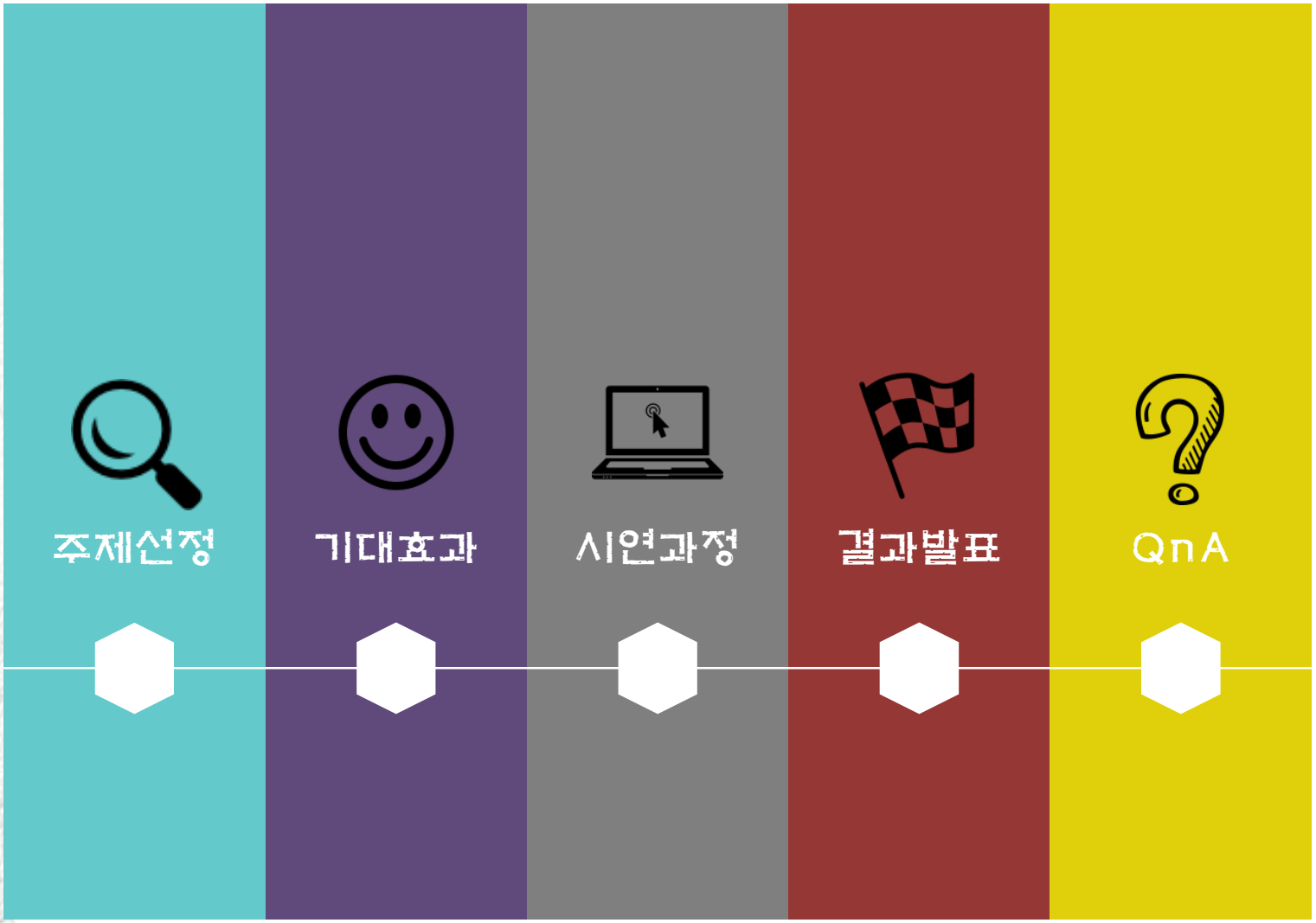
# B3 <Sherlock>

## 빅데이터와 네트워크 연계를 통한 SNS 여론조작과 거짓정보 탐지

추지나 최세원 이시연 김유진 김슬기

# 목차



# 주제

빅데이터와 네트워크 연계를 통한 SNS 여론조작과 거짓정보 탐지

# 목적

Sns중에서도 트위터를 가지고 빅데이터와 연계를 하여 주요 단어를 추출한 다음 우리만의 사전을 만든다. 그 후에 그 사전을 기준으로 참 거짓패턴을 만들 수 있고 나아가 사용자 인증과 함께 여론 조작 방지할 수 있다.

