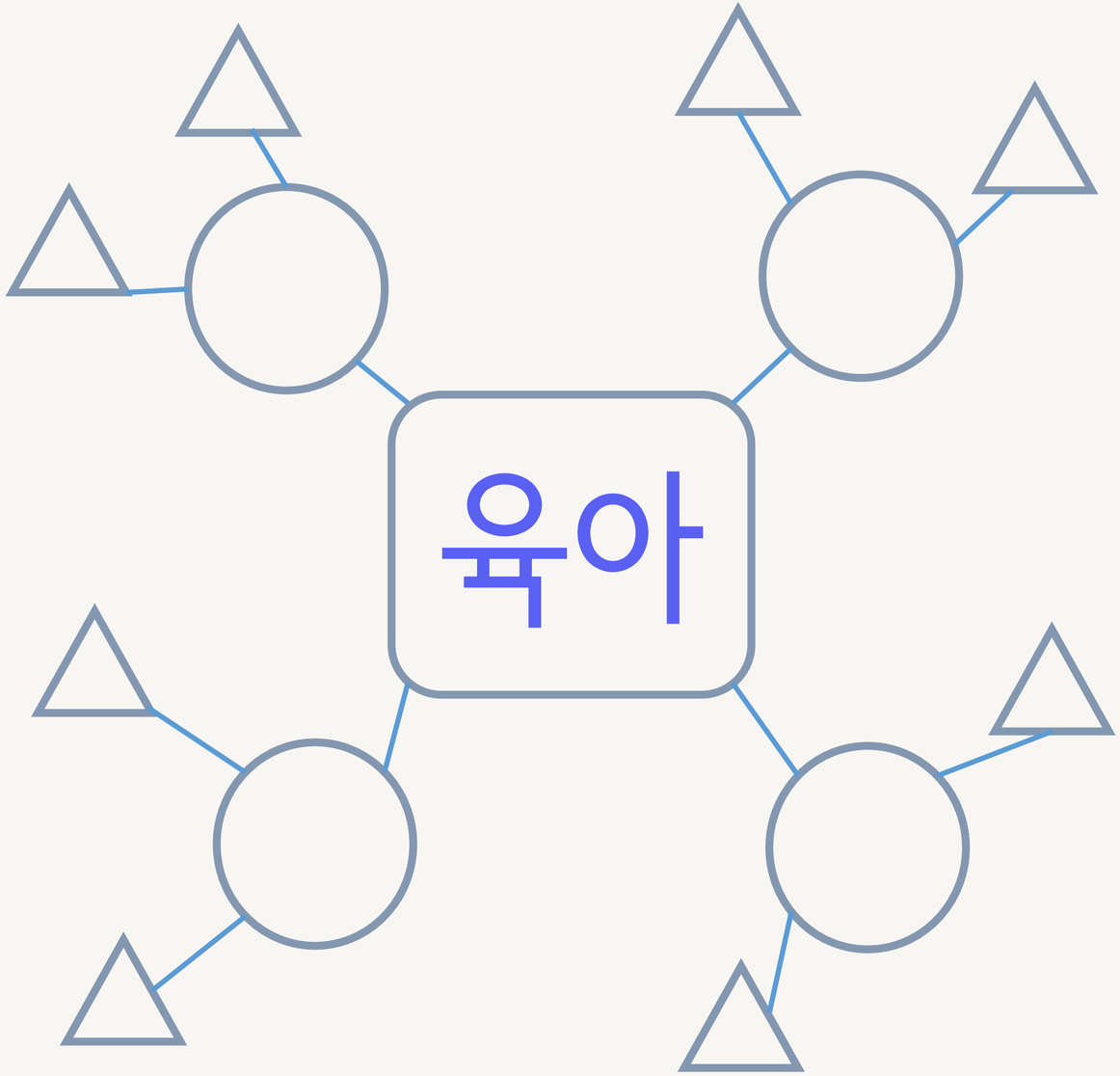
Web Crawling & Text Mining



최재명

연세 IT 미래 교육

목 차

Chapter 1.
관심 Keyword

[육아]

