트위터자료 추출 및 분석

Jinseog Kim

2018-02-27

접근승인을 위한 키발급

1. https://apps.twitter.com/app/new

접근승인을 위한 키발급 절차-1

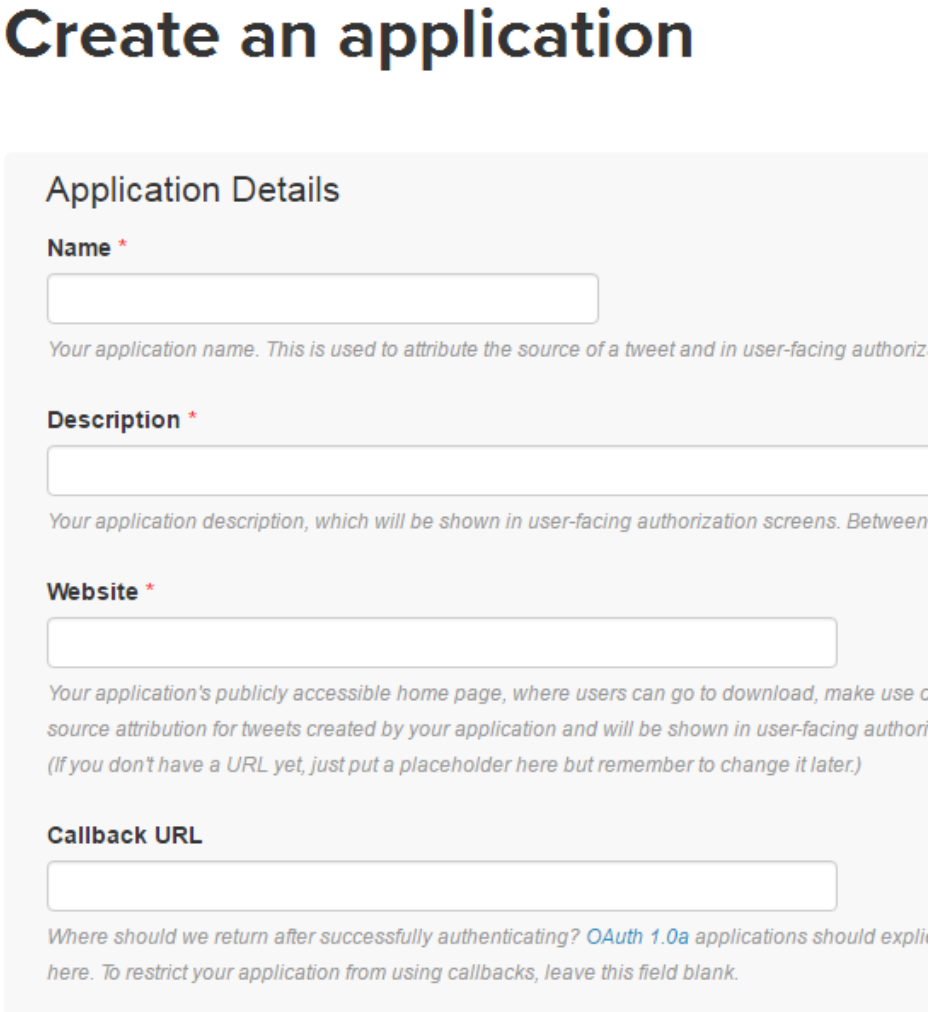

Figure 1: 키발급 절차-1

접근승인을 위한 키발급 절차-2