# 기대효과

**<작은 범위 보안>**

1. 자살예방을 위한 사전 대책
2. 취약한 SNS 사용자 인증을 파고든 사기 범죄 예방

**<큰 범위 보안>**

1. 실시간으로 미리 해킹대상이 될 수 있는
국가공인기관 및 사설보안업체에 알려줄 수 있다.

2. 사용자 인증을 할 수 있도록 예측을 위한 기계학습으로 구성되어 있기에
국정원 직원의 여론 조작 사건과 같은 일이 없도록 미리 사전에 사용자가
진짜 그 사용자인가, 아니면 여론 조작을 위해 사용하는
가짜 아이디 인가를 식별할 수 있게 해준다.

3. 계속 축적되는 데이터들을 통해 학습된 AI로 발전이 된다
네트워크를 이용하여 IDS를 통해 해당 사용자의 가까운 미래를
예측할 수 있는 스마트 생활환경을 이끌어나가는 시초가 될 수 있을 것이다.

# 시연과정

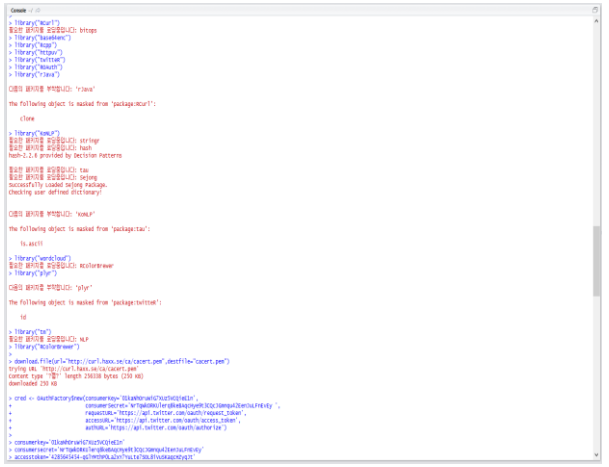
## 1) 트위터 연동

패키지 인스톨

데이터 추출

Consumer key secret 입력

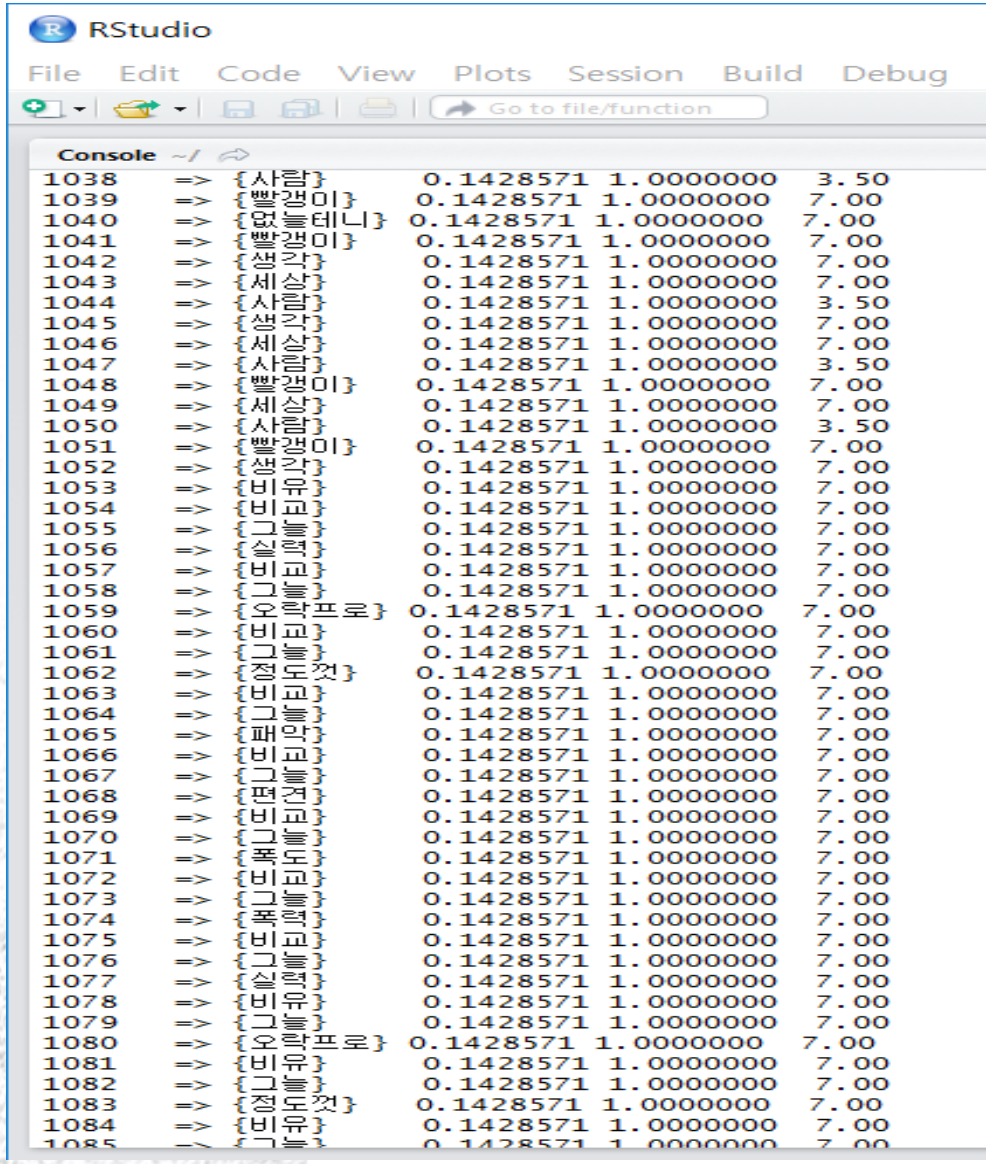
계정 불러오기



```
[[10]]
[1] "trytango: I.GWANGJU.U..나는 너를 폭동한다?! 맞나?"

[[12]]
[1] "trytango: 문죄인, 홍남의 꿈 어쩌고 떠드는데..홍남의 꿈?! 뭐? 국가전복? 폭동의 추억? 뭔데?

[[10]]
[1] "9ood2cu: RT @ShowMeTheGogi: 종북단체 언론노조가 '

[41]]
[1] "9ood2cu: RT @FreeUnify_LEE: 김영삼 전 대통령 영결식도 꼴난이니 5.18특별법은 폐지하고 광주사태 당시 발생한 폭도들에 대해 재조명해야 합니다."
```

```
Console ~/
> accesssecret='mqyXNWW5eBY8uMFNXMX4DGULFQsFpMtjHChvCEqbLUE1T'
> setup_twitter_oauth('01kaNhOruWiG7XuzSvCQieE1n','NrTqWkORKUlerq8keBAqcHye9t3CQcJG                    OLa2xY7YuLte7SOL8iVuSKagcHZygJt','mqyXNWW5eBY8uMFNXMX4DGULFQsFpMtjHChvCEqbLUE1T')
[1] "Using direct authentication"
Use a local file to cache OAuth access credentials between R sessions?
1: Yes
2: No

선택:
Enter an item from the menu, or 0 to exit
선택:
Enter an item from the menu, or 0 to exit
선택:
Enter an item from the menu, or 0 to exit
선택: 1
> searchTwitter("trytango",n = 30, lang = "ko", )
```

# 2) 임의 단어등급사전 만들기




```
1614    => {한반도가} 0.1428571 1.0000000  7.00
1615    => {정신}     0.1428571 1.0000000  7.00
1616    => {비판}     0.1428571 1.0000000  7.00
1617    => {흑백논리} 0.1428571 1.0000000  7.00
1618    => {정신}     0.1428571 1.0000000  7.00
1619    => {비판}     0.1428571 1.0000000  7.00
