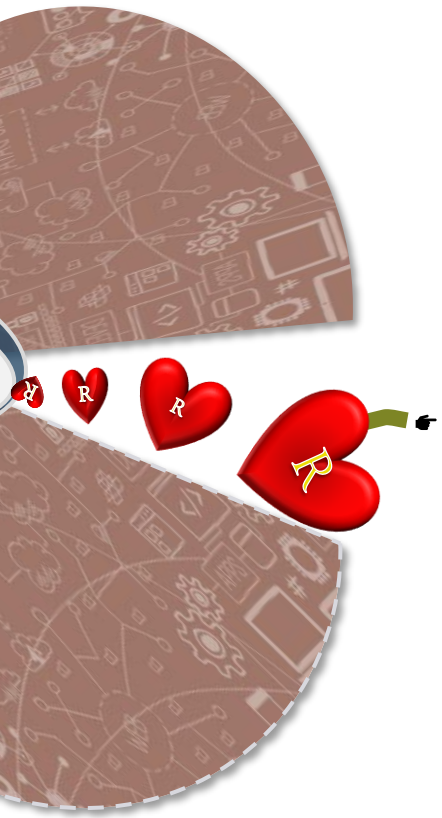
6. 트윗데이터 전처리



트윗데이터를 읽어온 후 기본 전처리 과정을 진행해보아요.

I. 트윗 데이터 읽어온 후 정제작업

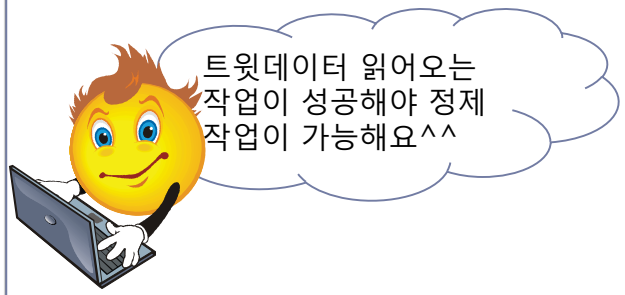
◆ 트위터에 접속하여 "present"라는 검색어를 입력하여 트윗데이터 1000건을 읽어온 후 불필요한 태그와 url을 제거한 후 단어단위로 빈도를 분석해보는 작업을 함께 해보죠.


```
install.packages(c("twitteR","dplyr"))
library(twitteR)
library(dplyr)
Sys.setenv(JAVA_HOME="c:\\program files\\java\\jdk1.8.0_141")
install.packages("KoNLP")
library(KoNLP)
api_key <-"fxd8VtG22tTN25KvQLiF5DjfE"
api_secret <-"nAhcVYDk4aGvlD5W4Hr5GizSsbKEafy2LmiUJ0PYbxpufTh6xY"
access_token <-"257325922-KAH1GdOi2Z4jwDRGDIgkvO24eJN3MwymrZ3gYoJH"
access_token_secret <-"P8ap67uUllyXc45A10PwSRhdzcaMpbAxBNDsT8ujBA7e4"
setup_twitter_oauth(api_key,api_secret,access_token,access_token_secret)
```

```
> library(dplyr)

다음의 패키지를 부착합니다: 'dplyr'

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

    id, location

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

    filter, lag

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

    intersect, setdiff, setequal, union

> api_key <- "fxd8VtG22tTN25KvQLiF5DjfE"
> api_secret <- "nAhcVYDk4aGvlD5W4Hr5GizSsbKEafy2LmiUJ0PYbxp
> access_token <- "257325922-KAH1GdOi2Z4jwDRGDIgkvO24eJN3Mwy
> access_token_secret <- "P8ap67uUllyXc45A10PwSRhdzcaMpbAxBN
> setup_twitter_oauth(api_key,api_secret,access_token,access
[1] "Using direct authentication"
>
```

HOW TO 1) 트위터에 검색어 "present"입력하여 1000개의 트윗을 가져와서 확인하는 작업을 그림처럼 실행해보아요.

```
> msg_tweet<-searchTwitter(enc2utf8("새해"),n=1000,lang="ko")
> head(msg_tweet)
[[1]]
[1] "xoskiz: RT @Stray_Kids: [SPOT KIDS]\n새해맞이 Stray Kids 랜덤 선물 TIME !\n\n호기심이 많은 삼공즈는\n$

[[2]]
[1] "dokdotea: RT @Jaemyung_Lee: &lt;예산은 충분한데도 성남은 자유한국당과 바른정당 시의원들이 극력반대입$

[[3]]
[1] "Wanna19991102: RT @_deockjil_: 재환초커 겸 새해 겸 옹넬환 예능겸 완깐 기원하면서 존잘밈과 합작으로 제$

[[4]]
[1] "gwirlgroups: RT @10_PRISTIN: [나영] 하이들~새해복많이받으세염ㅎㅎ\nwe무대한날 셀카 빠밤 https://t.co/$

[[5]]
[1] "hscw0526: @woo_gun1122 혹여 현생이 바쁘신거라면 그에 부응할수있는 성과를 꼭 이루셨으면 좋겠어요 :) 부$

[[6]]
[1] "Nuqlear__: RT @maboo_chu: 아침에 일어나면 밤사이에 육백명의 회원들로부터 단톡방에 몇백통의 냥짤이 올$

> class(msg_tweet)
[1] "list"
> length(msg_tweet)
[1] 1000
```



