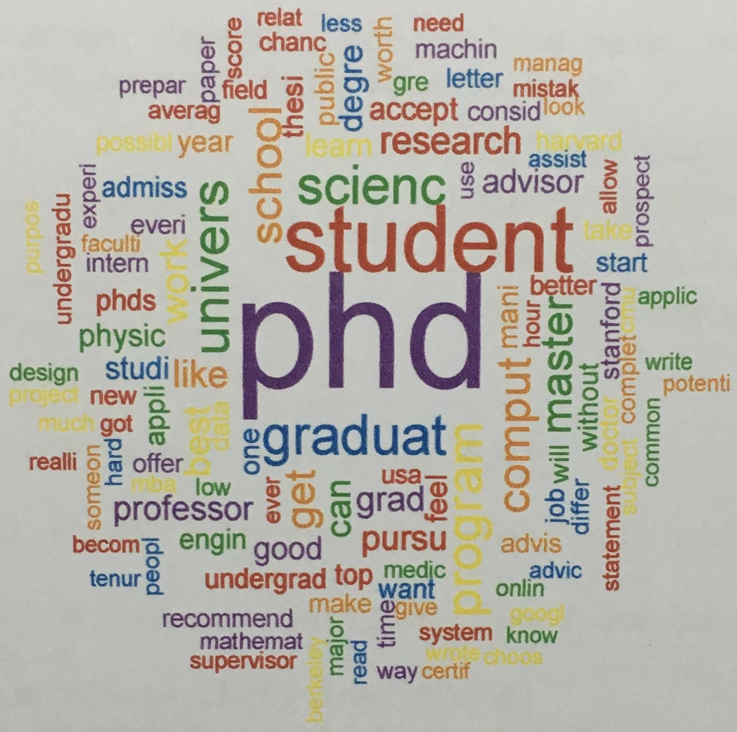
```
R version 3.3.2 (2016-10-31) -- "Sincere Pumpkin Patch"
Copyright (C) 2016 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin13.4.0 (64-bit)

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.

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.

During startup - Warning messages:
1: Setting LC_CTYPE failed, using "C"
2: Setting LC_COLLATE failed, using "C"
3: Setting LC_TIME failed, using "C"
4: Setting LC_MESSAGES failed, using "C"
5: Setting LC_MONETARY failed, using "C"
[R.app GUI 1.68 (7288) x86_64-apple-darwin13.4.0]

WARNING: You're using a non-UTF8 locale, therefore only ASCII characters will
work.
Please read R for Mac OS X FAQ (see Help) section 9 and adjust your system
preferences accordingly.
[Workspace restored from /Users/Christine/.RData]
[History restored from /Users/Christine/.Rapp.history]

> setwd("/Users/Christine/Documents/")
> library(tm)
Loading required package: NLP
> library(SnowballC)
> library(wordcloud)
Loading required package: RColorBrewer
> data1 <- readLines("quora_data.txt")
> head(data1)
[1] "Besides Stanford and CMU, which graduate schools have a prestigious
machine learning program?"
[2] "I am pre-med, should I transfer from UM-Dearborn to UM-Ann Arbor as an
undergrad?"
[3] "Why do people seem to hate their PhD?"
[4] "Why do graduate students allow themselves to be exploited like cheap
labor?"
[5] "Has anyone ever regretted getting a PhD?"
[6] "What are the main differences between a Masters and a PhD in computer
science?"
>
> docs <- Corpus(VectorSource(data1))
> docs
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 334
>
> toSpace<-content_transformer(function(x, pattern) gsub(pattern, " ", x))
```

#필요한 패키지 설치및 라이브러리 가동

#1) 제가 분석 대상으로 설정한 사이트는 미국의 지식인과 같은 사이트 'Quora'입니다.
'Graduate School Admission' 카테고리의 질문들을 분석함으로써
Graduate School 에 대해 미국 대학생들은 무엇을 물어보는지 알고 싶었습니다.

