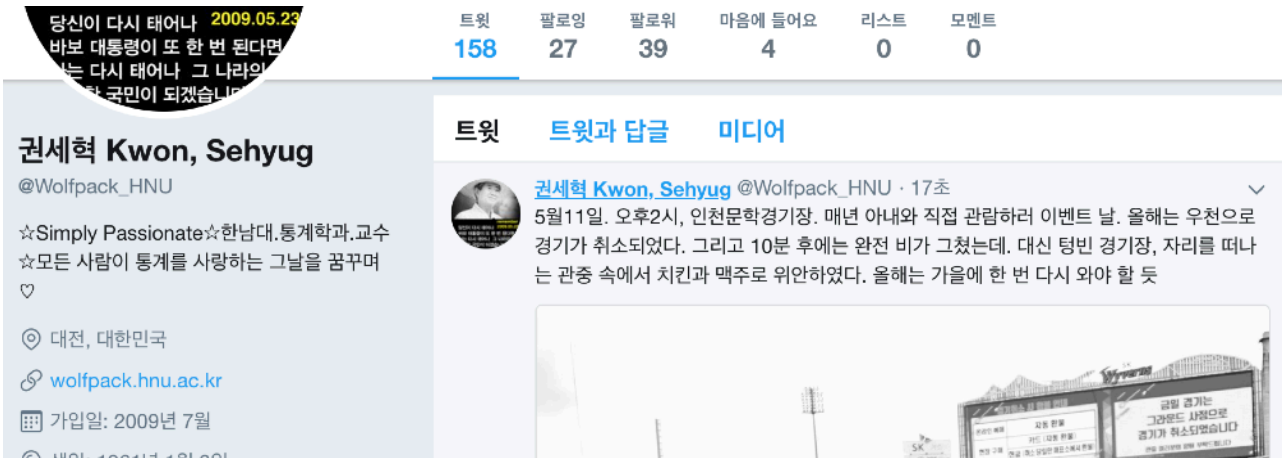
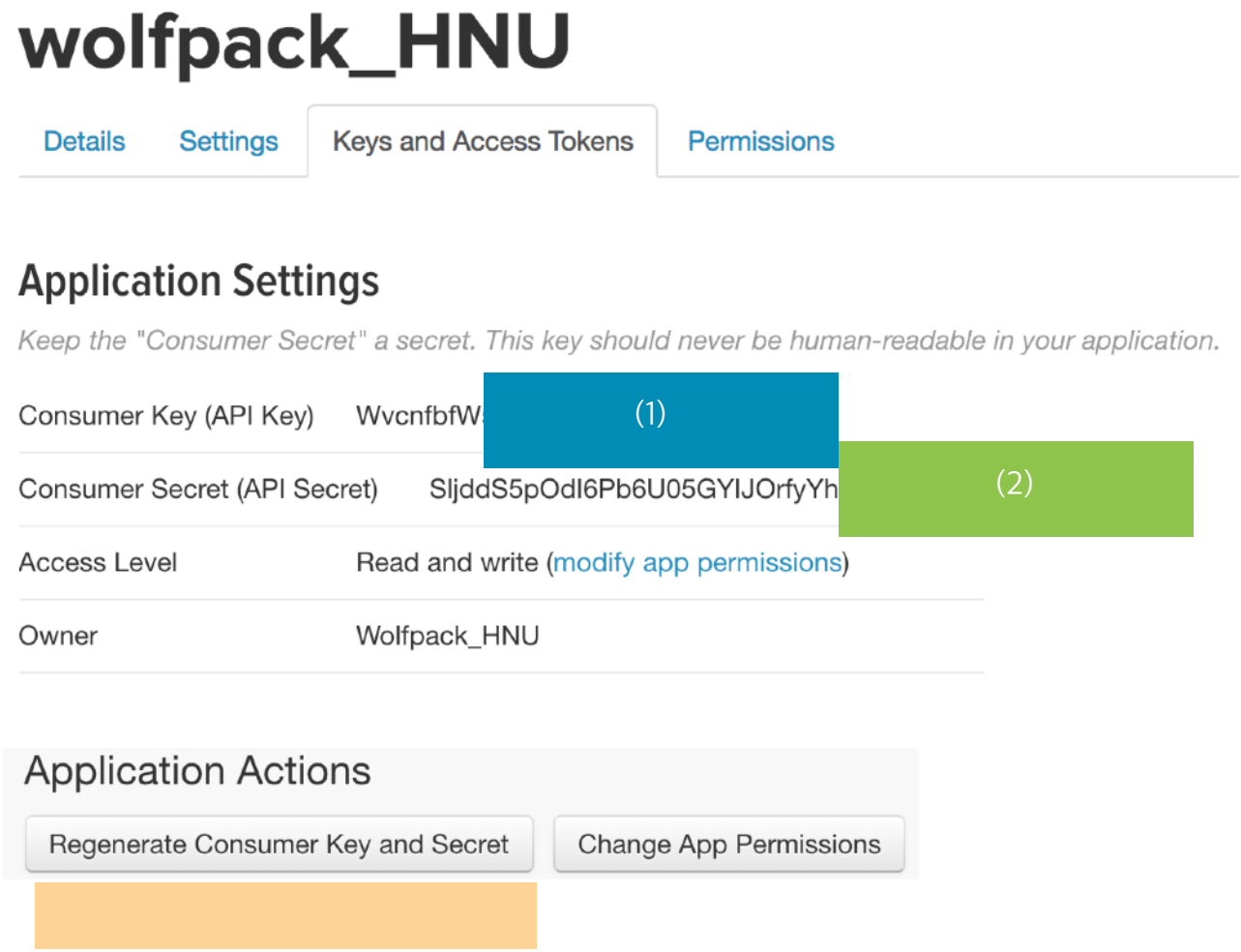