2) 트윗의 자료구조형태가 리스트이므로, 이를 데이터프레임 구조로 변경한 다음 중복된 트윗내용에 대해 제거하는 작업을 그림처럼 코딩해보죠.

```
> df<-twListToDF(msg_tweet)
> class(df)
[1] "data.frame"
> df1<-distinct(df,text)
>
```

트윗을 읽어오면 list형태인데요. 이를 데이터프레임으로 변경할 때 twListTiDF()함수를 이용해요.

3) 그림처럼 TEXT를 실행해보면 중복제거된 트윗 내용을 볼 수 있어요. 그럼, 여기서 우리가 분석할 실제 텍스트는?

```
txt<-df1$text
txt
```

```
[379] "RT @image_sadam: 크으으 넘들 새해가 밝았으니 새 마음 새 뜻으로 계연을 함 때려줘야 하지 않겠습니까?($
[380] "많이 배울게요~!! &lt; 새해다짐 중 최고민음직"                                        $
[381] "@Kadena__ 새해버프 기간 끝낫지안앗나유"                                             $
[382] "RT @bugissu: RT) 추첨 한분께 한세트\n\xed ½\xed°¢선물같은 종현이의 굿즈팩을 판매\xed ½\xed°¢\n$
[383] "RT @shadow0506: Another three hundred sixty five blank pages to fill 新年快□ 새해 복 많이 받으세요$
[384] "리퀘박스로 새해 소원 지정 https://t.co/blaaNpsLAC"                                   $
[385] "RT @SONAMOO_EuiJin: 새해 복 많이 받으세요\xed ½\xed¹\u008c\xed ¼\xed¿» 2018년에도 건강하고 행복$
[386] "RT @yousowol: 레진 코믹스 청원이 8만명이라는 굉장한 성적으로 종료되었습니다. 전부 여러 분들이 참여$
[387] "RT @min_x_minmin: 부족한 그림인데도 예쁘게 봐 주셔서 항상 감사드립니다:)\n올해도 잘 부탁드리겠습니$
[388] "RT @zzv0w0vzz: 종현이도 새해복 많이 받아:) https://t.co/hxZun3FFTk"                  $
[389] "RT @naturalraison: @kihone17 워아”이니”\xed ½\xed \u0080 워아이니 “WeAre이니” 갈기도 하고 좋$
[390] "RT @2moonorbit: RT♥ 새해 첫달 1월 이달도 저와 찬백하시죠 이벤트 (거창\n\n이번에 성인이 되신 분들도$
[391] "RT @suki_vixx: 녹화 다 끝나고 마무리 인사 한명씩\n\nㅇㄴㅇ 감기 조심하고 옷 따듯하게 입어요.\n'ㅅ' $
[392] "RT @HashTag_offcial: 다들 새해 첫 주말 잘 보내구 계신가요?! 저는 얼른 여러분을 보고 싶은 마음 뿐입$
[393] "RT @johahani: \xed ½\xed²\u0095\xed ½\xed°¨2018 하니 달력 EVENT\xed ½\xed°¨\xed ½\xed²\u0$
[394] "RT @cotton_peach_: [히카게] 새해 소원 https://t.co/XuNY9Yt59k #히나원_특별전력 #히나원_전력 @hinata$
[395] "RT @megatonagaci: 진짜 최고의 새해선물이다≡\n쇼 타임 뉴이스트 타임!!! ㅠㅠㅠㅠㅠㅠ https://t.co/ln$
[396] "RT @BAP_Fanstaff: 2018년 새해가 밝았습니다! BABY 여러분들 새해 복 많이 받으세요!\xed ½\xed¹\u008f$
[397] "새해떡 72개ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ"         $
[398] "@aristiaramoke 새해라서 그러면.. 더 심각하잖어요..!88 진짜 몸 챙기세요.. 밤 새고 그러시지 마시구..\$
[399] "원래 왓챠를 본 영화 기록하는 용도로 썼는데 데이터베이스도 빈약하고 별점을 반드시 표기해야 하는 것도$
[400] "@0alwayss0 히히 감사합니다 □ 새해에는 새로운 마음가짐으루 살고 싶은데 쉽지 않네욧 늘넘두 올해 딱 고$
[401] "RT @akauro_Petit: [RT이벤트] 새해를 맞이하여 아카시와 쿠로코가 새해선물을 보내드립니다□(□□ω□)□☆\$
[402] "백가희 선생님 새해 복 많이 받으시고 행복하세요...ㅠㅠ"                                $
[403] "@_white_snow_S2 코미님 생일축하드려요!!  오늘하루 즐겁게보내시고 오늘말고도 새해  쭈욱 행복한일만 $
[404] "RT @VIXX_Record_: 새해부터 뜨거운 관심 가져주시고 응원해주신 모든 분들 진심으로 감사드립니다. 앞으$
[405] "RT @gyeomuring: #새해_첫_트친소 #코스어_트친소\n\n안녕하세요!! 중부권 코스어 겨무링이라고 합니다\xe$
[406] "새해목표는 그림실력 전체적으로ㅈ높이기.."                                              $
[407] "RT @TYPEB_SCM: #심창민 #최강창민 #MyAngel\n2018 SMTOWN FRIENDS 새해 연탄봉사 \n\n울오빠 웃음소리 들$
[408] "젭깅이들아 안녕하새해요\xed ½\xed±\u008b\xed ¼\xed¿»\n펜밋 끝나구 홍콩여행중인데 어딜가든 너희$
[409] "@omari_cos 안녕하세요! 반갑습니다(*´ω`*) 찾아와주셔서 감사해요. 팔로 드렸스 니다. 새해 복 많이 받$
[410] "RT @dipidalidu: #정용화  180106 새해 복 많이 받으세요 \xed ½\xed²\u0095 https://t.co/doHcpPer6L" $
[411] "진짜 최고의 새해선물이다≡\n쇼 타임 뉴이스트 타임!!! ㅠㅠㅠㅠㅠㅠ https://t.co/ln9jDYCIjo"      $
[412] "RT @BZ_House: 크리스마스 대소동 下편\n새해가 되어서야 끝난 하편...^_T 얼렁뚱땅엔딩 죄성합미댜..!!\n$
[413] "RT @Let_s_Taek: 예쁘고 착하고 아름다운 택운이도\n새해 복 많이 받아□□\n@JUNGTW_LEO https://t.co/XV0lR$
[414] "새해버프주삼..."                                                                    $
[415] "RT @GOT7Official: [REPLAY]\nHappy New Year I GOT7□\xed ½\xed°¥\nhttps://t.co/lOWaMlTQ6K\n아가새 $
[416] "RT @dodam_cos_: #새해 첫 트친소 \n#코스어 트친소 \n\n슴살 성인 된 기념 뱜라가 넘 조용—해서 여는 트$
```

