

--- Data Visualization with R

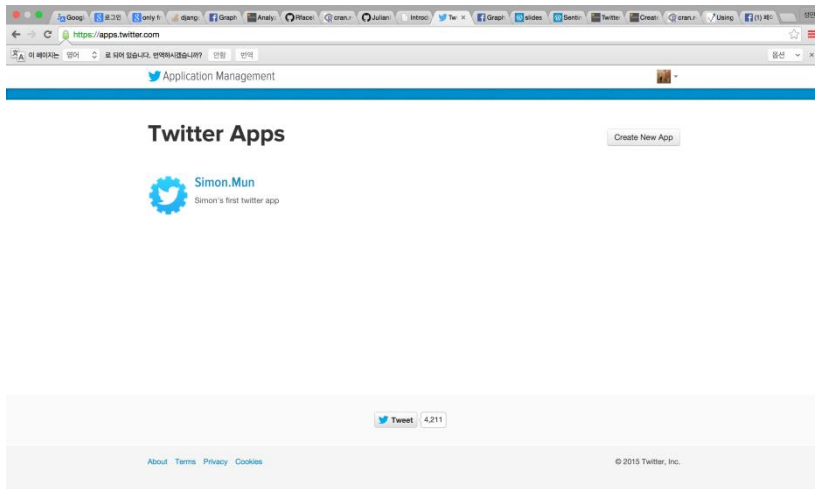
Text Analysis using twitter

Text Analysis using twitter

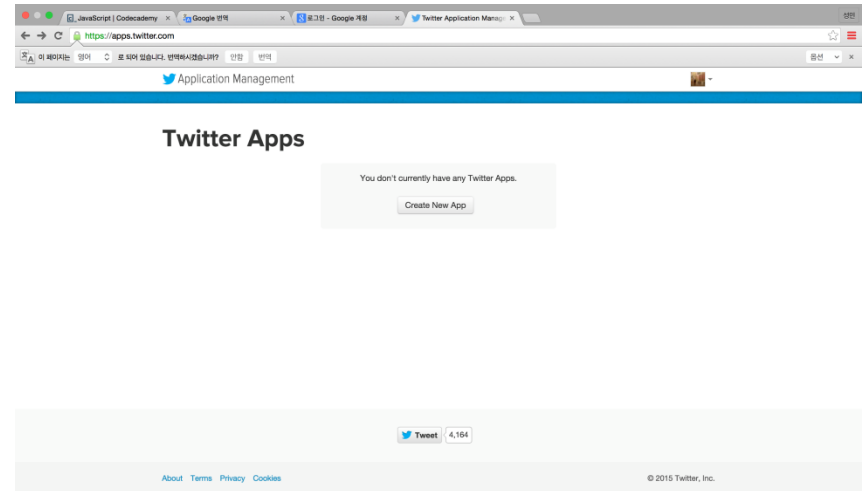
● 트위터 계정 생성

- <https://apps.twitter.com>

1) Process



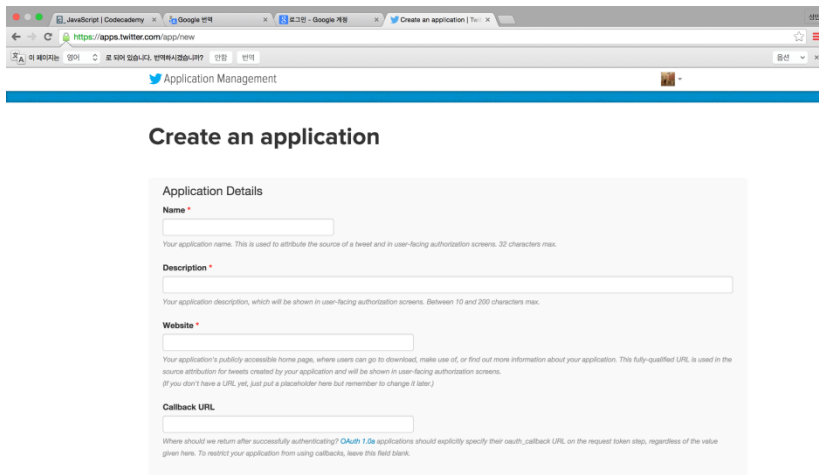
2) Process



Text Analysis using twitter

- 트위터 계정 생성
 - <https://apps.twitter.com>

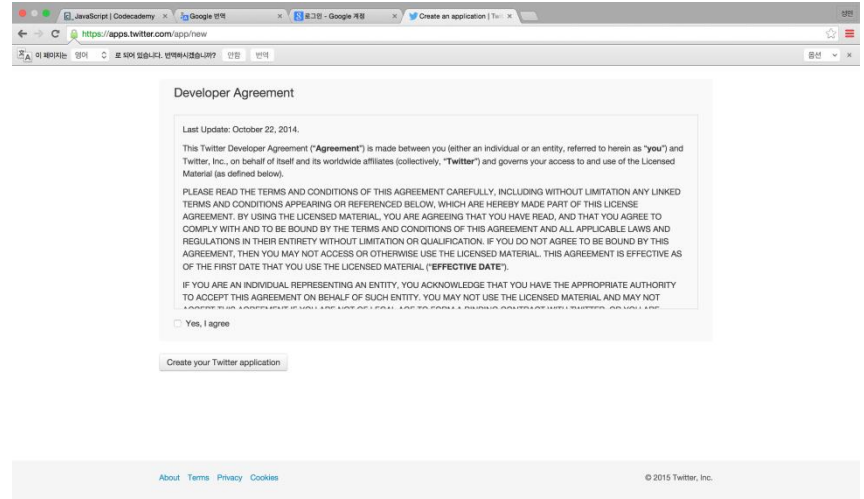
3) Process



The screenshot shows the 'Create an application' page on the Twitter Developer portal. The page has a blue header with the Twitter logo and 'Application Management'. Below the header, the main heading is 'Create an application'. The form is titled 'Application Details' and contains the following fields:

- Name ***: A text input field for the application name. Below it, a note states: 'Your application name. This is used to attribute the source of a tweet and in user-facing authorization screens. 32 characters max.'
- Description ***: A text input field for the application description. Below it, a note states: 'Your application description, which will be shown in user-facing authorization screens. Between 10 and 200 characters max.'
- Website ***: A text input field for the application's website. Below it, a note states: 'Your application's publicly accessible home page, where users can go to download, make use of, or find out more information about your application. This fully-qualified URL is used in the source attribution for tweets created by your application and will be shown in user-facing authorization screens. (If you don't have a URL, yet, just put a placeholder here but remember to change it later.)'
