질병에 관하여

배상일

2023-02-13

코로나로 인해 질병에 대한 경각심을 가지게 된 현재, 주로 어떤 질병이 건강에 큰 영향을 주는지 기사를 통해 조사해보았습니다.

작업을 위한 라이브러리 세팅

```
library(ggplot2)
library(wordcloud)
```

```
## 필요한 패키지를 로딩중입니다: RColorBrewer
```

```
library(wordcloud2)
library(KoNLP)
```

```
## Checking user defined dictionary!
```

```
library(rvest)
library(stringr)
```

크롤링을 하기 위한 주머니 생성

```
title <- c()
#press <- c()
time <- c()
body <- c()
#url <- c()
```

크롤링

```
p_url <- "https://search.naver.com/search.naver?where=news&sm=tab_pge&query=%EC%9A%B0%EB%A6%AC%EB%82%98%EB%9D%BC%20%EC%A7%88%EB%B3%91%20%EC%82%AC%EB%A7%9D%EB%A5%A0&sort=0&photo=0&field=0&pd=0&ds=&de=&cluster_rank=30&mynews=0&office_type=0&office_section_code=0&news_office_checked=&nso=so:r,p:all,a:all&start="
for(i in 1:50){
  turl <- paste0(p_url,10*i+1)
  t_css <- ".news_tit"
  #p_css <- ".info.press"
  b_css <- ".dsc_txt_wrap"

  hdoc <- read_html(turl)
  t_node <- html_nodes(hdoc,t_css)
  #p_node <- html_nodes(hdoc,p_css)
  b_node <- html_nodes(hdoc,b_css)

  title_part <- html_text(t_node)
  #p_part <- html_text(p_node)
  #time_part <- str_sub(pt_part,-5)
  #press_part <- str_sub(pt_part,end = -9)
  b_part <- html_text(b_node)
  body_part <- gsub("₩n",'',b_part)
  body_part <- str_trim(body_part,side = "both")

  title <- c(title,title_part)
  body <- c(body,body_part)


}
news <- data.frame(title,body)
```