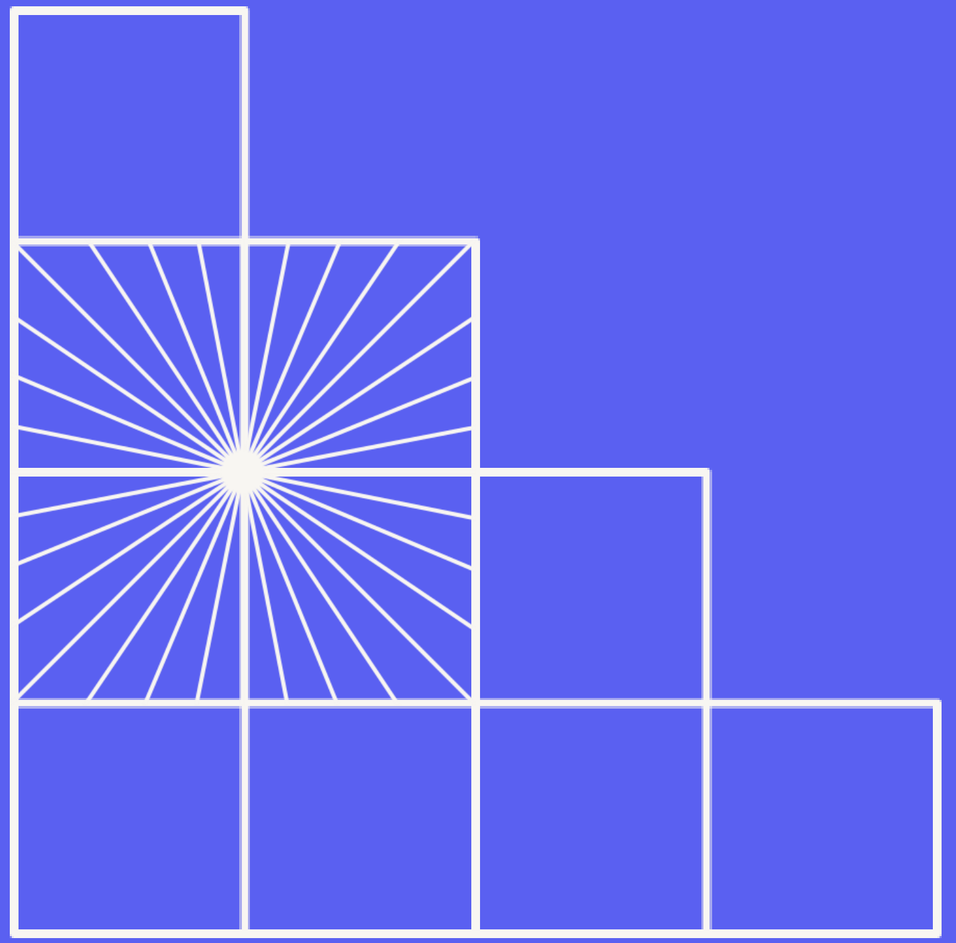
Keyword 선정 이유

- 두 아이의 아버지
- 좋은 아버지
- 삶의 이유

Chapter 2.

Web Site검색

[NAVER, DAUM]



어디서?

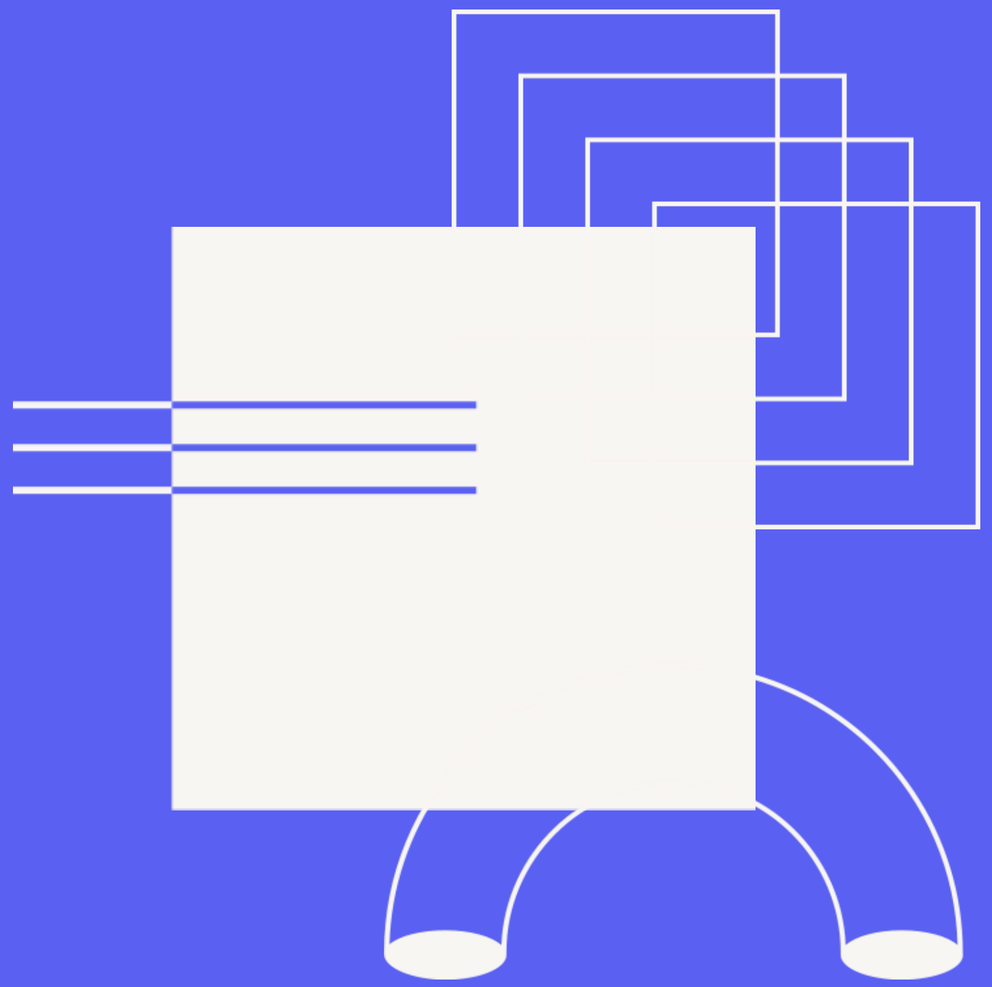
한국인이 가장 많이 사용하는 검색엔
진인 [NAVER], [DAUM] 선정