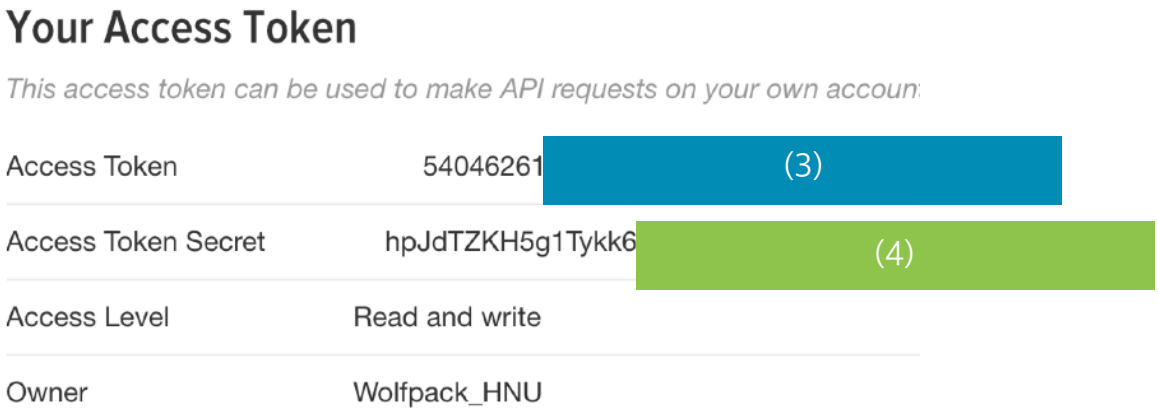
사전 준비