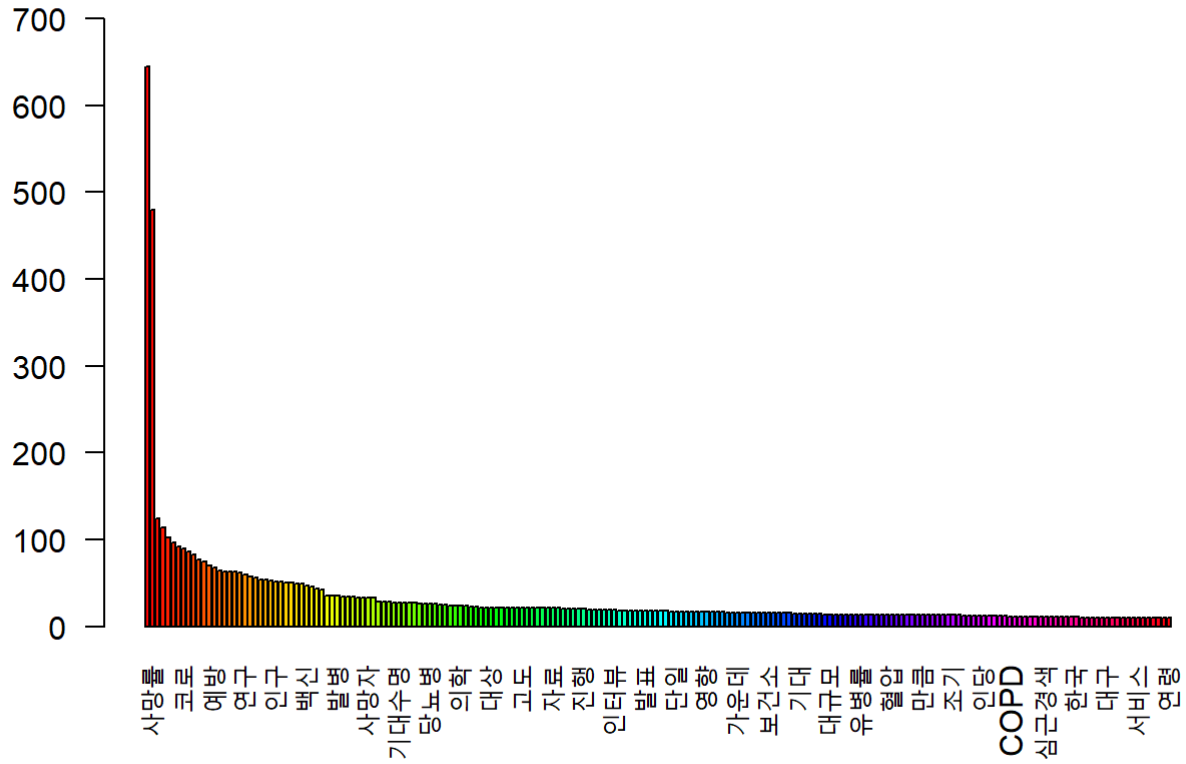
크롤링 추출 자료 전처리

추출한 무작위 단어들을 주제에 맞게 제거하고 추출 기준을 설정 했습니다.

```
txt <- sapply(news,extractNoun,USE.NAMES = F)
txt <- unlist(txt)
txt <- gsub('[http|kr|www|news|https|html|php|co|10|idxno|view|우리나라|aV|ay22|23|28|4.|95]','',txt)
txt <- gsub('[^ㄱ-힣A-Za-z]','',txt)

count <- Filter(function(x){nchar(x)>=2 & nchar(x)<=5},txt)
word <- table(count)

txt2 <- gsub('[질병|사망률|코로|때문|청장|필요|66|76|발생|환자|OECD|이상|국가|건강|치료|교수|인터뷰|연구|=&=]','',txt)
txt2 <- gsub('[^ㄱ-힣A-Za-z]','',txt2)

count1 <- Filter(function(x){nchar(x)>=2 & nchar(x)<=5},txt2)
word1 <- table(count1)
```

막대 그래프 시각화

```
kk <- head(sort(word,decreasing=T),200)
tt <- barplot(kk,col = rainbow(200),ylim = c(0,700),las=2)


kk1 <- head(sort(word1,decreasing=T),200)
tt1 <- barplot(kk,col = rainbow(200),ylim = c(0,700),las=2)
```



Wordcloud 시각화

질병 사망률에 대한 최근 기사들의 연관 검색어를 추출한 결과입니다.

```
#display.brewer.all()
palate <- brewer.pal(9,"Set1")
#palate
wordcloud(
  names(kk),
  freq = kk,
  scale = c(5,0.8),
  #rot.per = 0.25,
  min.freq = 2,
  random.order = F,
  random.color = T,
  colors = palate
)
```

워드클라우드2

연관 검색어 중 기사에서 언급이 많이 된 질병들을 추출한 결과입니다.

```
wordcloud2(data = kk1,
           size = 0.3,
           shape = 'pentagon')
```


뇌졸중
결핵
예방
비만
감염
감소
폐암
위험
접종
간암
만성
백신
검진
세계
부담
확진
평균
진단
통계
결과
당뇨
지속
여성
기준
관련
혈관
중풍
기대명
일본뇌염
지난해
영아
정은경
하기
보험
입원
위암
대비
후진
의학
고혈압
트럼프
극복
심각
바이러스
관본부
지역
단일
시작
의원
전체
진행
기록
뇌혈관
보소
운데
폐혈증
정도
시행
일반
고령
임신부
개선
대한
부산
분석
고도
회원
섭취
조기
대규모
유방암
무엇
지표
커피
방법
효과
폭염
영향
캠페