무엇을?

- 블로그와 카페 검색을 통해 대중 관련 관심사 파악
- 뉴스 검색을 통한 관련 이슈 파악

어떻게?

- R 프로그램을 활용한 연관 Keyword 추출
- 추가 연관 검색 실행

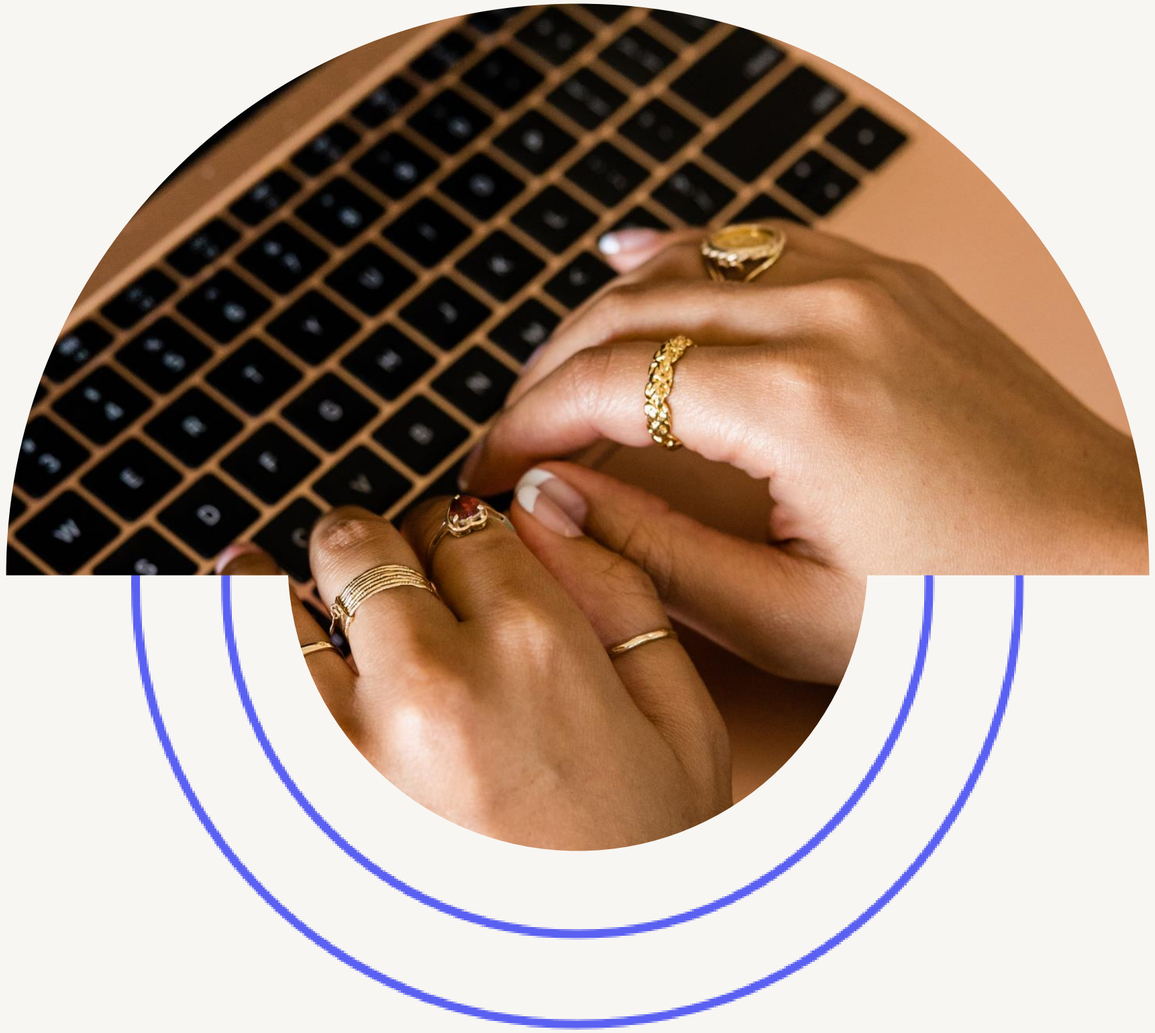


Chapter 3.

