개요

PDF 문서에서 자주 사용되는 단어(키워드)를 분류한 Text mining 결과를 시각적으로 표현하는 것을 word cloud라 한다.

예제 (PDF) 문서

한글개역(신약) 성경 [http://wolfpack.hnu.ac.kr/Big_Data/data/한글성경_신약.pdf] - 한글 성경 전체를 워드 클라우드를 위한 단어 빈도 작성 시 메모리 문제로 작업 종료 오류가 나서 신약 성경만 사용한 분석 결과임

(마1:8)아사는 여호사밧을 낳고 여호사밧은 요람을 낳고 요람은 웃시야를 낳고
(마1:9)웃시야는 요담을 낳고 요담은 아하스를 낳고 아하스는 히스기야를 낳고
(마1:10)히스기야는 므낫세를 낳고 므낫세는 아몬을 낳고 아몬은 요시야를 낳고
(마1:11)바벨론으로 이거할 때에 요시야는 여고냐와 그의 형제를 낳으니라
(마1:12)바벨론으로 이거한 후에 여고냐는 스알디엘을 낳고 스알디엘은 스룹바벨을 낳고
(마1:13)스룹바벨은 아비훗을 낳고 아비훗은 엘리아김을 낳고 엘리아김은 아소르를 낳고

[txt 성경] [http://wolfpack.hnu.ac.kr/Big_Data/data/개역개정.txt]

태초에 하나님이 천지를 창조하시니라
땅이 혼돈하고 공허하며 흑암이 깊음 위에 있고 하나님의 영은 수면 위에 운행하시니라
하나님이 이르시되 빛이 있으라 하시니 빛이 있었고
빛이 하나님이 보시기에 좋았더라 하나님이 빛과 어둠을 나누사
하나님이 빛을 낮이라 부르시고 어둠을 밤이라 부르시니라 저녁이 되고 아침이 되니 이는 첫째 날이니라
하나님이 이르시되 물 가운데에 궁창이 있어 물과 물로 나뉘라 하시고
하나님이 궁창을 만드사 궁창 아래의 물과 궁창 위의 물로 나뉘게 하시니 그대로 되니라
하나님이 궁창을 하늘이라 부르시니라 저녁이 되고 아침이 되니 이는 둘째 날이니라
하나님이 이르시되 천하의 물이 한 곳으로 모이고 뭍이 드러나라 하시니 그대로 되니라
하나님이 뭍을 땅이라 부르시고 모인 물을 바다라 부르시니 하나님이 보시기에 좋았더라
하나님이 이르시되 땅은 풀과 씨 맺는 채소와 각기 종류대로 씨 가진 열매 맺는 나무를 내라 하시니 그대로 되어
땅이 풀과 각기 종류대로 씨 맺는 채소와 각기 종류대로 씨 가진 열매 맺는 나무를 내니 하나님이 보시기에 좋았더라

PDF 데이터 텍스트 데이터로 읽기 : pdf_text() 함수

```
# Read text data from text file.
txt <- readLines("http://wolfpack.hnu.ac.kr/Big_Data/data/개역개정.txt",
encoding="UTF-8")
head(txt,3)
library(pdftools)
txt.pdf <- pdf_text("http://wolfpack.hnu.ac.kr/Big_Data/data/한글성경_신
약.pdf") #PDF -> text
head(txt.pdf,1)
```

```
> head(txt,3)
[1] "태초에 하나님이 천지를 창조하시니라"
[2] "땅이 혼돈하고 공허하며 흑암이 깊음 위에 있고 하나님의 영은 수면 위에 운행하시니라"
[3] "하나님이 이르시되 빛이 있으라 하시니 빛이 있었고"
```

