2021.02.17

[텍스트 마이닝(Text Mining)]

1. 텍스트 마이닝이란?

:문자로 된 데이터에서 가치 있는 정보를 얻어내는 분석 기법

2. 텍스트 마이닝 절차

(1) 패키지 설치 및 로드

```
install.packages("rJava"); install.packages("memoise"); install.packages("KoNLP")
```

(2) 사전(dictionary) 설정 :**useNIADic()**

(3) 텍스트 마이닝을 할 데이터 준비(불러오기): **readLines**("불러올 파일명") ex) txt <- readLines("1.txt")

(4) 불필요한 문자들 제거하기 (아래의 함수들 중 하나를 사용한다)

- **gsub**(oldStr, newStr, string)
- **str_replace_all**(string, oldStr, newStr) => stringr 로드 필요

(5) 명사 추출: **extractNoun**( ) ex) nouns <- extraNoun(txt)

(6) 워드 카운트(단어 빈도표) 만들기

1) 추출한 명사들의 list를 vector로 변환 ex) wordcount <- table(unlist(nouns)
2) 데이터 프레임으로 변환 (및 변수 수정) ex) df_word <- as.data.frame(wordcount, stringAsFactor = F)
3) 추출할 단어들의 글자 수 설정: **filter**(데이터프레임명, **nchar(word)** >= 2); *두글자 이상인 단어들만 필터링*

(7) 워드 클라우드 만들기

1) 필요한 패키지 설치 및 로드: **install.package("wordcloud")** / **library(worldcloud)**
2) 단어에 입힐 색상 설정: **brewer.pal**(8, "팔레트set이름") *=> RColorBrewer 패키지 설치 필요*
3) 워드 클라우드 생성

```
set.seed(1234)                    # 난수 고정
wordcloud(words = df_word$word,   # 불러올 단어들
          freq = df_word$freq,    # 빈도
          min.freq = 5,           # 최소 단어 빈도
          max.words = 200,        # 나타낼 단어의 수
          random.order = F,       # 고빈도 단어들을 중앙에 배치
          rot.per = 0.1,          # 단어 회전 비율
          scale = c(3, 0.3),      # 단어 크기의 범위
          colors = )              # 단어 색 지정
```