International Journal of Trend in Research and Development, Volume 4(5), ISSN: 2394-9333
www.ijtrd.com

# Text Mining and Big Data Analysis in the Relational Database with R

Gye-Soo, Kim
Professor, Department of Business Administration, Semyung University, Korea

***Abstract:*** During the last decade text mining and big data analysishas become a widely used discipline utilizing statistical and machine learning methods. We present the r package which provides a framework for text mining applications within R. We give a survey on text mining facilities in R and explain how typical application tasks can be carried out using our framework. We present techniques for count-based analysis methods, text clustering, and text classification. This article will provide compelling practical implication and possible strategy based on text mining. Graphic results can move customer mind and interest.

***Keywords:*** *Text Mining, R, Count-Based Evaluation, Text Clustering, Text Classification, Graphic Results.*

## I. INTRODUCTION

Big data analytics is the use of advanced analytic techniques against very large, diverse data sets that include different types such as structured/unstructured and streaming, and different sizes from terabytes to zettabytes.Analyzing big data allows analysts, researchers, and business users to make better and faster decisions using data that was previously inaccessible or unusable. Twenty-first century innovation is increasingly about 'open innovation' a multi-player game where connections and the ability to find, form and deploy creative relationships is of the essence (Bessant and Tidd, 2011).

In theses day, text mining encompasses a vast field of theoretical approaches and methods with one thing in common: text as input information. This allows various definitions, ranging from an extension of classical data mining to texts to more sophisticated formulations like "the use of large online text collections to discover new facts and trends about the world itself."

If you work in analytics or data science, like we do, you are familiar with the fact that data is being generated all the time at ever faster rates. Analysts are often trained to handle tabular or rectangular data that is mostly numeric, but much of the data proliferating today is unstructured and text-heavy. Many analysts who work in analytical fields are not trained in even simple interpretation of natural language.

We developed the tidytext (Silge and Robinson, 2016) R package because we were familiar with many methods for data wrangling and visualization, but couldn't easily apply these same methods to text. We found that using tidy data principles can make many text mining tasks easier, more effective, and consistent with tools already in wide use. Treating text as data frames of individual words allows us to manipulate, summarize, and visualize the characteristics of text easily and integrate natural language processing into effective workflows we were already using.

This thesis serves as an introduction of text mining using the r package and other package tools in R. After text mining and bigdata analysis, this article will provide compelling practical implication and possible strategy based on text mining.

## II. R AND TEXTMINING

The benefit of text mining comes with the large amount of valuable information latent in texts which is not available in classical structured data formats for various reasons: text has always been the default way of storing information for hundreds of years, and mainly time, personal and cost contraints prohibit us from bringing texts into well structured formats (like data frames or tables).R provides details on other ways to use computational linguistics. There are several areas that authors may want to explore in more detail according to theirneeds(Silge, Robinson, 2017).

Pulling data off of social media networks typically involves the use of the network's API, which requires both coding time and skill. The social media oriented packages in R are intended to do this heavy lifting for us, which is one of the upsides of using R packages for social media data analysis. The packages provide a relatively easy interface with social media data, without requiring a lot of coding. And we can structure the data we pull into formats amenable to other analyses in R, such as a sentiment analysis of Tweets or a visualization using popular R graphics packages. The main downside, however, is that we are limited to the functionality of the R package. Fortunately, however, most of the packages provide great functionality. I'll provide an R script file to get us started and point you toward additional resources.

- Clustering, classification, and prediction: Machine learning on text is a vast topic that could easily fill its own volume. In theses day, many more machine learning algorithms can be used in dealing with text.
- Word embedding: One popular modern approach for text analysis is to map words to vector representations, which can then be used to examine linguistic relationships between words and to classify text. Such representations of words are not tidy in the sense but have found powerful applications in machine learning algorithms.
- More complex tokenization: The R (tidytext) package trusts the tokenizers package to perform tokenization, which itself wraps a variety of tokenizers with a consistent interface, but many others exist for specific applications.
- Languages other than English: Some of our users have had success applying tidytext to their text mining needs for languages other.

And general text mining process is the following figure.

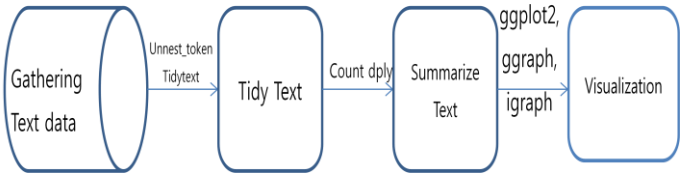


Figure 1: A flowchart of a typical text analysis process

## III. VISUALIZING A NETWORK WITH GGRAPH

Network science offers a language through which different disciplines can seamlessly interact with each other. Indeed, cell biologists, brain scientists and computer scientists alike are faced with the task of characterizing the wiring diagram behind their system, extracting information from incomplete and noisy datasets, and understanding their systems' robustness to failures or attacks. We live in a web-based view of nature, society, and complex businessenvironment. In this circumstance, we need to knownew framework for understanding issues ranging from democracy on the Web to the vulnerability of the Internet and the spread of deadly viruses(Barabási, 2002). Networks are present everywhere.