- Callback URL**: A text input field for the callback URL. Below it, a note states: 'Where should we return after successfully authenticating? OAuth 1.0a applications should explicitly specify their oauth_callback URL on the request token step, regardless of the value given here. To restrict your application from using callbacks, leave this field blank.'

4) Process



The screenshot shows the 'Developer Agreement' page on the Twitter Developer portal. The page has a blue header with the Twitter logo and 'Create an application'. Below the header, the main heading is 'Developer Agreement'. The page contains the following text:

Last Update: October 22, 2014.

This Twitter Developer Agreement ("Agreement") is made between you (either an individual or an entity, referred to herein as "you") and Twitter, Inc., on behalf of itself and its worldwide affiliates (collectively, "Twitter") and governs your access to and use of the Licensed Material (as defined below).

PLEASE READ THE TERMS AND CONDITIONS OF THIS AGREEMENT CAREFULLY, INCLUDING WITHOUT LIMITATION ANY LINKED TERMS AND CONDITIONS APPEARING OR REFERENCED BELOW, WHICH ARE HEREBY MADE PART OF THIS LICENSE AGREEMENT. BY USING THE LICENSED MATERIAL, YOU ARE AGREEING THAT YOU HAVE READ, AND THAT YOU AGREE TO COMPLY WITH AND TO BE BOUND BY THE TERMS AND CONDITIONS OF THIS AGREEMENT AND ALL APPLICABLE LAWS AND REGULATIONS IN THEIR ENTIRETY WITHOUT LIMITATION OR QUALIFICATION. IF YOU DO NOT AGREE TO BE BOUND BY THIS AGREEMENT, THEN YOU MAY NOT ACCESS OR OTHERWISE USE THE LICENSED MATERIAL. THIS AGREEMENT IS EFFECTIVE AS OF THE FIRST DATE THAT YOU USE THE LICENSED MATERIAL ("EFFECTIVE DATE").

IF YOU ARE AN INDIVIDUAL REPRESENTING AN ENTITY, YOU ACKNOWLEDGE THAT YOU HAVE THE APPROPRIATE AUTHORITY TO ACCEPT THIS AGREEMENT ON BEHALF OF SUCH ENTITY. YOU MAY NOT USE THE LICENSED MATERIAL AND MAY NOT ACCESS OR OTHERWISE USE THE LICENSED MATERIAL.