PDF는 페이지 단위, txt 파일은 행(라인) 단위 읽어 들이기 - 본 강의 예제는 PDF 파일 기준임

```
> head(txt.pdf,1)
[1] "(마1:8)아사는 여호사밧을 낳고 여호사밧은 요람을 낳고 요람은 웃시야를 낳고\n(마1:9)웃시야는 요담을 낳고
요담은 아하스를 낳고 아하스는 히스기야를 낳고\n(마1:10)히스기야는 므낫세를 낳고 므낫세는 아몬을 낳고 아몬은 요시
야를 낳고\n(마1:11)바벨론으로 이거할 때에 요시야는 여고냐와 그의 형제를 낳으니라\n(마1:12)바벨론으로 이거한 후
에 여고냐는 스알디엘을 낳고 스알디엘은 스룹바벨을 낳고\n(마1:13)스룹바벨은 아비훗을 낳고 아비훗은 엘리아김을 낳고
 엘리아김은 아소르를 낳고\n(마1:14)아소르는 사독을 낳고 사독은 아킴를 낳고 아킴은 엘리웃을 낳고\n(마1:15)엘리
웃은 엘르아살을 낳고 엘르아살은 맛단을 낳고 맛단은 야곱을 낳고\n(마1:16)야곱은 마리아의 남편 요셉을 낳았으니 마
리아에게서 그리스도라 칭하는 예수가 나시니라\n(마1:17)그런즉 모든 대 수가 아브라함부터 다윗까지 열네 대요 다윗부
```

텍스트 문장 자연어 처리 : KoNLP 라이브러리 – Corpus() 함수. 말뭉치 만들기

```
library(rJava)
library(KoNLP)
# Use Korean dictionary.
useSejongDic().  #세종단어사전 이용

library(tm)
docs <- Corpus(VectorSource(txt.pdf)) #말뭉치 만들기 - VectorSource 문자벡터 말뭉치
inspect(docs[1]) #말뭉치 작성 확인 - 페이지 지정
```

- inspect() 함수 : 만들어진 말뭉치(docs 오브젝트) 1 페이지 결과 보기

```
> inspect(docs[1]) #말뭉치 작성 확인 - 페이지 지정
<<SimpleCorpus>>
Metadata:  corpus specific: 1, document level (indexed): 0
Content:  documents: 1

[1] (마1:8)아사는 여호사밧을 낳고 여호사밧은 요람을 낳고 요람은 웃시야를 낳고\n(마1:9)웃시야는 요담을 낳고 요담은 아하스를 낳고 아하스는 히스기야를 낳고\n(마1:10)히스기야는 므낫세를 낳고 므낫세는 아몬을 낳고 아몬은 요시야를 낳고\n(마1:11)바벨론으로 이거할 때에 요시야는 여고냐와 그의 형제를 낳으니라\n(마1:12)바벨론으로 이거한 후에 여고냐는 스알디엘을 낳고 스알디엘은 스룹바벨을 낳고\n(마1:13)스룹바벨은 아비훗을 낳고 아비훗은
```

특수문자 빈칸(space) 변환 : content_transformer() 함수

```
#tm_map() 함수 활용 : content_transformer() 함수
#특수문자 “/”, “@”, “|”,"\n" 을 space(공백)으로 변환 혹은 대체

toSpace<-content_transformer(function (x , pattern ) gsub(pattern,' ', x))
docs <- tm_map(docs, toSpace, ':')
inspect(docs[1]) #특수문자  제외 확인

toPeriod<-content_transformer(function (x , pattern ) gsub(pattern,'.', x))
docs <- tm_map(docs, toPeriod, '\n')
inspect(docs[1]) #  마침표 변환 확인
```

- toSapce 오브젝트는 content_transformer() 함수를 이용하여 적용되는 특수문자(:)를 ' '(공백)으로 변환 하였고. - toPeriod는 특수문자 '\n'을 특수문자 '.'로 변경하였음

```
> inspect(docs[1]) #특수문자
<<SimpleCorpus>>
Metadata:  corpus specific: 1, document level (indexed): 0
Content:  documents: 1

[1] (마1 8)아사는 여호사밧을 낳고 여호사밧은 요람을 낳고 요람은 웃시야를 낳고\n(마1
9)웃시야는 요담을 낳고 요담은 아하스를 낳고 아하스는 히스기야를 낳고\n(마1 10)히스기야
는 므낫세를 낳고 므낫세는 아몬을 낳고 아몬은 요시야를 낳고\n(마1 11)바벨론으로 이거할
때에 요시야는 여고냐와 그의 형제를 낳으니라\n(마1 12)바벨론으로 이거한 후에 여고냐는
스알디엘을 낳고 스알디엘은 스룹바벨을 낳고\n(마1 13)스룹바벨은 아비훗을 낳고 아비훗은
```

```
> inspect(docs[1]) #마침표 변환 확인
<<SimpleCorpus>>
Metadata:  corpus specific: 1, document level (indexed): 0
Content:  documents: 1

[1] (마1 8)아사는 여호사밧을 낳고 여호사밧은 요람을 낳고 요람은 웃시야를 낳고.(마1 9
)웃시야는 요담을 낳고 요담은 아하스를 낳고 아하스는 히스기야를 낳고.(마1 10)히스기야는
므낫세를 낳고 므낫세는 아몬을 낳고 아몬은 요시야를 낳고.(마1 11)바벨론으로 이거할 때에
요시야는 여고냐와 그의 형제를 낳으니라.(마1 12)바벨론으로 이거한 후에 여고냐는 스알디엘
을 낳고 스알디엘은 스룹바벨을 낳고.(마1 13)스룹바벨은 아비후을 낳고 아비후은 엘리아김을
```