RT,@트윗태그, HTTP URL주소,숫자,특수문자 등등은 한 마디로 분석에서 제외시켜야 한다는 것죠.

정리) 정규표현식 이용한 문자열처리

➢ 정규표현식 – 문자열 처리를 유연하게 하기위한 기법
여러 프로그래밍 언어에서 공통적으로 문자열 매칭/치환에 사용

* 주요 정규표현식 기본정리

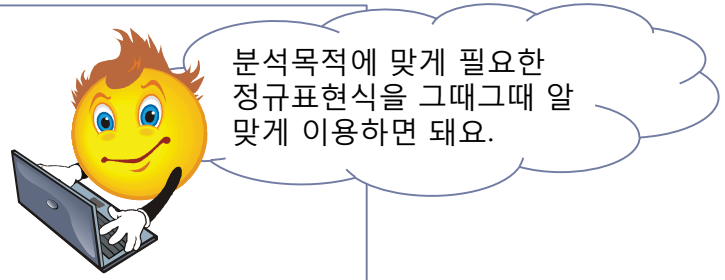
표현식	설명	\w	알파벳이나 숫자
^	문자열 시작	\W	알파벳이나 숫자를 제외한 문자
$	문자열 종료	\d	숫자[0~9]와 동일
*	앞 문자가 없을 수도 무한정 많을 수도 있음.	\D	숫자를 제외한 모든 문자
+	앞문자가 하나 이상	\특수문자	해당 특수문자를 의미
?	앞 문자 없거나 하나 있음		
[]	문자집합이나 범위를 나타내며, 두 문자 사이에는 – 기호로 범위표시		
{}	횟수 또는 범위를 나타냄		
()	소괄호 안의 문자를 하나의 문자로 인식		
|	패턴 안에서 or 연산을 수행할 때 사용		
\s	공백문자		
\S	공백문자가 아닌 나머지 문자		

* 표현식을 유연하게 – 조금 더 향상된 문자열 클래스 정리

표현식	설명
[:digit:]	숫자 의미. [0-9] 혹은 \d
[:alpha:]	영문 대소문자 표현 [A-z] 또는 [A-Za-z]
[:alnum:]	영문, 숫자 표현 [A-z0-9] 또는\w
[:space:]	공백, 탭, 개행문자 등을 표시
[:punct:]	특수문자 표시 !@#$%^&*();:<>,.{}
[:graph:]	영문자 특수문자 모두를 표시
[:cntrl:]	\n, \r과 같은 제어문자를 표시