☐ Yes, I agree

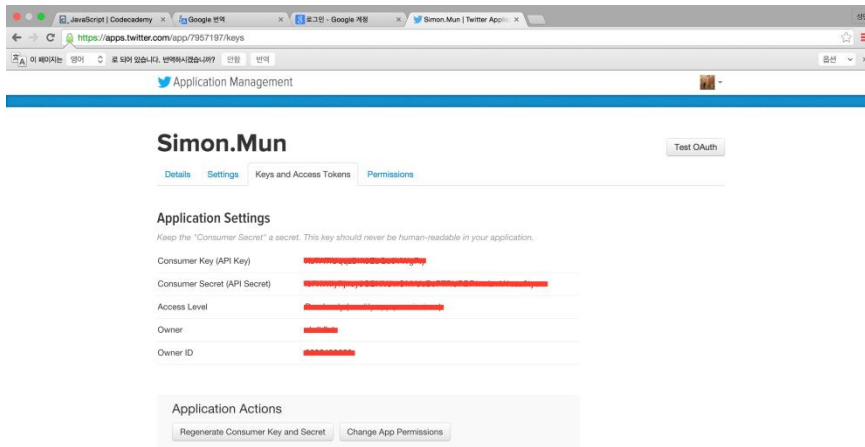
[Create your Twitter application](#)

At the bottom of the page, there are links for 'About', 'Terms', 'Privacy', and 'Cookies', and a copyright notice: '© 2015 Twitter, Inc.'

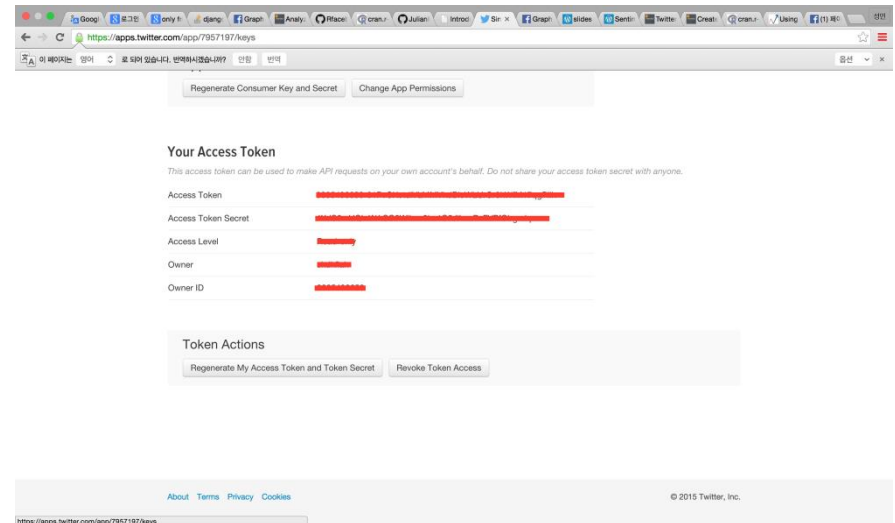
Text Analysis using twitter

- 트위터 계정 생성
 - <https://apps.twitter.com>

5) Process



6) Process



Text Analysis using twitter

● 라이브러리 설치

```
> library(bitops)
> library(RCurl)
> library(RJSONIO)
> library(twitter)
> library(ROAuth)
> library(RColorBrewer)
> library(devtools)
> install_github("twitter", username="geoffjentry")
Downloading github repo geoffjentry/twitteR@master
Installing twitteR
'/Library/Frameworks/R.framework/Resources/bin/R' --vanilla CMD INSTALL \

'/private/var/folders/28/g8cf_pvx46sSphqwr6qq7jw0000gn/T/Rtmp8qGMiY/devtoolscb924cc3a7ae/geoffj
entry-twitteR-563a23c' \
  --library='/Library/Frameworks/R.framework/Versions/3.1/Resources/library' \
  --install-tests

* installing *source* package 'twitteR' ...
** R
** inst
** preparing package for lazy loading
Creating a generic function for 'as.data.frame' from package 'base' in package 'twitteR'
** help
*** installing help indices
** building package indices
** testing if installed package can be loaded
* DONE (twitteR)
Reloading installed twitteR

Attaching package: 'twitteR'

The following object is masked from 'package:plyr':

    id

The following objects are masked from 'package:dplyr':

    id, location

경고메시지:
Username parameter is deprecated. Please use geoffjentry/twitteR
```

GitHub에서 twitteR패키지의 최신버전을 다운로드한다.