워드 클라우드 불필요 단어 제외 : tm_map() 함수

```
# 텍스트 대문자를 소문자로 변환
docs <- tm_map(docs, content_transformer(tolower))
# 텍스트 숫자를 제거
docs <- tm_map(docs, removeNumbers)
# 텍스트 english 제거
docs <- tm_map(docs, removeWords, stopwords("english"))
# 텍스트 물음표 제거
docs <- tm_map(docs, removePunctuation)
# 텍스트 빈칸(white space) 제거
docs <- tm_map(docs, stripWhitespace)
# Text stemming : 동사 기본형으로 표준화 : do, done, doing -> do로 표준화
docs <- tm_map(docs, stemDocument)

# 사용자 지정 제거 단어 설정 및 문서(텍스트)에서 제거
my_custom_stopwords <- c('불필요단어1','단어2')
docs <- tm_map(docs, removeWords, my_custom_stopwords)
inspect(docs[1]) #불필요 단어 제외 확인

docs <- tm_map(docs, removeWords, my_custom_stopwords)
inspect(docs[2]) #불필요 단어 제외 확인
```

```
> inspect(docs[1]) #불필요 단어 제외 확인
<<SimpleCorpus>>
Metadata:  corpus specific: 1, document level (indexed): 0
Content:  documents: 1

[1] 마 아사는 여호사밧을 낳고 여호사밧은 요람을 낳고 요람은 웃시야를 낳고마 웃시야는 요
담을 낳고 요담은 아하스를 낳고 아하스는 히스기야를 낳고마 히스기야는 므낫세를 낳고 므낫세
는 아몬을 낳고 아몬은 요시야를 낳고마 바벨론으로 이거할 때에 요시야는 여고냐와 그의 형제
를 낳으니라마 바벨론으로 이거한 후에 여고냐는 스알디엘을 낳고 스알디엘은 스룹바벨을 낳고
마 스룹바벨은 아비훗을 낳고 아비훗은 엘리아김을 낳고 엘리아김은 아소르를 낳고마 아소르는
사독을 낳고 사독은 아킴를 낳고 아킴은 엘리우을 낳고마 엘리우은 엘르아살을 낳고 엘르아살은
```

문장 단어 카운트 및 결과 보기

```
#문서번호와 단어간의 사용여부 또는 빈도 카운트
dtm <- TermDocumentMatrix(docs)
#생성된 document term matrix에서 1-10번 빈도 높은단어의
#100번째에서 150번째 문서의 분포를 확인
#Sparsity          : 공백 수 (%로 표현)
inspect(dtm[1:10,100:150])

findFreqTerms(dtm, lowfreq = 500) #500회 이상 사용된 단어 출력
findFreqTerms(dtm, 500,600) #500에서 600회 사이로 사용된 단어 출력
```

- TermDocumentMatrix() 문장 내 단어 빈도 행렬 표현
- 빈도가 높은 단어들의 문서내 분포

```
> inspect(dtm[1:10,100:150])
<<TermDocumentMatrix (terms: 10, documents: 51)>>
Non-/sparse entries: 165/345
Sparsity           : 68%
Maximal term length: 5
Weighting          : term frequency (tf)
Sample             :
             Docs
Terms         102 104 107 108 111 121 122 125 128 136
  가다가        0   0   0   0   0   0   0   0   0   0
  가라사대마    0   0   0   0   0   0   0   0   0   0
  가로되        7   6   5   5   6   7   8   6   3   9
```

```
> findFreqTerms(dtm, lowfreq = 500) #500회 이상 사용된 단어 출력
 [1] "가로되"    "것을"      "그"        "내"        "또"
 [6] "모든"      "예수께서"  "이"        "이는"      "내가"
[11] "너희가"
> findFreqTerms(dtm, 500,600) #500에서 600회 사이로 사용된 단어 출력
[1] "가로되" "내"      "또"      "이는"    "너희가"
```

