텍스트마이닝 과제3

데이터테크전공 20173204 곽명빈

2020-11-05

문제 1

- 관심 있는 영화를 한편 골라 다음이나 네이버에서 100명 정도의 관객 영화평을 [movie.txt] 에 저장하시오.
- 정리한 파일을 불러와서 명사 키워드를 추출하고 단어구름을 그려보시오.
- 키워드 선정과정에서 필요한 경우 적절한 전처리 (pre-processing)를 하는 것을 권장합니다.

```
library(KoNLP)
```

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

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(wordcloud)
```

```
## Loading required package: RColorBrewer
```

```
library(stringr)
```

데이터(영화: 나는 내일, 어제의 너와 만난다)

```
txt <- readLines("movie1.txt")
head(txt,20)  ## 데이터 20개만 보이기
```

```
##  [1] "관람객 개인적으로 일본실사화영화중 인생작 이라고 불릴만한 영화를 찾은것같습니다 ㅠㅠ"
##  [2] "관람객 1번보면 마지막에 울고 2번보면 처음부터 운다"
##  [3] "아무렇지않게 본 첫장면이마지막에 이렇게 가슴 아프게할지 몰랐습니다.살면서 이토록 깊고 진한 여운을 주는 영화는 처음입니다.논리적으로 따지며 보기 보다는감정적으로 이해하려는 마음으로 감상하는게 좋습니다."
##  [4] "일본 로맨스는 시공간을 초월한 사랑을 너무나도 아름답게 그려낸다."
##  [5] "이 영화는 두번 봐야합니다.... 영화를 본사람들은 무슨뜻인지 알거에요 ㅜㅜ"
##  [6] "조정석도 나오네요 일본어 배우러가야겠습니다."
##  [7] "관람객 20세 남자 혼자서 보고왔어요. 정말 영화관에서 그렇게 울어본게 처음이었고, 개인적으로는 혼자가서 정말 더 몰입하고 볼수 있었던 것 같아요. 처음엔 ₩"에이 이런 전개야?₩" 하면서 의아해하다가, 정말로 ..."
##  [8] "관람객 두번은 못 볼 것 같아요 바로 울 것 같아서.. 올해 본 영화중 최고인 듯"
##  [9] "영상이 진짜 예쁘고 배경음악도 진짜 좋고 고마츠 나나가 정말 사랑스럽게 나오는 영화! 일본에서 개봉당일 보러가서 후유증 때문에 교토가고 아직도 해피앤드 매일 듣는중ㅠㅠ 어쩌면 초반5분이 가장슬픈영화! 판타지 로맨스 ..."
## [10] "관람객 보면서, 느낀건 마치 서로의 관계는 시소와 같아서 한쪽이 내려가면 다른쪽은 올라간다는걸 느꼈어요.남자의관점과 여자의 관점을 보여주니 더이해가 갔어요."
## [11] "관람객 개인적으로 일본실사화영화중 인생작이라고불릴만한 영화를 찾은것같습니다 ㅠㅠ"
## [12] "로맨스 중에는 노트북이 제 마음속 1위였는데 오늘 이후로 2위로 바꼈어요"
## [13] "25살의 타카토시는 15살의 에미를 찾아가 그림을 선물하지만... 25살의 에미는 15살의 타카토시를 찾지않고 30살이 되어서야 10살의 타카토시를 찾아가 타코야끼를 먹으며 사진상자만 선물한다. 25살때 얼마나 보고..."
## [14] "초반 이해는 힘들었지만 어느순간 눈물이 고여있음을 느낍니다"
## [15] "차라리 만나지 않았더라면 아프지 않았을텐데. 그럼에도 만남을 택해버린 그런 사랑."
## [16] "마지막 여주 지하철씬 넘 슬픔 ㅠ 마지막 처음 ㅠ"
## [17] "불법으로 인터넷에서 감상했습니다 죄송합니다 다음주 시간나는대로 휴지챙겨서 영화관가겠습니다 정말 죄송합니다"
## [18] "관람객 주어진 열악한 상황 속에서 최선을 다해 사랑하는 이야기."
## [19] "이해가안되는건 연애경험이없거나 아직헤어져보지않은사람들인가봄.... 아직도 많이 좋아하는데 아무것도아닌사이로 돌아가야한다는게 참 너무나슬픔.."
## [20] "관람객 가볍게들어가서 울고나왔다"
```

```
nouns <- sapply(txt, extractNoun, USE.NAMES = F)

nouns_unlist <- unlist(nouns)