1) Twitter 계정 만들기 : https://twitter.com



2) twitter Apps Key 만들기 : https://apps.twitter.com





트윗 가져오기

- lang= : 언어 지정, 디폴트는 영어임
- since=, until= 날짜 지정
- geocode='위도, 경도, 반경' : 지도상의 위도 경우, 반경은 km, mi 둘 다 가능
- n= : 가져오는 트윗 수 , 최대 3,200 가능
- userTimeline() : 최근 게시물 가져오기

```
#R=Twitter mining with R
library(twitteR)    #install.packages("twitteR")
library(ROAuth)     #install.packages("ROAuth")
consumerKey <- "WvcnfbfW5Hj*****"     #(1)
consumerSecret<-"SljddS5pOdI6Pb6U05GYIJOrfyYh*********"  #(2)
accesstoken <- "54046261-SDH8eu34JoW0J***************"  #(3)
accesstokensecret <- "hpJdTZKH5g1Tykk6t***********"    #(4)
setup_twitter_oauth(consumerKey, consumerSecret, accesstoken,
accesstokensecret)
#Search keyword in Twitter
#미국 뉴욕- 1000km 반경 - 키워드 '남북회담' 최대개수 1,000개 (5월1일~17일)
#반경을 1,000km 설정에도 불구하고 실제 가져온 트윗 수는 227
tweets.usa<-searchTwitter('Korea',n=1000, lang="ko", since ="2018-05-01",
until="2018-05-17", geocode='43.4060924,-77.697743,1000km')
#서울 100km-키워드: 남북회담 (5월1일~17일)
tweets.korea<-searchTwitter('남북회담',n=1000, lang="ko", since ="2018-05-01" ,
until="2018-05-17", geocode='37.566535,126.977969, 100km')
#청와대
tweets.blue<-userTimeline(user='TheBlueHouseKR',n=1000)
```

'남북회담' 미국 뉴욕은 반경 1,000km, n=1000 지정했으나 236개 트윗만 있음

```
> tweets.usa<-searchTwitter('남북회담',n=1000, lang="ko",
+ since ="2018-05-01", until="2018-05-18",
+ geocode='43.4060924,-77.6977438,1000km') #미국뉴욕 1000km반경 - Korea
Warning message:
In doRppAPICall("search/tweets", n, params = params, retryOnRateLimi
mit,  :
  1000 tweets were requested but the API can only return 236
```

트윗 데이터 포멧으로 변환하기

실제 불러들인 트윗은 데이터 프레임 형식이 아님. 이를 분석 가능한 데이터 형식으로 바꾸어 주는 함수 twListToDF() 이용한다.

```
is.data.frame(tweets.usa)  #데이터 프레임 여부 체크
tweets.usa.df<-twListToDF(tweets.usa) #트윗 리스트 - 데이터 프레임 변환
is.data.frame(tweets.usa.df)
names(tweets.usa.df) #트윗 변수 정보 출력
head(tweets.usa.df,3) #트윗 데이터 첫 3개 출력
tweets.korea.df<-twListToDF(tweets.korea)
head(tweets.korea.df,3)
tweets.blue.df<-twListToDF(tweets.blue)
head(tweets.blue.df,3)
```

```
> is.data.frame(tweets.usa)  #데이터 프레임 여부 체크
[1] FALSE
> tweets.usa.df<-twListToDF(tweets.usa) #트윗 리스트 - 데이터 프레임 변환
> is.data.frame(tweets.usa.df)
[1] TRUE
> names(tweets.usa.df) #트윗 변수 정보 출력
 [1] "text"          "favorited"     "favoriteCount" "replyToSN"
 [5] "created"       "truncated"     "replyToSID"    "id"
 [9] "replyToUID"    "statusSource"  "screenName"    "retweetCount"
[13] "isRetweet"     "retweeted"     "longitude"     "latitude"
```