1620    => {구분}     0.1428571 1.0000000  7.00
1621    => {정신}     0.1428571 1.0000000  7.00
1622    => {비판}     0.1428571 1.0000000  7.00
1623    => {김영삼}   0.1428571 1.0000000  3.50
1624    => {정신}     0.1428571 1.0000000  7.00
1625    => {비판}     0.1428571 1.0000000  7.00
1626    => {한반도가} 0.1428571 1.0000000  7.00
1627    => {통일}     0.1428571 1.0000000  7.00
1628    => {비판}     0.1428571 1.0000000  7.00
1629    => {흑백논리} 0.1428571 1.0000000  7.00
1630    => {통일}     0.1428571 1.0000000  7.00
1631    => {비판}     0.1428571 1.0000000  7.00
1632    => {구분}     0.1428571 1.0000000  7.00
1633    => {통일}     0.1428571 1.0000000  7.00
1634    => {비판}     0.1428571 1.0000000  7.00
1635    => {김영삼}   0.1428571 1.0000000  3.50
1636    => {통일}     0.1428571 1.0000000  7.00
1637    => {비판}     0.1428571 1.0000000  7.00
1638    => {흑백논리} 0.1428571 1.0000000  7.00
1639    => {한반도가} 0.1428571 1.0000000  7.00
1640    => {비판}     0.1428571 1.0000000  7.00
1641    => {구분}     0.1428571 1.0000000  7.00
1642    => {한반도가} 0.1428571 1.0000000  7.00
1643    => {비판}     0.1428571 1.0000000  7.00
1644    => {김영삼}   0.1428571 1.0000000  3.50
1645    => {한반도가} 0.1428571 1.0000000  7.00
1646    => {비판}     0.1428571 1.0000000  7.00
1647    => {구분}     0.1428571 1.0000000  7.00
1648    => {흑백논리} 0.1428571 1.0000000  7.00
1649    => {비판}     0.1428571 1.0000000  7.00
1650    => {김영삼}   0.1428571 1.0000000  3.50
1651    => {흑백논리} 0.1428571 1.0000000  7.00
1652    => {비판}     0.1428571 1.0000000  7.00
1653    => {김영삼}   0.1428571 1.0000000  3.50
1654    => {구분}     0.1428571 1.0000000  7.00
1655    => {비판}     0.1428571 1.0000000  7.00
1656    => {전두환도} 0.1428571 1.0000000  7.00
```

