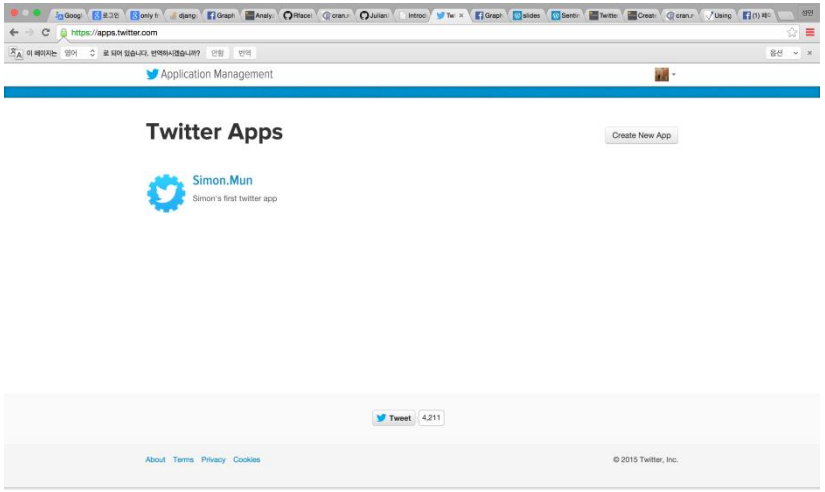
Data Visualization with R

Text Analysis using twitter

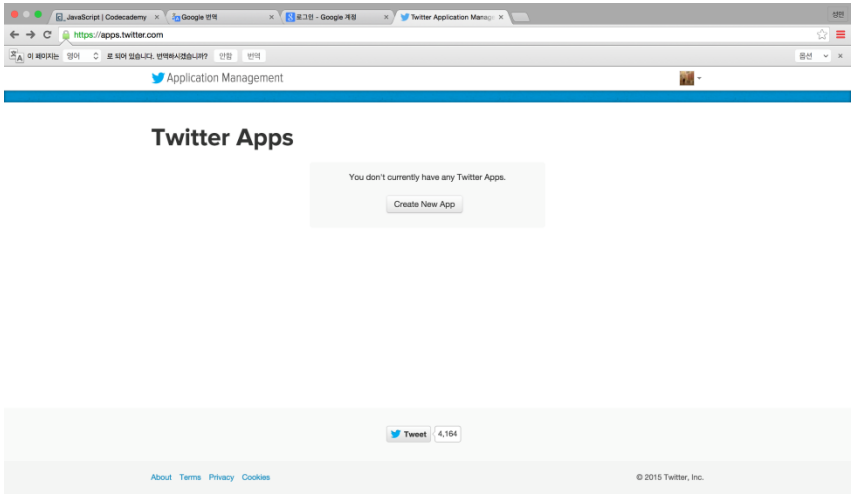
Text Analysis using twitter

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

1) Process



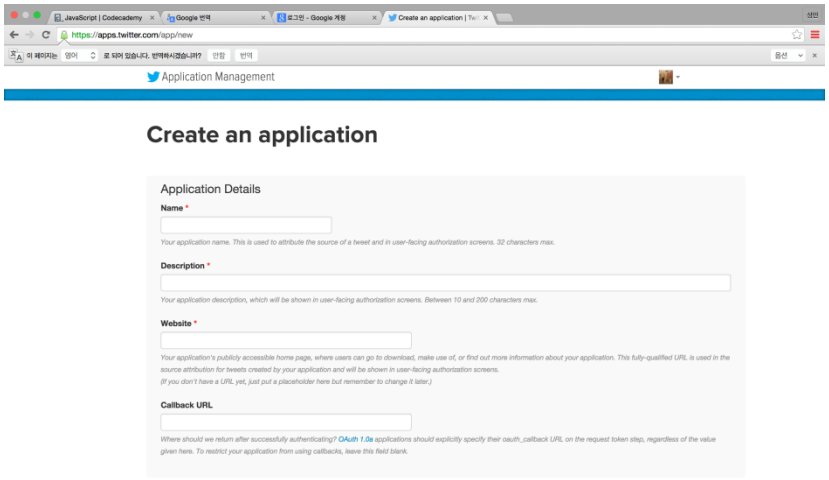
2) Process



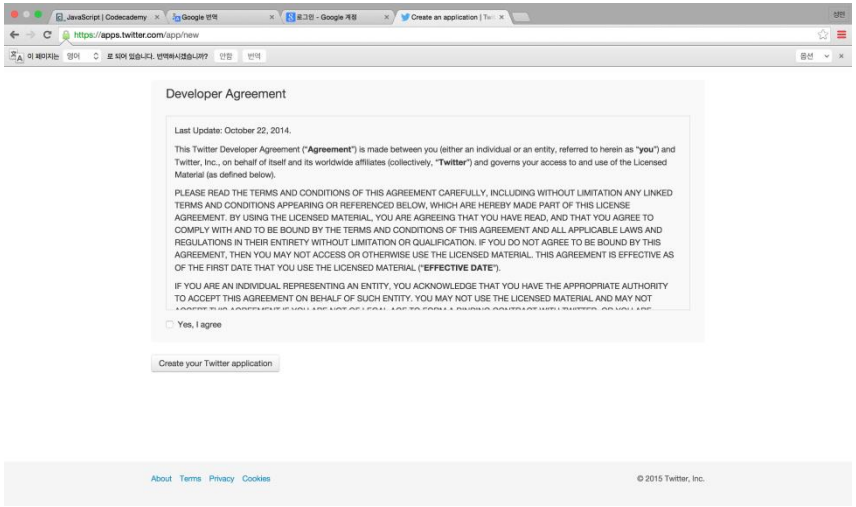
Text Analysis using twitter