Text Analysis using twitter

- 유저 정보 입력

```
> api_key <- "[REDACTED]"
>
> api_secret <- "[REDACTED]"
>
> access_token <- "[REDACTED]"
>
> access_token_secret <- "[REDACTED]"
>
> setup_twitter_oauth(api_key,api_secret,access_token,access_token_secret)
[1] "Using direct authentication"
```

- <https://apps.twitter.com>에서 로그인 후 제공받은 api_key, api_secret, access_token, access_token_secret을 입력한다.

Text Analysis using twitter

● 긍부정 분류함수

```
> score.sentiment = function(sentences, pos.words, neg.words, .progress='none')
+ {
+   require(plyr)
+   require(stringr)
+
+   # we got a vector of sentences. plyr will handle a list or a vector as an "l" for us
+   # we want a simple array of scores back, so we use "l" + "a" + "ply" = laply:
+   scores = laply(sentences, function(sentence, pos.words, neg.words) {
+
+     # clean up sentences with R's regex-driven global substitute, gsub():
+     sentence = gsub('[[:punct:]]', '', sentence)
+     sentence = gsub('[[:cntrl:]]', '', sentence)
+     sentence = gsub('\\d+', '', sentence)
+     # and convert to lower case:
+     sentence = tolower(sentence)
+
+     # split into words. str_split is in the stringr package
+     word.list = str_split(sentence, '\\s+')
+     # sometimes a list() is one level of hierarchy too much
+     words = unlist(word.list)
+
+     # compare our words to the dictionaries of positive & negative terms
+     pos.matches = match(words, pos.words)
+     neg.matches = match(words, neg.words)
+
+     # match() returns the position of the matched term or NA
+     # we just want a TRUE/FALSE:
+     pos.matches = !is.na(pos.matches)
+     neg.matches = !is.na(neg.matches)
+
+     # and conveniently enough, TRUE/FALSE will be treated as 1/0 by sum():
+     score = sum(pos.matches) - sum(neg.matches)
+
+     return(score)
+   }, pos.words, neg.words, .progress=.progress )
+
+   scores.df = data.frame(score=scores, text=sentences)
+   return(scores.df)
+ }
```

Score.sentiment함수를 입력하여 준다.

Text Analysis using twitter

- Greece에 관련된 텍스트 1000개 크롤링

```
> Greece.tweets = searchTwitter("Greece" , n = 1000)
```

- Greece에 관련된 텍스트만 추출

```
> library(plyr)
>
> Greece.text = laply(Greece.tweets,function(t)t$getText())
```

- 긍부정 단어가 들어있는 사전 불러오기

```
> getwd()
[1] "/Users/Seongmin_M/Downloads"
>
> setwd("/Users/Seongmin_M/Downloads")
>
> pos.words= scan("positive-words.txt",what="character",comment.char=";")
Read 2006 items
>
> neg.words = scan("negative-words.txt",what="character",comment.char=";")
Read 4783 items
```

Text Analysis using twitter

- 텍스트가 깨지지 않게 문자 인코딩 방식을 UTF-8로 변환

```
> Greece.text = Greece.text[!Encoding(Greece.text)=="UTF-8"]
```

- Greece에 관련 텍스트를 긍부정 단어 사전을 사용하여 분류하기

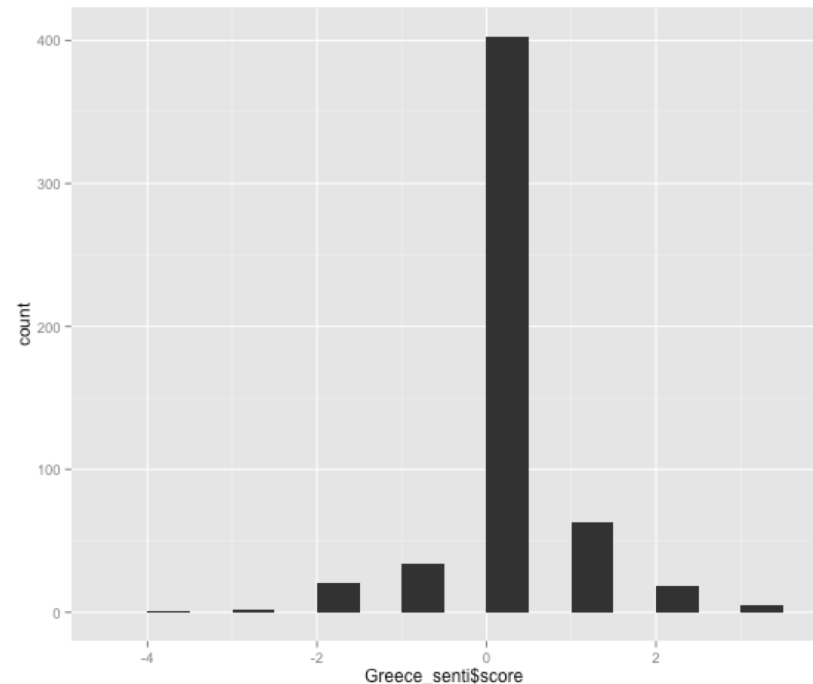
```
> Greece_senti =score.sentiment(Greece.text,pos.words,neg.words,.progress='text')
```

```
|=====| 100%
```

