통계계산프로그래밍 과제 2

2019131238 정예린

1. 스티브잡스의 스탠포드 졸업식 연설문(한글)에 대해 워드클라우드를 작성하시오.

```
#텍스트파일 불러오기
> raw_sj<-readLines("C:/Users/Chung Yelin/Desktop/sj.txt", encoding="UTF-8")
#전처리
> library(tm)
> library(SnowballC)
> sj<-Corpus(VectorSource(raw_sj)) #말뭉치 생성
> inspect(sj)
> sj<-tm_map(sj, removeNumbers) #숫자 제거
> sj<-tm_map(sj, removePunctuation) #특수문자 제거
> sj<-tm_map(sj, stripWhitespace) #공백문자 제거
#불용어 제거
> library(stringr)
> sj_noun<-sapply(sj, extractNoun, USE.NAMES=F)
> sj_unlist<-unlist(sj_noun)
> head(sj_unlist)
[1] "오늘" "세계" "최고" "대학" "중"   "한"
> sj_unlist<-gsub("그것","",sj_unlist)
> sj_unlist<-gsub("가지","",sj_unlist)
> sj_unlist<-gsub("들이","",sj_unlist)
> sj_unlist<-gsub("여러분","",sj_unlist)
> sj_unlist<-gsub("여러분들은","",sj_unlist)
> sj_unlist<-gsub("여러분들도","",sj_unlist)
> sj_unlist<-gsub("우리","",sj_unlist)
> sj_unlist<-gsub("당신","",sj_unlist)
> sj_unlist<-gsub("개월","",sj_unlist)
> sj_unlist<-gsub("오늘","",sj_unlist)
> sj_unlist<-gsub("그때","",sj_unlist)
> sj_unlist<-gsub("뭔가","",sj_unlist)
> sj_unlist<-gsub("것들은","",sj_unlist)
> sj_unlist<-gsub("아무","",sj_unlist)
> sj_unlist<-gsub("않았다","",sj_unlist)
> sj_unlist<-gsub("있습니","",sj_unlist)
```

```
> sj_unlist<-gsub("때문","",sj_unlist)
> sj_unlist<-gsub("겁니","",sj_unlist)
> sj_unlist<-gsub("Unwed","",sj_unlist)
> sj_unlist<-Filter(function(x){nchar(x)>=2}, sj_unlist) #두 글자 이상의 단어만 출력
> wordcount<-table(sj_unlist)
> wordcount_top<-head(sort(wordcount, decreasing=T),10)
> wordcount_top
sj_unlist
대학 인생 애플 사랑 입양 생각 시작 졸업 죽음 Stay
  16   11    9    8    8    7    7    7    7    5
#워드클라우드 생성
> library(wordcloud)
> library(wordcloud2)
> library(RColorBrewer)
> table<-table(sj_unlist)
> table.2<-sort(table, decreasing=TRUE)
> table.2a<-table.2[table.2>1]
> windows(width=4, height=4); par(mar=c(2,2,2,2))
> set.seed(12345)
> wordcloud(words=names(table), freq=table, scale=c(5,0.5), min.freq=1,
colors=rainbow(10), random.color=FALSE, random.order=FALSE, rot.per=0.25)
> wordcloud2(table.2a)
```

wordcloud

wordcloud2

2. 강의노트_텍스트마이닝 05, 06에서 진행한 과정을 한글위키의 “동학농민혁명” 키워드에 대해 적용하시오.

