 0 0 0 0 0 0 0 0 0 0 0 0
21 0 0 0 0 0 0 0 0 0 0 0 0 0 0
22 0 0 0 0 0 0 0 0 0 0 0 0 0 0
23 0 0 0 0 0 0 0 0 0 0 0 0 0 0
24 0 0 0 0 0 0 0 0 0 0 0 0 0 0
25 0 0 0 0 0 0 0 0 0 0 0 0 0 0
26 0 0 0 0 0 0 0 0 0 0 0 0 0 0
27 0 0 0 0 0 0 0 0 0 0 0 0 0 0
28 0 0 0 0 0 0 0 0 0 0 0 0 0 0
29 0 0 0 0 0 0 0 0 0 0 0 0 0 0
30 0 0 0 0 0 0 0 0 0 0 0 0 0 0
31 0 0 0 0 0 0 0 0 0 0 0 0 0 0
32 0 0 0 0 0 0 0 0 0 0 0 0 0 0
33 0 0 0 0 0 0 0 0 0 0 0 0 0 0
34 0 0 0 0 0 0 0 0 0 0 0 0 0 0
35 0 0 0 0 0 0 0 0 0 0 0 0 0 0
36 0 0 0 0 0 0 0 0 0 0 0 0 0 0
37 0 0 0 0 0 0 0 0 0 0 0 0 0 0
38 0 0 0 0 0 0 0 0 0 0 0 0 0 0
39 0 0 0 0 0 0 0 0 0 0 0 0 0 0
40 0 0 0 0 0 0 0 0 0 0 0 0 0 0
41 0 0 0 0 0 0 0 0 0 0 0 0 0 0
42 1 1 1 1 1 1 1 1 1 3 1 1 1 1
> |
<
```

- Type `d=dist(mat)`

Explanation: The `dist` command calculates the similarity between each line of text and every other line of text. Two lines of text that have high overlap between words are said to be “close” in distance. Two lines of text that do not have many words in common are said to be “far” in distance. We store the distances between lines of text in the variable `d`.

```
R Console
20 0 0 0 0 0 0 0 0 0 0 0 0 0
21 0 0 0 0 0 0 0 0 0 0 0 0 0
22 0 0 0 0 0 0 0 0 0 0 0 0 0
23 0 0 0 0 0 0 0 0 0 0 0 0 0
24 0 0 0 0 0 0 0 0 0 0 0 0 0
25 0 0 0 0 0 0 0 0 0 0 0 0 0
26 0 0 0 0 0 0 0 0 0 0 0 0 0
27 0 0 0 0 0 0 0 0 0 0 0 0 0
28 0 0 0 0 0 0 0 0 0 0 0 0 0
29 0 0 0 0 0 0 0 0 0 0 0 0 0
30 0 0 0 0 0 0 0 0 0 0 0 0 0
31 0 0 0 0 0 0 0 0 0 0 0 0 0
32 0 0 0 0 0 0 0 0 0 0 0 0 0
33 0 0 0 0 0 0 0 0 0 0 0 0 0
34 0 0 0 0 0 0 0 0 0 0 0 0 0
35 0 0 0 0 0 0 0 0 0 0 0 0 0
36 0 0 0 0 0 0 0 0 0 0 0 0 0
37 0 0 0 0 0 0 0 0 0 0 0 0 0
38 0 0 0 0 0 0 0 0 0 0 0 0 0
39 0 0 0 0 0 0 0 0 0 0 0 0 0
40 0 0 0 0 0 0 0 0 0 0 0 0 0
41 0 0 0 0 0 0 0 0 0 0 0 0 0
42 1 1 1 1 1 1 1 1 3 1 1 1 1
> d=dist(mat)
> |
<
```

- Type `hc=hclust(d)`

Explanation: The command `hclust` clusters our lines of text two lines at a time, starting with the two most similar, then finding the line most similar to those two, then finding the line most similar to those three, and so on. We store the result in a variable `hc`.

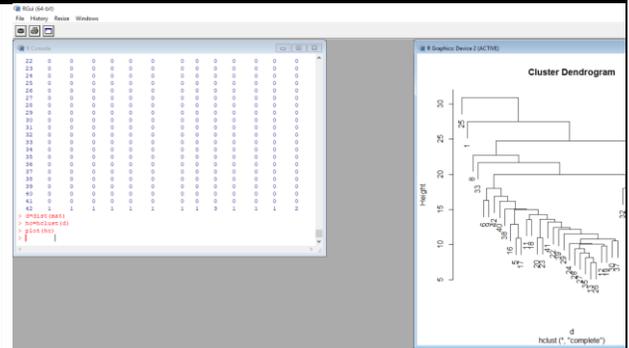
The problem is that we can't see the cluster until we plot it. We do that next.

```
R Console
21 0 0 0 0 0 0 0 0 0 0 0 0 0
22 0 0 0 0 0 0 0 0 0 0 0 0 0
23 0 0 0 0 0 0 0 0 0 0 0 0 0
24 0 0 0 0 0 0 0 0 0 0 0 0 0
25 0 0 0 0 0 0 0 0 0 0 0 0 0
26 0 0 0 0 0 0 0 0 0 0 0 0 0
27 0 0 0 0 0 0 0 0 0 0 0 0 0
28 0 0 0 0 0 0 0 0 0 0 0 0 0
29 0 0 0 0 0 0 0 0 0 0 0 0 0
30 0 0 0 0 0 0 0 0 0 0 0 0 0
31 0 0 0 0 0 0 0 0 0 0 0 0 0
32 0 0 0 0 0 0 0 0 0 0 0 0 0
33 0 0 0 0 0 0 0 0 0 0 0 0 0
34 0 0 0 0 0 0 0 0 0 0 0 0 0
35 0 0 0 0 0 0 0 0 0 0 0 0 0
36 0 0 0 0 0 0 0 0 0 0 0 0 0
37 0 0 0 0 0 0 0 0 0 0 0 0 0
38 0 0 0 0 0 0 0 0 0 0 0 0 0
39 0 0 0 0 0 0 0 0 0 0 0 0 0
40 0 0 0 0 0 0 0 0 0 0 0 0 0
41 0 0 0 0 0 0 0 0 0 0 0 0 0
42 1 1 1 1 1 1 1 1 3 1 1 1 1
> d=dist(mat)
> hc=hclust(d)
> |
<
```

## PLOT THE DENDROGRAM

- Type `plot(hc)`

Explanation: This plots a tree diagram from the hierarchical cluster (`hc`) known as a dendrogram.



- You can resize the dendrogram by resizing the plot window (drag the borders!). This lets you see the clustering a lot better.