# 3) 단어별 사용자 아이디 검색

## 필요한 패키지 로딩 및 부착

```
> library("ROAuth")
> library("twitteR")
> library("base64enc")
> library("KoNLP")
필요한 패키지를 로딩중입니다: rJava
필요한 패키지를 로딩중입니다: stringr
필요한 패키지를 로딩중입니다: hash
hash-2.2.6 provided by Decision Patterns

필요한 패키지를 로딩중입니다: tau
필요한 패키지를 로딩중입니다: Sejong
Successfully Loaded Sejong Package.
Checking user defined dictionary!


다음의 패키지를 부착합니다: 'KoNLP'

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

    is.ascii
```

## 트위터 권한 얻어오기

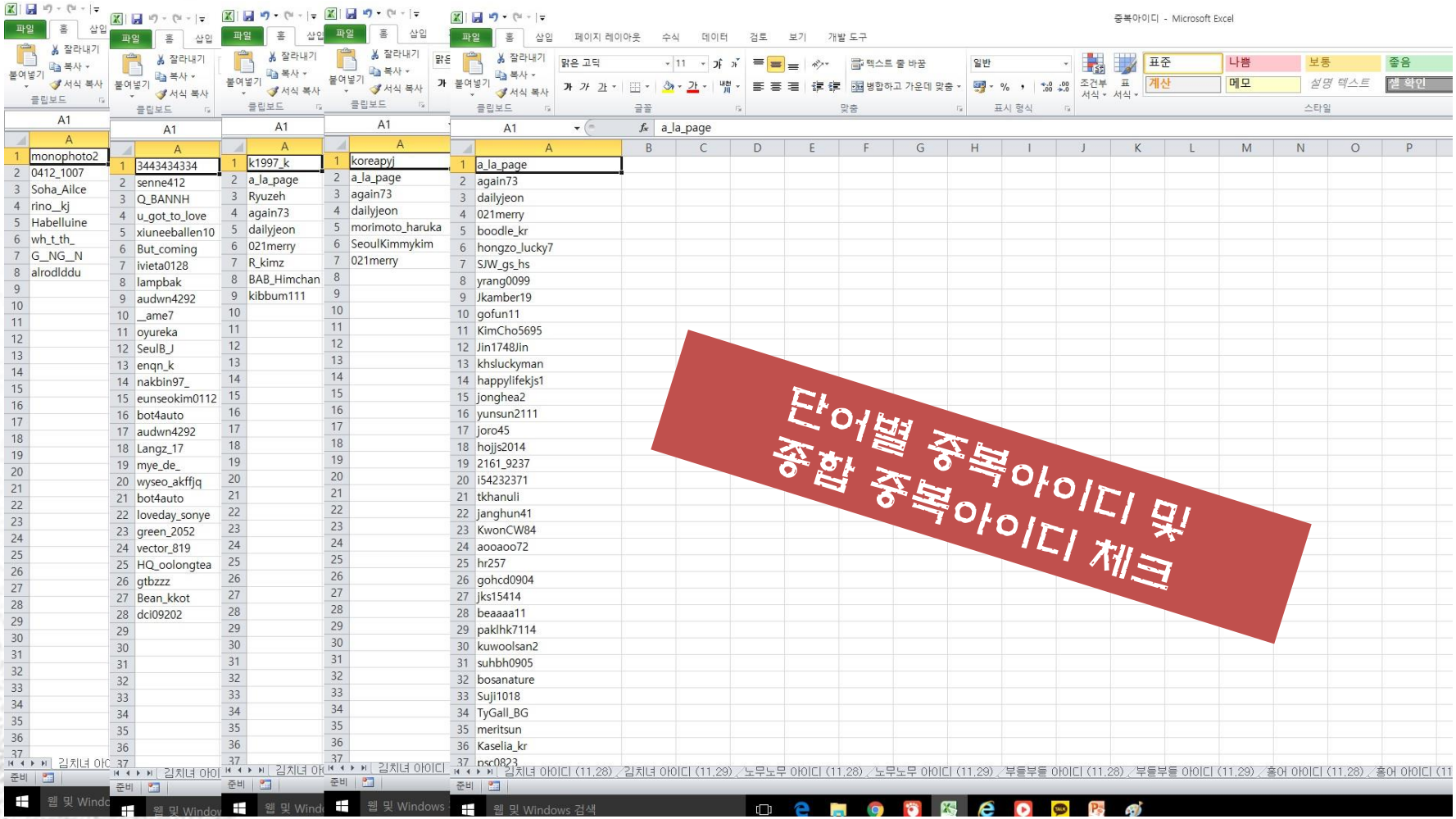
```
> consumerKey <- "K4XH53hnX3eb0pht9ppwSmsnN"
> consumerSecret <- "Mk2u64TpRLicGk4OeEYr9u1D14OecJH3EP8MJ
VGOe4PIxECYpX"
> accessToken <- "1015005583-GXUOm9xYdNPpJJvYFslOHwzysxYPW
VUVvvSOnS4"
> accessTokenSecret <- "U6SGIKPDblFfltbvhpd2jOdYNKRx1jZTD9
QskxcJXDjLR"
> setup_twitter_oauth(consumerKey, consumerSecret, accessT
oken, accessTokenSecret)
[1] "Using direct authentication"
Use a local file to cache OAuth access credentials between
 R sessions?
1: Yes
2: No

선택: 1
>
```