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

3) Process



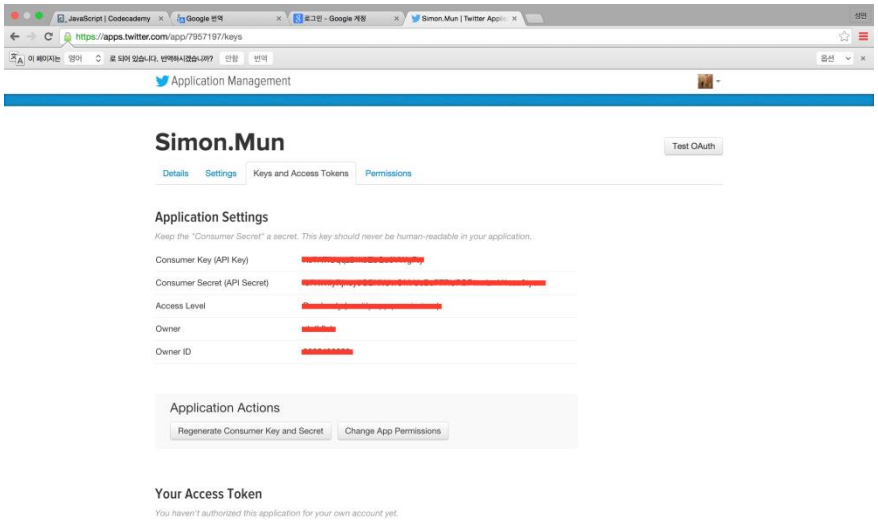
4) Process



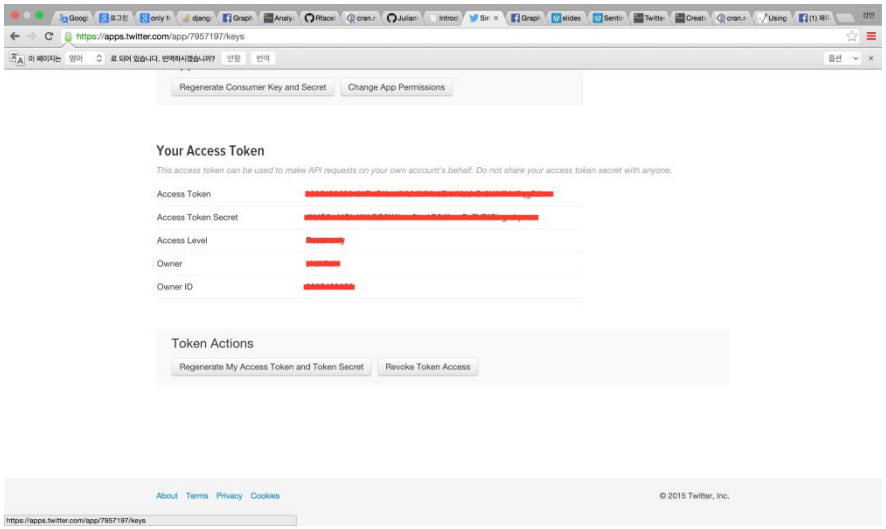
Text Analysis using twitter

- 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_pvx46s5phqgwr6qq7jw0000gn/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 최신버젼을 다운로드한다.