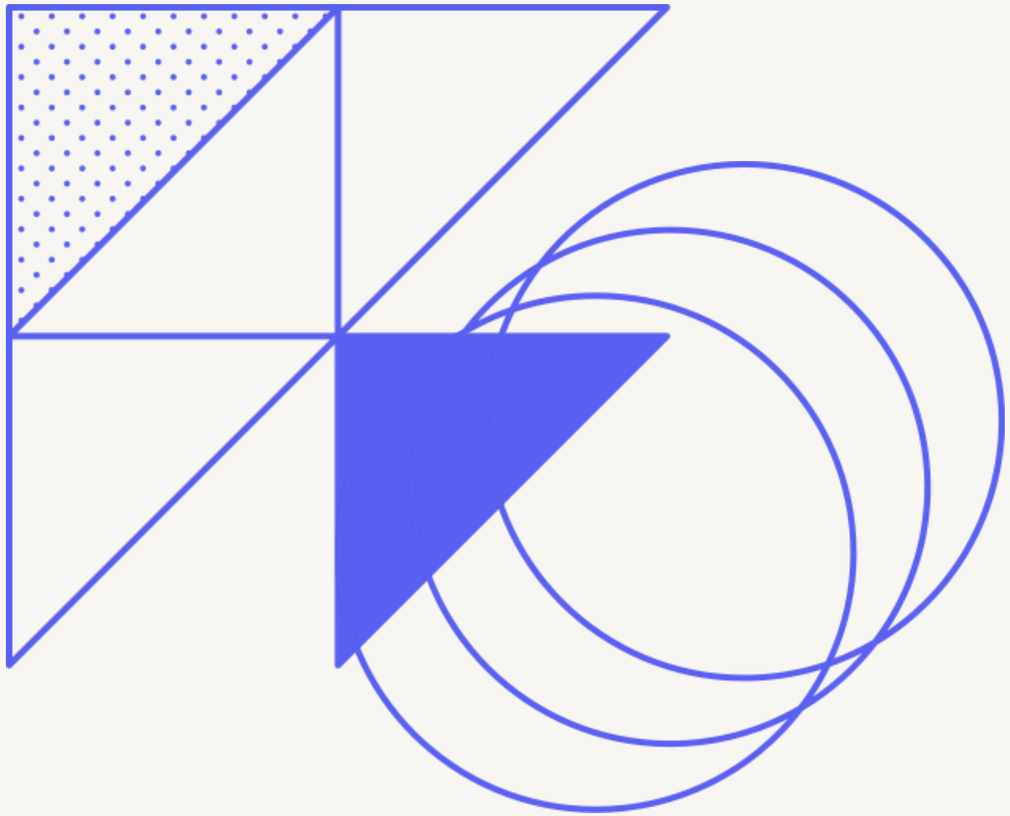
R code 소개

Package 설명

- library(rvest)
  - html 관련 함수
- library(stringr)
  - str_sub( ) : 원하는 부분을 추출하는 함수
- library(KoNLP)
  - 명사 추출
- library(wordcloud)
- library(wordcloud2)
  - 시각화



