12. 텍스트 마이닝(Text mining) # src: 7_R/12_텍스트마이닝(워드클라우드) 참고

1) 텍스트 마이닝; 문자로 된 비정형 데이터에서 가치있는 정보를 얻어내는 분석 기법

(1) 분석 절차: ① 형태소 분석 → ② 단어 추출 → ③ **빈도표** 생성 → ④ 시각화

2) 텍스트 마이닝 준비하기

(1) 패키지 설치 및 로드 #JAVA 설치 및 JAVA_HOME 환경변수 설정 필요

① rJava, memoise, KoNLP, devtools, hash, tau, Sejong #별도로 KoNLP 0.80.2.tar.gz 다운로드必

② 우측하단 Packages->install 이용(devtools)하여 KoNLP 패키지 설치(KoNLP 0.80.2.tar.gz)

#scala-library-2.11.8.jar 다운로드必(위치: C:/Users/Home/Documents/R/win-library/4.0/KoNLP/java)

③ 로드: library(KoNLP)

④ 사전 설정: useNIADic()

3) 단어 추출 및 빈도표 생성

(1) 전처리: 데이터 준비 및 특수문자 제거

① 데이터 준비: 데이터명 <-readLines("파일명") ┌#library(stringr) 필요

② 특수문자 제거: 데이터명 <-str_replace_all(데이터명, "₩₩W", " ") or gsub(oldStr, newStr, string)

(2) **단어(명사) 추출**: nouns <- extractNoun(데이터명)

(3) **단어별 빈도표 생성**: wordcount <-table(unlist(nouns)) #추출한 명사를 list 문자열 벡터로 변환

4) 시각화; 시각화 위한 데이터 프레임 변환, 변수명 수정, 2글자 이상 단어 추출(nchar()) 등

(1) 데이터 프레임으로 변환: df_word <-as.data.frame(wordcount, stringsAsFactors=F)

(2) ggplot2 활용한 그래프 생성, wordcloud 활용 등

5) 워드 클라우드 #install.package("wordcloud"); library(wordclod) 필요

```
Pal <- brewer.pal(9, "Blues")[5:9]          #색상 목록 생성
set.seed(1234)                              #난수 고정
wordcloud(words = df_word$word,             #단어
          freq = df_word$freq,              #빈도
          min.freq = num,                   #최소 단어 빈도 ex) 2
          max.words = num,                  #표현 단어 수   ex) 200
          random.order = F,                 #고빈도 단어 중앙 배치
          rot.per = 0.1                     #회전 단어 비율
          scale = c(num1, num2)             #단어 크기 범위(num1>num2) ex) 2, 0.3
          colors = pal)                     #색깔 목록, 색상 팔레트 이름 기입 가능
```