- 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.com에서 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

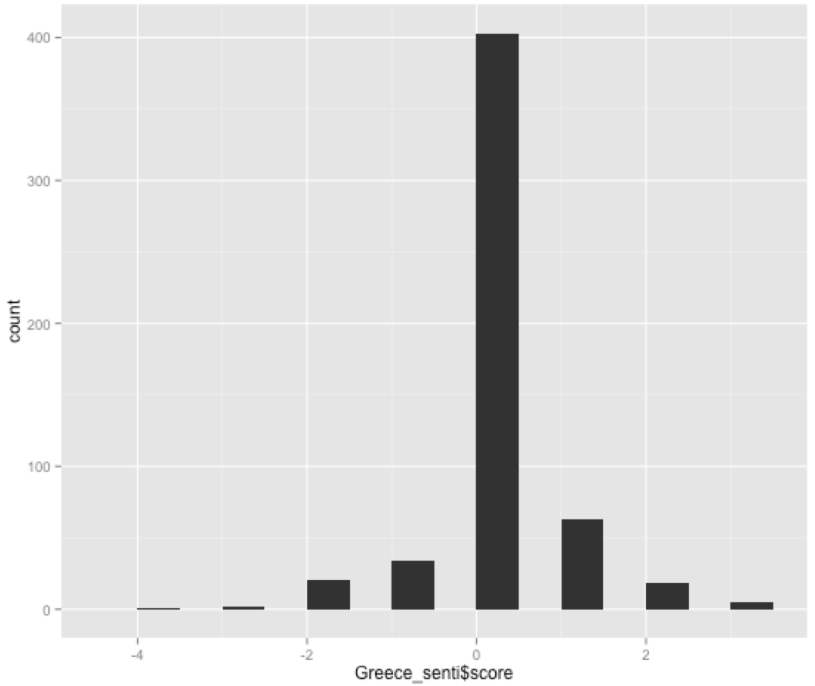
```
> Greece.text = Greece.text[!Encoding(Greece.text)=="UTF-8"]
```

- Greece에 관련 텍스트를 긍부정 단어 사전을 사용하여 분류하기

```
> Greece_senti =score.sentiment(Greece.text,pos.words,neg.words,.progress='text')
  |=================================================================================================| 100%
```

- 히스토그램 생성

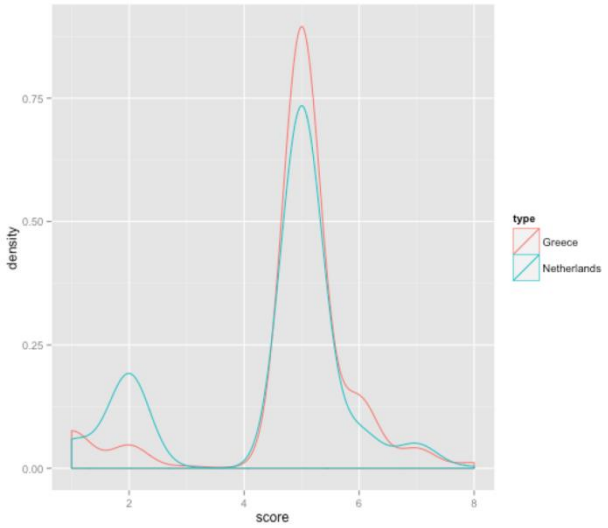
```
> library(ggplot2)
> qplot(Greece_senti$score,binwidth=0.5)
```



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
```

Text Analysis using twitter

- 워드 클라우드 생성

  - Corpus생성

```
>
> Greece.text.corpus <- Corpus(VectorSource(Greece.text))
```

  - Tm_map을 활용하여 Stop words 삭제

```
> Greece.text.corpus <- tm_map(Greece.text.corpus, function(x)removeWords(x,stopwords()))
```

  - 워드클라우드 생성

```
> install.packages("wordcloud")
URL 'http://cran.rstudio.com/bin/macosx/contrib/3.1/wordcloud_2.5.tgz'을 시도합니다
Content type 'application/x-gzip' length 117239 bytes (114 Kb)
URL을 열었습니다
==================================================
downloaded 114 Kb


The downloaded binary packages are in
        /var/folders/28/g8cf_pvx46s5phqgwr6qq7jw0000gn/T//RtmpTkkSha/downloaded_packages
>
> library(wordcloud)
필요한 패키지를 로딩중입니다: RColorBrewer
>
> wordcloud(dalta.text.corpus,min.freq = 2, scale=c(7,0.5),colors=brewer.pal(8, "Dark2"),  rando
m.color= TRUE, random.order = FALSE, max.words = 150)
```