불러온 트윗은 데이터 프레임 아님

데이터 프레임으로 변환

16개 변수가 있음 - 글 내용은 text 변수에 저장되어 있음

```
> head(tweets.usa.df,1) #트윗 데이터 첫 3개 출력

text
1 남북 대치상황은  곧\n자유한국당의 존립기반,\n북풍과 각종공작정치로\n자신들의 실정과\n부정부패를 \n조중동으로
하여금\n덮어 버리는 저질 패륜정치꾼들..,\n\n태영호 강연..민주 "남북회담 연기에 빌미"·한국 "헌법… https://t.
co/hqgr7KU4fT
  favorited favoriteCount replyToSN             created truncated replyToSID
1     FALSE             0      <NA> 2018-05-17 22:56:17      TRUE       <NA>
                  id replyToUID
1 997249476585652225       <NA>
                                                                        statusSource
1 <a href="http://twitter.com/download/android" rel="nofollow">Twitter for Android</a>
  screenName retweetCount isRetweet retweeted longitude latitude
1    coreain             0     FALSE     FALSE        NA       NA
```

```
> head(tweets.korea.df,1)

text
1 태영호 강연..민주 "남북회담 연기에 빌미"·한국 "헌법상 자유" | 다음뉴스 https://t.co/pUSyNYJAfs
  favorited favoriteCount replyToSN             created truncated replyToSID
1     FALSE             0      <NA> 2018-05-17 22:53:34     FALSE       <NA>
                  id replyToUID
1 997248793190912001       <NA>
                                                    statusSource screenName
1 <a href="http://twitter.com" rel="nofollow">Twitter Web Client</a>    k000042
  retweetCount isRetweet retweeted longitude latitude
1            0     FALSE     FALSE      <NA>     <NA>
```

```
> head(tweets.blue.df,1)

text
1 문재인 대통령은 5월21일, 22일 양일간 미국을 공식 실무 방문하여 트럼프 대통령과 5월22일 백악관에서 정상회담 등
 일정을 가질 예정입니다. 이와 관련해 남관표 국가안보실 제2차장이 브리핑을 가졌습니다. https://t.co/ubGuYpiP
Ha
  favorited favoriteCount replyToSN             created truncated replyToSID
1     FALSE           988      <NA> 2018-05-18 07:02:33     FALSE       <NA>
                  id replyToUID
1 997371846587641856       <NA>
                                                                    statusSource
1 <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>
      screenName retweetCount isRetweet retweeted longitude latitude
1 TheBlueHouseKR          963     FALSE     FALSE        NA       NA
```

불 필요한 글자 제외

불러 들인 트윗 내용 중 불필요한 문장을 제외한다. 트윗 문자의 특성을 고려함

```
tweets.text<-tweets.korea.df$text  #분석 대상 트윗 내용
# 불필요한 문자를 필터링
tweets.text <- gsub("\n", "", tweets.text)
tweets.text <- gsub("\r", "", tweets.text)
tweets.text <- gsub("RT", "", tweets.text)
tweets.text <- gsub("H3", "", tweets.text)
tweets.text <- gsub("h3", "", tweets.text)
tweets.text <-gsub("http","",tweets.text)
```

한글 자연어 처리 : 트윗 문장 단어 처리 : Map(), sapply() 함수 이용

- KoNLP() 한글 자연어 처리 함수 라이브러리 - 영어는 NLP임
- 문장에서 단어(명사) 분리하는 함수 : sapply(), Map() 어느 함수를 이용하여도 결과는 동일
- useSejongDic() 은 세종사전 사용하여 명사 단어를 선택하게 된다.

```
# Apply extract Noun. : koNLP(Korean Natural Lanaguage Processing)
library(rJava)
library(KoNLP)
library(plyr)
useSejongDic()  #사용할 사전 설정

# 문장에서 단어(명사) 분리 - Map 이용
tweets.nouns<-Map(extractNoun, tweets.text)
head(tweets.nouns,1)
tweets.word<-unlist(tweets.nouns, use.name=F)
head(tweets.word,5)

#sapply 함수 이용
txt.nouns<-sapply(tweets.text,extractNoun,USE.NAMES = F)
head(txt.nouns,1)
txt.word<-unlist(txt.nouns)
head(txt.word,5)
```

```
> head(tweets.nouns,1)
$...
 [1] "남북"          "대치상황"      "곧자유한국당"
 [4] "존립기반"      "북풍"          "각종공작정치로자신들"
 [7] "실정"          "과부"          "정부"
[10] "패"            "조중동으로"    "하여금덮어"
[13] "저질"          "패륜"          "정치꾼"
[16] "들"            ",태영호"       "강연"
[19] "민주"          "남북회담"      "연기"
[22] "빌미\"·한국"   "헌법"          "s"
[25] "t"             "co/hqgr7KU4f"  "T"
```

```
> head(tweets.word,5)
[1] "남북"      "대치상황"  "곧자유한국당" "존립기반"  "북풍"
```

```
> head(txt.nouns,1)
[[1]]
 [1] "남북"          "대치상황"      "곧자유한국당"
 [4] "존립기반"      "북풍"          "각종공작정치로자신들"
 [7] "실정"          "과부"          "정부"
[10] "패"            "조중동으로"    "하여금덮어"
[13] "저질"          "패륜"          "정치꾼"
[16] "들"            ",태영호"       "강연"
[19] "민주"          "남북회담"      "연기"
[22] "빌미\"·한국"   "헌법"          "s"
[25] "t"             "co/hqgr7KU4f"  "T"
```

```
> head(txt.word,5)
[1] "남북"      "대치상황"  "곧자유한국당" "존립기반"  "북풍"
```