R code

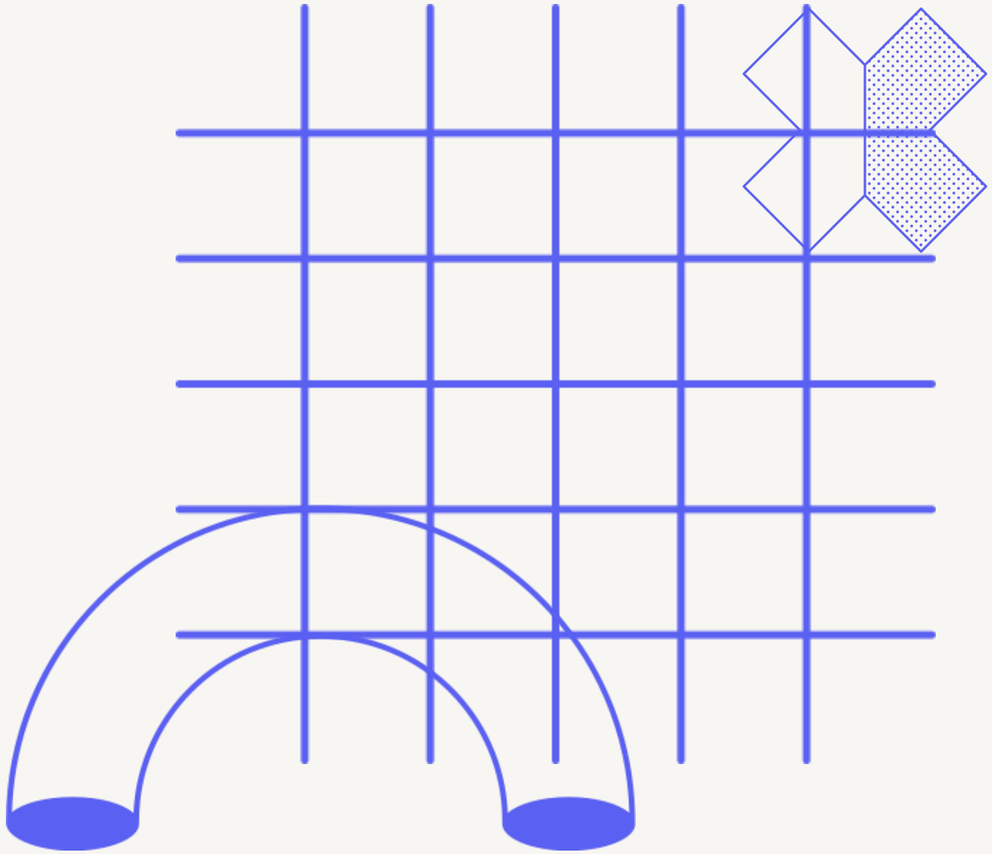


```
#주머니 만들기
title=c()
body=c()

#URL 따오기
url='https://search.daum.net/search?w=fusion&DA=PGD&enc=utf8&q=%EC%9C%A
1%EC%95%84&p=‘

# Web Searching
for(i in 1:100){
  turl=paste0(url,i)
  t_css=".tit_main“
  b_css="#twcColl .desc“
  hdoc=read_html(turl)
  t_node=html_nodes(hdoc,t_css)
  b_node=html_nodes(hdoc,b_css)
  title_part=html_text(t_node)
  b_part=html_text(b_node)
  body_part=gsub("₩n","",b_part)
  body_part=str_trim(body_part,side="both")
  url_part=html_attr(t_node,"href")
  title=c(title,title_part)
  body=c(body,body_part) }
blog=data.frame(title,body)
head(blog)
```

R code



단어 선택 정리

```
a=sapply(news,extractNoun,USE.NAMES=F)
a=unlist(a)
a=gsub("육아","",a)
a=gsub("[^ㄱ-힣]","",a
)a=gsub("들","",a)
count=Filter(function(x){nchar(x)>=2},a)
word=table(count
)b=head(sort(word,decreasing = T),20)
print(b)
```

막대그래프

```
c=barplot(b,col=rainbow(30),ylim=c(0,450),las=2)text(c,b,paste0(b,"개
"),pos=3,col=2,cex=1)
```

WordCloud

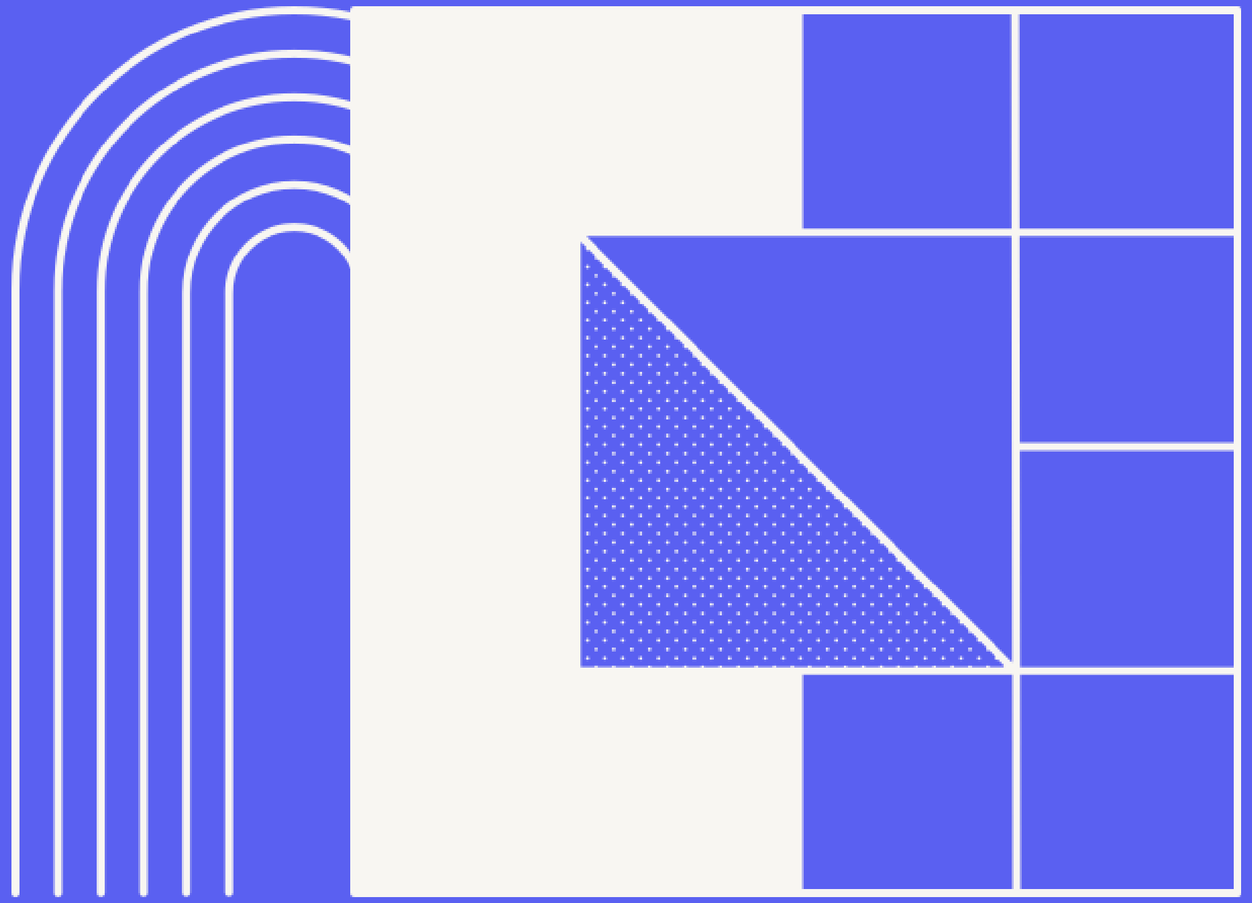
```
palate=brewer.pal(9,"Set1")
wordcloud(names(b),
          freq = b,
          scale = c(6,0.5),
          rot.per = 0.25,
          min.freq = 2,
          random.order = F,
          colors = palate)
```