#총 334개의 질문을 수집하였습니다

```
> docs <- tm_map(docs, toSpace, "/")
> docs <- tm_map(docs, toSpace, "@")
> docs <- tm_map(docs, toSpace, "\\|")
> docs <- tm_map(docs, content_transformer(tolower))
> docs <- tm_map(docs, removeNumbers)
> docs <- tm_map(docs, removeWords, stopwords("english"))
> docs <- tm_map(docs, removePunctuation)
> docs <- tm_map(docs, stripWhitespace)
> docs <-tm_map(docs, removeWords, c("<a1><b0>", "<80>"))
> docs <-tm_map(docs, stemDocument)
> docs                                    # tm_map을 통해 필요없는 부분 ("$?"과 같은 Punctuation등)을 빼거나
<<VCorpus>>                               대문자를 소문자로 바꾸는등 데이터 정제 완료
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 334
>
> dtm <- TermDocumentMatrix(docs)
> m <- as.matrix(dtm)
> View(m)                                 # RStudio에서는 문제없이 표가 뜨는데 R로만 돌리면 에러메시지가 뜹니다
Error in check_for_XQuartz() :
  X11 library is missing: install XQuartz from xquartz.macosforge.org
> nrow(m)
[1] 699
> str(m)
 num [1:699, 1:334] 0 0 0 0 0 0 0 0 0 0 ...
 - attr(*, "dimnames")=List of 2                    #데이터를 단어별로 수집
  ..$ Terms: chr [1:699] "<80>" "<a1><b0>" "abolish" "abroad" ...
  ..$ Docs : chr [1:334] "1" "2" "3" "4" ...
> v <-sort(rowSums(m),decreasing=TRUE)
> View(v)                                 #RStudio에서는 문제없이 표가 뜨는데 R로만 돌리면 에러메시지가 뜹니다
Error in check_for_XQuartz() :
  X11 library is missing: install XQuartz from xquartz.macosforge.org
>
> d <-data.frame(word = names(v),freq=v)
> head(d, 10)                             #데이터를 단어의 빈도수별로 정제
           word freq
phd         phd  157
student student   80
graduat graduat   41
scienc   scienc   38
univers univers   38
program program   37
comput   comput   31
school   school   30
master   master   28
get         get   26
>
> wordcloud(words = d$word, freq = d$freq, min.freq=4, max.words=200,
random.order=F, random.color=T, rot.per=0.35, colors=brewer.pal(6, "Set1"))
>                                                   #워드클라우드 생성
```


phd
student
graduat
univers
program
comput
master
scienc
school
research
professor
relat
less
need
chanc
machin
score
paper
worth
degre
public
thesi
manag
gre
letter
prepar
field
mistak
averag
accept
consid
look
possibl
year
learn
harvard
assist
admiss
experi
use
advisor
allow
prospect
purpos
everi
faculti
take
undergradu
intern
start
work
better
applic
phds
mani
hour
stanford
physic
write
design
studi
like
potenti
project
new
appli
without
complet
cmu
doctor
much
got
best
data
common
one
will
realli
offer
hard
usa
someon
mba
low
ever
get
can
grad
feel
job
differ
subject
statement
becom
engin
good
pursu
advis
peopl
tenur
undergrad
top
medic
advic
want
onlin
make
recommend
give
googl
time
mathemat
system
know
major
read
wrote
choos
supervisor
berkeley
way
certif

완성된 워드클라우드

```
R version 3.3.2 (2016-10-31) -- "Sincere Pumpkin Patch"
Copyright (C) 2016 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin13.4.0 (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.

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.
```