자연어 처리 후 불필요 단어 제거 : gsub() 함수

단어 처리 후 불필요한 단어를 제거한다. 이 작업은 단어 빈도분석 후 작업자가 반복적으로 수작업으로 진행하게 된다. 예를 들어 '키워드'가 남북회담이었으므로 남북회담이 가장 빈번히 나와 이를 제외하였고, 다시 분석 결과 문재, 남북, 회담의 단어가 많아 차례로 삭제하였음

```
# 워드 클라우드 사용하지 않은 단어 제거
tweets.word<- gsub("[[:punct:]]","", tweets.word)
tweets.word<- gsub("[0-9]+[A-Za-z]", "", tweets.word) #숫자+알파벳 제거
tweets.word<- gsub("남북회담", "", tweets.word) #빈도 분석 후 불필요 단어 제거
tweets.word<- gsub("문재", "", tweets.word)
tweets.word<- gsub("남북", "", tweets.word)
tweets.word<- gsub("회담", "", tweets.word)
tweets.word<- gsub("co", "", tweets.word)
#단어 길이 2개 이상 선택
tweets.word<- Filter(function(x){nchar(x)>=2}, tweets.word)
```

```
> head(tweets.word,5)
[1] "대치상황"                "곧자유한국당"              "존립기반"
[4] "북풍"                    "각종공작정치로자신들"
```

단어 빈도 카운트 - table() 함수

```
# 단어별 카운팅, 상위 10개 단어 선택
tweets.count<-table(tweets.word)
head(sort(tweets.count, decreasing=T), 10)
```

table() 함수를 사용하여 단어의 빈도를 계산 - 상위 10개 단어 빈도와 함께 출력

```
> head(sort(tweets.count, decreasing=T), 10)
tweets.word
      고위급       대통령         남자         도시 맵짜다맵짜다       북한서
         122          107          103          103          103          103
      소탈한       이미지      이미지와       카리스마
         103          103          103          103
```

단어 빈도 표현 wordcloud() 함수

```
library(RColorBrewer)
library(wordcloud)
pal <- brewer.pal(7,"Paired") # 컬러 세팅
wordcloud(names(tweets.count),freq=tweets.count,scale=c(4,0.5),min.freq=5,
          random.order=F,rot.per=.1,colors=pal,family='AppleGothic')
```

최소빈도 5개(미국) - 10개(한국) 단어만 표현

서울 트윗

뉴욕 트윗

[청와대 트윗 텍스트 마이닝 결과]

```
> tweets.nouns<-Map(extractNoun, tweets.text)
> head(tweets.nouns,1)
$...
 [1] "문재"       "대통령"     "5월21일"    "22"          "양일간"     "미국을"
 [7] "공식"       "실무"       "방문"       "트럼프"      "대통령"     "5월22일"
[13] "백악관"     "정상회담"   "등"         "일정"        "예정"       "이"
[19] "관련"       "해"         "남관표"     "국가안보"    "제2차장이"  "브리핑"
[25] "s"          "t"          "co"         "ubGuYpiPH"   "a"

> tweets.word<-unlist(tweets.nouns, use.name=F)
> head(tweets.word,5)
[1] "문재"     "대통령"   "5월21일" "22"       "양일간"
```