1. 키워드 지정
2. 최근 일주일 동안 올라온 트위터 중 키워드가 포함된 트위터 1000개 획득
3. 결과들 중 아이디만 뽑아내기
4. 결과 개수 확인
5. 결과 출력

```
> keyword<-enc2utf8("노무노무")
> result<-searchTwitter(keyword, lang="ko", n=1000)
Warning message:
In doRppAPICall("search/tweets", n, params = params, retryOnRateLimit = retryOn
RateLimit,  :
  1000 tweets were requested but the API can only return 380
> result.df<-twListToDF(result)
> result.text<-result.df$screenName
> length(result.text)
[1] 380
```

```
> result.text[1:380]
  [1] "han79bd"          "JP_In"            "wjunbek"
  [4] "cargonoir0207"    "p0rar1s_r4ps0dy"  "fungky88"
  [7] "han79bd"          "Esther_Jang_"     "Fantasia_with_"
 [10] "risrisblue"       "yhk9900"          "charisma749"
 [13] "Jenewarer"        "ramamoo140619"    "vudghkxhddlf123"
 [16] "acquasong"        "SNS_Youjeen"      "jeongdojeon"
 [19] "Kcta3441Crime7"   "PakSujoung"       "dragonoflake"
 [22] "bts000321"        "whlee21"          "iqlimasaphira_"
 [25] "inmo1074"         "Happygaegul"      "sol625625"
 [28] "jthan091"         "mkmbcc"           "kangsan129"
 [31] "histy_00"         "_jtlko07"         "orientalhope"
 [34] "sky966"           "wwwno1room"       "k1997_k"
 [37] "dksalwk___dks"    "yeonhwa9074"      "Nielsemo"
 [40] "Balemyeni"        "heyitsme_jirah"   "pfomtu"
 [43] "cozsy"            "dong_ba_"         "gohcd0904"
 [46] "ranger2848"       "bobsbs200"        "pigret65"
 [49] "ymj31632204"      "choihsk78"        "tngkwls93"
 [52] "hye_orange0439"   "sungjoo1106"      "lsk1955"
 [55] "K_JHoon"          "treenymph1"       "dowon1023"
 [58] "Balemyeni"        "sunklee48"        "haveaniceday60"
 [61] "hs641102"         "rmafbs1"          "sjr0114"
 [64] "happylifekjs1"    "cdh5034"          "leeyoungsik0910"
 [67] "genialjh745"      "changwookim622"   "swsoon7"
 [70] "beaaaa11"         "58ajd"            "Tartuca7219"
 [73] "seogaden"         "SangSang2411"     "ooparts0924"
 [76] "BMbstar17"        "syoung0"          "yrang0099"
 [79] "snag0328"         "nothingtohurt"    "aqpr"
 [82] "winbio4"          "Kimhaksoo228"     "TerraYoon"
 [85] "jdnday1576"       "GOP_DEM"          "hoseongkr"
 [88] "paklhk7114"       "jonghea2"         "LucyCheong"
 [91] "gojangkwan"       "acuhealtcm"       "Pfeila"
 [94] "751fa0b191fe494"  "hope_findaway"    "changbogo100"
 [97] "kecchang"         "hangukin56"       "suddenlly"
[100] "twsomeone"        "wolfnfox0"        "andrewhong2"
[103] "Kslee0601"        "TaoEq"            "Kunhong_Kim"
[106] "JSK135"           "jal_sal_ja"       "commonster1137"
[109] "sana_93"          "Hansarangnim"     "twsomeone"
[112] "cjttkfkd22"       "vkfehrkdtks37"    "emilnolde11"
```