문장 내 함께 사용된 단어들의 연관분석

#findAssocs - 설정 단어(work)와 연관성이 (상관계수) 기준 이상 높은 단어
findAssocs(dtm, terms = "죄", corlimit = 0.4)
findAssocs(dtm, terms = "요한", corlimit = 0.75)

- 단어 '죄' 와 함께 쓰인 빈도가 0.4 이상인 단어 : 사한다

```
> findAssocs(dtm, terms = "죄", corlimit = 0.4)
$죄
    사함을     사하는     세관에   쓸데없고     의원이
      0.55       0.41       0.41       0.41       0.41
자에게라야 중풍병자를
      0.41       0.41
```

- 단어 '요한' 와 함께 쓰인 빈도가 0.4 이상인 단어 : 살아났다, 걸으니 - 기적

```
> findAssocs(dtm, terms = "요한", corlimit = 0.8)
$요한
  가두었으니막 가운데서살아났도다         거기서는
          0.82             0.82             0.82
        걸으니           것은막       경계하시고
          0.82             0.82             0.82
          고막       고치더라막           고치실
          0.82             0.82             0.82
    곳에서든지           권능도       놀라거늘막
          0.82             0.82             0.82
    느끼면서도             달게         달리다굼
          0.82             0.82             0.82
    드러난지라       들어가거든       들어가사막
          0.82             0.82             0.82
```

품사 단어 말뭉치 만들기

```
#품사 단어 선택 후 자연어 처리 방법
txt.cnt<-SimplePos22(docs) #하위분류 22 품사 단어 추출
head(txt.cnt,1)
# 품사 테그 중 필요한 품사만 선택
# 영어, 숫자 모두 필요하면 [ㄱ-ㅎ|가-힣|a-z|A-Z|0-9].
# 보통명사 NC, 고유명사 NP 둘 다 필요하면 [NC|NP]
library(stringr)
txt.cnt0<-str_match_all(txt.cnt,'([ㄱ-ㅎ|가-힣]+)/[NC]')
head(txt.cnt0,1)
txt.cnt0<-unlist(txt.cnt0) #단어를 벡터 포맷으로 리스트 (flatten list)
head(txt.cnt0,5)
txt.cnt00<-Filter(function(x){nchar(x)>3},txt.cnt0) #단어 2글자(/N 포함 4) 이상
head(txt.cnt00,5)
```

docs 데이터는 페이지 5의 tm_map() 함수가 적용된 것임

```
[[1]]$낳고
[1] "낳/PV+고/EC"

[[1]]$므낫세는
[1] "므낫세는/NC"

[[1]]$아몬을
[1] "아몬을/NC"
```