##데이터 전처리
nouns_unlist <- Filter(function(x){nchar(x)>=2}, nouns_unlist)
nouns_unlist<- gsub('[~!@#$%&*()_+=?<>]','',nouns_unlist)
nouns_unlist <- gsub("₩₩[","",nouns_unlist)
nouns_unlist <- gsub('[ㄱ-ㅎ]','',nouns_unlist)
nouns_unlist<- gsub('(ㅜ|ㅠ)','',nouns_unlist)
nouns_unlist <- gsub("₩₩d+","",nouns_unlist)

head(nouns_unlist, 30) ##30개 추출
```

```
##  [1] "관람객"           "일본"           "실사화"           "영화"
##  [5] "인생"             "영화"           "관람객"           "번보면"
##  [9] "마지막"           "번보면"         "첫장면이마지막에" "가슴"
## [13] "여운"             "영화"           "처음"             "논리"
## [17] "보다는감정적으로" "이해"           "하려"             "마음"
## [21] "감상"             "일본"           "로맨스"           "시공간"
## [25] "초월"             "사랑"           "영화"             "영화"
## [29] "사람들"           "무슨뜻인지"
```

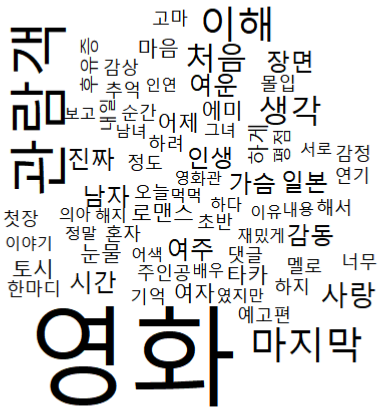
```
wordcount <- table(nouns_unlist)

wordcount_top <-head(sort(wordcount, decreasing = T),100)

wordcount_top      # 많이나온 단어
```

```
## nouns_unlist
##         영화       관람객                 마지막         이해         생각
##           65           32           18           18           18           14
##         처음         사랑         남자         시간         인생         진짜
##           13           12           10           10           10           10
##         가슴         감동         여주         일본         장면         여운
##            9            9            9            9            9            8
##         눈물         마음         어제         하게       로맨스         에미
##            7            7            7            7            6            6
##         여자         타카         토시         기억         오늘         첫장
##            6            6            6            5            5            5
##         추억         평점         하려         하지         해서         혼자
##            5            5            5            5            5            5
##         감상         감정         고마         너무         댓글         멜로
##            4            4            4            4            4            4
##         몰입         순간         연기       예고편         정도       주인공
##            4            4            4            4            4            4
##         초반       한마디       후유증         그녀         남녀         내용
##            4            4            4            3            3            3
##         내일         먹먹         배우         보고         서로         어색
##            3            3            3            3            3            3
##       였지만       영화관         의아       이야기         이유         인연
##            3            3            3            3            3            3
##       재밌게         정말         하다         해지         가안         감사
##            3            3            3            3            2            2
##         감성     감정이입         계산         고메       나와 남길정도로..
##            2            2            2            2            2            2
##       남배우         너와       노트북         대사         때문         말하
##            2            2            2            2            2            2
##         면의         발특         배려         버스       번보면         번째
##            2            2            2            2            2            2
##         부족       분위기       비디오       비현실     비현실적         사람
##            2            2            2            2            2            2
##       상대방         상상         선물         세계
##            2            2            2            2
```

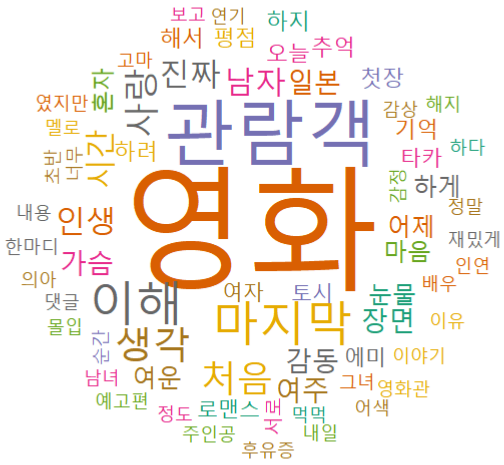
```
wordcloud(names(wordcount_top), wordcount_top)
```



워드클라우드