# 4) 중복 아이디 체크



# 5) 주요 단어 추출(단어사전-근접중심성)

## 1. 패키지 인스톨

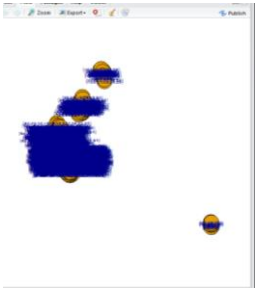
```
> library(KoNLP)
필요한 패키지를 로딩중입니다: rJava
필요한 패키지를 로딩중입니다: stringr
필요한 패키지를 로딩중입니다: hash
hash-2.2.6 provided by Decision Patterns
```

## 2. 코드 실행

```
> f <- file("dddd.txt", encoding="UTF-8")
> fl <- readLines(f)
> close(f)
> tran <- Map(extractNoun, fl)
> tran <- unique(tran)
> tran <- sapply(tran, unique)
> tran <- sapply(tran, function(x) {Filter(function(y) {nchar(y) <= 4 && nchar(y) > 1 && is.hangul(y)},x)} )
> tran <- Filter(function(x){length(x) >= 2}, tran)
> names(tran) <- paste("Tr", 1:length(tran), sep="")
> wordtran <- as(tran, "transactions")
> .
```

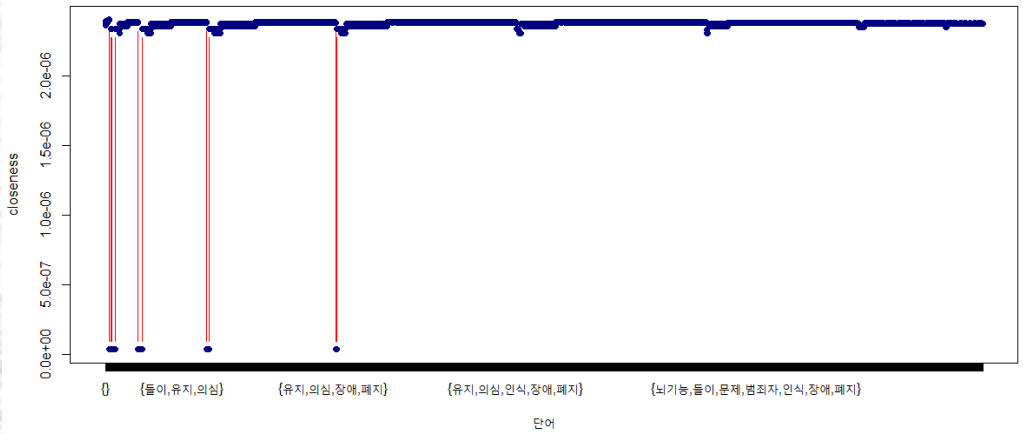
```
> #co-occurance table
> wordtab <- crossTable(wordtran)
> ares <- apriori(wordtran, parameter=list(supp=0.05, conf=0.05))
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport support minlen maxlen
       0.05    0.1    1 none FALSE            TRUE    0.05      1     10
 target   ext
  rules FALSE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 0

Warning in apriori(wordtran, parameter = list(supp = 0.05, conf = 0.05)) :
  You chose a very low absolute support count of 0. You might run out of memory! Increase minimum support.

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[70 item(s), 6 transaction(s)] done [0.00s].
sorting and recoding items ... [70 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 5 6 7 8 9 10 done [0.03s].
writing ... [614888 rule(s)] done [0.15s].
creating S4 object  ... done [0.63s].
> inspect(ares)
```