#1) 정형데이터 분석을 위해서는 유니크한 데이터베이스에서 세계각 나라별로 R&D에 종사하는 Personnel 수를 정리한 데이터를 받았습니다 나라는 Austria부터 Korea까지 15개국으로 정제했고 (좀더 확실한 시각화를 위해 연도는 비교하기쉽도록 2011년, 2013년을 선정했습니다

```
During startup - Warning messages:
1: Setting LC_CTYPE failed, using "C"
2: Setting LC_COLLATE failed, using "C"
3: Setting LC_TIME failed, using "C"
4: Setting LC_MESSAGES failed, using "C"
5: Setting LC_MONETARY failed, using "C"
[R.app GUI 1.68 (7288) x86_64-apple-darwin13.4.0]

WARNING: You're using a non-UTF8 locale, therefore only ASCII characters will
work.
Please read R for Mac OS X FAQ (see Help) section 9 and adjust your system
preferences accordingly.
[Workspace restored from /Users/Christine/.RData]
[History restored from /Users/Christine/.Rapp.history]

> data <- read.csv("/Users/Christine/Downloads/rnd_3.csv", header=T)
Warning message:
In read.table(file = file, header = header, sep = sep, quote = quote,  :
  incomplete final line found by readTableHeader on '/Users/Christine/
Downloads/rnd_3.csv'
> head(data)
  Country Austria Belgium Canada Czech.Republic Denmark Finland France
1    2013   66186   67899 226620          61976   58246   52972 418141
2    2011   61171   62895 239920          55697   57585   54526 402492
  Germany Greece Hungary Iceland Ireland  Italy  Japan  Korea
1  588615  42188   38163    2766   24129 246764 865523 401444
2  575099  36913   33960    3244   21591 228094 869825 361374
> summary(data)
```