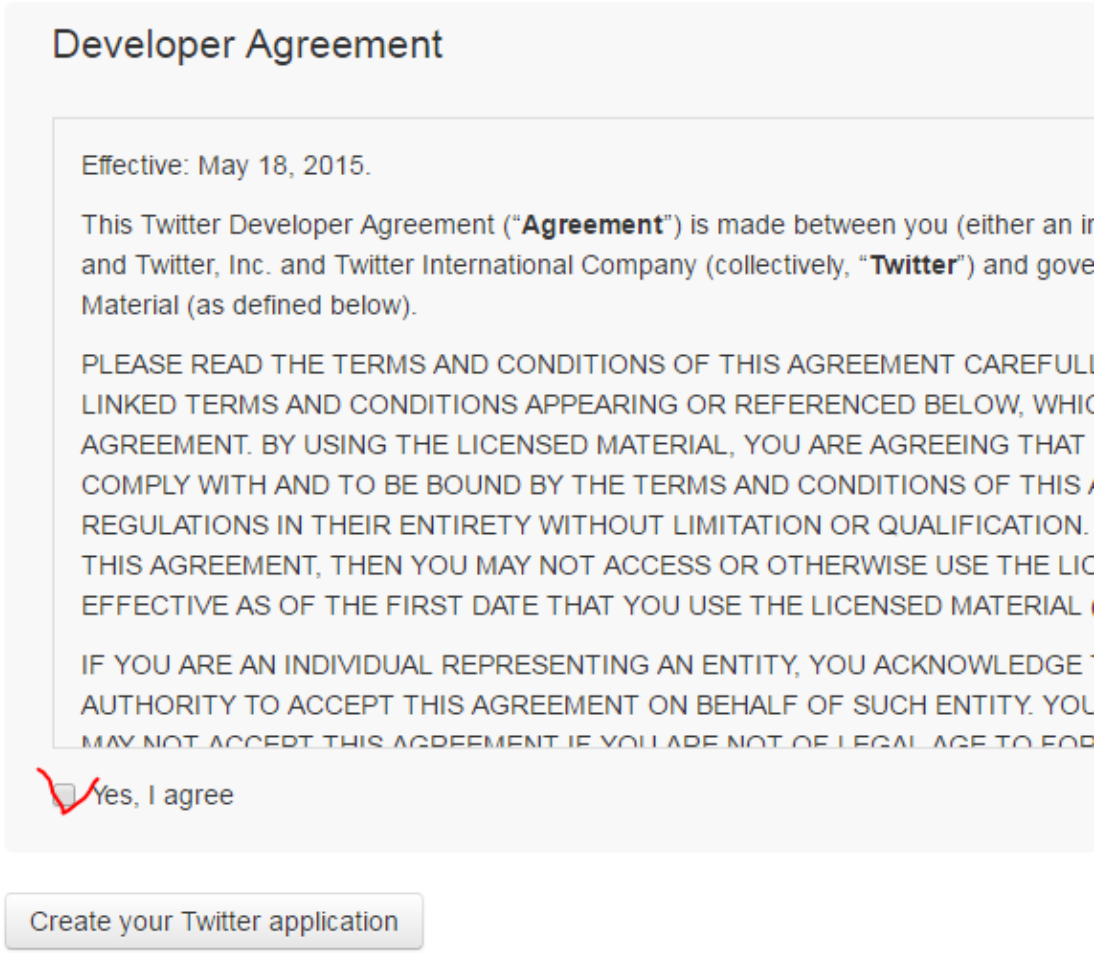

Figure 2: 키발급 절차-2

접근승인을 위한 키발급 절차 - Consumer Key(API Key) & Secret

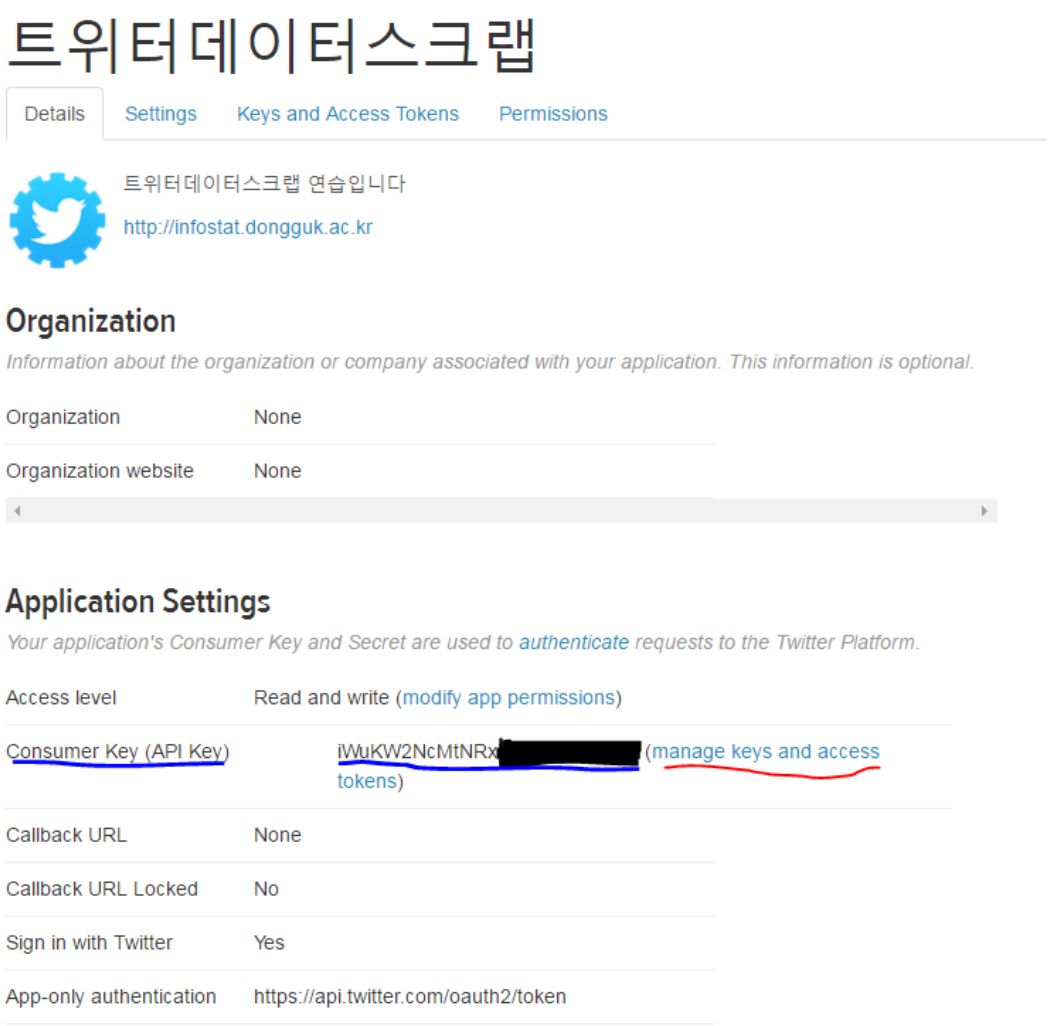

Figure 3: 키발급 절차-3