#텍스트파일 불러오기(html 파일을 읽어올 때 incomplete final line 오류가 떠서 readr 라이브러리를 추가로 설치)

```
> library(readr)
>t<-readr::read_lines('https://ko.wikipedia.org/wiki/%EB%8F%99%ED%95%99_%EB%86%8D%EB%AF%BC_%ED%98%81%EB%AA%85')
> d<-htmlParse(t, asText=TRUE) #HTML에서 텍스트 읽어오기
> clean_doc<-xpathSApply(d,"//p",xmlValue) #R 데이터형으로 변환

#전처리 수행
> doc<-Corpus(VectorSource(clean_doc))
> inspect(doc)
<<SimpleCorpus>>
Metadata:  corpus specific: 1, document level (indexed): 0
Content:  documents: 198

  [1] \n\t\t로그아웃한 편집자를 위한 문서 더 알아보기\n\t
  [2] 농민혁명(東學農民革命)[1], 동학 혁명(東學革命), 동학 운동(東學運動), 동학 농민운동(東學農民運動) 또는 동학 농민 전쟁(東學農民戰爭)으로 불리우기 시작한 동학난(東學亂)은 1894년 동학 지도자들과 동학 교도 및 농민들에 의해 일어난 백성의 무장 봉기를 가리킨다. 크게 1894년 음력 1월의 고부 봉기(1차)와 음력 4월의 전주성 봉기(2차)와 음력 9월의 전주·광주 궐기(3차)로 나뉜다.\n  #이후 생략
#DTM 구축
> dtm<-DocumentTermMatrix(doc)
> dim(dtm)
[1]  198 4762
> inspect(dtm)
<<DocumentTermMatrix (documents: 198, terms: 4762)>>
Non-/sparse entries: 7315/935561
Sparsity           : 99%
Maximal term length: 18
Weighting          : term frequency (tf)
Sample             :
     Terms
Docs  1894년 그 그러나 농민 농민군은 농민군의  동학 이  전봉준은 함께
```

102	0	1	0	0	1	0	0	0	1	0
105	0	0	0	0	4	0	1	0	1	0
139	1	1	2	0	0	0	1	0	1	0
140	0	0	1	3	0	0	3	0	0	0
143	0	3	0	0	0	0	0	0	0	0
145	0	3	0	0	0	0	0	0	0	0
24	0	0	0	0	0	0	0	0	0	0
55	0	1	0	0	2	2	0	0	2	0
85	0	0	1	0	0	0	0	0	0	0
94	1	0	0	0	0	0	1	0	0	0

```
> m = as.matrix(dtm)
> v = sort(colSums(m), decreasing = TRUE)
> d = data.frame(word = names(v), freq = v)
> d1 = d[1:100, ] #100개 단어만 표시
> wordcloud2(d1)
```

'그러나', '그', '이', '한' 등이 중요 자리 차지 : 영어 텍스트 마이닝을 한글에 적용한 한계

```
#KoNLP를 이용한 형태소 분석
> library(KoNLP)
> useSystemDic()
> useSejongDic()
> useNIADic() #1213109 words dictionary
> nouns<-extractNoun(clean_doc)
> head(nouns)
[[1]]
[1] "로그아웃" "한"        "편집자"    "문"        "아보"
```

```
[6] "기"

[[2]]
 [1] "농민혁명(東學農民革命)[1]" "동학"
 [3] "혁명(東學革命)"             "동학"
 [5] "운동(東學運動)"             "동학"
 [7] "농민"                       "운동(東學農民運動)"
 [9] "동학"                       "농민"                    #이후 생략

> head(SimplePos22(clean_doc), 1) #Pos22 단계까지 형태소 분석, 일부 추출
[[1]]
[[1]]$로그아웃한
[1] "로그아웃한/NC"

[[1]]$편집자를
[1] "편집자/NC+를/JC"

[[1]]$위한
[1] "위하/PV+ㄴ/ET"

[[1]]$문서
[1] "문/NC+서/JC"

[[1]]$더
[1] "더/MA"

[[1]]$알아보
[1] "알아보/NC"

[[1]]$기
[1] "기/NB“
#KoNLP 워드클라우드 생성하기
> mnous<-unlist(nouns)
> mnous<-Filter(function(x){nchar(x)>=2}, mnous) #두 글자 이상의 단어만 출력
> mnous_freq<-table(mnous)
> v<-sort(mnous_freq, decreasing=TRUE)
```

> wordcloud2(v)