읽어들인 문서 모든 단어에 대하여 품사셋 붙음

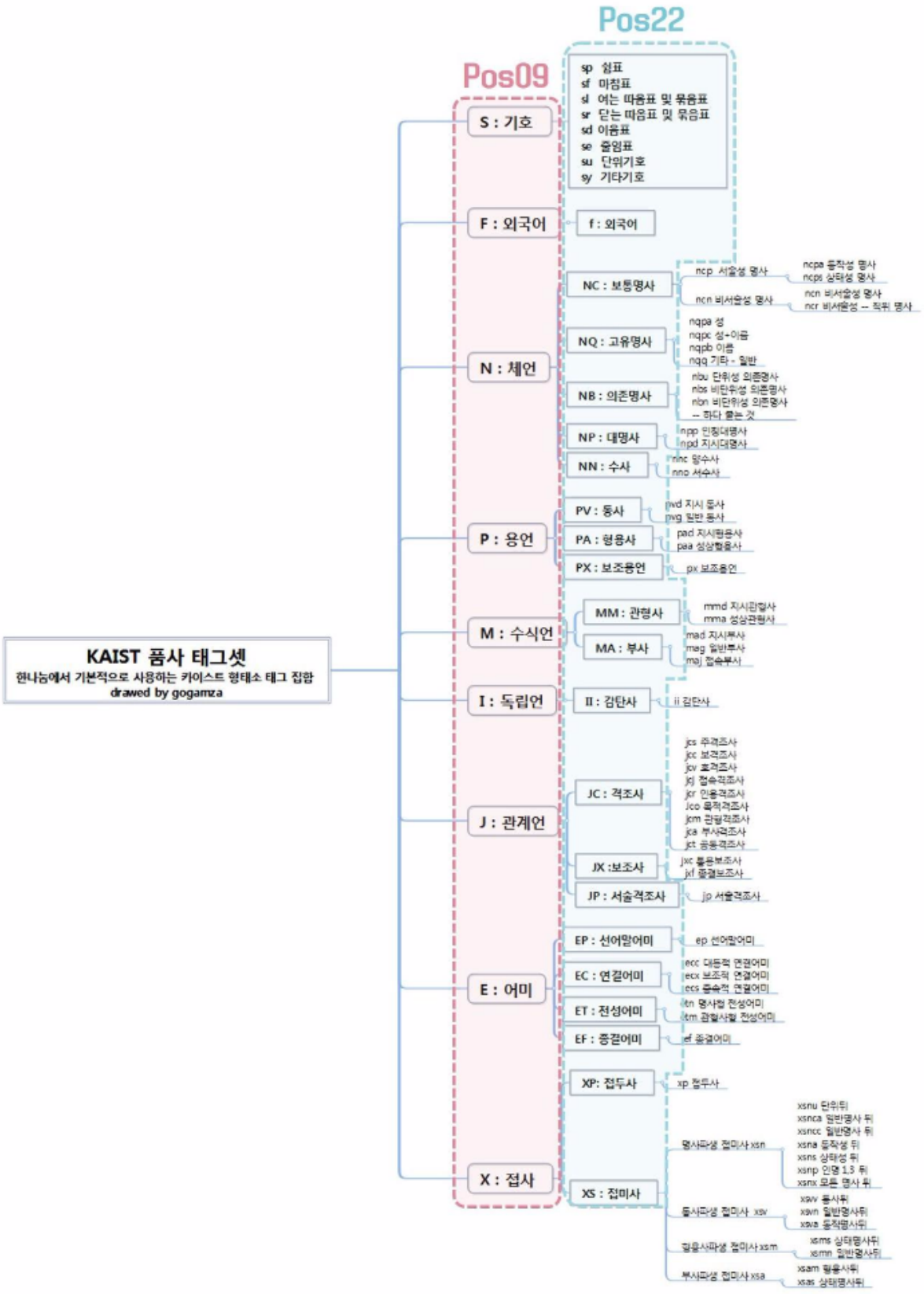
명사(NC) 품사 단어만 선택

```
> head(txt.cnt0,1)
[[1]]
      [,1]              [,2]
 [1,] "마/N"            "마"
 [2,] "아사/N"          "아사"
 [3,] "여호사밧을/N"    "여호사밧을"
 [4,] "여호사밧은/N"    "여호사밧은"
 [5,] "요람/N"          "요람"
 [6,] "요람/N"          "요람"
      ...시야를/N"      "웃시야를"
```

단어 벡터 포맷 리스트, 글자 2개 이상 단어

```
> txt.cnt0<-unlist(txt.cnt0) #단어를 벡터 포맷으로 리스트 (flatten list)
> head(txt.cnt0,5)
[1] "마/N"          "아사/N"        "여호사밧을/N" "여호사밧은/N" "요람/N"
> txt.cnt00<-Filter(function(x){nchar(x)>3},txt.cnt0) #단어 2글자(/N 포함 4) 이상
> head(txt.cnt00,5)
[1] "아사/N"        "여호사밧을/N" "여호사밧은/N" "요람/N"        "요람/N"
```