- 히스토그램 생성

```
> library(ggplot2)
```

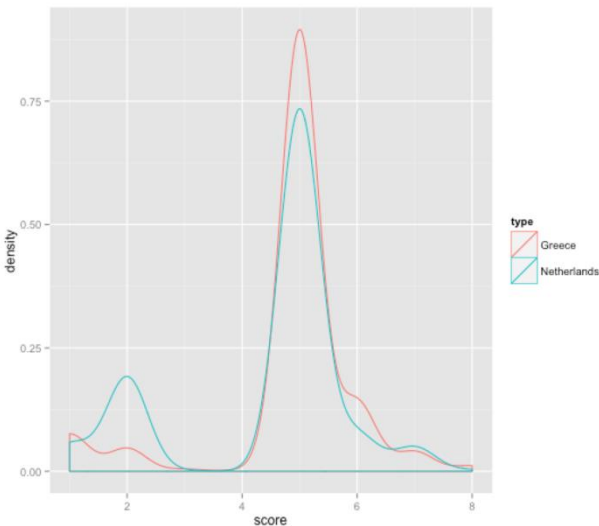
```
> qplot(Greece_senti$score,binwidth=0.5)
```



Text Analysis using twitter

● Greece와 Netherlands간의 긍부정 비교

```
> a<-dim(Greece_senti)[1]
>
> b<-dim(Netherlands_senti)[1]
>
> country<-rbind(as.data.frame(cbind(type=rep("Greece",a),score=Greece_senti[,1])),as.data.frame(
cbind(type=rep("Netherlands",b),score=Netherlands_senti[,1])))
>
> country$type<-factor(country$type)
>
> country$score<-as.integer(country$score)
>
> ggplot(country,aes(x=score,colour=type))+geom_density()
```



트위터 텍스트를 활용하여 두 나라간 긍부정 반응을 비교한 결과 그리스에 비해 네덜란드에 대해 더 긍정적 반응을 보이는 것을 확인 하였다.

Text Analysis using twitter

- 워드 클라우드 생성

- 모든 문자 소문자로 변환

```
> Greece.text <- tolower(Greece.text)
>
```

- Rt를 빈공간으로 바꾸기(삭제)

```
> Greece.text <- gsub("rt", "", Greece.text)
>
```

- 유저이름 삭제(@||w+)

```
> Greece.text <- gsub("@\\w+", "", Greece.text)
>
```

- 문장 부호 제거

```
> Greece.text <- gsub("[[:punct:]]", "", Greece.text)
>
```

- 링크 제거

```
> Greece.text <- gsub("http\\w+", "", Greece.text)
>
```

Text Analysis using twitter

- 워드 클라우드 생성

- 탭 제거

```
> Greece.text <- gsub("[ \\t]{2,}", "", Greece.text)
>
```

- 시작 부분의 문자 제거

```
> Greece.text <- gsub("^ ", "", Greece.text)
>
```

- 끝 부분의 문자 제거

```
> Greece.text <- gsub(" $", "", Greece.text)
```

- TM라이브러리 설치

```
> install.packages("tm")
```

URL 'http://cran.rstudio.com/bin/macosx/contrib/3.1/tm_0.6.tgz'을 시도합니다

Content type 'application/x-gzip' length 647048 bytes (631 Kb)

URL을 열었습니다

```
=====
downloaded 631 Kb
```

The downloaded binary packages are in

/var/folders/28/g8cf_pvx46s5phqgwr6qq7jw0000gn/T//RtmpTkkSha/downloaded_packages

```
>
```

```
> library("tm")
```

필요한 패키지를 로딩중입니다: NLP

```
> wordcloud(dalta.text.corpus,min.freq = 2, scale=c(7,0.5),colors=brewer.pal(8, "Dark2"), random.color= TRUE, random.order = FALSE, max.words = 150)
```