WordCloud2

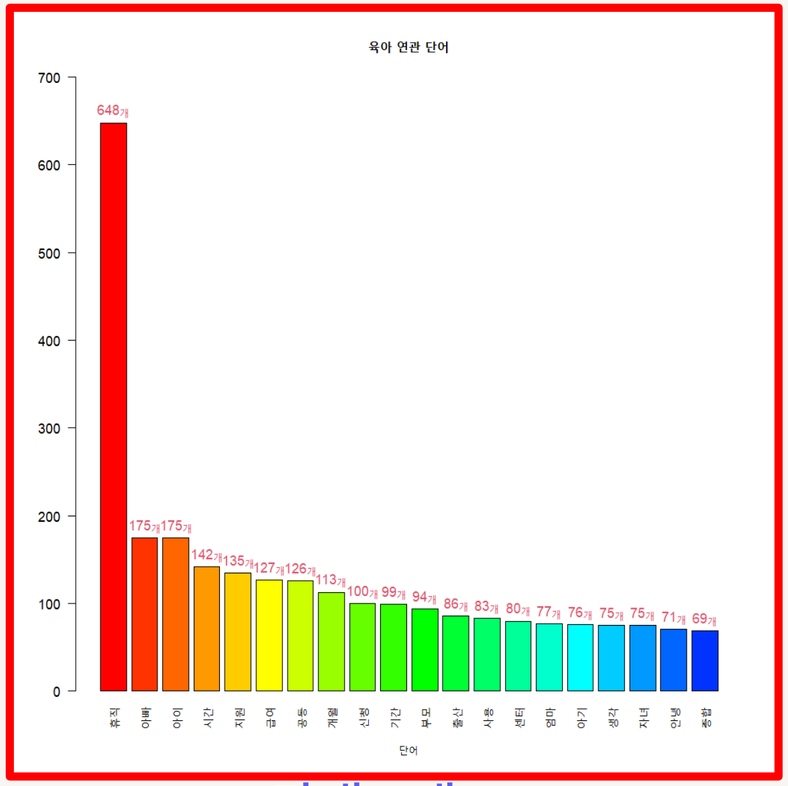
```
wordcloud2(data=b,size=0.5)
```

Chapter 4.