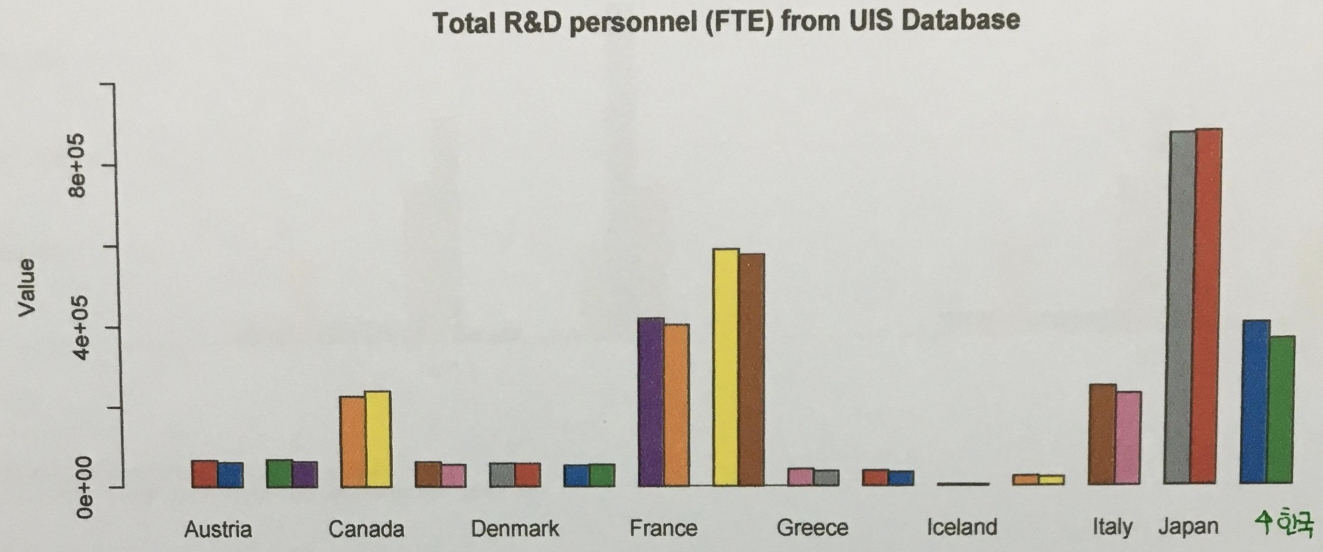
#데이터의 기본적인 형태들

```
    Country        Austria         Belgium         Canada
 Min.   :2011   Min.   :61171   Min.   :62895   Min.   :226620
 1st Qu.:2012   1st Qu.:62425   1st Qu.:64146   1st Qu.:229945
 Median :2012   Median :63678   Median :65397   Median :233270
 Mean   :2012   Mean   :63678   Mean   :65397   Mean   :233270
 3rd Qu.:2012   3rd Qu.:64932   3rd Qu.:66648   3rd Qu.:236595
 Max.   :2013   Max.   :66186   Max.   :67899   Max.   :239920
 Czech.Republic     Denmark         Finland         France
 Min.   :55697   Min.   :57585   Min.   :52972   Min.   :402492
 1st Qu.:57267   1st Qu.:57750   1st Qu.:53360   1st Qu.:406404
 Median :58836   Median :57916   Median :53749   Median :410316
 Mean   :58836   Mean   :57916   Mean   :53749   Mean   :410316
 3rd Qu.:60406   3rd Qu.:58081   3rd Qu.:54138   3rd Qu.:414229
```

```
Max.   :61976   Max.   :58246   Max.   :54526   Max.   :418141
   Germany          Greece          Hungary         Iceland
Min.   :575099   Min.   :36913   Min.   :33960   Min.   :2766
1st Qu.:578478   1st Qu.:38232   1st Qu.:35011   1st Qu.:2886
Median :581857   Median :39550   Median :36062   Median :3005
Mean   :581857   Mean   :39550   Mean   :36062   Mean   :3005
3rd Qu.:585236   3rd Qu.:40869   3rd Qu.:37112   3rd Qu.:3124
Max.   :588615   Max.   :42188   Max.   :38163   Max.   :3244
   Ireland          Italy           Japan           Korea
Min.   :21591   Min.   :228094   Min.   :865523   Min.   :361374
1st Qu.:22226   1st Qu.:232762   1st Qu.:866598   1st Qu.:371392
Median :22860   Median :237429   Median :867674   Median :381409
Mean   :22860   Mean   :237429   Mean   :867674   Mean   :381409
3rd Qu.:23494   3rd Qu.:242096   3rd Qu.:868750   3rd Qu.:391426
Max.   :24129   Max.   :246764   Max.   :869825   Max.   :401444
> str(data)
'data.frame':  2 obs. of  16 variables:
 $ Country       : int  2013 2011
 $ Austria       : int  66186 61171
 $ Belgium       : int  67899 62895
 $ Canada        : int  226620 239920
 $ Czech.Republic: int  61976 55697
 $ Denmark       : int  58246 57585
 $ Finland       : int  52972 54526
 $ France        : int  418141 402492
 $ Germany       : int  588615 575099
 $ Greece        : int  42188 36913
 $ Hungary       : int  38163 33960
 $ Iceland       : int  2766 3244
 $ Ireland       : int  24129 21591
 $ Italy         : int  246764 228094
 $ Japan         : int  865523 869825
 $ Korea         : int  401444 361374
>
> library(RColorBrewer)                    #①그래프1
> pal <- brewer.pal(9, "Set1")
>
> barplot(as.matrix(data[1:2,2:16]), main=paste("Total R&D personnel (FTE) from
UIS Database"), ylab="Value", beside=T, xlim=c(0,45),
ylim=c(0,1000000),col=pal)
> legend(20,20,2013,cex=0.8,fill=pal,bg="white")
>
> install.packages("googleVis")
Installing package into '/Users/Christine/Library/R/3.3/library'
(as 'lib' is unspecified)
--- Please select a CRAN mirror for use in this session ---
Error in if (res > nrow(m)) { : argument is of length zero
> library(googleVis)
Creating a generic function for 'toJSON' from package 'jsonlite' in package
'googleVis'

Welcome to googleVis version 0.6.1

Please read the Google API Terms of Use
before you start using the package:
https://developers.google.com/terms/

Note, the plot method of googleVis will by default use
```