## 3. 데이터 추출



| | | | | | |
|---|---|---|---|---|---|
| 697 | {비교} | => | {정도껏} | 0.1666667 | 1.0000000 | 6 |
| 698 | {정도껏} | => | {비교} | 0.1666667 | 1.0000000 | 6 |
| 699 | {비교} | => | {취지} | 0.1666667 | 1.0000000 | 6 |
| 700 | {취지} | => | {비교} | 0.1666667 | 1.0000000 | 6 |
| 701 | {비교} | => | {패악} | 0.1666667 | 1.0000000 | 6 |
| 702 | {패악} | => | {비교} | 0.1666667 | 1.0000000 | 6 |
| 703 | {비교} | => | {편견} | 0.1666667 | 1.0000000 | 6 |
| 704 | {편견} | => | {비교} | 0.1666667 | 1.0000000 | 6 |
| 705 | {비교} | => | {폭도} | 0.1666667 | 1.0000000 | 6 |
| 706 | {폭도} | => | {비교} | 0.1666667 | 1.0000000 | 6 |
| 707 | {비교} | => | {폭력} | 0.1666667 | 1.0000000 | 6 |
| 708 | {폭력} | => | {비교} | 0.1666667 | 1.0000000 | 6 |

## 4. 근접중심성(그래프)




```
> closen <- closeness(ruleg)
> plot(closen, col="red",xaxt="n", lty="solid", type="b", xlab="단어", ylab="closeness")
> points(closen, pch=16, col="navy")
> axis(1, seq(1, length(closen)), V(ruleg)$name, cex=5)
```

# 6) 주요단어 확인하기(워드 클라우드 이용)

1. 일베가 쓴 트위터에서 글 추출 후 데이터파일 만들기
   그 파일에서 주요 단어 100개 추출

```
> f <- file("be2.txt",blocking=F)
> txtLines <- readLines(f)
> nouns <- sapply(txtLines,extractNoun,USE.NAMES=F)
> head(unlist(nouns), 20)
 [1] "복면가왕은"    "편견"          "실력"          "취지"          "광화"          "문"            "폭도"
 [8] "들"            "복면"          "익명"          "그늘"          "폭력"          "패악"          "질"
[15] "감추려는거다" "비유"          "때"            "오락프로"      "비교"          "핑계"
> head(unlist(nouns), 50)
 [1] "복면가왕은"    "편견"          "실력"          "취지"          "광화"          "문"            "폭도"
 [8] "들"            "복면"          "익명"          "그늘"          "폭력"          "패악"          "질"
[15] "감추려는거다" "비유"          "때"            "오락프로"      "비교"          "핑계"          "정도껏"
[22] "대"           "씨뱅이들"      "범죄자"        "들이"          "복면쓰고"      "복면가왕"      "폐지"
[29] "유지"         "복면"          "문제"          "범죄자"        "문제라는걸"    "인식"          "것"
[36] "뇌기능"       "장애"          "의심"          "됨"            "조계사"        "좌표"          "수도권"
[43] "포병"         "대"            "전달"          "십자"          "포화"          "김진태"        "의원"
[50] "조계사"
> head(unlist(nouns), 100)
  [1] "복면가왕은"    "편견"          "실력"          "취지"          "광화"          "문"
  [7] "폭도"          "들"            "복면"          "익명"          "그늘"          "폭력"
 [13] "패악"          "질"            "감추려는거다"  "비유"          "때"            "오락프로"
 [19] "비교"          "핑계"          "정도껏"        "대"            "씨뱅이들"      "범죄자"
 [25] "들이"          "복면쓰고"      "복면가왕"      "폐지"          "유지"          "복면"
 [31] "문제"          "범죄자"        "문제라는걸"    "인식"          "것"            "뇌기능"
 [37] "장애"          "의심"          "됨"            "조계사"        "좌표"          "수도권"
 [43] "포병"          "대"            "전달"          "십자"          "포화"          "김진태"
 [49] "의원"          "조계사"        "측에"          "사과"          "뜻"            "점"
 [55] "불교인"        "이유"          "하나"          "국가"          "합법"          "적"
 [61] "공권력"        "무력"          "국회의원"      "이"            "지적"          "이상"
 [67] "한"            "거"            "박종철"        "검안"          "박원순"        "시장"
 [73] "아들"          "병역"          "의혹"          "감정"          "참여"          "출처"
 [79] "YTN"           "TV"            "|"             "네이버"        "뉴스"          "IS"
 [85] "운"            "IS그렇게"      "탱그로"        "나"            "너"            "폭동한다"
 [91] "문죄인"        "호남"          "꿈"            "떠드는데…호남" "꿈"            "뭐"
 [97] "국가전복"      "폭동"          "추억"          "뭔데"
> write(unlist(nouns),"be2_1.txt")
> revised <- read.table("be2_2")
```