A network theorist will recognize super-spreaders as hubs, nodes with an exceptional number of links in the contact network on which a disease spreads. As hubs appear in many networks, super-spreaders have been documented in many infectious diseases, social network service and social network field.

Communities play a particularly important role in our understanding of how specific biological functions are encoded in cellular networks. Understanding community is the axiom of marketing. To uncover the community structure of large real networks we need algorithms whose running time grows polynomially with N. Hierarchical clustering, the topic of this section, helps us achieve this goal.

Analysts may be interested in visualizing all of the relationships among words simultaneously, rather than just the top few at a time. As common visualization, Analysts can arrange the words into a network, or "graph." Here we'll be referring to a "graph" not in the sense of the visualization, but as a combination of connected nodes. A graph can be constructed from a tidy object since it has three variables:

Table 1: Network object

| |
|---|
| **from**: the node an edge is coming from<br>**to**: the node an edge is going towards<br>**weight**: A numeric value associated with each edge |

To uncover the community structure of large real networks,Analyst need to know algorithms. We are trying to research decision making structure in a company. This company focuses on horizontal organization culture, creativity, and continuous improvement for development. The igraph package has many powerful functions for manipulating and analyzing networks. We researched decision making method and made the relational database in 2016 and 2017. One way to create an igraph object from tidy data is the graph_from_data_frame() function, which takes a data frame of edges with columns for "from", "to", and edge attributes (in this case n). The following is algorithm of decision making process in one company. Data is a relational data base.

```
library(ggplot2)
library(ggraph)
library(igraph)
uni=read.csv("D:/kistep/data/uni.csv")
head(uni)
graph <- graph_from_data_frame(uni)
p <- ggraph(graph, layout = 'kk') +
geom_edge_link(aes(colour = factor(year))) +
geom_node_point() +
ggtitle('Decision Making Structure')
p
p + theme_graph()
p + theme_graph(background = 'grey20', text_colour = 'white')
set_graph_style()
p

layout <- createLayout(graph, layout = 'drl')
ggraph(layout) +
geom_edge_link(aes(colour = factor(year))) +
geom_node_point()
p
```