카이스트 품사셋


KAIST 품사 태그셋
한나눔에서 기본적으로 사용하는 카이스트 형태소 태그 집합
drawed by gogamza
Pos09
Pos22
S : 기호
sp 쉼표
sf 마침표
sl 여는 따옴표 및 묶음표
sr 닫는 따옴표 및 묶음표
sd 이음표
se 줄임표
su 단위기호
sy 기타기호
F : 외국어
f : 외국어
N : 체언
NC : 보통명사
ncp 서술성 명사
ncpa 동작성 명사
ncps 상태성 명사
ncn 비서술성 명사
ncn 비서술성 명사
ncr 비서술성 -- 직위 명사
NQ : 고유명사
nqpa 성
nqpc 성+이름
nqpb 이름
nqq 기타 - 일반
NB : 의존명사
nbu 단위성 의존명사
nbs 비단위성 의존명사
nbn 비단위성 의존명사
-- 하다 붙는 것
NP : 대명사
npp 인칭대명사
npd 지시대명사
NN : 수사
nnc 양수사
nno 서수사
P : 용언
PV : 동사
pvd 지시 동사
pvg 일반 동사
PA : 형용사
pad 지시형용사
paa 성상형용사
PX : 보조용언
px 보조용언
M : 수식언
MM : 관형사
mmd 지시관형사
mma 성상관형사
MA : 부사
mad 지시부사
mag 일반부사
maj 접속부사
I : 독립언
II : 감탄사
ii 감탄사
J : 관계언
JC : 격조사
jcs 주격조사
jcc 보격조사
jcv 호격조사
jcj 접속격조사
jcr 인용격조사
jco 목적격조사
jcm 관형격조사
jca 부사격조사
jct 공동격조사
JX :보조사
jxc 통용보조사
jxf 종결보조사
JP : 서술격조사
jp 서술격조사
E : 어미
EP : 선어말어미
ep 선어말어미
EC : 연결어미
ecc 대등적 연결어미
ecx 보조적 연결어미
ecs 종속적 연결어미
ET : 전성어미
etn 명사형 전성어미
etm 관형사형 전성어미
EF : 종결어미
ef 종결어미
X : 접사
XP: 접두사
xp 접두사
XS : 접미사
명사파생 접미사 xsn
xsnu 단위뒤
xsnca 일반명사 뒤
xsncc 일반명사 뒤
xsna 동작성 뒤
xsns 상태성 뒤
xsnp 인명 1,3 뒤
xsnx 모든 명사 뒤
동사파생 접미사 xsv
xsvv 동사뒤
xsvn 일반명사뒤
xsva 동작명사뒤
형용사파생 접미사 xsm
xsms 상태명사뒤
xsmn 일반명사뒤
부사파생 접미사 xsa
xsam 형용사뒤
xsas 상태명사뒤

말뭉치 만들기 - 단어 빈도 계산 Corpus() 함수, TermDocumentMatrix()함수

```
library(tm)
words <- Corpus(VectorSource(txt.cnt0)) #말뭉치 만들기
#워드 클라우드 1차 분석 후 제외 단어 추가
stopword <- c(stopwords('english'),'것','고','이','서','요','그','하','나','자',
              '말','내','중','저','지','한','가라사대','내가','우리','저희',
              '너','자기','두','또','안','줄','바','리','라','앞','수','후',
              '하나','대','만')
words <- tm_map(words,removeWords,stopword)
dtm <- TermDocumentMatrix(words)
findFreqTerms(dtm, lowfreq = 100)
dtm.mat <- as.matrix(dtm)
v <- sort(rowSums(dtm.mat),decreasing=TRUE)
word.freq <- data.frame(word = names(v),freq=v)

#상위 빈도 20개 빈도표 출력
head(word.freq, 20)
```