접근승인을 위한 키발급 절차 - Consumer Key & Access token

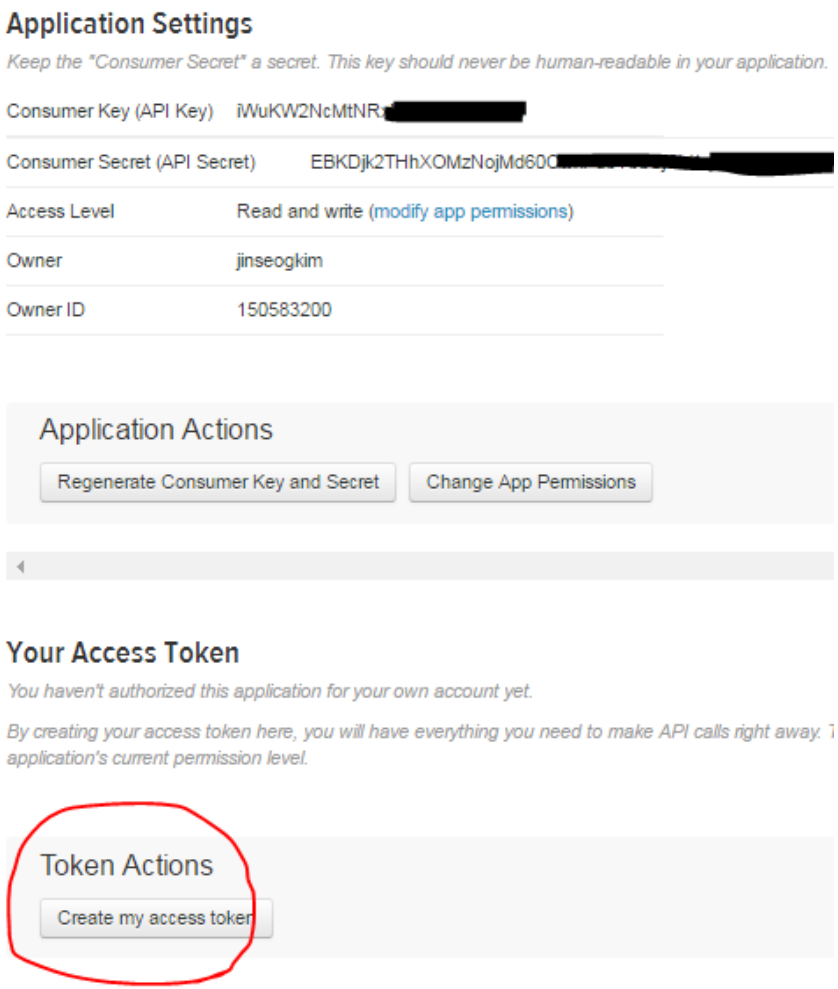

Figure 4: 키발급 절차-4

접근승인을 위한 키발급 절차 - Consumer Key & Access token

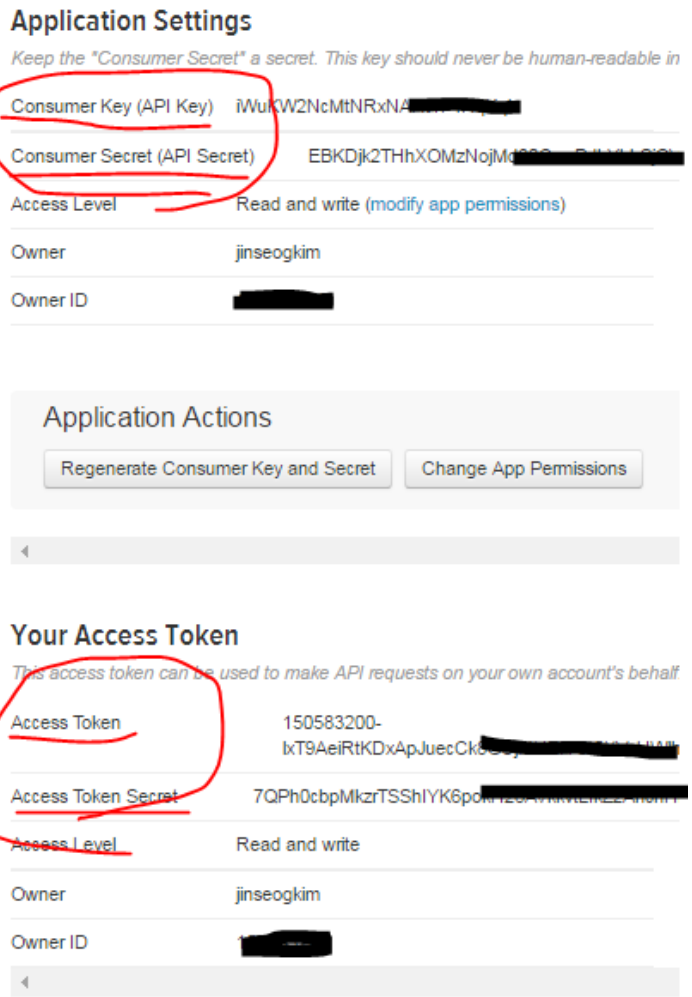

Figure 5: 키발급 절차-5

R 에서 트위터데이터의 수집

- **패키지 설치 및 로드**

```
#install.packages("twitteR")
#install.packages("ROAuth")
library(twitteR) # twitter R client
library(ROAuth) # for authentication
```