시각화



KeyWord 육아 [in Daum 통합웹]

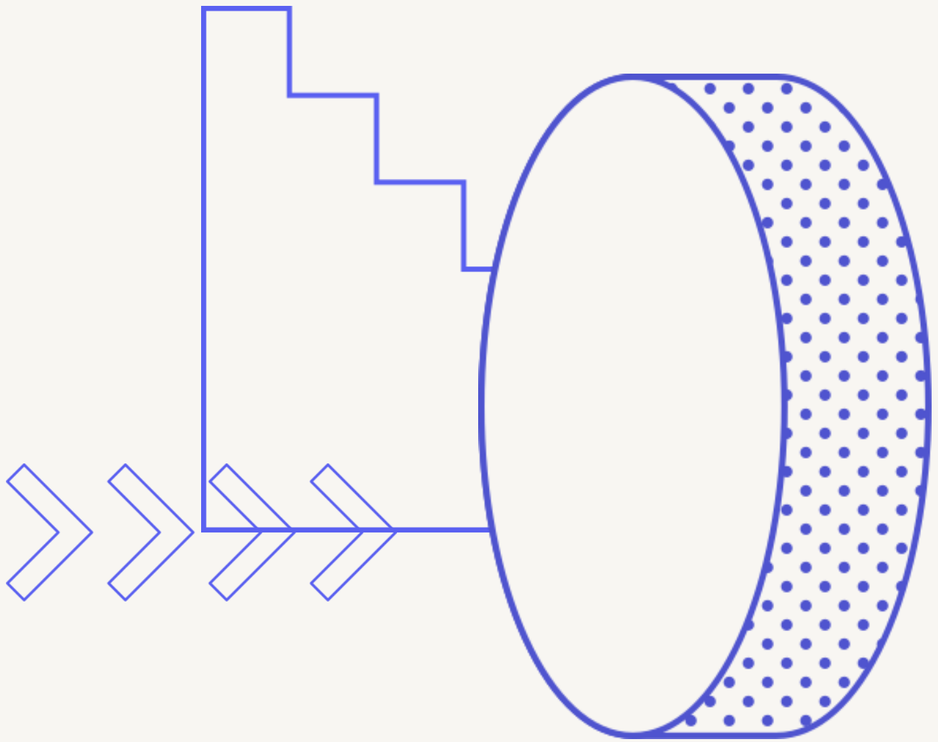


막대그래프

WordCloud

WordCloud2

최다 빈도 단어 : [휴직]

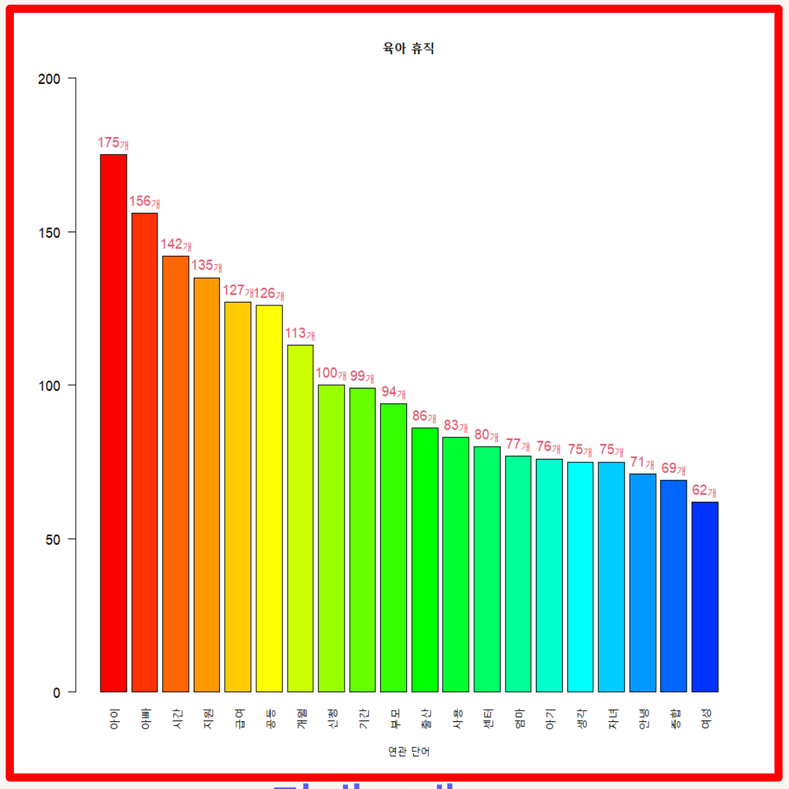


650개

16%

- 추출 단어 내 [육아] 키워드 삭제로 인해 [육아 휴직]으로 추정
- 부모들의 육아에서 가장 관심 있는 것은 육아휴직

KeyWord 육아휴직 [in NAVER 뉴스]