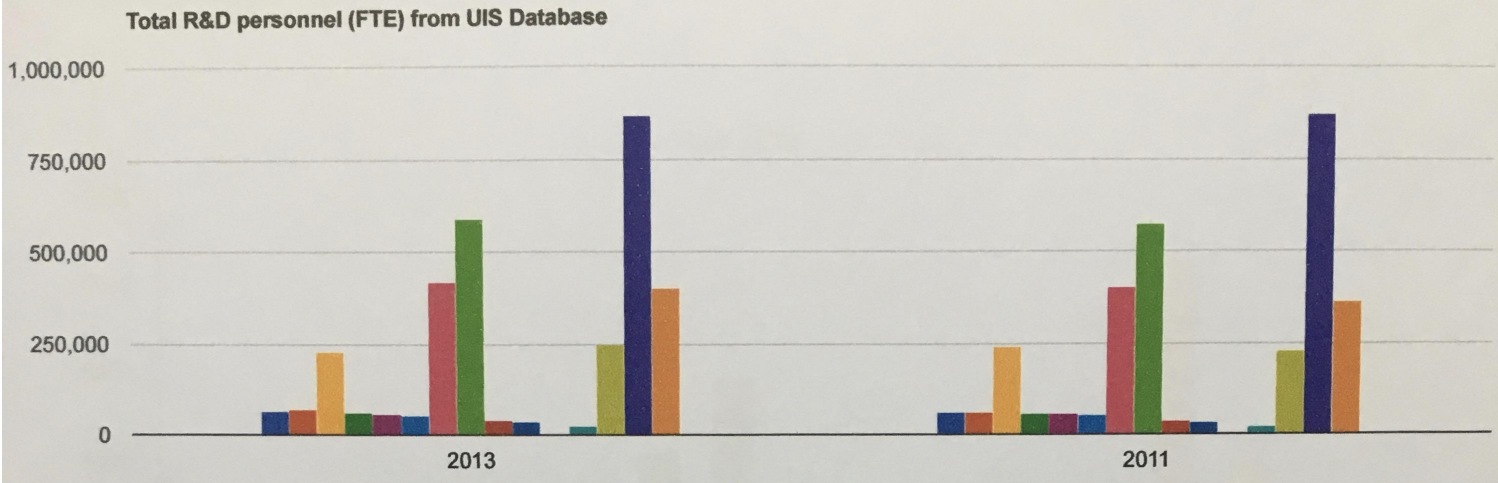
```
the standard browser to display its output.

See the googleVis package vignettes for more details,
or visit http://github.com/mages/googleVis.

To suppress this message use:
suppressPackageStartupMessages(library(googleVis))

> google <- gvisColumnChart(data,options=list(title="Total R&D personnel (FTE)
from UIS Database",height=400,weight=500))     #② 구글Vis를 활용한 그래프 2
> plot(google)
starting httpd help server ... done
>
> data1<-read.csv("/Users/Christine/Downloads/rnd.csv", header=T)
>
> v1<-data[1,2:16]*0.01
> v2<-data[2,2:16]*0.01
> name <- data1$Country
> time <- data1$TIME
> year_2013 <- barplot(as.matrix(v1),main="Year 2013", beside=T, axes=T,
cex.names=0.85, las=2, col=pal, ylim=c(0,10000))
> year_2011 <- barplot(as.matrix(v2),main="Year 2011", beside=T, axes=T,
cex.names=0.85, las=2, col=pal, ylim=c(0,10000))
>                                          #③ 2011, 2013년별로 시각화한 그래프 3
>
```

#①



플다임으로 R&D에 종사하는 사람들을 표시한 표인데, xlim의 한계로 나라이름들이 다 보이지는 않으나
가장 Personnel이 많은 나라는 일본임을 알수있다. 그 다음은 독일, 프랑스, 한국으로 분석된다.
과제수행자의 실력이 아직 미흡하여 y축 눈금을 어떻게 다뤄야할지 알지 못하였고, 숫자가 너무 크고 차이가 많아 (ex. 아이슬란드와 일본)
abline을 차마 그리지 못했다는 데에 한계점이 있는 그래프이다.

googleVis 그래프 2



미찬가지로 R&D에 종사하는 Personnel (FTE)을 표시한 표이다.