twitteR을 이용한 트위터데이터의 수집

- twitteR 사용 인증(authentication)

```
consumer_key <- "fZuqjfR*****";
consumer_secret <- "YS2NTH*****";
access_token <- "150583200*****";
access_secret <- "ekQDwvCX*****";
```

```
setup_twitter_oauth(consumer_key,consumer_secret,
                    access_token,access_secret)
```

twitteR을 이용한 트위터데이터의 수집

- R 에서 트위터데이터의 수집

```
keyword <- "big data"
tout <- searchTwitter(keyword, n=100) # list
toutDF <- twListToDF(tout)
```

twitteR을 이용한 트위터데이터의 수집

- R 에서 트위터데이터의 수집

```
  ### list with following components ####
  # text: The text of the status
  # screenName:  Screen name of the user who posted this status
  # id:  ID of this status
  # replyToSN:  Screen name of the user this is in reply to
  # replyToUID:  ID of the user this was in reply to
  # statusSource:  Source user agent for this tweet
  # created:  When this status was created
  # truncated:  Whether this status was truncated
  # favorited:  Whether this status has been favorited
  # retweeted:  TRUE if this status has been retweeted
  # retweetCount:  The number of times this status has been retweeted
  ##
#  tout[[1]]$getCreated()
#  tout[[1]]$getFavoriteCount()
#  tout[[1]]$getFavorited()
#  tout[[1]]$getId()
```

twitteR을 이용한 트위터데이터의 수집

- **R 에서 트위터데이터의 수집**

```
> names(toutDF)
 [1] "text"          "favorited"     "favoriteCount" "replyToSN"     "created"
 [6] "truncated"     "replyToSID"    "id"            "replyToUID"    "statusSource
[11] "screenName"    "retweetCount"  "isRetweet"     "retweeted"     "longitude"
[16] "latitude"
```

예제 R코드

```
#install.packages("twitteR");install.packages("ROAuth")
library(twitteR);library(ROAuth)
```

```
setup_twitter_oauth(consumer_key="**********",
consumer_secret="**********T4VWqGcXuH",
access_token="150583200-P**********L",
access_secret="ekQD**********v9woUa3")
  # OAuth authentication for a twitteR session
  # - consumer_key: The consumer key supplied by Twitter
  # - consumer_secret  : The consumer secret supplied by Twitter
  # - access_token  : The access token supplied by Twitter
  # - access_secret : The access secret supplied by Twitter
```

예제 R코드

- keyword: “빅데이터”
- 수집 자료: 1000

```
keyword <- "빅데이터"
tout <- searchTwitter(keyword, n=1000)
toutDF <- twListToDF(tout)
head(toutDF$text)
```

텍스트 기초 처리

```
load(file="twit_bigdata.RData")
require(stringr)
l <- grepl("가수|브랜드|평판", toutDF$text)
text <- toutDF$text[!l]
x <- gsub("[^A-Za-z가-힣[:space:][:digit:][:punct:]]", "", text)
x <- gsub("@|\n|RT", "", x)
x <- gsub("", " ", x)
x <- gsub("[[:punct:]]", " ", x)
x <- gsub("[[:digit:]]", "", x)
x <- tolower(x)
x <- gsub("[a-z]", "", x)
x <- gsub("댓글|https|co|녀모두|11위", "", x)
x <- gsub("출처|하였습니다|지난달|위에서|기사", "", x)

x <- str_trim(x)
#x
```

텍스트 기초 처리

- 명사 추출, 워드 카운팅

```
library(KoNLP)
x1 <- lapply(x, extractNoun)
x2 <- lapply(x1, function(x) x[nchar(x)>1])

x3 <- do.call(c, x2)
o <- table(x3)
```

워드 클라우드

```
library(wordcloud)
pal <- brewer.pal(8,"Dark2")
wordcloud(names(o), o, min.freq=3, random.order=F,random.color=T,colors=pal)
```