막대그래프

WordCloud

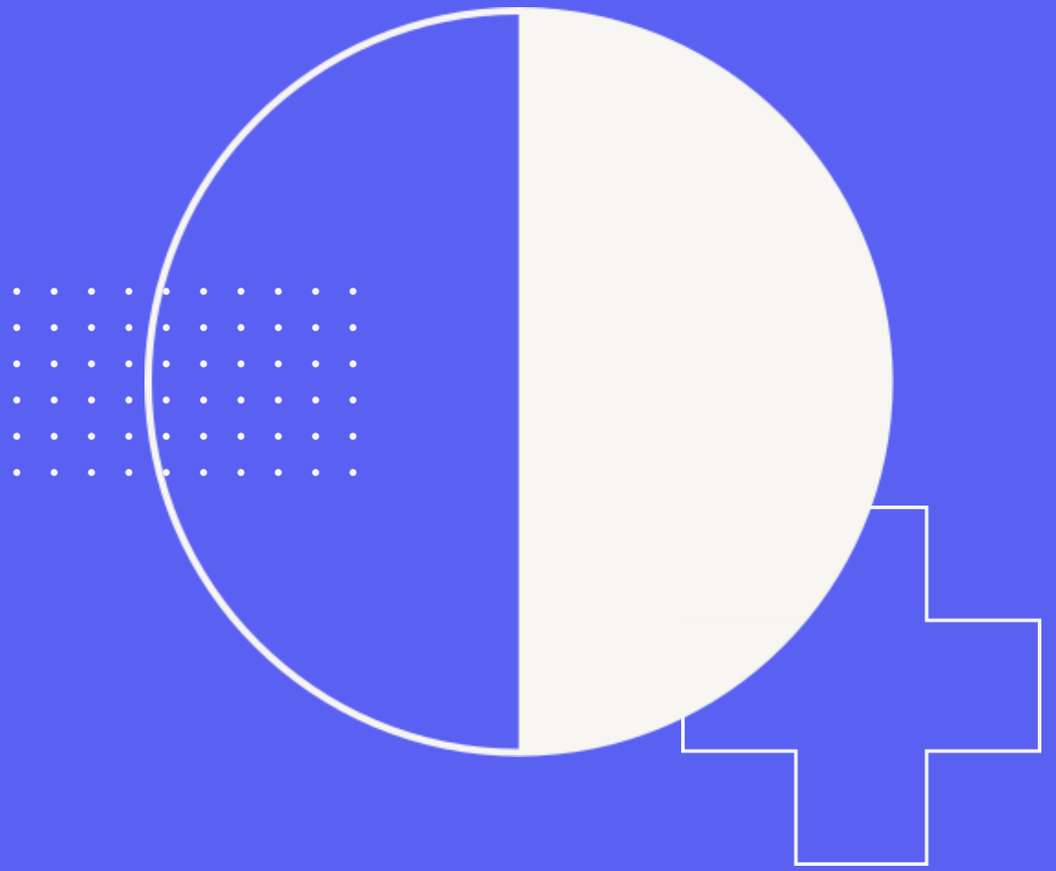
WordCloud2

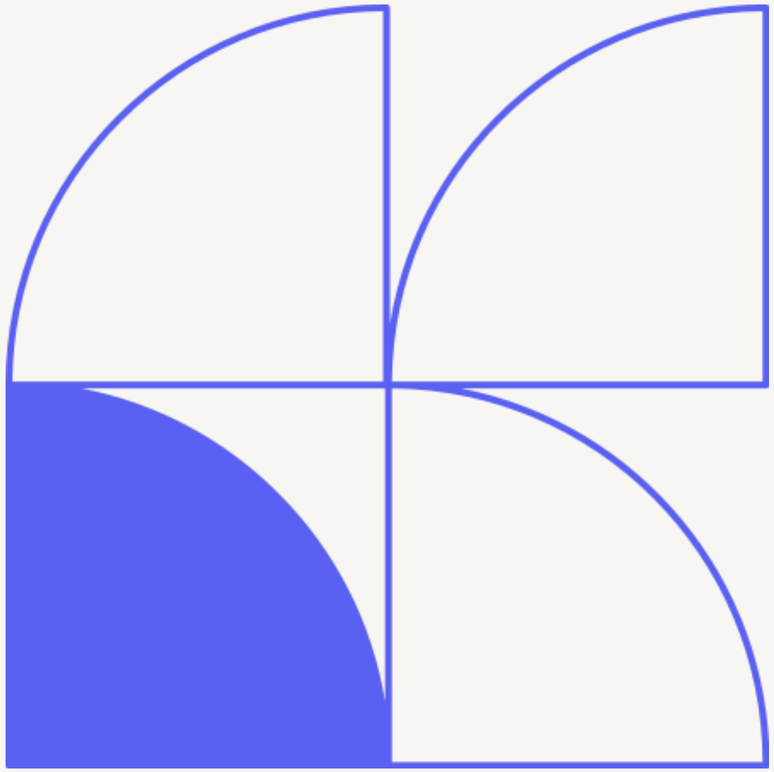
빈도 단어 : [아빠], [공동], [급여], [지원]

- 아빠의 육아참여도가 높아짐에 따라 아빠의 육아휴직도 이슈화
- 육아휴직 후 급여에 많은 관심
- 부모 공동 육아가 관건
- 육아휴직 후 생활에 대한 지원에도 많은 관심

육아는 부모 모두의 해야할 일 입니다.

어떤 부모로 기억되고 싶으신가요?





감사합니다!