4) 정규표현식을 이용하여 불필요한 문자,특수문자등을 제거하는 작업을 그림처럼 작업해보죠.

```
> txt1<-gsub('RT','',txt)
> txt1<-gsub('https.* *','',txt1)
> txt1<-gsub('@[[:graph:]]*','',txt1)
> txt1<-gsub('[\t\r\n]+','',txt1)
> txt1<-gsub('[[:punct:]]+','',txt1)
> txt1<-gsub('[A-Za-z]+','',txt1)
> txt1<-gsub('\\d+','',txt1)
> txt1<-gsub('[ㅏㅑㅜㅠㅗㅛㅋㅍ]*','',txt1)
> head(txt1,10)
 [1] "   새해맞이   랜덤 선물  호기심이 많은 삼공즈는멤버들을 위한 실용템을 선택과연 누구에게 선물이 갈지$
 [2] "  예산은 충분한데도 성남은 자유한국당과 바른정당 시의원들이 극력반대입니다 교복지원도 반대 급식지원$
 [3] "  재환초커 겸 새해 겸 옹벨환 예능겸 완깐 기원하면서 존잘밈과 합작으로 제작한 드림 재환 도무송 나눔해
 [4] "  나영 하이들새해복많이받으세염ㅎㅎ무대한날 셀카 빠밤 "
 [5] " 혹여 현생이 바쁘신거라면 그에 부응할수있는 성과를 꼭 이루셨으면 좋겠어요  부디 조금만 방황하시고 많
 [6] "  아침에 일어나면 밤사이에 육백명의 회원들로부터 단톡방에 몇백통의 냥짤이 올라와 있는것임ㄱ 거기서는
 [7] "  일주일이나 지났지만 새해복 많이 받으세요□□愼□□□□□□□□□ "                                $
 [8] "  □□愼□□□□□□□□□년 황금개의 해 새해 복 많이 받으세요□□愼□□□□□□□□□ "          $
 [9] "  급작스럽지만 새해 맞이  이벤트 해욥  해주신 분들 중 분께 금뚜껑이 넝뭉행 시즌 세트스티커 매  도무$
[10] "   망고와 약속한 새해 소망을 열심히 지키고 있을 러브들에게 겨울 망고를 선물합니다 □□愼□□□□□□$
>
>
```



5) 유의미한 명사형 단어들을 추출하고 이들을 벡터형태로 변경하여 각 단어마다 음절이 2 음절 이상인 단어만을 추출하는 동작을 그림처럼 코딩해요.

```
> txt_word<-sapply(txt1,extractNoun,USE.NAMES=F)
> txt_data<-unlist(txt_word)
>
> txt_r<-Filter(function(x){nchar(x)>=2},txt_data)
> head(txt_r,20)
 [1] "새해맞이"        "랜덤"          "선물"          "호기심"        "삼공즈는멤버들을"
 [6] "실용템을"        "선택과"        "누구"          "선물"          "갈지내일공개"
[11] "스트레이키"      "예산"          "충분"          "한데"          "자유한국당과"
[16] "바른정당"        "시의원"        "들이"          "극력반대입니다" "교복"
>
```

여러분의 분석 목적에 따라 음절의 개수 조건을 달리해도 되겠죠?

6) 가장 많이 언급된 단어들만 간추려서 그 결과를 알아보죠. 그림처럼 코딩해보죠.

```
> txt_r<-table(txt_r)
> class(txt_r)
[1] "table"
> head(sort(txt_r,decreasing=T),20)
txt_r
    새해     행복   이벤트     선물     기념     맞이     트친     마음 받으세요     올해     우리
     326       41       30       26       22       22       22       21       20       18       18
    추첨 새해맞이     인사     진짜   받으세     가요     건강     오늘   대제전
      18       17       17       17       16       14       14       14       13
> |
```

단어의 빈도수를 계산

언급이 많이 된 순서대로 20개를 추출하여 화면에 표시

7) 빈도결과를 시각화 하여 표시하면 더 낫겠죠? 가장 많이 사용하는 wordcloud 형태로 결과를 표시해보죠.

```
> install.packages(c("wordcloud","RColorBrewer"))
경고: 패키지 'wordcloud', 'RColorBrewer'가 사용중이므로 설치되지 않을 것입니다
> library(wordcloud)
> library(RColorBrewer)
> choice_pal<-brewer.pal(8,'Dark2')
> wordcloud(names(txt_r),freq=txt_r,scale=c(6,1),rot.per=0.25,min.freq=7,colors=choice_pal)
> |
```