```
pal <- brewer.pal(8, "Dark2") ##단어의 색

wordcloud(names(wordcount_top),
          wordcount_top,
          scale=c(5,0.5),
          random.order = FALSE,
          random.color = TRUE,
          colors = pal,
          family = "")
```



문제 2

- [서울시 착한가격 업소.xlsx] 의 "업소명" 변수를 이용하여 정규표현식을 익히고자합니다. 한글이름 엑셀파일을 여는데 문제가 있는 학생들을 위해 [seoul.xlsx] 로도 올려두었습니다.
- 업소명 정보는 아래의 코드와 같이 불러와집니다.

  1) 업소명에 '분식'이 들어간 업소를 모두 찾으시오.
  2) 업소명에 숫자가 들어간 업소를 모두 찾으시오.
  3) 업소명에 '헤어'가 들어간 업소와 '미용실'이 들어간 업소를 찾아서 개수를 비교해보시오.

```
library(readxl)
library(ggplot2)
seoul <- read_excel("seoul.xlsx")
```

#분식이 들어간 업소

```
##2-1번
address <- seoul$'업소명'
head(address, 50)
```

```
##  [1] "망우찜쌈밥"                    "아폴로헤어크리닉"
##  [3] "상록수 미용실"                 "열린미용실"
##  [5] "미림17분칼라"                  "노랑머리미용실"
##  [7] "덕성이발관"                    "헤어디자인하우스"
##  [9] "목우촌 부추삼겹살"             "박막례청진동해장국"
## [11] "행운미용실"                    "재희분식"
## [13] "오백냥분식"                    "토방 닭 한마리"
## [15] "호계대중사우나"                "행복한미용실"
## [17] "재경헤어라인"                  "한독세탁"
## [19] "왕세숫대야냉면(행복을파는집)" "돌마리유황오리"
## [21] "길거리야"                      "장미 미용실"
## [23] "강헤어컬렉션"                  "벧엘 칼국수"
## [25] "보라매25시해장국"              "예성미용실"
## [27] "목화미용실"                    "행운의스튜디오"
## [29] "헤어포유"                      "중화루"
## [31] "유성자헤어아트"                "해피분식"
## [33] "윤희미용실"                    "머리잘하는집"
## [35] "에스터미용실"                  "으뜸크리닝"
## [37] "금강숯불생고기"                "한우마당"
## [39] "흥부농장"                      "그린세탁"
## [41] "자매식당"                      "김밥나라"
## [43] "헤어포인트"                    "명신미용실"
## [45] "다비다식당"                    "유경희헤어샵"
## [47] "이계임헤어모드"                "한일세탁"
## [49] "코닥스튜디오"                  "스타머리방"
```

```
grep('분식', address, value = T)
```

```
##  [1] "재희분식"          "오백냥분식"        "해피분식"          "다사랑분식"
##  [5] "학생회관분식"      "박리분식"          "무진분식"          "홍가네왕만두분식"
##  [9] "중앙한분식"        "강남분식"          "사랑분식"          "서울분식"
## [13] "한아름분식"        "명가분식"          "모아분식"          "짱구분식"
## [17] "건대종합분식"      "분식나라"          "한분식"            "허브분식"
## [21] "개봉분식"          "닷컴분식"          "일억조분식"        "다동분식"
## [25] "미진분식"          "자매분식"          "이모네분식"        "명동분식"
## [29] "쌍둥이분식"        "김밥분식"
```

#숫자가 들어간 업소

```
##2-2번
grep('₩d', address, value = T)
```

```
##  [1] "미림17분칼라"              "보라매25시해장국"
##  [3] "21세기이발"                "제2연출헤어모드"
##  [5] "머리못하는집 목동 89호점" "21C헤어미넷"
##  [7] "24시헤어샵"                "G7크린랜드"
##  [9] "5080 실버전용미용실"       "백제231"
## [11] "머리못하는집 목동 95호점" "카츠3.3"
```

#헤어가 들어간 업소와 미용실이 들어간 업소 비교

```
##2-3번
hair <- grep('헤어', address)  #'헤어'가 들어간 업소
hairshop <- grep('미용실', address) #'미용실'이 들어간 업소
length(hair)    #헤어가 들어간 업소의 갯수
```

```
## [1] 51
```

```
length(hairshop)# 미용실이 들어간 업소의 갯수
```

```
## [1] 77
```

```
# barplot
a <- length(hair)
b <- length(hairshop)
al=c(a, b)
barplot(al,
        main = "두 변수 비교",
        ylim = c(0,80),
        col =c("#2EFEF7", "#81F7BE"),
        names.arg = c("hair", "hairshop"),
        width = 0.00001)
```