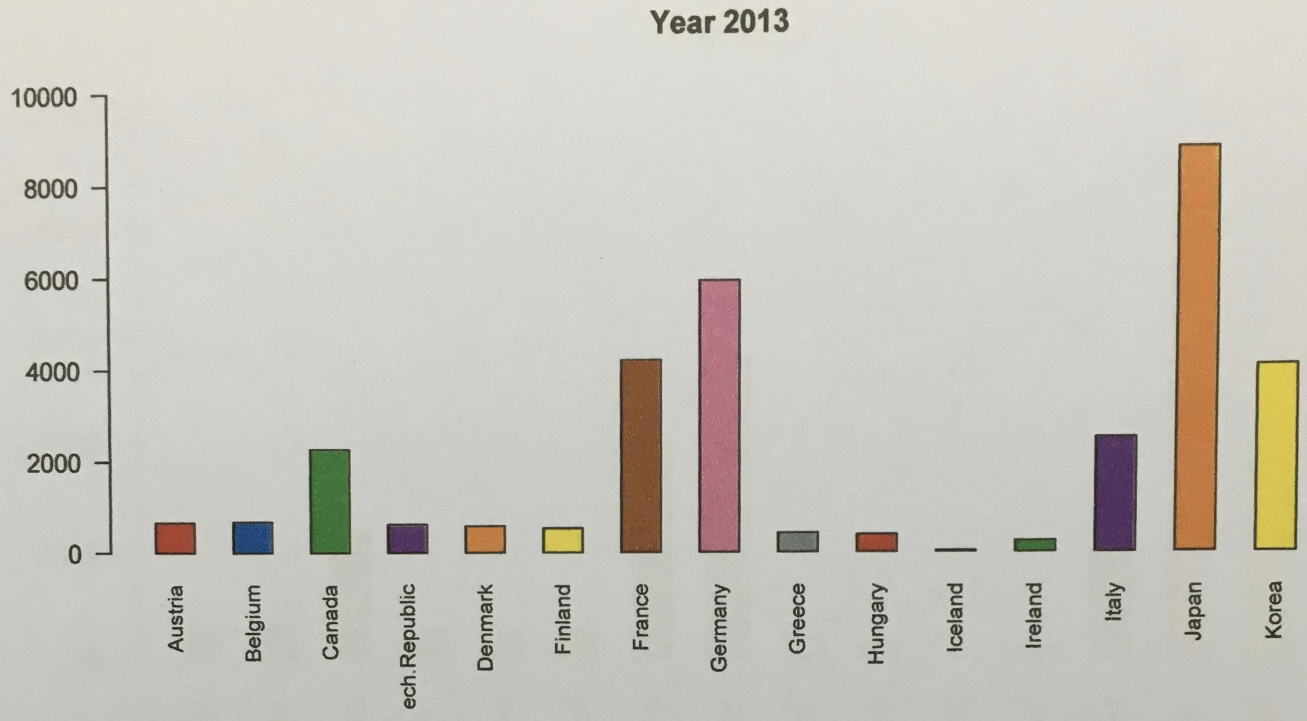
옆에 색깔(국가)을 나타내는 색깔들에 대한 설명이 나오지 않았지만 2013, 2011 연도 두가지를 한번에 보기좋게 제시함으로써

앞서 그렸던 #① 그래프보다 훨씬 깔끔한 시각화를 경험할 수 있다.

다만 이 그래프에서 세로축은 250,000명 이하의 나라들을 보기에는 힘들게 설정되어 있어 좀더 확실한 분석과 시각화를 위해서는

데이터를 Personnel이 많은 나라/적은나라로 다시 정제하여 따로따로 그려봐야 할 것으로 보인다.

#③ 그래프 3-1 (2013년)



마찬가지로 FTE R&D Personnel Total을 표시한 표이다.
실제 Personnel 값에 0.1을 곱한 수치라 y축이 보기좋게 표시되었는데, 0.1을 곱하지 않고도
보기좋게 시각화를 할 수 있는 방법이나, 앞서 그린 그래프들에서 지적했던 한계점들을 여전히 지적해보아야할 필요성이 느껴진다.

#③ 그래프 3-2 (2011년)