Figure 2: Syntax for visual map

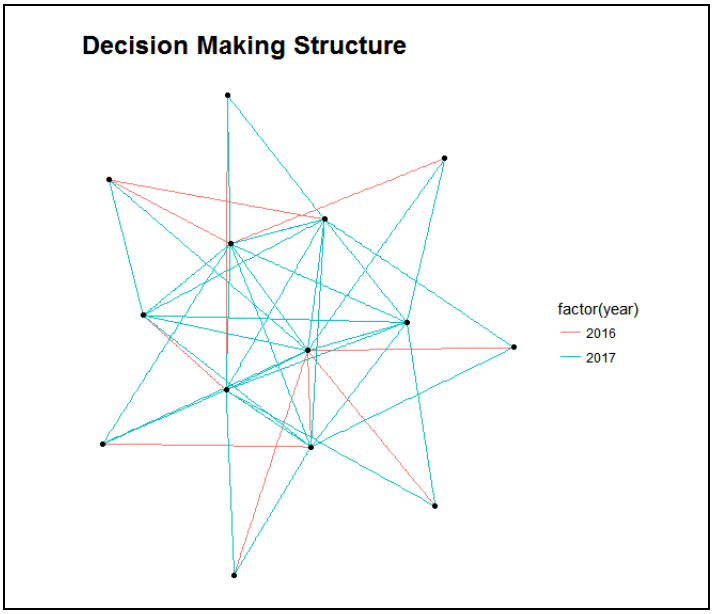


Figure 3: Decision Making Structure

This company is trying to deal with an uncertain world by constantly trying new organization culture. To sustain and create value, this company adopted innovation culture and established horizontal organization culture more than 2016. And analyst can confirm who is hub of decision making.

```
library(ggplot2)
library(ggraph)
library(igraph)
# We use a friendship network
uni=read.csv("D:/kistep/data/uni.csv")
friendGraph<-
graph_from_data_frame(uni)
V(friendGraph)$degree <-
degree(friendGraph, mode = 'in')
graph2016 <-
subgraph.edges(friendGraph,
which(E(friendGraph)$year ==2016),
F)
graph2017<-
subgraph.edges(friendGraph,
which(E(friendGraph)$year ==2017),
F)
V(friendGraph)$pop.increase<-
degree(graph2017, mode = 'in') >
   degree(graph2016, mode = 'in')
ggraph(friendGraph, 'igraph',
algorithm = 'kk') +
geom_edge_fan(aes(alpha
= ..index..)) +
eom_node_point(aes(size = degree,
colour = pop.increase)) +
cale_edge_alpha('relations with',
guide = 'edge_direction') +
scale_colour_manual('Improved',
values = c('firebrick', 'forestgreen')) +
scale_size('# Relations') +
facet_wrap(~year) +
ggforce::theme_no_axes()
```

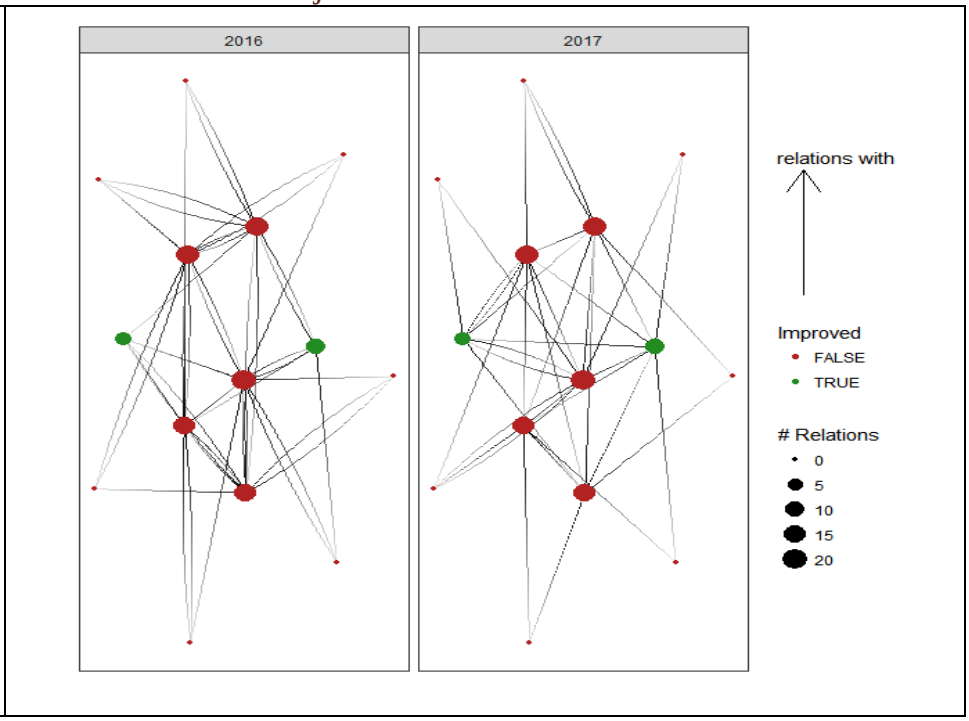


Figure 4: Decision Making Structure in 2016, 2017

## CONCLUSION

Network is Science. Network science is a new discipline.If analysts know about secret of network in system he or she can establish various strategies. Most networks facilitate transfer along their links: transfer of trust, knowledge, habits or information (social networks), electricity (power grid), money (financial networks), goods (trade networks). To understand these phenomena, we must understand how the network topology affects these dynamical processes. Usually products and ideas spread by being adapted by the hubs. Hub is the highly connected nodes of the consumer network.Hub is main activity and role model in decision making.

This thesis demonstrated how treating text as data frames enables analyst to manipulate, summarize, and visualize characteristics of text. Analysts also learn how to integrate natural language processing (NLP) into effective workflows. Practical code examples and data explorations will help reader generate real insights from literature, news, and social media. In this thesis, we discussed the forces that have led to the emergence of this new research field and its impact on science, technology, and society.We should seek to develop models and theories to explain why, when, and where we do the things we do with some regularity.

The innovation of organization is not a solo act. Successful players work hard to build links across boundaries inside the organizations and to the many external agencies who can play a part in the innovation process-suppliers, customers, source of finance, skilled resources and of the knowledge, etc.

### *References*

[1] Barabási.A.-L. (2002), Linked: The New Science of Networks, basic book.
[2] Barabási.A.-L.(2005),The origin of bursts and heavy tails in human dynamics. Nature, 435:207-11.
[3] Bessant, J., Tidd, J.(2011), Innovation and Entrepreneurship, 2nd edition, Wiley.
[4] Silge, J., Robinson, D. (2017), Welcome to Text Mining with R, OREILLY.