## 2. 추출한 단어에서 중복되는 단어 추출 (중복되는 단어가 **중요한 단어**임)

```
> revised <- read.table("be2_1.txt")
> nrow(revised)
[1] 343
> wordcount <- table(revised)
> length(wordcount)
[1] 281
> sort(wordcount, decreasing=T)
revised
        김영삼         대통령             한              나              들              것              대          비판
              5              5              5              4              4              3              3             3
         사람              이              적              전           감정              거             김           꿈
              3              3              3              2              2              2              2             2
         노숙              님           대한             때             뜻           멀쩡             뭐          민국
              2              2              2              2              2              2              2             2
       박종철         범죄자             법           복면             분         빨갱이         사무실          선
              2              2              2              2              2              2              2             2
         세금           세상           아들              자           정치         조계사         지도자          친위
              2              2              2              2              2              2              2             2
       쿠테타       흑백논리     [전두환이냐              |              …       13대총선              2            21
              2              2              1              1              1              1              1             1
           23             IS           IS그렇게             TV            YTN           각하              간        간접세
              1              1              1              1              1              1              1             1
   감정위원장         감정인     감추려는거다         개념도           거기           검안           경제       고문치사
              1              1              1              1              1              1              1             1
         고인             곳           공권력           광화           교육           구분           구호          국가
              1              1              1              1              1              1              1             1
     국가전복         국무총리           국회         국회의원           균형              그           그늘          글로벌
              1              1              1              1              1              1              1             1
         기억           기일           기틀     김영삼대통령         김영삼도       김영삼이냐         김종필        김진태
              1              1              1              1              1              1              1             1
      깊은애도           나중             내              너         네이버         뇌기능           누구          뉴스
              1              1              1              1              1              1              1             1
         능사       단식투쟁           단체             당           대학           돼니             됨          들이
              1              1              1              1              1              1              1             1
떠드는데…호남              로           막론             말           명령           명복         모독죄          목숨
              1              1              1              1              1              1              1             1
         무력           무료           무엇              문           문제     문제라는걸         문죄인          문화
              1              1              1              1              1              1              1             1
         뭔데           미국           박군         박원순           법원           법정         법제정          병신
              1              1              1              1              1              1              1             1
         병역           병원       복면가왕       복면가왕은       복면쓰고           부산         불교인          비교
              1              1              1              1              1              1              1             1
         비난     비난하려거든           비유            빠돌이           사건           사과         사람들          사회
              1              1              1              1              1              1              1             1
```

# 3. 워드 클라우드 형성

```
                1                 1                 1                 1                 1                 1                 1                 1
             포화              폭도              폭동            폭동한다              폭력                 표              핑계               ㅎㅎ
                1                 1                 1                 1                 1                 1                 1                 1
             하나            한국사          한반도가          한홍구에게              합법                 해              호남                 화
                1                 1                 1                 1                 1                 1                 1                 1
             활용
                1
> library(wordcloud)
> library(RColorBrewer)
> pal <- brewer.pal(9,"Set1")
> wordcloud(names(wordcount),freq=wordcount,
+           scale=c(5,1),rot.per=0.25,min.freq=1,
+           random.order=F,random.color=T,colors=pal)
There were 50 or more warnings (use warnings() to see the first 50)
>
```

# 결과발표

⇨ 일간베스트 사용 확정 아이디 추출

a_la_page again73 dailyjeon 021merry boodle_kr
hongzo_lucky7 SJW_gs_hs yrang0099 Jkamber19 gofun11
KimCho5695 Jin1748Jin khsluckyman happylifekjs1
Jonghea2 yunsun2111 joro45 hojjs2014 2161_9237
l54232371 tkhanuli janghun41 KwonCW84 aooaoo72
Hr257 gohcd0904 jks15414 beaaaa11 paklhk7114
Kuwoolsan2 suhbh0905 bosanature Suji1018 TyGall_BG
Meritsun Kaselia_kr psc0823 Hosung__Bot jwb6894
PakSujoung rmafbs1 But_coming audwn4292 bot4auto

⇨ 일간베스트 단어 등급

# QnA