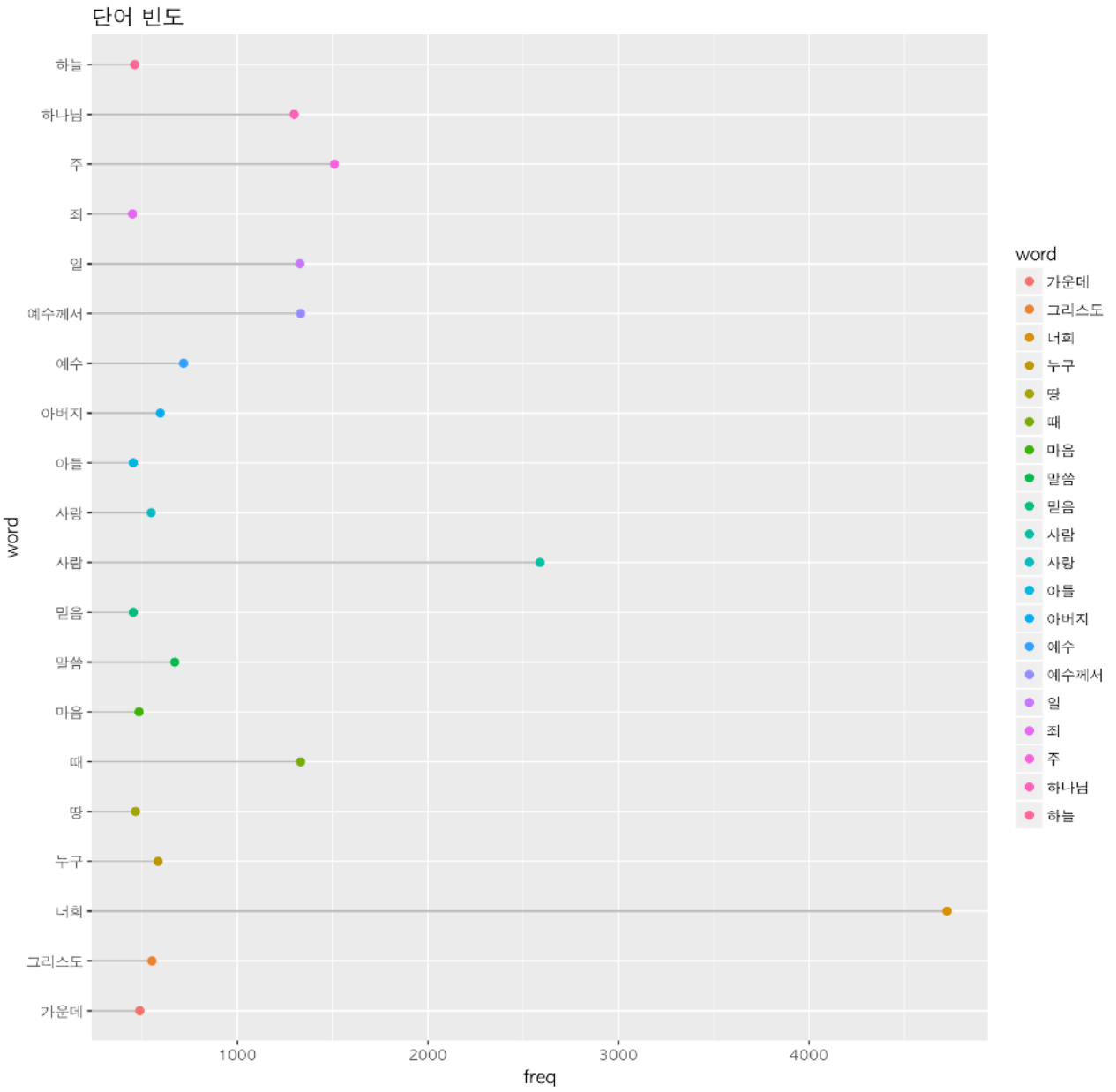
```
> #상위 빈도 20개 빈도표 출력
> head(word.freq, 20)
              word freq
너희          너희 4724
사람          사람 2588
주              주 1510
때              때 1332
예수께서 예수께서 1332
일              일 1328
하나님      하나님 1298
예수          예수  718
말씀          말씀  672
```

워드 cloud 표현

```
#Generate the Word cloud
#word cloud 라이브러리
library(SnowballC)
library(RColorBrewer)
library(wordcloud)

wordcloud(words = word.freq$word, freq = word.freq$freq,
          min.freq = 100, random.order=F, rot.per=0.35, colors=brewer.pal(8,
'Dark2'), family="AppleGothic")

#Frequency Table
require(ggplot2)
ggplot(word.freq[1:20,],aes(x=freq,y=word,col=word)) +
  geom_segment(aes(yend=word),xend=0,color='grey') +
  geom_point(size=2,aes(color=word)) +
  ggtitle('단어 빈도')
```


단어 빈도
하늘
하나님
주
죄
일
예수께서
예수
아버지
아들
사랑
사람
믿음
말씀
마음
때
땅
누구
너희
그리스도
가운데
word
1000
2000
3000
4000
freq
word
가운데
그리스도
너희
누구
땅
때
마음
말씀
믿음
사람
사랑
아들
아버지
예수
예수께서
일
죄
주
하나님
하늘